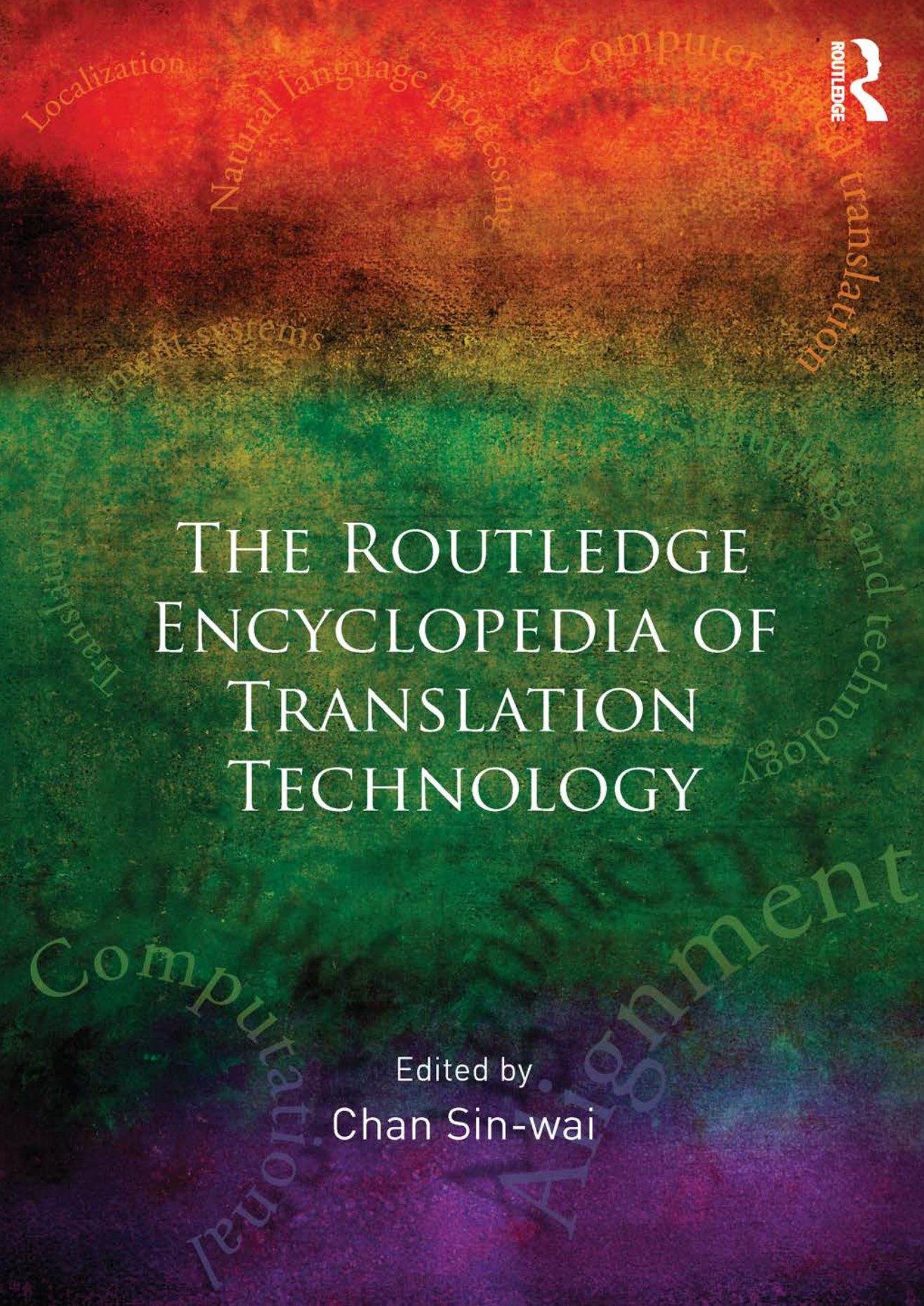


The Routledge Encyclopedia of Translation Technology

Edited by

Chan Sin-wai

THE ROUTLEDGE ENCYCLOPEDIA OF TRANSLATION TECHNOLOGY

The Routledge Encyclopedia of Translation Technology provides a state-of-the art survey of the field of computer-assisted translation. It is the first definitive reference to provide a comprehensive overview of the general, regional and topical aspects of this increasingly significant area of study.

The *Encyclopedia* is divided into three parts:

- Part One presents general issues in translation technology, such as its history and development, translator training and various aspects of machine translation, including a valuable case study of its teaching at a major university.
- Part Two discusses national and regional developments in translation technology, offering contributions covering the crucial territories of China, Canada, France, Hong Kong, Japan, South Africa, Taiwan, the Netherlands and Belgium, the United Kingdom and the United States.
- Part Three evaluates specific matters in translation technology, with entries focused on subjects such as alignment, bitext, computational lexicography, corpus, editing, online translation, subtitling and technology and translation management systems.

The Routledge Encyclopedia of Translation Technology draws on the expertise of over 50 contributors from around the world and an international panel of consultant editors to provide a selection of articles on the most pertinent topics in the discipline. All the articles are self-contained, extensively cross-referenced, and include useful and up-to-date references and information for further reading.

It will be an invaluable reference work for anyone with a professional or academic interest in the subject.

Chan Sin-wai is Professor in the Department of Translation at The Chinese University of Hong Kong. His research interests include computer translation, translation studies and lexicography.

Praise for this edition:

"It is unique in that it provides both a truly encyclopedic overview of the field of computer-assisted translation and an in-depth discussion of the various aspects of this young discipline by some of the best-known experts in each of the disciplines this book covers. I was particularly pleased by the fact that so much emphasis was placed on teaching translation technology, which clearly is an area that deserves this type of attention."
Uawe Muegge, *Monterey Institute of International Studies, USA*

"This is an immensely useful compendium of information about the history, techniques and world-wide distribution of translation technologies over the last 70 years. The articles are authored by international specialists whose wide-ranging expertise is brought together here for the first time. The individual articles are almost without exception at the cutting edge of knowledge, so will be of value to researchers and other specialists as well as students … Part Two (on national and regional developments) is particularly innovative and valuable in setting translation technology in its social and political context in different societies around the world."
Andrew Rothwell, *Swansea University, UK*

THE ROUTLEDGE ENCYCLOPEDIA OF TRANSLATION TECHNOLOGY

Edited by Chan Sin-wai

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2015
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Routledge encyclopedia of translation technology / edited by Sin-Wai Chan.
pages cm
1. Translating and interpreting--Technological innovations--Encyclopedias. 2. Translating and interpreting--Encyclopedias. I. Chan, Sin-wai, editor. II. Title: Encyclopedia of translation technology.
P306.97.T73R68 2015
418'.020285--dc23
2014018223

ISBN: 978-0-415-52484-1 (hbk)
ISBN: 978-1-315-74912-9 (ebk)

Typeset in Bembo
by Saxon Graphics Ltd, Derby

CONTENTS

FIGURES

TABLES

CONSULTANT EDITORS

Lynne Bowker
School of Translation and Interpretation
University of Ottawa, Canada

David Farwell
TALP Research Centre, Universitat Politecnica de Catalunya
Barcelona, Spain

W. John Hutchins
University of East Anglia, the United Kingdom

Alan K. Melby
Department of Linguistics and English Language
Brigham Young University, the United States

William S-Y. Wang
Department of Electronic Engineering
The Chinese University of Hong Kong, Hong Kong

CONTRIBUTORS

Lars Ahrenberg is Professor of Computational Linguistics at the Department of Computer and Information Science, Linköping University and head of its NLP research since 1995. He received a BA in General Linguistics and Mathematics from Stockholm University and entered computational linguistics at Uppsala University in the beginning of the 1980s. Dr Ahrenberg's research spans several areas of NLP including dialogue systems, grammar formalisms and syntactic analysis, corpus linguistics, terminology extraction, and translation technologies. He has published more than 15 papers on methods and uses of word alignment. His current research interests include machine translation, translation evaluation and the growing interface of translation technology and translation studies. Dr Ahrenberg has been editor-in-chief of the *Northern-European Journal of Language Technology* since 2012, and is a member of the Association for Computational Linguistics and the European Association for Machine Translation.

Bai Xiaojing is an associate professor at the Department of Foreign Languages and Literatures of Tsinghua University. She received her PhD in Computer Software and Theory at the Institute of Computational Linguistics, Peking University, in 2004. Her main research interest is in computational linguistics, specifically, corpus linguistics, computer-aided translation studies, and educational technology. She published papers in *Chinese Translators Journal* and *Computer-assisted Foreign Language Education*, and was one of the contributors to *China Translation Yearbook*. She is currently a member of the Professional Committee of Computer-Assisted Language Learning, China English Language Education Association, and a council member of Chinese Information Processing Society of China. She is also a member of the Center for Translation and Interdisciplinary Studies at Tsinghua University.

Lynne Bowker holds a BA and MA in Translation (University of Ottawa, Canada), an MSc in Computer Applications for Education (Dublin City University, Ireland) and a PhD in Language Engineering (University of Manchester Institute of Science and Technology, United Kingdom). She is currently a Full Professor at the School of Translation and Interpretation at the University of Ottawa, where her teaching and research focus on the use of computer tools in the disciplines of translation and terminology. Her publications include *Computer-Aided Translation Technology* (University of Ottawa Press, 2002) and *Working with Specialized Language: A Practical Guide to Using Corpora* (Routledge, 2002), as well as many articles and book chapters

on various aspects of computer-aided translation. Additionally, she serves on the editorial board of the *International Journal of Corpus Linguistics* (John Benjamins Publishing Company) and the *International Journal of Lexicography* (Oxford University Press), as well as on the advisory board for *The Interpreter and Translator Trainer* (Routledge).

Sylviane Cardey, PhD and State Thesis in linguistics, Member of the Institut universitaire de France, is Tenured Professor (Classe exceptionelle) of Linguistics and Natural Language Processing in the University of Franche-Comté, Besançon, France. She is Director of the Centre de Recherche en Linguistique et Traitement Automatique des Langues Lucien Tesnière and has created the Natural Language Processing (NLP) programme at Besançon and the European Master Mundus Course in NLP and HLT (Human Language Technology). Her research is fundamentally oriented and principally concerns the development of her systemic linguistics analysis model and its associated calculus. This work has led to applications in which quality is paramount. The principal applied research domains are in machine translation (between languages of the same and different origins or families) where she has been lead partner in collaborative projects in France and internationally, sense-mining and controlled languages. She has several international collaborations in particular in safety critical environments, for example with enterprises in the aircraft manufacturing and food-processing industries. She has supervised 37 PhDs, has 126 international papers and a book, *Modelling Language* (2013), has created the international conference series XTAL, and is a national and international expert. Sylviane Cardey is chevalier de la Légion d'honneur.

Chan Sin-wai is Professor at the Department of Translation, the Chinese University of Hong Kong. He is also Director of the Master of Arts in the Computer-aided Translation Programme and Director of the Centre for Translation Technology. His teaching and research interests lie mainly in the areas of translation studies, translation technology, and bilingual lexicography. He is the Chief Editor of *Journal of Translation Technology*, published by the Chinese University of Hong Kong. He has published 40 books in 50 volumes, mainly dictionaries and scholarly monographs, and translated works in different fields. He edited *An Encyclopaedia of Translation*, revised *Longman Dictionary of English Language and Culture* (bilingual edition), authored *A Dictionary of Translation Technology* and *A Chinese–English Dictionary of the Human Body*. His book translations from Chinese into English include *An Exposition of Benevolence*, *Palaces of the Forbidden City*, *Letters of Prominent Figures in Modern China*, *Paintings and Calligraphy of Jao Tsung-I*, *Stories by Gao Yang*, *An Illustrated History of Printing in Ancient China*, *Famous Chinese Sayings Quoted by Wen Jiabao*, and *Selected Works of Cheng Siwei: Economic Reforms and Development in China*, Volume 2. He also translated *My Son Yo Yo* from English into Chinese. His most recent co-edited books include *Style, Wit and Word-Play* (2012) and *The Dancer and the Dance* (2013), both published by Cambridge Scholars Publishing.

Freddy Y. Y. Choi created the C99 text segmentation algorithm during his PhD and has worked on a range on applied natural language processing projects with BBN Technologies, the European Commission, BBC, Rolls-Royce, Finmeccanica, and the UK Government over the last 20 years. He has a keen interest in making cutting-edge but imperfect artificial intelligence solutions work in real world applications.

Ian Castor Chow received his PhD in computational linguistics from City University of Hong Kong. He is currently a lecturer at the Department of Translation, the Chinese University of Hong Kong. Before joining The Chinese University of Hong Kong, he was an instructor

and research fellow at the Department of Chinese, Translation and Linguistics, City University of Hong Kong. His publications mainly focused on Systemic Functional Linguistics with the interoperability of WordNet, FrameNet and other linguistic resources. Currently, his teaching concentrates on the areas of translation technologies, computer-aided translation, localization and terminology. His research interests are computational linguistics, linguistic resource and ontology engineering, terminology, discourse analysis and computer-aided translation.

Jennifer DeCamp is the Chief Scientist for Human Language Technology in MITRE Corporation's Department of Social, Behavioral, and Linguistic Sciences in their Center for Connected Government. She works with MITRE's five Federally Funded Research and Development Centers (FFRDCs) to provide consulting to the US Government on translation processes and technology. Jennifer has worked with translation for over 30 years, including as a translator, an editor, a professor of translation, an industry programme manager and a developer. She has worked extensively with the American Translators Association (ATA), American Standards for Time and Materials (ASTM), the International Organization for Standardization (ISO), and other organizations to develop standards for improving translation and interpretation, particularly for advancing technology in these areas.

Christophe Declercq was born in Antwerp and graduated as a translator (Dutch/English/Russian) from Lessius in 1998. After lecturing at Imperial College London for 11 years, he moved to University College London in October 2013. He lectures in translation technology and localization. Christophe has been an evaluator for multilingual ICT projects under FP7. Other than a freelance translator, he also lectures in British culture at the University of Antwerp. He has been a visiting lecturer at universities in Belgium, France, the Netherlands and the United Kingdom. Christophe has written several articles and chapters on translation and language technology.

Jorge Díaz Cintas is the Director of the Centre for Translation Studies (CenTraS) at University College London. He is the author of numerous articles, special issues and books on audiovisual translation. He is the founding Chief Editor of *New Trends in Translation Studies* and an expert board member of the EU LIND-Web initiative.

Keiran J. Dunne is a professor of Translation Studies, chair of the Department of Modern and Classical Language Studies and a member of the faculty in the Institute for Applied Linguistics at Kent State University, where he teaches graduate courses on computer-assisted translation, localization, project management and the language industry. Drawing upon more than 15 years' experience as a French localization and technical translation subcontractor for Fortune 500 companies and other corporate clients, his research interests include localization, project management, quality management, terminology management and the industrialization of translation. He is the editor of the collective volume *Perspectives on Localization* (2006) and the co-editor of the collective volume *Translation and Localization Project Management: The Art of the Possible* (2011). He is a member of the editorial advisory boards of the journal *Translation Spaces* and of the American Translators Association Scholarly Monograph Series.

David Farwell was a senior ICREA research scientist at the Catalonia Polytechnic University in Barcelona, Spain, from 2002 through 2011 and served as Director of the University's Centre for Language and Speech Applications and Technologies from 2009 through 2011. From 1985 through 2006, Dr Farwell was a Research Scientist at the Computing Research Laboratory,

New Mexico State University in Las Cruces, New Mexico. He received a PhD in Linguistics in 1985 from the University of Illinois-Urbana. In 2012, he retired from active research. Dr Farwell was particularly active in the areas of Machine Translation, Natural Language Processing, semantic representation and reasoning and Computational Pragmatics. During that time, he was Principal Investigator for three National Science Foundation projects, two US Defense Department projects and two industry-funded projects. In addition, he participated in four European Commission Information and Communication Technologies projects, four Spanish Ministry of Science and Innovation projects and an additional five US Defense Department projects. He has published over well 100 articles in conference proceedings, journals and edited volumes, including some 25 on Pragmatics-based Machine Translation. Additionally, he has given several invited talks, panel presentations, tutorials and seminars as well as chaired or participated in numerous conference organizing and programme committees.

Mikel L. Forcada was born in Caracas (Venezuela) in 1963 and is married with two children. He graduated in Science in 1986 and got his PhD in Chemistry in 1991. Since 2002 he is full professor ('Catedràtic') of Computer Languages and Systems at the Universitat d'Alacant. Professor Forcada is also secretary of the European Association for Machine Translation, president of the Special Interest Group for Speech and Language Technologies for Minority Languages (SaLTMiL) of the International Speech Communication Association, and book review editor of the international journal *Machine Translation*. From the turn of the millennium on, Professor Forcada's interests have mainly focused on the field of translation technologies, but he has worked in fields as diverse as quantum chemistry, biotechnology, surface physics, machine learning (especially with neural networks) and automata theory. He is the author of 17 articles in international journals (three on machine translation), 48 papers in international conferences (25 on machine translation) and six book chapters (three on machine translation).

Professor Forcada has headed several publicly and privately funded projects and has led the development of the machine translation systems interNOSTRUM (Spanish–Catalan) and Traductor Universia (Spanish–Portuguese). More recently (2004), he started the free/open-source machine translation platform Apertium (with more than 26 language pairs), where he is currently the president of the project management committee. He is also administrator in three more free software projects (Bitextor, Orthoepikon, Tagaligner) and co-founder of start-up company Prompsit Language Engineering (2006). Professor Forcada has participated in the scientific committees of more than 20 international conferences and workshops. Recently (2009–2010) he has been an ETS Walton Visiting Professor in the machine translation group at Dublin City University.

Ignacio Garcia is a senior lecturer at the School of Humanities and Communication Arts, University of Western Sydney, where he teaches in translation English–Spanish and translation technologies. He has published in academic and professional journals on translation memory, translation memory and machine translation integration, post-editing, and uses of machine translation for language learning. His current interest is on the deployment of digital technology to assist bilinguals to translate and everyone to interact in unfamiliar linguistic environments. He has also taught and published on Spanish and Latin American studies, having completed his PhD on this area.

Federico Gaspari has a background in translation studies and holds a PhD in machine translation from the University of Manchester (United Kingdom). He was a lecturer in Italian language and technical and specialized translation at the Universities of Manchester and Salford

(United Kingdom) before joining the Universities of Bologna (Forlì campus) and Macerata in Italy, where he teaches English language and linguistics, specialized translation and translation technology, with a focus on machine translation, post-editing and online translation tools.

He has held postdoctoral research fellowships at the University of Bologna at Forlì (Italy), where he worked on corpus-based projects investigating translation universals and the features of translated and non-native/L2 English as mediated language varieties resulting from contact with Italian. He is a postdoctoral researcher at the Centre for Next Generation Localisation at Dublin City University (Ireland), where he is involved in international EU-funded projects concerning (human and machine) translation quality evaluation. He has published mostly in the areas of translation technology, machine translation (including evaluation), corpus-based translation studies and English corpus linguistics, and is a regular presenter at international conferences in these fields. He is writing a book exploring the impact of the Internet on translators and translation.

Marissa Griesel completed a Bachelor's degree in Computational Linguistics as well as a Master's degree in Applied Linguistic Theory (cum laude) at North-West University (NWU), South Africa. The title of her thesis was 'Syntactic Reordering as Pre-processing in the Development of an English-Afrikaans Machine Translation System'. As part of her MA degree, she spent three months as an exchange student at the University of Tilburg, the Netherlands. She worked as research assistant (2005–2007), project manager (2008) and later computational linguist (2009–2012) at the NWU's Centre for Text Technology (CTexT), and was involved in a wide range of projects, from compiling speech corpora to developing educational software. For example, she played an important role in the Autshumato project – the first machine translation system for various South African languages – and published numerous papers about the project. She also lectured in numerous HLT courses, and led workshops for freelance translators in using computer assisted translation (CAT) tools. She is currently enrolled as a PhD student at the NWU, investigating various ways of fast-tracking natural language processing for South African languages.

Stephen Helmreich worked as a computational linguist at the Computing Research Laboratory of New Mexico State University in Las Cruces, New Mexico, from 1988 through 2007, serving as Deputy Director from 2003 to 2007. He continued to work as a principal investigator in the Department of Information Sciences and Security Systems of the Physical Science Laboratory (also at New Mexico State University) until his retirement in 2009. He received a PhD in Linguistics in 1996 from the University of Illinois-Urbana/Champaign.

Dr Helmreich's areas of interest include Machine Translation, particularly interlingual approaches, computational morphology, and computational approaches to metaphor, metonymy and other types of non-literal language. He was a Principal Investigator on seven projects, and co-principal investigator on three more, including five NSF grants and five defense-related and funded projects. He also served as a board member of the Association of Machine Translation in the Americas and as treasurer from 2001 to 2006. He is the author or co-author of over 40 papers. With Dr David Farwell he has published a series of papers on Pragmatics-based Machine Translation. They have also edited a series of workshop proceedings on Interlingual Machine Translation.

W. John Hutchins is retired from the University of East Anglia (Norwich, United Kingdom), where he worked in the university library. He is the author of books and articles on linguistics, information retrieval and in particular on machine translation. Principal works include: *Machine*

Translation: Past, Present, Future (Chichester: Ellis Horwood, 1986), *An Introduction to Machine Translation* [with Harold Somers] (London: Academic Press, 1992), and editor of *Early Years in Machine Translation: Memoirs and Biographies of Pioneers* (Amsterdam and Philadelphia: John Benjamins Publishing Company 2000). He was editor of *MT News International* from 1991 until 1997, and the *Compendium of Translation Software*, from 2000 until 2011. Since 2004 he has compiled the *Machine Translation Archive* (http://www.mt-archive.info), an electronic repository of publications on MT and related topics – now containing over 10,000 items. He has been a speaker at many machine translation conferences, the *MT Summit* conferences, the EAMT workshops and conferences, the *Translating and Computer* conferences and the AMTA conferences. He was president of the European Association for Machine Translation, 1995–2004, and president of the International Association for Machine Translation, 1999–2001. His website containing most of his publications is at: http://www.hutchinsweb.me.uk.

Hitoshi Isahara is Professor of Information and Media Center at Toyohashi University of Technology, Japan. He received the BE, ME, and PhD degrees in electrical engineering from Kyoto University, Kyoto, Japan, in 1978, 1980, and 1995, respectively. His research interests include natural language processing and lexical semantics. Until December 2009, he was working at the National Institute of Information and Communications Technology (NICT), Japan, as a Leader of the Computational Linguistics Group and a Director of the Thai Computational Linguistics Laboratory (TCL). He was a Guest Professor at Kobe University Graduate School of Engineering, Japan, and part-time lecturer at Kyoto University Graduate School of Human and Environmental Studies and Doshisha University. He was a President of the International Association for Machine Translation (IAMT), and a President of the Asia-Pacific Association for Machine Translation (AAMT). He is working for ISO/TC37/SC4 and a project co-leader of one of its working items. He is a board member of GSK (Gengo Shigen Kyokai, linguistic resource association), Japan.

Kit Chunyu is an associate professor currently teaching and researching in computational linguistics and machine translation at the City University of Hong Kong. He obtained his B. Eng. degree in computer science and technology from Tsinghua University (1985), MSc in computational linguistics from Carnegie Mellon University (1994) and PhD in computer science from the University of Sheffield (2001). He also holds an MA degree in applied linguistics from the Chinese Academy of Social Sciences (1988) and an M.Phil. in Linguistics from the City University of Hong Kong (1993). His research interests include Chinese language processing, text mining, computational terminology and poetry. He has published over 100 research papers in academic journals (including *Information Sciences*, *Journal of Artificial Intelligence Research*, *International Journal of Corpus Linguistics*, *Journal of Computer Science and Technology*, *Law Library Journal*, *Machine Translation*, and *Terminology*) and international conferences (including *ACL*, *COLING*, *CoNLL*, *EMNLP* and *IJCAI*).

Olivia Kwong Oi Yee graduated from the University of Hong Kong with a first degree in Psychology and obtained her PhD in Computational Linguistics from the University of Cambridge with a doctoral thesis on automatic word sense disambiguation. She is currently Assistant Professor in the Department of Linguistics and Translation of the City University of Hong Kong. Her research interests cover natural language processing, corpus linguistics, and psycholinguistics, involving mainly English and Chinese. She has worked in many related areas including lexical semantics, lexical resources, corpus development, bilingual corpus

processing, translation lexicon extraction, name transliteration, mental lexicon and lexical access, and sentiment analysis, amongst others. Her work often focuses on multi-disciplinary and empirical approaches, and has appeared in international journals and important international conferences. Recently she has published a monograph titled *New Perspectives on Computational and Cognitive Strategies for Word Sense Disambiguation*, which presents the topic from an integrated viewpoint. She has also been actively involved in the organization of many international conferences and has served on the editorial board of the journal *Computational Linguistics*.

Lee Tan is an associate professor at the Department of Electronic Engineering and the Director of the DSP and Speech Technology Laboratory, the Chinese University of Hong Kong. He was a visiting researcher at the Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Sweden, during 1997–1998. Tan Lee's research covers many different areas of speech processing and spoken language technologies, including automatic speech and speaker recognition, text-to-speech, spoken dialogue systems, prosody modelling, language identification and music signal processing. He is a member of IEEE Signal Processing Society and a member of the International Speech Communication Association (ISCA). He was the Chairman of IEEE Hong Kong Chapter of Signal Processing during 2005–2006 and the General Chair of the 2012 Oriental COCOSDA Conference. Currently he is an Associate Editor of the *EURASIP Journal on Advances in Signal Processing*, and a Workgroup Chair of the ISCA Special Interest Group on Chinese Spoken Language Processing. Tan Lee received the Chinese University of Hong Kong Vice-Chancellor's Exemplary Teaching Award in 2005.

Li Lan is a fellow of the Chartered Institute of Linguists, United Kingdom, with M.Phil and PhD degrees in Applied Linguistics from the University of Exeter. She works as an Associate Professor at the Department of English, Hong Kong Polytechnic University. Her teaching and research interests cover corpus linguistics, semantics, lexicology, comparative study and sociolinguistics.

Liu Qun is a professor of Machine Translation and a principal investigator at CNGL Centre for Global Intelligent Content at Dublin City University in Ireland. He is also an adjunct professor at the Institute of Computing Technology at the Chinese Academy of Sciences in Beijing, China. He obtained his Master's degree from the same institute in 1992 and his PhD degree from Peking University in 2004. His research interests focus on machine translation and natural language processing, especially on Chinese language processing, statistical machine translation models, approaches and evaluation methods. He was the co-author of the open source Chinese lexical analysis system ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). He is the author or co-author of more than 150 research publications in peer-reviewed journals and conferences.

Liu Yang is an associate professor in the Department of Computer Science and Technology at Tsinghua University, China. His research interests include natural language processing, Chinese information processing and machine translation. He has published more than 20 papers in leading natural language processing journals and conferences, such as *Computational Linguistics*, conferences of the Association for Computational Linguistics (ACL), the Association for the Advancement of Artificial Intelligence (AAAI), Empirical Methods in Natural Language Processing (EMNLP), and Computational Linguistics (COLING). He won COLING/ACL 2006 Meritorious Asian NLP Paper Award. He will serve as ACL 2014 Tutorial Co-chair and ACL 2015 Local Arrangement Co-chair.

Arle Lommel is a recognized expert in the fields of localization and translation. Widely published in the field, he focuses on both the technical and business aspects of the industry and how they relate to each other. As director of standards at LISA, he was responsible for submission of the TBX standard to ISO (now ISO 30042) and for driving standards development at LISA. He holds a PhD in Folkloristics and is trained as a linguist. He has worked extensively in the fields of translation theory and technology, standards development, translation and localization business, ethnographic research, and linguistics. He is currently a senior consultant at the German Research Center for Artificial Intelligence (DFKI) in Berlin, where he is focused on work related to the assessment and improvement of machine translation quality.

Elliott Macklovitch is a linguist who has been actively involved in machine and machine-aided translation since 1977, when he joined the TAUM group at the University of Montreal; this is the group that developed the MÉTÉO System, long considered to be one of the most successful applications in the history of machine translation. From 1986 to 1997, he worked at the Centre for Information Technology Innovation, an Industry Canada research lab, where he was responsible for the Translator's Workstation project. He returned to the University of Montreal in 1997, as Coordinator of the RALI Laboratory. The RALI has developed a number of innovative tools for translators, including the *TransSearch* system, which is now successfully marketed in Canada. Between 2000 and 2004, he served two terms as President of AMTA, the Association for Machine Translation in the Americas. Since leaving the RALI in 2009, he has worked as independent consultant in machine translation. His hobby is literary translation.

Alan K. Melby is Professor of Linguistics at Brigham Young University. He has been involved in translation technology since 1970, when he started working on a machine translation project. In the late 1970s, the team he was part of, and other teams around the world, saw the need to rethink how humans and computers interact in producing translations. The result was translation memory and other productivity tools that are widely used. In the 1980s, he saw the need for data exchange standards and has been involved in the development of TMX and TBX. He is currently the chair of the Standards Committee of the International Federation of Translators. His interests include translation quality assessment, philosophy of language as applied to translation, and the use of structured translation specifications. He served for 16 years on the board of directors of the American Translators Association and is a certified translator (French to English).

Lucía Morado Vázquez is a post-doctorate at the Faculty of Translation and Interpretation, University of Geneva, Switzerland. She joined the multilingual information-processing department TIM/ISSCO in 2012 to work as a researcher and lecturer in the localization field. Lucía obtained a PhD in localization at the Localisation Research Centre (LRC), based in the Computer Science and Information Systems department at the University of Limerick, Ireland. Her PhD research was conducted in association with the Centre for Next Generation Localisation. She also holds a BA in translation and interpreting from the University of Salamanca, Spain. Since 2009, she has been a voting member of the XLIFF Technical Committee and the XLIFF Promotion and Liaison Subcommittee since its establishment. Lucía's research interests are standards of localization, localization training and translation memory metadata.

Nie Jian-Yun is a professor at the Department of Computer Science and Operations Research, University of Montreal. His main research areas are information retrieval and natural language

processing. He is a graduate from Southwest University in China, and he obtained a PhD from the University Joseph Fourier of Grenoble, France. During his career of more than two decades, he has published widely in journals and conferences in these areas. In particular, he obtained the Best Paper award at SIGIR 1999 for his work on cross-language information retrieval using parallel web pages. Jian-Yun Nie has been invited to deliver keynote and invited speeches at a number of conferences. He is regularly invited to serve on programme committees, and is a member of editorial board of several journals. More information about Professor Nie can be obtained from his website http://www.iro.umontreal.ca/~nie.

Qian Duoxiu is an associate professor at Beihang University, Beijing, China. She is also Head of the Department of Translation and Interpretation and Deputy Dean of the School of Foreign Languages of the University. Concurrently, she has been an executive secretary of the Corpus Linguistics Society of China since 2010. Her publications have appeared in *Target: International Journal of Translation Studies*, *META: Journal of Translation Studies*, *Translation Review*, *Machine Translation*, *The ATA Chronicle*, *Terminology*, *Chinese Translators Association Journal*, *Journal of Translation Studies*, and other internationally refereed journals. She has also published a coursebook entitled *Computer-aided Translation* (in Chinese, Beijing: Foreign Language Teaching and Research Press, 2011) and a monograph entitled *Computer-aided Quality Assessment in Scientific and Technical Translation – The Pharmacopoeia of the People's Republic of China as a Case Study* (Changchun: Jilin University Press, 2008). Her research interests include translation theory and practice in the Chinese context and computer-aided translation. She can be reached at qianduoxiu@126.com.

Felipe Sánchez-Martínez, PhD, is a lecturer at the Departament de Llenguatges i Sistemes Informàtics (Universitat d'Alacant, Spain) and a member of the European Association for Machine Translation since 2009. He has published 10 articles in international journals, 5 articles in Spanish journals, a book chapter and more than 30 papers in international conferences. His main field of research is MT and the application of unsupervised corpus-based methods to build some of the modules and linguistic resources needed by rule-based MT systems; his doctoral dissertation focused on the unsupervised learning of part-of-speech taggers and structural transfer rules. He has also worked on the hybridization of MT systems and on the selection of the best MT system to use to translate a given sentence using only source-language information. He has participated in the design and development of the Apertium rule-based machine translation platform, where he is currently a member of the Project Management Committee, and in several projects funded by the Spanish government. Most of his undergraduate and graduate teaching involves translation and language technologies.

Rolf Schwitter is a senior lecturer in the Centre for Language Technology at Macquarie University in Sydney, Australia. His main research interests are natural and formal language processing, in particular controlled natural languages, knowledge representation, automated reasoning and the Semantic Web. Rolf received a PhD in computational linguistics from the University of Zurich. He designed and implemented controlled natural languages and tools for controlled natural language processing for a number of research projects and as a contractor for NICTA, Australia's Information Communication Technology Research Centre of Excellence and for the DSTO, Australia's Defence Science and Technology Organisation. Contact him at Macquarie University, Department of Computing, Centre for Language Technology, NSW, 2109 Australia; Rolf.Schwitter@mq.edu.au; web.science.mq.edu.au/~rolfs.

Shih Chung-ling is Professor of English at National Kaohsiung First University of Science and Technology in Taiwan. Her research focuses and interests are broad, ranging through machine translation (MT), computer-aided translation (CAT), corpus-based translation studies, translation studies, cultural studies, English teaching, and literature studies. Because of her MT and CAT research, she was invited to give a lecture at the School of Foreign Languages of China University of Geosciences in China. Also she was a keynote speaker at the International Conference organized by Guangdong University of Foreign Studies and hosted by Zhongnan University of Economics and Law in China. In addition to these, she has given lectures at some universities in Taiwan, transmitting the concept of controlled web cultural writing for effective multilingual MT application. She has completed a series of researches with the grant of Taiwan's NSC and part of the results was posted on her teaching website. Her research results can also be found in some books such as *Helpful Assistance to Translators: MT and TM* (2006), *Real-time Communication through Machine-enabled Translation: Taiwan's Oracle Poetry* (2011), *Translation Research Models and Application* (2012) and others. Over 30 papers in her name have been published in various journals in Taiwan and abroad.

Mark Shuttleworth has been involved in teaching translation technology since 1996, first at the University of Leeds, then at Imperial College London, and currently at University College London where he works as a Senior Lecturer. As and when time permits he is active as a translator. His publications include the *Dictionary of Translation Studies*, which appeared in 1997 and which was translated into Chinese in 2005, as well as works on metaphor in translation, translation technology, translator training and medical translation. He studied at the Universities of Oxford and Birmingham, and has recently completed a PhD at the University of London. He is a fluent speaker of Russian, German, Polish and French and has some knowledge of a number of other languages.

Antal van den Bosch (1969) held research positions at the experimental psychology labs of Tilburg University, The Netherlands and the Université Libre de Bruxelles, Belgium (1993–1994), obtained his PhD in computer science at the Universiteit Maastricht, the Netherlands (1994–1997) and held several positions at Tilburg University (1997–2011), where he was appointed full professor in computational linguistics and AI in 2008. In 2011 he took on a full professorship in language and speech technology at Radboud University Nijmegen, the Netherlands. His research interests include memory-based natural language processing and modelling, machine translation, text analytics applied to historical texts and social media, and proofing tools. He is a member of the Netherlands Royal Academy of Arts and Sciences.

Leonoor van der Beek (1978) combines Journalism with Natural Language Processing. With an MA in the first and a PhD in the latter, she went to work in the search industry (Ask.com, RightNow technologies) to improve the search experience by applying natural language processing techniques. In 2011, she published a book on the rise of Natural Language Processing in the Netherlands and Flanders (*Van Rekenmachine tot Taalautomaat*). This book, commissioned by the University of Groningen, is based on interviews with pioneers in the field and gives an insight into the early battles for automation of the translation process. Van der Beek currently works on automated processing of structured and unstructured data for the Dutch e-commerce giant bol.com.

Gerhard B. van Huyssteen completed his PhD in Linguistics in 2000 at the North-West University (NWU) in South Africa. He is currently appointed as research professor at NWU,

and has published more than 50 local and international scholarly articles/papers/chapters in books. Although his core research focuses on Afrikaans morphology, most of his research was done on the development of numerous human language technology (HLT) resources and applications for various South African languages. Van Huyssteen is therefore best known for his contribution as a linguist in the development of real-world text-based computer applications and core computer technologies for Afrikaans and other South African languages. He serves on the expert panel of the South African National Centre for HLT, and was (co-)guest editor of two special issues of journals focusing on HLT in South Africa. Van Huyssteen is also actively involved in various initiatives to improve spelling skills of language users.

William S-Y. Wang is Research Professor at the Chinese University of Hong Kong, based in its Department of Electronic Engineering. He is also Professor Emeritus of the University of California at Berkeley (where he was Professor of Linguistics for 30 years), Honorary Professor at Peking University and at Beijing Language and Culture University. He is Editor of the *Journal of Chinese Linguistics*, which he founded in 1973. In 1992, he was elected President of the International Association of Chinese Linguistics at its formation. In the same year he was elected to the Academia Sinica in Taiwan. In 2013, he was appointed Director of the new Joint Research Centre for Language and Human Complexity at The Chinese University of Hong Kong.

Wang's central interest is in language within an evolutionary perspective. He has published some 200 papers and 10 books in diverse areas of theoretical and applied linguistics. These have appeared in general magazines, such as *American Scientist*, *Nature*, *Proceedings of the National Academy of Sciences (USA)*, *Scientific American*, etc., in specialized journals, such as *Brain and Language*, *Diachronica*, *Language*, *Lingua*, *Language and Cognitive Processes Neuropsychologia*, *Journal of Phonetics*, etc., and in various encyclopedias. His writings have been translated into many languages. He has lectured widely in America, Asia, and Europe.

Kara Warburton is a classically educated terminologist, holding MA and PhD degrees in Terminology (Université Laval, Canada, and City University of Hong Kong). But she also has over 20 years of experience managing terminology in various production-oriented settings, including 15 years at IBM and consulting engagements for a dozen or so enterprises, among them the World Bank and ISO. Having also a BA in Translation (Université Laval) and a BA in Education (Dalhousie University, Canada), she has taught courses at universities in Canada and Hong Kong and provided training to language professionals in over 50 companies. Kara has been actively contributing to international standards and best practices in the field of terminology for over ten years, and currently holds the position of International Chair of ISO Technical Committee 37, which sets standards for terminology and other types of language resources. She offers consultancy services in terminology management through Termologic (www.termologic.com).

Jonathan J. Webster is Professor of the Department of Chinese, Translation and Linguistics; and Director, The Halliday Centre for Intelligent Applications of Language Studies, City University of Hong Kong. He is also the Managing Editor of the International Linguistic Association's journal *WORD*, the Editor of *Linguistics and the Human Sciences* and the forthcoming *Journal of World Languages* (2014).

Cecilia Wong Shuk Man received her PhD in Linguistics from the City University of Hong Kong. She taught at the City University of Hong Kong before joining the Hong Kong

Polytechnic University in 2004. Since then, she has been teaching part-time at The Chinese University of Hong Kong. She has written several research articles on ontology processing and text analysis. Her research interests are computer translation, discourse analysis and computational linguistics.

Billy Wong Tak-ming is a research fellow in the University Research Centre, the Open University of Hong Kong. He received his PhD in Computational Linguistics from City University of Hong Kong. He has taught computer-aided translation and linguistics at City University of Hong Kong, Hong Kong Polytechnic University and the Chinese University of Hong Kong. His research interests include corpus linguistics and evaluation of machine translation, in particular its automated metrics and theoretical studies of the relationship between human and automatic measures.

Sue Ellen Wright is a professor of German and a member of the Kent State University Institute for Applied Linguistics, where she teaches computer applications for translators and German-to-English technical translation. She has served as chair of the American Translators Association (ATA) Terminology Committee and is ATA certified for German-to-English translation. She is active as a terminology trainer and consultant for companies and institutions implementing terminology management in localization environments.

Sue Ellen Wright is engaged in the national and international standards community (ASTM International and the International Organization for Standardization) and chairs the U.S. mirror committee (Technical Advisory Group) for ISO Technical Committee 37, *Terminology and language and content resources*. Together with Professor Gerhard Budin of the University of Vienna she compiled the *Handbook for Terminology Management* and is the author of many articles on applied terminology management in industry. She is chair of the TC 37 Data Category Registry. She was the recipient of the Eugen Wüster Prize awarded by the International Information Centre for Terminology (Infoterm), the Center for Translation Studies (University of Vienna), and the Department of Planned Languages and Esperanto Museum (Austrian National Library) in 2010.

Yu Shiwen is a professor of the Key Laboratory of Computational Linguistics (Peking University), Ministry of Education of PRC and the Institute of Computational Linguistics at Peking University. He has been working in Peking University since he graduated from the Department of Mathematics of Peking University in 1964. His most representative academic achievement is the *Comprehensive Language Knowledge Base* (CLKB) based on the *Grammatical Knowledge-base of Contemporary Chinese*. CLKB has made a great contribution to the development of Chinese information processing and won the second prize of the National Science and Technology Progress Award of PRC in 2011. Professor Yu also won the Lifetime Achievement Award of Chinese Information Processing Society of China in 2011.

Federico Zanettin is Associate Professor of English Language and Translation at the University of Perugia, Italy. His research activity has focused mainly on two areas of translation studies, corpus-based studies and the translation of comic books, on which he has published and lectured widely. His publications include *Translation Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies* (2012), *Corpora in Translator Education* (co-editor, with Silvia Bernardini and Dominic Stewart, 2003), and *Comics in Translation* (editor, 2008).

Jost Zetzsche is an English-to-German translator, a localization and translation consultant, and a widely published author on various aspects of translation. Originally from Hamburg, Germany, he earned a PhD in Sinology with a dissertation on the history of Chinese Bible translation. He writes regular columns in the official publication of the American Translators Association and its British counterpart, the Institute of Translation and Interpreting; his computer guide for translators, *A Translator's Tool Box for the 21st Century*, is now in its tenth edition; and he writes a technical newsletter for translation professionals. In 2012, Penguin published his co-authored *Found in Translation*, a book about translation and interpretation for the general public. Zetzsche has presented workshops, keynotes, and other presentations at many conferences.

Zhang Min, a distinguished professor and director of the Research Institute of Human Language Technology at Soochow University, received his BA and PhD in computer science from Harbin Institute of Technology in 1991 and 1997 respectively. From 1997 to 1999, he worked as a postdoctoral research fellow at the Korean Advanced Institute of Science and Technology in South Korea. He began his academic and industrial career as a researcher at Lernout & Hauspie Asia Pacific (Singapore) in September 1999. He joined Infotalk Technology (Singapore) as a researcher in 2001 and became a senior research manager in 2002. He joined the Institute for Infocomm Research (Singapore) in 2003. He has co-authored more than 150 papers in leading journals and conferences. He was the recipient of several awards in China and overseas, including the '2013 Shuang Chuang Program (1000 Talent Program of Jiangsu Province)', the second prize of the Ministry of Aerospace of China in 1995 and 1997, the 2008/2009 research achiever of the Institute for Infocomm Research and the 2002 National Infocomm Award (group) at Singapore. He is the Vice President of COLIPS (2011–2013), a steering committee member of PACLIC (2011–now), an executive member of AFNLP (2013−2014) and a member of ACL (2006–).

Zhang Xiaojun is a postdoctoral researcher at CNGL (Centre for Global Intelligent Content) at Dublin City University in Ireland. He is also an associate professor at Shaanxi Normal University in China. He has over 60 publications in translation, linguistics and natural language processing, is the co-translator of the Chinese version of the book *Statistical Machine Translation* written by Philipp Koehn and a member of the Association of Computational Linguistics (ACL) since 2010.

Zhang Yihua is a professor of linguistics and applied linguistics and the director of the Center for Lexicographical Studies in Guangdong University of Foreign Studies, and concurrently the vice-president of the China Association for Lexicography and chairman of Chinalex Bilingual Committee. Since 1998 he has authored numerous publications in lexicography, including 105 academic papers, 9 academic works, 2 translation works and 9 dictionaries. Among these, *English-Chinese Medical Dictionary* won the First Prize of the Fifth National Dictionary Award and *Illustrated English-Chinese Dictionary for Primary School Learners* won the Second Prize of the Fifth National Dictionary Award; and his monographs entitled *Semantics and Lexicographical Definition*, *Computational Lexicography*, *Contemporary Lexicography*, *Meaning, Cognition and Definition,* as well as a number of published papers have made a great impact on the teaching and research of lexicography in China. His book, *Contemporary Lexicography*, won The Fourth National Excellent Achievement Award of Humanities and Social Science Research Achievement in Universities.

PREFACE

Chan Sin-wai

Introduction

In recent decades, as a result of the rapid advances in computer science and other related disciplines, such as computational linguistics and terminology studies, translation technology has become a norm in translation practice, an important part of translation studies, a new paradigm of translation pedagogy, and a major trend in the industry. It is generally recognized that translation technology has become popular both in Asia and in the West. It is widely used by translation companies as an indispensable tool to conduct their business with high productivity and efficiency, by international corporations as a foundation for their global language solutions, by professional translators as a core component of their personal workstations, and by occasional users as an important means of multilingual information mining. The advent of translation technology has totally globalized translation and drastically changed the way we process, teach, and study translation. Translation technology has, in short, brought fundamental changes and additional dimensions to all aspects of the contemporary world of translation. And yet there is a total lack of encyclopedic references for such an emerging and important area, apart from *A Dictionary of Translation Technology*, which I authored, and published in Hong Kong some nine years ago (Chan 2004). The time has really come for us to sum up what has been done so far and what needs to be done in the future through the publication of the first encyclopedia on this important subject.

Definition of translation technology

The scope of this encyclopedia covers as far as possible all the concepts in the field and all the changes that translation technology has brought to it. This scope determines the way we define translation technology. According to Lynne Bowker, translation technology refers to different types of technology used in human translation, machine translation, and computer-aided translation, covering the general tools used in computing, such as word processors and electronic resources, and the specific tools used in translating, such as corpus-analysis tools and terminology management systems (Bowker 2002: 5–9). A broader definition is given in *A Dictionary of Translation Technology*, which describes translation technology as 'a branch of translation studies that specializes in the issues and skills related to the computerization of translation' (Chan 2004: 258). This means that translation technology is inclusive of both

computer-aided translation and machine translation. As machine translation serves basically as an aid to human translation without human intervention, it is considered to be a form of computer-aided translation. In this encyclopedia, translation technology covers both computer-aided translation and machine translation.

Aims of the *Encyclopedia*

The main purpose of preparing this *Routledge Encyclopedia of Translation Technology*, as mentioned above, is to produce a comprehensive reference for scholars and specialists engaged in the study of computer-aided translation and machine translation and for general readers who are interested in knowing, learning and using new concepts and skills of translation technology in translation practice. To meet the aspirations of these two groups of users, the contents of all the chapters in this encyclopedia are both academic and general, and brief and self-contained, depending on the nature of the topic. Useful references and resources are given in each chapter to show the scholarship that has been attained in relevant areas and what essential works are available for readers to delve deeper into the areas they are interested in.

To achieve the above purposes, we have invited 49 scholars and specialists working at academic institutions or private organizations in different parts of the world to contribute chapters to this encyclopedia. Their national or regional affiliations include, in alphabetical order, Australia, Canada, China, France, Germany, Hong Kong, Ireland, Italy, Japan, the Netherlands, Singapore, South Africa, Spain, Sweden, Taiwan, the United Kingdom, and the United States. Chapters by scholars in these countries are in general comprehensive, informative, prospective, and well documented. As the first definitive encyclopedia of translation technology, this book aims to serve as an authoritative reference to scholars and students of both machine translation and computer-aided translation and lay a solid foundation for translation technology's rapid growth in the future.

Coverage of this *Encyclopedia*

This encyclopedia is structured in such a way as to facilitate the understanding of translation technology in all its aspects. It is divided into three parts: Part I, 'General Issues in Translation Technology', covers the general issues relating to both computer-aided translation and machine translation; Part II, 'The National/Regional Developments of Translation Technology' contains chapters on the history and growth of translation technology in countries and regions where this technology is researched, developed, and widely used; and Part III, 'Specific Topics in Translation Technology', has 18 chapters on the various aspects of machine translation and computer-aided translation, including topics such as alignment, concordancing, localization, online translation, and translation memory.

Part 1: General issues in translation technology

Part 1 has four chapters on computer-aided translation and nine chapters on machine translation. The first four chapters cover the general issues relating to the history, major concepts, major systems and translator training of computer-aided translation. This Part begins with a history of computer-aided translation in the last five decades by Chan Sin-wai of the Chinese University of Hong Kong, who divides the entire history of translation technology (1967–2013) into four periods: the period of germination (1967–1983), the period of steady growth (1984–1992), the period of rapid growth (1993–2003), and the period of global development (2004–2013). The

second chapter, by the same author, is about the seven major concepts in computer-aided translation that shape the development of functions in translation technology. These concepts are: simulativity, emulativity, customizability, compatibility, controllability, productivity and collaborativity. The third chapter in this part is by Ignacio Garcia of Sydney University in Australia, who writes on computer-aided translation systems. He is of the view that computer-aided translation systems are software applications aimed at increasing translators' productivity while maintaining an acceptable level of quality. The use of these systems, restricted to technical translation in the 1990s, has extended now to most types of translation, and most translators, including non-professionals, could benefit from using them. The fourth chapter is by Lynne Bowker of Montreal University, who discusses translator training in the context of computer-aided translation. She believes that just as translation has been affected by the use of computers, so too has the way in which translators are trained. Her chapter explores questions such as which types of tools are relevant for translators, what do translators need to learn about technologies, who should be responsible for teaching translators about computer aids, and when should technologies be introduced into the curriculum. The answers are not always clear cut, but solutions and best practices are emerging.

The second half of Part I has chapters on various general aspects of machine translation, including its general aspects, history, systems, evaluation criteria, approaches and teaching. It begins with a chapter by Liu Qun and Zhang Xiaojun, both of Dublin City University in Ireland. Their chapter introduces the technology of machine translation (MT), also known as automatic translation. It defines machine translation, outlines its history, and describes its various approaches, evaluation methods and applications. The second chapter is by W. John Hutchins, formerly of the University of East Anglia. He provides a history of machine translation from the 'pioneering' research and the early operational systems (1950s and 1960s) to the dominance of rule-based systems (1967 to 1989) and finally to the emergence of corpus-based systems (in particular statistical approaches), translation memories, evaluation methods and current applications.

The next five chapters cover five major approaches to machine translation: example-based machine translation, open-source machine translation, pragmatics-based machine translation, rule-based machine translation and statistical machine translation. Billy Wong Tak-ming of the Open University of Hong Kong and Jonathan Webster of the City University of Hong Kong collaborated on the writing of the chapter on example-based machine translation. This chapter presents an overview of example-based machine translation (EBMT), covering its history, the major issues related to translation examples, and the fundamental stages of translation for an EBMT system. The suitability issue is also discussed, showing the types of translation that are deemed suitable for EBMT, and how it interoperates with other MT approaches. The second chapter on machine translation is by Mikel L. Forcada of the Universitat d'Alacant in Spain. It defines free/open-source (FOS) software and reviews its licensing and implications for machine translation (MT) as data-intensive software, and the types of free/open-source MT systems and users and their use in business and research. It also surveys the existing free/open-source MT systems and looks into the challenges their systems face in the future. The third chapter in this Part is on pragmatics-based machine translation by David Farwell, formerly of the Catalonia Polytechnic University in Barcelona, Spain, and Stephen Helmreich of the New Mexico State University in the United States. They hold the view that pragmatics-based machine translation relies on reasoning to determine speech act content and on beliefs ascription for modelling the participants in the translation process (source text author and audience, translator, intended audience of translation). The theoretical framework and computational platform are presented along with an analysis of their benefits and shortcomings. Rule-based machine translation is the

subject of the fourth chapter in this part, written by Yu Shiwen of Peking University and Bai Xiaojing of Tsinghua University in China. According to the authors, the rule-based method had been dominant in machine translation research for several decades, and it is still functioning in present-day MT systems. Despite its difficulties and problems, this method is now gaining a new momentum, as the significance of linguistic research has been realized more than ever and the formalization of linguistic knowledge is growing and maturing. The fifth chapter in this part is on statistical machine translation, which is currently the most popular approach in the field. According to the authors of this chapter, Liu Yang of Tsinghua University and Zhang Min of Soochow University in China, statistical machine translation (SMT) is a machine translation paradigm that generates translations based on a probabilistic model of the translation process, the parameters of which are estimated from parallel text. Modelling, training and decoding are three fundamental issues in SMT, which has evolved from early word-based approaches to recent phrase-based and syntax-based approaches in the past decades. As a data-driven technology, it will continue to develop in the era of big data.

The last two chapters on machine translation are on its evaluation and teaching. Kit Chunyu of the City University of Hong Kong and Billy Wong Tak-ming introduce the key issues and basic principles of computer(-aided) translation evaluation, covering its historical evolution, context-dependent multi-dimensional nature, and existing methodologies. The major evaluation approaches, including both manual and automatic, are presented with a full discussion of their strengths and weaknesses. Cecilia Wong Shuk Man, who has taught machine translation at the Chinese University of Hong Kong for a number of years, recounts her experience in teaching machine translation at the MA in Computer-aided Translation Programme of the Chinese University.

Part 2: The national/regional developments of translation technology

The importance of studying translation technology from a global perspective cannot be overemphasized. Part 2 of this encyclopedia contains chapters that describe the development of computer-aided translation and machine translation in some of the countries and regions where translation technology is studied, developed and used. A number of countries in different regions have been selected to illustrate the development and application of translation technology in different social and cultural situations and at different levels of technological advancement. These countries and regions include Belgium, France, the Netherlands and the United Kingdom in Europe; China, Hong Kong, Japan, Singapore and Taiwan in Asia; South Africa in Africa; and Canada and the United States in North America.

The first chapter, on China, is written by Qian Duoxiu of Beihang University in Beijing, China. Her chapter outlines the growth of translation technology in China from 1946 to the present, covering its major participants, achievements, applications, mainstream tools and prospects. The second chapter is by Elliott Macklovitch, former president of the Association for Machine Translation in the Americas. His contribution traces the evolution of translation technology in Canada, from the emergence of the first computerized aids (dedicated word processors), through the well-known success of the MÉTÉO system, to the development of innovative translator support tools like *TransType*. It also assesses the current use of cutting-edge technologies such as statistical MT and automatic dictation for the production of high-quality translation. Sylviane Cardey of the University of Franche-Comté in France writes on translation technology in France. According to the author, machine translation in France started at the time of the Cold War. She traces the development of machine translation in France from the 1950s to the present, focusing, in the latter part of her chapter, on six research centres and companies,

the technologies they used and the systems they produced. Two approaches to MT in France clearly stand out, one being based on linguistic methods and the other on statistical methods.

Chan Sin-wai, Ian Chow, and Billy Wong Tak-ming jointly authored the chapter on translation technology in Hong Kong, focusing on the research projects, course offerings, research centres at the local tertiary institutions and the use of translation technology in Hong Kong's translation industry. The situation of Japan is described by Hitoshi Isahara of Toyohashi University of Technology in Japan. He gives a historical overview of research and development of machine translation systems in Japan, and then goes on to describe one of the latest government-funded MT projects, research activities related to pre- and post-editing, the development of linguistic resources for MT systems, and research on evaluation of MT systems. The only chapter on Africa is by Gerhard van Huyssteen and Marissa Griesel, both of North-West University in South Africa. Their chapter centres on the development of translation technology in South Africa. From Africa we turn back to Asia and introduce the situation of translation technology in Taiwan, written by Shih Chung-ling of the National Kaohsiung First University of Science and Technology. This chapter describes the findings regarding translation technological (TT) development and the use of translation technology in Taiwan's translation industry, university education and academic research. The author also makes some suggestions to address the inadequacy in the use of translation technology in industry and academia through on-the-job training, joint lectures and regular conferences, and highlights the need to incorporate elements of translation technology in translation research in response to the changing situation of the field of translation.

Leonoor van der Beek of RightNow Technologies and Antal van den Bosch of Radboud University, Nijmegen in the Netherlands write on translation technology in the Netherlands and Belgium. Their chapter highlights the development of the Eurotra, METAL, DLT, and Rosetta systems and examines the current state of translation in Dutch. According to the authors, researchers from Belgium and the Netherlands participated in large national and international projects in the 1980s. Disappointing results led to a decade of silence, but in the 2000s new research projects embraced statistical and example-based machine translation. Dutch, with an estimated 25 million native speakers, is a source or target language in about 15 per cent of current translation technology products. The chapter on translation technology in the United Kingdom is by Christophe Declercq of London University. He examines this topic under the headings of peculiar relations, education, devolution, and translation technology companies. And the last chapter is on translation technology in the United States, jointly written by Jennifer DeCamp, the Chief Scientist for Human Language Technology in MITRE Corporation, and Jost Zetzsche, a German-American translator, Sinologist and writer who lives in Oregon. To them, while the history of translation technology development in the United States has been highly international, there are certain unique features in the country that differentiate it from its development elsewhere. These include the extensive investment by the US military, the wide gulf between machine translation researchers and human translators, and the extensive involvement of religious groups in the development and use of translation technology. They provide a list of major events in the history of translation technology in the United States, highlighting the special features in each decade.

Part 3: Specific topics in translation technology

Whereas topics in the first two parts are on the whole of a more general nature, the 18 topics in Part 3, written by 21 scholars, are more specific to translation technology. These chapters have been arranged in alphabetical order for easy reference, beginning with Alignment and

ending with Translation Management Systems. Alignment, the first chapter of this part, is written by Lars Ahrenberg of Linköping University, Sweden. His chapter covers the main algorithms and systems for sentence alignment and word alignment. The focus is on statistical properties of bi-texts and the way these properties are exploited for alignment in generative as well as discriminative and heuristic models. In addition, this chapter provides an overview of standard evaluation metrics for alignment performance, such as precision, recall and Alignment Error Rate. Bi-text is the topic discussed by Alan Melby, Yves Savourel and Lucia Morado Vázquez. Zhang Yihua of the Guangdong University of Foreign Studies, writes on computational lexicography, an area which is closely related to translation technology. Computational lexicography, according to the author, has gone through decades of development, and great achievements have been obtained in building and using corpora, which contribute enormously to the development of lexicographical databases and computer-aided dictionary writing and publishing systems. Computer lexicography is also closely associated with machine translation as all MT systems have electronic dictionaries. The topic of concordancing is covered by Federico Zanettin of the University of Perugia, Italy. This chapter provides an historical overview of concordances and concordancers, describes how different types of corpus resources and tools can be integrated into a computer-assisted translation environment, and examines a set of parameters, including data search and display options, which may be used to evaluate concordancing applications. Controlled languages, a topic of considerable interest to translation technologists, are described by Rolf Schwitter of Macquarie University in Australia. Controlled languages are subsets of natural languages which use a restricted vocabulary and grammar in order to reduce or eliminate ambiguity and complexity. Some of these controlled languages are designed to improve communication between humans. Some of them make it easier for non-native speakers to read technical documentation. Some aim to improve the quality of machine translation, and another group of controlled languages serve as high-level interface languages to semantic systems where automated reasoning is important.

It is generally recognized that corpus is important both in lexicography and translation technology. The chapter on corpus is written by Li Lan of Hong Kong Polytechnic University. Her chapter introduces the important advances in corpus-based translation studies, presents detailed information on standard monolingual and bilingual corpora, and argues that both can help translators to establish equivalence, terminology and phraseology between languages. In addition, corpus-based quantitative and qualitative methods can help to verify, refine or clarify translation theories. The topic of editing in translation technology is authored by Christophe Declercq, who also writes a chapter on translation technology in the United Kingdom in Part 2. He covers a number of areas in this topic, including language and translation technology, 'traditional' translation technology and editing, cognitive processes and editing, forms of editing, revision and proof-reading, post-editing and machine translation, and post-editing guidelines. The topic of information retrieval and text mining is covered in the chapter co-authored by Kit Chunyu of the City University of Hong Kong and Nie Jianyun of the University of Montreal in Canada. They discuss the main operations in information retrieval and issues in text mining. Sue Ellen Wright, of Kent State University in the United States, explores the issues of language codes in the next chapter of this book. Her colleague at the same university, Keiran J. Dunne, writes on the topic of localization, which covers most of the essential points of this subject. Olivia Kwong Oi Yee of the City University of Hong Kong contributes a chapter on natural language processing. According to her, the primary concern of natural language processing is the design and implementation of computational systems for analysing and understanding human languages to automate certain real-life tasks demanding

human language abilities. It is typically a multidisciplinary endeavour, drawing on linguistics, computer science, mathematics and psychology amongst others, with a particular focus on computational models and algorithms at its core. The chapter on online translation is written by Federico Gaspari of the University of Bologna, Italy. It concerns key aspects of online translation, focusing on the relationship between translators and the Web, with a review of the latest trends in this area. A wide range of Internet-based resources, tools and services for translators are presented, highlighting their key features and discussing their pros and cons.

Felipe Sánchez-Martínez of the Universitat d'Alacant in Spain writes on part-of-speech tagging. Part-of-speech tagging is a well-known problem and a common step in natural language processing applications; part-of-speech taggers try to assign the correct part of speech to all words of a given text. This chapter reviews the main approaches to part-of-speech tagging and their use in machine translation. Segmentation is a topic discussed by Freddy Choi. His chapter introduces text segmentation, covers all the elements that make up a working algorithm, key considerations in a practical implementation, and the impact of design decisions on the performance of a complete machine translation solution. The narrative offers a survey of existing design options and recommendations for advancing the state-of-the-art and managing current limitations. Lee Tan of the Chinese University of Hong Kong writes on speech translation. According to the author, speech translation is an advanced computer-based technology that enables speech communication between people who speak different languages. A speech translation system is an integration of speech recognition, machine translation and speech synthesis. The latest systems are available as smartphone applications. They can perform translation of naturally spoken sentences and support multiple languages. Jorge Díaz Cintas of University College London writes on subtitling and technology. His chapter highlights some of the most significant technological milestones that have been reached in the field of subtitling and considers more recent developments in this arena, such as machine translation and cloud subtitling. Kara Warburton, an experienced terminologist residing in Hong Kong, writes on terminology management. Her chapter provides an introduction to Terminology as a field of applied linguistics and as a strategic pursuit in information technology. It covers relations to lexicology, basic concepts and principal theories, methods and workflows for managing terminologies, uses of terminology, connections with corpora, terminology databases, and standards and best practices. Alan Melby and Sue Ellen Wright discuss translation memory and the computer-aided translation in the translation environment tools, sub-segment identification, advantages of a translation memory, how to create, use, and maintain a translation memory, history of translation memory, and the future developments and industry impact of translation memory. The last chapter in this volume is on translation management systems, written by Mark Shuttleworth of University College London. He traces the history of translation management, studies its common features, and estimates the future of technology in the field of translation. Computerized translation management systems have been in existence since the late 1990s. They were introduced in order to enable translation companies and individual translators to remain in control of ever-increasing volumes of content and to facilitate the monitoring of business, process and language aspects of translation and localization projects.

Conclusion

With five leading scholars in the field serving as Consultant Editors and around 50 eminent specialists contributing their chapters to this volume, this encyclopedia, the first of its kind, is a valuable and definitive reference in the field of translation technology. It is hoped that

specialists and general readers will find this encyclopedia informative and useful, while professionals will find the knowledge they gain from this volume helpful in translation practice.

References

Bowker, Lynne (2002) *Computer-aided Translation Technology: A Practical Introduction*, Ottawa: University of Ottawa Press.

Chan, Sin-wai (2004) *A Dictionary of Translation Technology*, Hong Kong: The Chinese University Press.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Andrea Hartill, Publisher at Routledge, for giving me an opportunity to fulfill my wish to edit the first encyclopedia of translation technology. Without her support and encouragement, this volume would not be able to see the light of day. Isabelle Cheng, who is in charge of the project, is the nicest and most helpful editor that I have ever worked with. Their contribution to the publication of this volume is invaluable and unsurpassed.

My most sincere gratitude goes to Professor David Pollard, former Professor of Translation and Chairman of the Department of Translation of The Chinese University of Hong Kong. It was through David's recommendation that resulted in the publication of this volume.

My thanks are due to all the five Consultant Editors of this encyclopedia, namely, Professor Lynne Bowker, Professor David Farwell, Dr. W. John Hutchins, Professor Alan K. Melby, and Professor William S-Y. Wang, whose support to this volume is indispensable in its completion. Alan Melby, in particular, deserves a special note of thank not only for his contribution of two chapters, but also for recommending prominent scholars in the field to contribute chapters to this volume.

Last but not least, I would like to thank Miss Florence Li Wing Yee, my colleague at the Department of Translation of The Chinese University of Hong Kong, for her dedicated and tireless efforts to prepare this volume for publication, and Miss Sara Román Galdrán for preparing the index.

Chan Sin-wai

The Editor and Publisher would like to thank the following copyright holders for permission to use material featured in the following chapters.

Figure "Dashboard of SDL-Trados 2014", featured in Chan, Sin-wai, *Computer-aided Translation: Major Concepts*. Used with kind permission of SDL PLC.

Figures "A KWIC concordance of the word 'translation' (from the Sketch Engine)" and "A search for the lemma have immediately followed by a verb (from the Sketch Engine)", featured in Zanettin, Federico, *Concordancing*. Used with kind permission of Lexical Computing Ltd.

Figure "Left-side concordance tree of the word 'translation' (from Luz's TEC browser)", featured in Zanettin, Federico, *Concordancing*. Used with kind permission of Saturnino Luz.

Figure "'verb + one's way + preposition' constructions in the 155 billion word Google Books Corpus of American English (from Davies' corpus.byu.edu)", featured in Zanettin, Federico, *Concordancing*. Used with kind permission of Mark Davies.

Figure "Parallel concordance ordered according to target language (from Barlow's ParaConc)", featured in Zanettin, Federico, *Concordancing*. Used with kind permission of Michael Barlow.

Screenshots "Detail of the Editor Environment of SDL Trados Studio 2011 (SP1), with 3+1+4+1 Units" and "Overview of SDL Products and Segment status in SDL Trados Studio 2011", featured in Declercq, Christophe, *Editing in Translation Technology*. Copyright © 2014 SDL plc. All Rights Reserved

Tables and/or figures "Editing Stages in an Overall Quality Assurance Approach (Makoushina and Kockaert 2008: 3)" ," How EN15038 could possibly set editing apart from review, revision and proof-reading", "Light and Full Post-editing of Raw MT Output" (O'Brien 2010: 5) and "TAUS Post-editing Guidelines versus Quality Assurance in SDL Trados Studio 2011", featured in Declercq, Christophe, *Editing in Translation Technology*. Used with kind permission of EUATC.

Screenshot "Various translation workflows possible in XTM Cloud", featured in Declercq, Christophe, *Editing in Translation Technology*. Used with kind permission of XTM International Ltd.

Figure "Detail of the Editor Environment of SDL Trados Studio 2011 (SP1)", featured in Declercq, Christophe, *Editing in Translation Technology*. Source text used by kind permission of Golazo media.

Figure "Anatomy of a Unicode Locale ID", featured in Wright, Sue Ellen, *Language Codes and Language Tags*. Image © S.R. Loomis and M. Davis, 2010. Used with kind permission.

Figure "Sub-languages drop-down menu, MultiTerm™ 2011", featured in Wright, Sue Ellen, *Language Codes and Language Tags*. Used with kind permission of SDL PLC.

Figure "Localization of a Sample Application Named Scribble Using a Visual Localization Tool. The Left-hand Pane Displays the Resource Tree, the Middle Pane Displays the Selected Resource in WYSIWYG Mode, and the Right-hand Pane Displays the Corresponding Source and Target Strings in Tabular Format", featured in Dunne, Keiran, *Localization*. Used with kind permission of SDL PLC.

Figure "Interface of the Professional Subtitling Program WinCAPS Qu4antum", featured in Díaz Cintas, Jorge, *Technological Strides in Subtitling*. Used with kind permission of Screen Systems.

Figure "A Workflow for Prescriptive Terminology", featured in Warburton, Kara, *Terminology Management*. Created for the Starter Guide for Terminology SIG by TerminOrgs. Used with kind permission of TerminOrgs http://www.terminorgs.net.

Figure "Side-by-side Segment Display", featured in Melby, Alan K and Wright, Sue Ellen, *Translation Memory*. Used with kind permission of SDL PLC.

Figure "Specifying Client and Subject Area in Déjà Vu X2 Workgroup", featured in Shuttleworth, Mark *Translation Management Systems*. Used with kind permission of Atril.

Figure "Selecting a Workflow in XTM 7.0", featured in Shuttleworth, Mark, *Translation Management Systems*. Used with kind permission of XTM International.

Figure "Screenshot of the OTM 5.6.6 iPhone interface (image taken from http://www.lsp.net/otm-mobile-devices.html)", featured in Shuttleworth, Mark, *Translation Management Systems*. Used with kind permission of OTM.

While we have made every effort to contact copyright holders of material used in this volume, we would be grateful to hear from any we were unable to reach.

This page intentionally left blank

PART I

General issues of translation technology

This page intentionally left blank

1

THE DEVELOPMENT OF TRANSLATION TECHNOLOGY

1967–2013

Chan Sin-wai

THE CHINESE UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Introduction

The history of translation technology, or more specifically computer-aided translation, is short, but its development is fast. It is generally recognized that the failure of machine translation in the 1960s led to the emergence of computer-aided translation. The development of computer-aided translation from its beginning in 1967 as a result of the infamous ALPAC report (1966) to 2013, totalling 46 years, can be divided into four periods. The first period, which goes from 1967 to 1983, is a period of germination. The second period, covering the years between 1984 and 1993, is a period of steady growth. The third period, which is from 1993 to 2003, is a decade of rapid growth. The last period, which includes the years from 2004 to 2013, is a period of global development.

1967–1983: A period of germination

Computer-aided translation, as mentioned above, came from machine translation, while machine translation resulted from the invention of computers. Machine translation had made considerable progress in a number of countries from the time the first computer, ENIAC, was invented in 1946. Several events before the ALPAC report in 1966 are worth noting. In 1947, one year after the invention of the computer, Warren Weaver, President of the Rockefeller Foundation and Andrew D. Booth of Birkbeck College, London University, were the first two scholars who proposed to make use of the newly invented computer to translate natural languages (Chan 2004: 290–291). In 1949, Warren Weaver wrote a memorandum for peer review outlining the prospects of machine translation, known in history as 'Weaver's Memorandum'. In 1952, Yehoshua Bar-Hillel held the first conference on machine translation at the Massachusetts Institute of Technology, and some of the papers were compiled by William N. Locke and Andrew D. Booth into an anthology entitled *Machine Translation of Languages: Fourteen Essays*, the first book on machine translation (Locke and Booth 1955). In 1954, Leon Dostert of Georgetown University and Peter Sheridan of IBM used the IBM701 machine to make a public demonstration of the translation of Russian sentences into English, which marked a milestone in machine translation (Hutchins 1999: 1–16; Chan 2004:

125–226). In the same year, the inaugural issue of *Mechanical Translation*, the first journal in the field of machine translation, was published by the Massachusetts Institute of Technology (Yngve 2000: 50–51). In 1962, the Association for Computational Linguistics was founded in the United States, and the journal of the association, *Computational Linguistics*, was also published. It was roughly estimated that by 1965, there were eighteen countries or research institutions engaged in the studies on machine translation, including the United States, former Soviet Union, the United Kingdom, Japan, France, West Germany, Italy, former Czechoslovakia, former Yugoslavia, East Germany, Mexico, Hungary, Canada, Holland, Romania, and Belgium (Zhang 2006: 30–34).

The development of machine translation in the United States since the late 1940s, however, fell short of expectations. In 1963, the Georgetown machine translation project was terminated, which signifies the end of the largest machine translation project in the United States (Chan 2004: 303). In 1964, the government of the United States set up the Automatic Language Processing Advisory Committee (ALPAC) comprising seven experts to enquire into the state of machine translation (ALPAC 1966; Warwick 1987: 22–37). In 1966, the report of the Committee, entitled *Languages and Machines: Computers in Translation and Linguistics*, pointed out that 'there is no immediate or predictable prospect of useful machine translation' (ALPAC 1966: 32). As machine translation was twice as expensive as human translation, it was unable to meet people's expectations, and the Committee recommended that resources to support machine translation be terminated. Its report also mentioned that 'as it becomes increasingly evident that fully automatic high-quality machine translation was not going to be realized for a long time, interest began to be shown in machine-aided translation' (ibid.: 25). It added that machine translation should shift to machine-aided translation, which was 'aimed at improved human translation, with an appropriate use of machine aids' (ibid.: iii), and that 'machine-aided translation may be an important avenue toward better, quicker, and cheaper translation' (ibid.: 32). The ALPAC report dealt a serious blow to machine translation in the United States, which was to remain stagnant for more than a decade, and it also made a negative impact on the research on machine translation in Europe and Russia. But this gave an opportunity to machine-aided translation to come into being. All these events show that the birth of machine-aided translation is closely related to the development of machine translation.

Computer-aided translation, nevertheless, would not be possible without the support of related concepts and software. It was no mere coincidence that translation memory, which is one of the major concepts and functions of computer-aided translation, came out during this period. According to W. John Hutchins, the concept of translation memory can be traced to the period from the 1960s to the 1980s (Hutchins 1998: 287–307). In 1978, when Alan Melby of the Translation Research Group of Brigham Young University conducted research on machine translation and developed an interactive translation system, ALPS (Automated Language Processing Systems), he incorporated the idea of translation memory into a tool called 'Repetitions Processing', which aimed at finding matched strings (Melby 1978; Melby and Warner 1995: 187). In the following year, Peter Arthern, in his paper on the issue of whether machine translation should be used in a conference organized by the European Commission, proposed the method of 'translation by text-retrieval' (Arthern 1979: 93). According to Arthern,

> This information would have to be stored in such a way that any given portion of text in any of the languages involved can be located immediately … together with its translation into any or all of the other languages which the organization employs.
>
> (Arthern 1979: 95)

In October 1980, Martin Kay published an article, entitled 'The Proper Place of Men and Machines in Language Translation', at the Palo Alto Research Center of Xerox. He proposed to create a machine translation system in which the display on the screen is divided into two windows. The text to be translated appears in the upper window and the translation would be composed in the bottom one to allow the translator to edit the translation with the help of simple facilities peculiar to translation, such as aids for word selection and dictionary consultation, which are labelled by Kay as a *translator amanuensis* (Kay 1980: 9–18). In view of the level of word-processing capacities at that time, his proposal was inspiring to the development of computer-aided translation and exerted a huge impact on its research later on. Kay is generally considered a forerunner in proposing an interactive translation system.

It can be seen that the idea of translation memory was established in the late 1970s and the 1980s. Hutchins believed that the first person to propose the concept of translation memory was Arthern. As Melby and Arthern proposed the idea almost at the same time, both could be considered as forerunners. And it should be acknowledged that Arthern, Melby, and Kay made a great contribution to the growth of computer-aided translation in its early days.

The first attempt to deploy the idea of translation memory in a machine translation system was made by Alan Melby and his co-researchers at Brigham Young University, who jointly developed the Automated Language Processing Systems, or ALPS for short. This system provided access to previously translated segments which were identical (Hutchins 1998: 291). Some scholars classify this type of full match a function of the first generation translation memory systems (Gotti *et al.* 2005: 26–30; Kavak 2009; Elita and Gavrila 2006: 24–26). One of the major shortcomings of this generation of computer-aided translation systems is that sentences with full matching were very small in number, minimizing the reusability of translation memory and the role of translation memory database (Wang 2011: 141).

Some researchers around 1980 began to collect and store translation samples with the intention of redeploying and sharing their translation resources. Constrained by the limitations of computer hardware (such as its limited storage space), the cost of building a bilingual database was high, and with the immaturity in the algorithms for bilingual data alignment, translation memory technology had been in a stage of exploration. As a result, a truly commercial computer-aided translation system did not emerge during the sixteen years of this period and translation technology failed to make an impact on translation practice and the translation industry.

1984–1992: A period of steady growth

The eight years between 1984 and 1992 are a period of steady growth for computer-aided translation and for some developments to take place. Corporate operation began in 1984, system commercialization, in 1988, and regional expansion, in 1992.

Company operation

It was during this period that the first computer-aided translation companies, Trados in Germany and Star Group in Switzerland, were founded in 1984. These two companies later had a great impact on the development of computer-aided translation.

The German company was founded by Jochen Hummel and Iko Knyphausen in Stuttgart, Germany, in 1984. Trados GmbH came from TRAnslation and DOcumentation Software. This company was set up initially as a language service provider (LSP) to work on a translation project they received from IBM in the same year. As the company later developed computer-

aided translation software to help complete the project, the establishment of Trados GmbH is regarded as the starting point of the period of steady growth in computer-aided translation (Garcia and Stevenson 2005: 18–31; http://www.lspzone.com).

Of equal significance was the founding of the Swiss company STAR AG in the same year. STAR, an acronym of Software, Translation, Artwork, and Recording, provided manual technical editing and translation with information technology and automation. Two years later, STAR opened its first foreign office in Germany in order to serve the increasingly important software localization market and later developed STAR software products, GRIPS and Transit for information management and translation memory respectfully. At the same time, client demand and growing export markets led to the establishment of additional overseas locations in Japan and China. The STAR Group still plays an important role in the translation technology industry (http://www.star-group.net).

It can be observed that during this early period of computer-aided translation, all companies in the field were either established or operated in Europe. This Eurocentric phenomenon was going to change in the next period.

System commercialization

The commercialization of computer-aided translation systems began in 1988, when Eiichiro Sumita and Yutaka Tsutsumi of the Japanese branch of IBM released the ETOC (Easy to Consult) tool, which was actually an upgraded electronic dictionary. Consultation of a traditional electronic dictionary was by individual words. It could not search phrases or sentences with more than two words. ETOC offered a flexible solution. When inputting a sentence to be searched into ETOC, the system would try to extract it from its dictionary. If no matches were found, the system would make a grammatical analysis of the sentence, taking away some substantive words but keeping the form words and adjectives which formed the sentence pattern. The sentence pattern would be compared with bilingual sentences in the dictionary database to find sentences with a similar pattern, which would be displayed for the translator to select. The translator could then copy and paste the sentence onto the Editor and revise the sentence to complete the translation. Though the system did not use the term translation memory and the translation database was still called a 'dictionary', it nevertheless had essentially the basic features of translation memory of today. The main shortcoming of this system is that as it needs to make grammatical analyses, its programming would be difficult and its scalability would be limited. If a new language were to be added, a grammatical analysis module would have to be programmed for the language. Furthermore, as the system could only work on perfect matching but not fuzzy matching, it drastically cut down on the reusability of translations (Sumita and Tsutsumi 1988: 2).

In 1988, Trados developed TED, a plug-in for text processor tool that was later to become, in expanded form, the first Translator's Workbench editor, developed by two people and their secretary (Garcia and Stevenson 2005: 18–31). It was around this time that Trados made the decision to split the company, passing the translation services part of the business to INK in the Netherlands, so that they could concentrate on developing translation software (http://www.translationzone.com).

Two years later, the company also released the first version of MultiTerm as a memory-resident multilingual terminology management tool for DOS, taking the innovative approach of storing all data in a single, freely structured database with entries classified by user-defined attributes (Eurolux Computers 1992: 8; http://www.translationzone.com; Wassmer 2011).

In 1991 STAR AG also released worldwide the Transit 1.0 ('Transit' was derived from the phrase 'translate it') 32-bit DOS version, which had been under development since 1987 and used exclusively for in-house production. Transit featured the modules that are standard features of today's CAT systems, such as a proprietary translation editor with separate but synchronized windows for source and target language and tag protection, a translation memory engine, a terminology management component and project management features. In the context of system development, the ideas of terminology management and project management began with Transit 1.0. Additional products were later developed for the implementation and automation of corporate product communications: TermStar, WebTerm, GRIPS, MindReader, SPIDER and STAR James (http://www.star-group.net).

One of the most important events in this period is obviously the release of the first commercial system, Trados, in 1992, which marks the beginning of commercial computer-aided translation systems.

Regional expansion

The year 1992 also marks the beginning of the regional expansion of computer-aided translation. This year witnessed some significant advances in translation software made in different countries. First, in Germany, Translator's Workbench I and Translator's Workbench II (DOS version of Trados) were launched within the year, with Workbench II being a standalone package with an integrated editor. Translator's Workbench II comprises the TW II Editor (formally TED) and MultiTerm 2. Translator's Workbench II was the first system to incorporate a 'translation memory' and alignment facilities into its workstation. Also of considerable significance was the creation by Matthias Heyn of Trados's T Align, later known as WinAlign, the first alignment tool on the market. In addition, Trados began to open a network of global offices, including Brussels, Virginia, the United Kingdom and Switzerland (Brace 1994; Eurolux Computers 1992; http://www.translationzone.com; Hutchins 1998: 287–307).

Second, in the United States, IBM launched its IBM Translation Manager / 2 (TM/2), with an Operating System/2 (OS/2) package that integrated a variety of translation aids within a Presentation Manager interface. TM/2 had its own editor and a translation memory feature which used fuzzy search algorithms to retrieve existing material from its translation database. TM/2 could analyse texts to extract terms. TM/2 came with lemmatizers, spelling lists, and other linguistic resources for nineteen languages, including Catalan, Flemish, Norwegian, Portuguese, Greek, and Icelandic. External dictionaries could also be integrated into TM/2, provided they were formatted in Standard Generalized Markup Language (SGML). TM/2 could be linked to logic-based machine translation (Brace 1992a). This system is perhaps the first hybrid computer-aided translation system that was integrated with a machine translation system (Brace 1993; Wassmer 2011).

Third, in Russia, the PROMT Ltd was founded by two doctorates in computational linguistics, Svetlana Sokolova and Alexander Serebryakov, in St. Petersburg in 1991. At the beginning, the company mainly developed machine translation (MT) technology, which has been at the heart of the @promt products. Later, it began to provide a full range of translation solutions: machine translation systems and services, dictionaries, translation memory systems, data mining systems (http://www.promt.com).

Fourth, in the United Kingdom, two companies specializing in translation software production were founded. First, Mark Lancaster established the SDL International, which served as a service provider for the globalization of software (http://www.sdl.com). Second,

ATA Software Technology Ltd, a London-based software house specializing in Arabic translation software, was established in 1992 by some programmers and Arabic software specialists. The company later developed a series of machine translation products (Arabic and English) and MT and TM hybrid system, Xpro7 and online translation engine (http://www.atasoft.com).

1993–2003: A period of rapid growth

This period, covering the years from 1993 to 2003, is a period of rapid growth, due largely to (1) the emergence of more commercial systems; (2) the development of more built-in functions; (3) the dominance of Windows operation systems; (4) the support of more document formats; (5) the support of more languages for translation; and (6) the dominance of Trados as a market leader.

(1) The emergence of more commercial systems

Before 1993, there were only three systems available on the market, including Translator's Workbench II of Trados, IBM Translation Manager / 2, and STAR Transit 1.0. During this ten-year period between 1993 and 2003, about twenty systems were developed for sale, including the following better-known systems such as Déjà Vu, Eurolang Optimizer (Brace 1994), Wordfisher, SDLX, ForeignDesk, Trans Suite 2000, Yaxin CAT, Wordfast, Across, OmegaT, MultiTrans, Huajian, Heartsome, and Transwhiz. This means that there was a six-fold increase in commercial computer-aided translation systems during this period.

Déjà Vu is the name of a computer-aided translation system developed by Atril in Spain after 1993. A preliminary version of Déjà Vu, a customizable computer-aided translation system that combined translation memory technology with example-based machine translation techniques, was initially developed by ATRIL in June to fulfil their own need for a professional translation tool. At first, they worked with machine translation systems, but the experiments with machine translation were extremely disappointing, and subsequent experiences with translation memory tools exposed two main shortcomings: all systems ran under MS-DOS and were capable of processing only plain text files. Then, ATRIL began considering the idea of writing its own translation memory software.

Déjà Vu 1.0 was released to the public in November 1993. It was with an interface for Microsoft Word for Windows 2.0, which was defined as the first of its kind. Version 1.1 followed soon afterwards, incorporating several performance improvements and an integrated alignment tool (at a time when alignment tools were sold as expensive individual products), and setting a new standard for the translation tool market (http://www.atril.com).

Déjà Vu, designed to be a professional translation tool, produced acceptable results at an affordable price. Déjà Vu was a first in many areas: the first TM tool for Windows; the first TM tool to directly integrate into Microsoft Word; the first 32-bit TM tool (Déjà Vu version 2.0); and the first affordable professional translation tool.

In the following year, Eurolang Optimizer, a computer-aided translation system, was developed by Eurolang in France. Its components included the translator's workstation, pre-translation server with translation memory and terminology database, and project management tool for multiple languages and users (Brace 1994).

In Germany, Trados GmbH announced the release of the new Windows version of Translator's Workbench, which could be used with standard Windows word processing packages via the Windows DDE interface (Brace 1994). In June 1994 Trados released

MultiTerm Professional 1.5 which was included in Translator's Workbench, which had fuzzy search to deliver successful searches even when words were incorrectly spelt, a dictionary-style interface, faster searches through use of new highly compressed data algorithms, drag and drop content into word processor and integrated programming language to create powerful layouts (http://www.translationzone.com).

In Hungary, Tibor Környei developed the WordFisher for Microsoft Word macro set. The programme was written in the WordBasic language. For translators, it resembled a translation memory programme, but provided a simpler interface in Word (Környei 2000).

In 1995, Nero AG was founded in Germany as a manufacturer of CD and DVD application software. Later, the company set up Across Systems GmbH as a division, which developed and marketed a tool of the same name for corporate translation management (CTM) that supported the project and workflow management of translations (Schmidt 2006; German 2009: 9–10).

During the first half of 1996, when Windows 95 was in its final stages of beta testing, Atril Development S.L. in Spain began writing a new version of Déjà Vu – not just porting the original code to 32 bits, but adding a large number of important functionalities that had been suggested by the users. In October, Atril released Déjà Vu beta v2.0. It consisted of the universal editor, Déjà Vu Interactive (DVI), the Database Maintenance module with an alignment tool, and a full-featured Terminology Maintenance module (Wassmer 2007: 37–38).

In the same year, Déjà Vu again was the first TM tool available for 32-bit Windows and shipped with a number of filters for DTP packages – including FrameMaker, Interleaf, and QuarkXPress – and provided extensive project management facilities to enable project managers to handle large, multi-file, multilingual projects.

In 1997, developments in France and Germany deserve mentioning. In France, CIMOS released Arabic to English translation software An-Nakel El-Arabi, with features like machine translation, customized dictionary and translation memory. Because of its deep sentence analysis and semantic connections, An-Nakel Al-Arabi could learn new rules and knowledge. CIMOS had previously released English to Arabic translation software (MultiLingual 1997). In Germany, Trados GmbH released WinAlign as a visual text alignment tool as the first fully-fledged 32-bit application in Trados. Mircosoft decided to base its internal localization memory store on Trados and consequently acquired a share of 20 per cent in Trados (http://www.translationzone.com).

The year 1998 marks a milestone in the development of translation technology in China and Taiwan. In Beijing, Beijing Yaxincheng Software Technology Co. Ltd. 北京雅信誠公司 was set up as a developer of translation software. It was the first computer-aided translation software company in China. In Taipei, the Inventec Corporation released Dr Eye 98 (譯典通) with instant machine translation, dictionaries and termbases in Chinese and English (http://www.dreye.com.tw).

In the same year, the activities of SDL and International Communications deserve special mention. In the United Kingdom, SDL began to acquire and develop translation and localization software and hardware – both for its own use in client-specific solutions, and to be sold as free-standing commercial products. At the end of the year, SDL also released SDLX, a suite of translation memory database tools. SDLX was developed and used in-house at SDL, and therefore was a mature product at its first offering (Hall 2000; MultiLingual 1998). Another British company, International Communications, a provider of localization, translation and multilingual communications services, released ForeignDesk v5.0 with the full support of Trados Translator's Workbench 2.0 and WinAlign, S-Tagger. Then, Lionbridge Technologies Inc. acquired it (known as Massachusetts-based INT'L.com at the transaction) and later in November 2001 decided to open-source the ForeignDesk suite free of charge under BSD

licence. ForeignDesk was originally developed by International Communications around 1995 (MultiLingual 2000).

In June 1999, Beijing YaxinCheng Software Technology Co. Ltd. established Shida CAT Research Centre (實達 CAT 研究中心), which later developed Yaxin CAT Bidirectional v2.5 (Chan 2004: 338). In June, SJTU Sunway Software Industry Ltd. acquired one of the most famous CAT products in China at the moment – Yaxin CAT from Beijing YaxinCheng Software Technology Co. Ltd., and it released the Yaxin CAT v1.0 in August. The release of this software signified, in a small way, that the development of computer-aided systems was no longer a European monopoly.

In France, the first version of Wordfast PlusTools suite of CAT (Computer-Assisted Translation) tools was developed. One of the developers was Yves A. Champollion, who incorporated Wordfast LLC later. There were only a few TM software packages available in the first version. It could be downloaded freely before 2002, although registration was required (http://www.wordfast.net/champollion.net).

In the United States, MultiCorpora R&D Inc. was incorporated, which was exclusively dedicated to providing language technology solutions to enterprises, governments, and language service providers (http://www.multicorpora.com).

In the United Kingdom, following the launch of SDL International's translation database tool, SDLX, SDL announced SDL Workbench. Packaged with SDLX, SDL Workbench memorized a user's translations and automatically offered other possible translations and terminology from a user's translation database within the Microsoft Word environment. In line with its 'open' design, it was able to work with a variety of file formats, including Trados and pre-translated RTF files (MultiLingual 1999).

The year 2000 was a year of activities in the industry. In China, Yaxin CAT v2.5 Bidirectional (English and Chinese) was released with new features like seventy-four topic-specific lexicons with six million terms free of charge, project analysis, project management, share translation memory online and simultaneous editing of machine output (Chen 2001).

In Germany, OmegaT, a free (GPL) translation memory tool, was publicly released. The key features of OmegaT were basic (the functionality was very limited), free, open-source, cross-operation systems as it was programmed in Java (http://www.omegat.org; Prior 2003).

In Ireland, Alchemy Software Development Limited announced the acquisition of Corel CATALYST™, which was designed to boost the efficiency and quality of globalizing software products and was used by over 200 software development and globalization companies worldwide (http://www.alchemysoftware.ie).

In the United Kingdom, SDL International announced in April the release of SDLX 2.0, which was a new and improved version of SDLX 1.03 (http://www.sdl.com). It also released SDL Webflow for managing multilingual website content (http://www.sdlintl.com).

In Germany, Trados relocated its headquarters to the United States in March and became a Delaware corporation.

In France, Wordfast v3.0 was released in September. The on-the-fly tagging and un-tagging of HTML (HyperText Markup Language) files was a major breakthrough in the industry. Freelance translators could translate HTML pages without worrying about the technical hurdles.

Not much happened in 2001. In Taiwan, Inventec Corporation released Dr Eye 2001, with new functions like online search engine, full-text machine translation from English to Chinese, machine translation from Japanese to Chinese and localization plug-in (Xu 2001). In the United Kingdom, SDL International released SDLX 4.0 with real-time translation, a flexible software licence and enhanced capabilities. In the United States, Trados announced the launch of Trados 5 in two versions, Freelance and Team (http://www.translationzone.com).

In contrast, the year 2002 was full of activities in the industry.

In North America, MultiCorpora R&D Inc. in Canada released MultiTrans 3, providing corpus-based translation support and language management solution. It also introduced a new translation technology called Advanced Leveraging Translation Memory (ALTM). This model provided past translations in their original context and required virtually no alignment maintenance to obtain superior alignment results. In the United States, Trados 5.5 (Trados Corporate Translation Solution™) was released. MultiCorpora released MultiTrans 3.0, which introduced an optional client-server add-on, so it could be used in a web-based, multi-user environment or as a standalone workstation. Version 3 supported TMX and was also fully Unicode compliant (Locke and Giguère 2002: 51).

In Europe and the United Kingdom, SDL International released its new SDLX Translation Suite 4, and then later that year released the elite version of the suite. The SDLX Translation Suite features a modular architecture consisting of five to eight components: SDL Project Wizard, SDL Align, SDL Maintain, SDL Edit and SDL TermBase in all versions, and SDL Analyse, SDL Apply and SDLX AutoTrans in the Professional and Elite versions (Wassmer 2003). In Germany, MetaTexis Software and Services released in April the first official version 1.00 of MetaTexis (http://www.metatexis.com).

In Asia, Huajian Corporation in China released Huajian IAT, a computer-aided translation system (http://www.hjtek.com). In Taiwan, Otek launched in July Transwhiz Power version (client/server structure), which aimed at enterprise customers (http://www.otek.com.tw). In Singapore, Heartsome Holdings Pte. Ltd. was founded to develop language translation technology (Garcia and Stevenson 2006: 77).

North America and Europe were active in translation technology in 2003.

In 2003, MultiCorpora R&D Inc. in Canada released MultiTrans 3.5 which had new and improved capabilities, including increased processing speed of automated searches, increased network communications speed, improved automatic text alignment for all languages, and optional corpus-based pre-translation. Version 3.5 also offered several new terminology management features, such as support for additional data types, additional filters, batch updates and added import and export flexibility, as well as full Microsoft Office 2003 compatibility, enhanced Web security and document analysis capabilities for a wider variety of document formats (MultiLingual 2003). In the United States, Trados 6 was launched in April and Trados 6.5 was launched in October with new features like auto concordance search, Word 2003 support and access to internet TM server (Wassmer 2004: 61).

In Germany, MetaTexis version 2.0 was released in October with a new database engine. And MetaTexis version 'Net/Office' was released with new features that supported Microsoft PowerPoint and Excel files, Trados Workbench, and could be connected with Logoport servers (http://www.metatexis.com).

In Russia, PROMT, a developer of machine translation products and services, released a new version @promt XT with new functions like processing PDF file formats, which made PROMT the first among translation software that supported PDF. Also, one of the editions, @promt Expert integrated translation memory solutions (Trados) and a proprietary terminology extraction system (http://www.promt.com).

In France, Atril, which was originally founded in Spain but which relocated its group business to France in the late 1990s, released Déjà Vu X (Standard, Professional, Workgroup and Term Sever) (Harmsen 2008). Wordfast 4, which could import and translate PDF contents, was also released (http://www.wordfast.net).

Some developers of machine translation systems also launched new versions with a translation memory component, such as LogoVista, An-Nabel El-Arabi and PROMT (http://www.

promt.com). Each of these systems was created with distinct philosophies in its design, offering its own solutions to problems and issues in the work of translation. This was aptly pointed out by Brace (1994):

> Eurolang Optimizer is based on an ambitious client / server architecture designed primarily for the management of large translation jobs. Trados Workbench, on the other hand, offers more refined linguistic analysis and has been carefully engineered to increase the productivity of single translators and small workgroups.

(2) The development of more built-in functions

Computer-aided translation systems of the first and second periods were usually equipped with basic components, such as translation memory, terminology management, and translation editor. In this period, more functions were developed and more components were gradually integrated into computer-aided translation systems. Of all the new functions developed, tools for alignment, machine translation, and project management were most significant. Trados Translator's Workbench II, for example, incorporated T Align, later known as WinAlign, into its workstation, followed by other systems such as Déjà Vu, SDLX, Wordfisher, and MultiTrans. Machine translation was also integrated into computer-aided translation systems to handle segments not found in translation memories. IBM's Translation Manager, for example, introduced its Logic-Based Machine Translation (LMT) to run on IBM mainframes and RS/6000 Unix systems (Brace 1993). The function of project management was also introduced by Eurolang Optimizer in 1994 to better manage translation memory and terminology databases for multiple languages and users (Brace 1992a).

(3) The dominance of Windows Operating System

Computer-aided translation systems created before 1993 were run either in the DOS system or OS/2 system. In 1993, the Windows versions of these systems were first introduced and they later became the dominant stream. For example, IBM and Trados GmbH released a Windows version of TM/2 and of Translator's Workbench respectively in mid-1993. More Windows versions came onto the market, such as the preliminary version of ATRIL's Déjà Vu 1.0 in June in Spain. Other newly released systems running on Windows include SDLX, ForeignDesk, Trans Suite 2000, Yaxin CAT, Across, MultiTrans, Huajian, and TransWhiz.

(4) The support of more document formats

Computer-aided translation systems of this period could handle more document formats directly or with filters, including Adobe InDesign, FrameMaker, HTML, Microsoft PowerPoint, Excel, Word, QuarkXPress, even PDF by 2003. Trados 6.5, for example, supported all the widely used file formats in the translation community, which allowed translators and translation companies to translate documents in Microsoft Office 2003 Word, Excel and PowerPoint, Adobe InDesign 2.0, FrameMaker 7.0, QuarkXPress 5, and PageMaker.

(5) The support of translation of more languages

Translation memory is supposed to be language-independent, but computer-aided translation systems developed in the early 1990s did not support all languages. In 1992, Translator

Workbench Editor, for example, supported only five European languages, namely, German, English, French, Italian and Spanish, while IBM Translation Manager / 2 supported 19 languages, including Chinese, Korean and other OS/2 compatible character code sets. This was due largely to the contribution of Unicode, which provided the basis for the processing, storage, and interchange of text data in any language in all modern software, thereby allowing developers of computer-aided translation systems to gradually resolve obstacles in language processing, especially after the release of Microsoft Office 2000. Systems with Unicode support mushroomed, including Transit 3.0 in 1999, MultiTerm and WordFisher 4.2.0 in 2000, Wordfast Classic 3.34 in 2001, and Tr-AID 2.0 and MultiTrans 3 in 2002.

(6) The dominance of Trados as a market leader

As a forerunner in the field, Trados became a market leader in this period. As observed by Colin Brace, 'Trados has built up a solid technological base and a good market position' in its first decade. By 1994, the company had a range of translation software, including Trados Translator's Workbench (Windows and DOS versions), MultiTerm Pro, MultiTerm Lite, and MultiTerm Dictionary. Its technology in translation memory and file format was then widely used in other computer-aided translation systems and its products were most popular in the industry. From the late 1990s, a few systems began to integrate Trados's translation memory into their systems. In 1997, ProMemoria, for example, was launched with its translation memory component provided by Trados. In 1998, International Communications released ForeignDesk 5.0 with the full support of Trados Translator's Workbench 2.0, WinAlign, and S-Tagger. In 1999, SDLX supported import and export formats such as Trados and tab-delimited and CSV files. In 2000, Trans Suite 2000 was released with the capacity to process Trados RTF file. In 2001, Wordfast 3.22 could directly open Trados TMW translation memories (Translator's Workbench versions 2 and 3). In 2003, PROMT XT Export integrated Trados's translation memory. In October 2003, MetaTexis 'Net/Office' 2.0 was released and was able to work with Trados Workbench.

2004–2013: A period of global development

Advances in technology have given added capabilities to computer-aided translation systems. During the last nine years, while most old systems have been upgraded on a regular basis, close to thirty new systems have been released to the market. This situation has offered a wider range of choices for buyers to acquire systems with different packages, functions, operation systems, and prices.

One of the most significant changes in this period is the addition of new computer-aided translation companies in countries other than those mentioned above. Hungary is a typical example. In 2004, Kilgray Translation Technologies was established by three Hungarian language technologists. The name of the company was made up of the founders' surnames: Kis Balázs (KI), Lengyel István (L), and Ugray Gábor (GRAY). Later, the company launched the first version of MemoQ, an integrated Localization Environment (ILE), in 2005. MemoQ's first version had a server component that enabled the creation of server projects. Products of Kilgray included MemoQ, MemoQ server, QTerm, and TM Repository (http://www.kilgray.com).

Another example is Japan. In Japan, Rozetta Corporation released TraTool, a computer-aided translation system with translation memory, an integrated alignment tool, an integrated terminology tool and a user dictionary. The product is still commercially available but no major improvement has been made since its first version (http://www.tratool.com).

Yet another example is Poland, where AidTrans Soft launched its AidTrans Studio 1.00, a translation memory tool. But the company was discontinued in 2010 (http://www.thelanguagedirectory.com/translation/translation_software).

New versions of computer-aided translation systems with new features are worth noting. In the United Kingdom, ATA launched a new Arabic Memory Translation system, Xpro7 which had the function of machine translation (http://www.atasoft.com). SDL Desktop Products, a division of SDL International, announced the launch of SDLX 2004. Its new features included TMX Certification, seamlessly integrating with Enterprise systems such as online terminology and multilingual workflow management, adaptation of new file formats, synchronized web-enabled TM, and Knowledge-based Translation (http://www.sdl.com). In the United States, Systran released Systran Professional Premium 5.0, which contained integrated tools such as integrated translation memory with TMX support, a Translator's Workbench for post-editing and ongoing quality analysis (http://www.systransoft.com). Multilizer Inc., a developer of globalization technologies in the United States, released a new version of Multilizer, which included multi-user translation memory with Translation Memory Manager (TMM), a standalone tool for maintaining Multilizer Translation Memory contents. TMM allowed editing, adding and deleting translations, and also included a briefcase model for working with translations off-line (http://www.multilizer.com).

In Ukraine, Advanced International Translations (AIT) started work on user-friendly translation memory software, later known as AnyMen, which was released in December 2008.

In 2005, translation technology moved further ahead with new versions and new functions.

In North America, MultiCorpora in Canada released MultiTrans 4**,** which built on the foundation of MultiTrans 3.7 and had a new alignment process that was completely automated (MultiLingual 2005d). Trados, incorporated in the United States, produced Trados 7 Freelance, which supported twenty additional languages, including Hindi. At an operating system level, Microsoft Windows 2000, Windows XP Home, Windows XP Professional, and Windows 2003 Server were supported. More file formats were now directly supported by TagEditor. MultiCorpora also introduced MultiTrans 4, which was designed to meet the needs of large organizations by providing the newest efficiencies for translators in the areas of text alignment quality, user-friendliness, flexibility and web access (http://www.multicorpora.com).

In Europe, Lingua et Machina, a memory translation tool developer, released SIMILIS v1.4, its second-generation translation tool. SIMILIS uses linguistic parsers in conjunction with the translation memory paradigm. This function allowed for the automatic extraction of bilingual terminology from translated documents. Version 1.4 brought compatibility with the Trados translation memory format (Text and TMX) and a new language, German (MultiLingual 2005b). In Switzerland, STAR Group released Transit XV Service Pack 14. This version extended its capabilities with a number of new features and support of 160 languages and language versions, including Urdu (India) and Urdu (Pakistan). It supported Microsoft Word 2003 files and had MySpell dictionaries (MultiLingual 2005a). PROMT released @promt 7.0 translation software, which supported the integrated translation memory, the first of its kind among PROMT's products (http://www.promt.com).

In the United Kingdom, SDL Desktop Products released the latest version of its translation memory tool SDLX 2005, which expanded the Terminology QA Check and automatically checked source and translations for inconsistent, incomplete, partial or empty translations, corrupt characters, and consistent regular expressions, punctuation, and formatting. Language support had been added for Maltese, Armenian and Georgian, and the system could handle more than 150 languages (MultiLingual 2005c). In June, SDL International acquired Trados for £35 million. The acquisition provided extensive end-to-end technology and service

solutions for global information assets (http://www.translationzone.com). In October, SDL Synergy was released as a new project management tool on the market.

In Asia, Huajian Corporation in China released in June Huajian Multilingual IAT network version (華建多語 IAT 網絡版) and in October Huajian IAT (Russian to Chinese) standalone version (http://www.hjtrans.com). In July, Beijing Orient Yaxin Software Technology Co. Ltd. released Yaxin CAT 2.0, which was a suite including Yaxin CAT 3.5, CAM 3.5, Server, Lexicons, Translation Memory Maintenance and Example Base. In Singapore, Heartsome Holdings Pte. Ltd. released Heartsome Translation Suite, which was composed of three programs: an XLIFF Editor in which source files were converted to XLIFF format and translated; a TMX Editor that dealt with TMX files; and a Dictionary Editor that dealt with TBX files (Garcia and Stevenson 2006: 77). In Taiwan, Otek released Transwhiz 9.0 for English, Chinese and Japanese languages (http://www.otek.com.tw).

Significant advances in translation technology were made in 2006 particularly in Europe, the United Kingdom, and the United States.

In Europe, Across Systems GmbH in Germany released in September its Corporate Translation Management 3.5, which marked the start of the worldwide rollout of Across software (MultiLingual 2006a). In the United Kingdom, SDL International released in February SDL Trados 2006, which integrated with Translators Workbench, TagEditor, SDLX editing environments and SDL MultiTerm. It included new support for Quark, InDesign CS2 and Java (http://www.sdl.com). In the United States, MultiCorpora launched TextBase TM concept (http://www.multicorpora.com). Apple Inc. released in August AppleTrans, a text editor specially designed for translators, featuring online corpora which represented 'translation memory' accessible through documents. AppleTrans helped users localize web pages (http://developer.apple.com). Lingotek, a language search engine developer in the United States, launched a beta version of a collaborative language translation service that enhanced a translator's efficiency by quickly finding meaning-based translated material for re-use. Lingotek's language search engine indexed linguistic knowledge from a growing repository of multilingual content and language translations, instead of web pages. Users could then access its database of previously translated material to find more specific combinations of words for re-use. Such meaning-based searching maintained better style, tone, and terminology. Lingotek ran completely within most popular web browsers, including initial support for Internet Explorer and Firefox. Lingotek supported Word, Rich Text Format (RTF), Open Office, HTML, XHTML and Excel formats, thereby allowing users to upload such documents directly into Lingotek. Lingotek also supported existing translation memory files that were TMX-compliant memories, thus allowing users to import TMX files into both private and public indices (MultiLingual 2006b).

In 2007, Wordfast 5.5 was released in France. It was an upgrade from Wordfast 4, in which Mac support was completely overhauled. This version continued to offer translators collaboration community via a LAN. Each Wordfast licence granted users the ability to search Wordfast's web-based TM and knowledge base, VLTM (http://www.wordfast.net). In Germany, a group of independent translators and programmers under the GNU GPL licence developed in October Anaphraseus, a computer-aided translation tool for creating, managing and using bilingual translation memories. Originally, Anaphraseus was developed to work with the Wordfast TM format, but it could also export and import files in TMX format (http://anaphraseus.sourceforge.net). In Hungary, Kilgray Translation Technologies released in January MemoQ 2.0. The main theme for the new version was networking, featuring a new resource server. This server not only stored translation memory and term bases, but also offered the possibility of creating server projects that allowed for the easy distribution of work among

several translators and ensured productivity at an early stage of the learning curve. Improvements on the client side included support for XML and Adobe FrameMaker MIF file formats; improvements to all other supported file formats; and support for the Segmentation Rule eXchange standard, auto-propagation of translated segments, better navigation and over a hundred more minor enhancements (Multilingual 2007). In Russia, MT2007, a freeware, was developed by a professional programmer Andrew Manson. The main idea was to develop easy-to-use software with extensive features. This software lacked many features that leading systems had. In the United Kingdom, SDL International released in March SDL Trados 2007, which had features such as a new concept of project delivery and supply chain, new one-central-view dashboard for new project wizard, PerfectMatch, automated quality assurance checker and full support for Microsoft Office 2007 and Windows Vista.

In the United States, MultiCorpora's Advanced Leveraging launched WordAlign which boasted the ability to align text at the individual term and expression level (http://www.multicorpora.com). MadCap Software Inc., a multi-channel content authoring company, developed in May MadCap Lingo, an XML-based, fully-integrated Help authoring tool and translation environment. MadCap Lingo offered an easy-to-use interface, complete Unicode support for all left-to-right languages for assisting localization tasks. Across Systems GmbH and MadCap Software announced a partnership to combine technical content creation with advanced translation and localization. In June, Alchemy Software Development Ltd. and MadCap Software, Inc. announced a joint technology partnership that combined technical content creation with visual TM technology.

In 2008, Europe again figured prominently in computer-aided translation software production. In Germany, Across Systems GmbH released in April Across Language Server 4.0 Service Pack 1, which contained various extensions in addition to authoring, such as FrameMaker 8 and SGML support, context matching, and improvements for web-based translations via crossWeb (MultiLingual 2008a). It also introduced in July its new Language Portal Solution (later known as Across Language Portal) for large-scale organizations and multinational corporations, which allowed customers operating on an international scale to implement Web portals for all language-related issues and for all staff levels that need to make use of language resources. At the same time Across released the latest update to the Across Language Server, offering many new functions for the localization of software user interfaces (http://www.across.net). In Luxembourg, Wordbee S.A. was founded as a translation software company focusing on web-based integrated CAT and management solutions (http://www.wordbee.com).

In Eastern Europe, Kilgray Translation Technologies in Hungary released in September MemoQ 3.0, which included a new termbase and provided new terminology features. It introduced full support for XLIFF as a bilingual format and offered the visual localization of RESX files. MemoQ 3.0 was available in English, German, Japanese and Hungarian (http://kilgray.com). In Russia, Promt released in March 8.0 version with major improvement in its translation engine, translation memory system with TMX files import support, and dialect support in English (UK and American), Spanish (Castilian and Latin American), Portuguese (Portuguese and Brazilian), German (German and Swiss) and French (French, Swiss, Belgian, Canadian) (http://www.promt.com). In Ukraine, Advanced International Translations (AIT) released in December AnyMen, a translation memory system compatible with Microsoft Word. In Uruguay, Maxprograms launched in April Swordfish version 1.0-0, a cross-platform computer-aided translation tool based on the XLIFF 1.2 open standard published by OASIS (http://www.maxprograms.com). In November, this company released Stingray version 1.0-0, a cross-platform document aligner. The translation memories in TMX, CSV or Trados

TXT format generated by Stingray could be used in most modern computer-aided translation systems (http://www.maxprograms.com).

In Ireland, Alchemy Software Development, a company in visual localization solutions, released in July Alchemy PUBLISHER 2.0, which combined visual localization technology with translation memory for documentation. It supported standard documentation formats, such as MS Word, XML, application platforms such as Windows 16/22/64x binaries, web-contents formats such as HTML, ASP, and all derivative content formats (http://www.alchemysoftware.ie).

In North America, JiveFusion Technologies, Inc. in Canada officially launched Fusion One and Fusion Collaborate 3.0. The launches introduced a new method of managing translation memories. New features included complete contextual referencing. JiveFusion also integrated Fusion Collaborate 3.0 with TransFlow, a project and workflow management solution by Logosoft (MultiLingual 2008b). In the United States, MadCap Software, Inc. released in February MadCap Lingo 2.0, which included the Darwin Information Typing Architecture standard, Microsoft Word and a range of standard text and language formats. In September, it released MadCap Lingo 3.0, which included a new project packager function designed to bridge the gap between authors and translators who used other translation memory system software and a new TermBase Editor for creating databases of reusable translated terms.

In Asia, Yaxin CAT 4.0 was released in China in August with some new features including a computer-aided project platform for project management and huge databases for handling large translation projects. In Taiwan, Otek released Transwhiz 10 for translating English, Chinese and Japanese languages, with fuzzy search engine and Microsoft Word workstation (http://www.otek.com.tw).

The year 2009 witnessed the development of Autshumato Integrated Translation Environment (ITE) version 1.0, a project funded by the Department of Arts and Culture of the Republic of South Africa. It was released by The Centre for Text Technology (CTexT®) at the Potchefstroom Campus of the North-West University and University of Pretoria after two years of research and development. Although Autshumato ITE was specifically developed for the eleven official South African languages, it was in essence language independent, and could be adapted for translating between any language pair.

In Europe, Wordfast released in January Wordfast Translation Studio, a bundled product with Wordfast Classic (for Microsoft Word) and Wordfast Pro (a standalone CAT platform). With over 15,000 licences in active use, Wordfast claimed itself the second most widely used translation memory tool (http://www.wordfast.net). In Germany, Across Systems GmbH released in May Across Language Server 5.0, which offered several options for process automation as well as for workflow management and analysis. Approximately fifty connections were available for interacting with other systems (MultiLingual 2009b). In September, STAR Group in Switzerland released Transit^{NXT} (Professional, Freelance Pro, Workstation, and Freelance). Service pack 1 for Transit NXT/TermStar NXT contained additional user interface languages for Chinese, Spanish, Japanese, and Khmer, enhanced alignment usability, support for QuarkXpress 7, and proofreading for internal repetitions.

In the United Kingdom, SDL announced in June the launch of SDL Trados® Studio 2009 in the same month, which included the latest versions of SDL MultiTerm, SDL Passolo Essential, SDL Trados WinAlign, and SDL Trados 2007 Suite. New features included Context Match, AutoSuggest, QuickPlace (http://www.sdl.com). In October, SDL released its enterprise platform SDL TM Server™ 2009, a new solution to centralize, share, and control translation memories (http://www.sdl.com).

In North America, JiveFusion Technologies Inc. in Canada released in March Fusion 3.1 to enhance current TMX compatibility and the capability to import and export to TMX while preserving the complete segment context (MultiLingual 2009a). In the United States, Lingotek introduced software-as-a-service collaborative translation technology which combined the workflow and computer-aided translation capabilities of human and machine translation into one application. Organizations could upload new projects, assign translators (paid or unpaid), check the status of current projects in real time and download completed documents from any computer with web access (MultiLingual 2009c).

In Asia, Beijing Zhongke LongRay Software and Technology Ltd. Co. in China released in September LongRay CAT 3.0 (standalone edition), a CAT system with translation memory, alignment, dictionary and terminology management and other functions (http://www.zklr.com). In November, Foshan Snowman Computer Co. Ltd. released Snowman version 1.0 in China (http://www.gcys.cn). Snowman deserves some mentioning because (1) it was new; (2) the green trial version of Snowman could be downloaded free of charge; (3) it was easy to use as its interface was user-friendly and the system was easy to operate; and (4) it had the language pair of Chinese and English, which caters to the huge domestic market as well as the market abroad.

Most of the activities relating to computer-aided translation in 2010 took place in Europe and North America.

In Germany, Across Systems GmbH released in August Across Language Server v. 5 Service Pack 1, which introduced a series of new functionalities and modes of operation relating to the areas of project management, machine translation, crowdsourcing and authoring assistance (http://new.multilingual.com). In October, MetaTexis version 3.0 was released, which imported filter for Wordfast Pro and Trados Studio translation memories and documents (http://www.metatexis.com). In France, Wordfast LLC released in July Wordfast Pro 2.4 (WFP) with over sixty enhancements. This system was a standalone environment that featured a highly customizable interface, enhanced batch processing functionality, and increased file format support (http://www.wordfast.net). In October, Wordfast LLC created an application to support translation on the iPhone and iPad in the Wordfast Anywhere environment (http://www.wordfast.net). In Hungary, Kilgray Translation Technologies released in February MemoQ 4.0, which was integrated with project management functions for project managers who wanted to have more control and enable translators to work in any translation tool. In October, the company released MemoQ 4.5, which had a rewritten translation memory engine and improvements to the alignment algorithm (http://www.kilgray.com). In France, Atril released in March TeaM Server, which allowed translators with Déjà Vu Workgroup to work on multinational and multisite translation projects on a LAN or over the Internet, sharing their translations in real-time, ensuring superior quality and consistency. TeaM Server also provided scalable centralized storage for translation memories and terminology databases. The size of translation repositories and the number of concurrent users were only limited by the server hardware and bandwidth (http://www.atril.com). In October, Atril released Déjà Vu X2 in four editions: Editor, Standard, Professional, and Workgroup. Its new features included DeepMiner data extraction engine, new StartView interface, and AutoWrite word prediction. In Switzerland, STAR Group released in October Transit NXT Service Pack 3 and TermStar NXT. Transit NXT Service Pack 3 contained the following improvements: support of Microsoft Office 2007, InDesign CS5, QuarkXpress 8 and QuarkXpress 8.1, and PDF synchronization for MS Word files.

In the United Kingdom, SDL released in March a new subscription level of its SDL Trados Studio, which included additional productivity tools for translators such as Service Pack 2, enabling translators to plug in to multiple automatic translation tools. The company also did a

beta launch of SDL OpenExchange, inviting the developer community to make use of standard open application programming interfaces to increase the functionality of SDL Trados Studio (Multilingual 2010a). In September, XTM International released XTM Cloud, which was a totally online Software-as-a-Service (SaaS) computer-assisted translation tool set, combining translation workflow with translation memory, terminology management and a fully featured translator workbench. The launch of XTM Cloud means independent freelance translators have access to XTM for the first time (http://www.xtm-intl.com). In Ireland, Alchemy Software Development Limited released in May Alchemy PUBLISHER 3.0**,** which supports all aspects of the localization workflow, including form translation, engineering, testing, and project management. It also provided a Machine Translation connector which was jointly developed by PROMT, so that documentation formats could be machine translated (http://www.alchemysoftware.ie; http://www.promt.com).

In North America, IBM in the United States released in June the open source version of OpenTM/2, which originated from the IBM Translation Manager. OpenTM/2 integrated with several aspects of the end-to-end translation workflow (http://www.opentm2.org). Partnering with LISA (Localization Industry Standards Association), Welocalize, Cisco, and Linux Solution Group e.V. (LiSoG), IBM aimed to create an open source project that provided a full-featured, enterprise-level translation workbench environment for professional translators on OpenTM/2 project. According to LISA, OpenTM/2 not only provided a public and open implementation of translation workbench environment that served as the reference implementation of existing localization industry standards, such as TMX, it also aimed to provide standardized access to globalization process management software (http://www.lisa.org; LISA 2010). Lingotek upgraded in July its Collaborative Translation Platform (CTP) to a software-as-a-service product which combined machine translation, real-time community translation, and management tools (MultiLingual 2010b). MadCap Software, Inc. released in September MadCap Lingo v4.0, which had a new utility for easier translation alignment and a redesigned translation editor. Systran introduced in December Desktop 7 Product Suite, which included the Premium Translator, Business Translator, Office Translator, and Home Translator. Among them, Premium Translator and Business Translator were equipped with translation memory and project management features.

In South America, Maxprograms in Uruguay released in April Swordfish II, which incorporated Anchovy version 1.0-0 as glossary manager and term extraction tool, and added support for SLD XLIFF files from Trados Studio 2009 and Microsoft Visio XML Drawings, etc. (http://www.maxprograms.com).

In 2011, computer-aided translation was active in Europe and America.

In Europe, ATRIL / PowerLing in France released in May Déjà Vu X2, a new version of its computer-assisted translation system, which had new features such as DeepMiner data mining and translation engine, SmartView Interface and a multi-file and multi-format alignment tool (MultiLingual 2011). In June, Wordfast Classic v6.0 was released with features such as the ability to share TMs and glossaries with an unlimited number of users, improved quality assurance, AutoComplete, and improved support for Microsoft Word 2007/2010 and Mac Word 2011 (http://www.wordfast.net). In Luxembourg, the Directorate-General for Translation of the European Commission released in January its one million segments of multilingual Translation Memory in TMX format in 231 language pairs. Translation units were extracted from one of its large shared translation memories in Euramis (European Advanced Multilingual Information System). This memory contained most, but not all, of the documents of the *Acquis Communautaire*, the entire body of European legislation, plus some other documents which were not part of the *Acquis*. In Switzerland, the STAR Group released

in February Service Pack 4 for Transit NXT and TermStar NXT. Transit NXT Service Pack 4 contained the following improvements: support of MS Office 2010, support of Quicksilver 3.5l, and preview for MS Office formats. In Eastern Europe, Kilgray Translation Technologies in Hungary released in June TM Repository, the world's first tool-independent Translation Memory management system (http://kilgray.com). Kilgray Translation Technologies later released MemoQ v 5.0 with the AuditTrail concept to the workflow, which added new improvements like versioning, tracking changes (to show the difference of two versions), X-translate (to show changes on source texts), the Post Translation Analysis on formatting tags (Kilgray Translation Technologies 2011).

In the United Kingdom, XTM International released in March XTM 5.5, providing both Cloud and On-Premise versions, which contained customizable workflows, a new search and replace feature in Translation Memory Manager and the redesign of XTM Workbench (http://www.xtm-intl.com).

In North America, MultiCorpora R&D Inc. in Canada released in May MutliTrans Prism, a translation management system (TMS) for project management, translation memory and terminology management (MultiCorpora 2011).

In 2012, the development of computer-aided translation in various places was considerable and translation technology continued its march to globalization.

In North America, the development of computer-aided translation was fast. In Canada, MultiCorpora, a provider of multilingual asset management solutions, released in June MultiTrans Prism version 5.5. The new version features a web editing server that extends control of the management of translation processes, and it can be fully integrated with content management systems. In September, Terminotix launched LogiTerm 5.2. Its upgrades and new features, including indexing TMX files directly in Bitext database, reinforced the fuzzy match window, and adjusted buttons (http://terminotix.com/news/newsletter). In December, MultiCorpora added new machine translation integrations to its MultiTrans Prism. The integration options include Systran, Google and Microsoft (http://www.multicorpora.com). In Asia, there was considerable progress in computer-aided translation in China. Transn Information Technology Co., Ltd. released TCAT 2.0 as freeware early in the year. New features of this software include the Translation Assistant (翻譯助理) placed at the sidebar of Microsoft Office, pre-translation with TM and termbase, source segment selection by highlighting (自動取句) (http://www.transn.com). In May, Foshan Snowman Computer Co. Ltd. released Snowman 1.27 and Snowman Collaborative Translation Platform (雪人 CAT 協同翻譯平臺) free version. The platform offers a server for a central translation memory and termbase so that all the users can share their translations and terms, and the reviewers can view the translations simultaneously with translators. It also supports online instant communication, document management and online forum (BBS) (http://www.gcys.cn). In July, Chengdu Urelite Tech Co. Ltd. (成都優譯信息技術有限公司), which was founded in 2009, released Transmate, including the standalone edition (beta), internet edition and project management system. The standalone edition was freely available for download from the company's website. The standalone edition of Transmate is targeted at freelancers and this beta release offers basic CAT functions, such as using TM and terminology during translation. It has features such as pre-translation, creating file-based translation memory, bilingual text export and links to an online dictionary website and Google MT (http://www.urelitetech.com.cn).

Heartsome Translation Studio 8.0 was released by the Shenzhen Office of Heartsome in China. Its new features include pre-saving MT results and external proofreading file export in RTF format. The new and integrated interface also allows the user to work in a single unified environment in the translation process (http://www.heartsome.net).

In Japan, Ryan Ginstrom developed and released Align Assist 1.5, which is freeware to align source and translation files to create translation memory. The main improvement of this version is the ability to set the format of a cell text (http://felix-cat.com). In October, LogoVista Corporation released LogoVista PRO 2013. It can support Microsoft Office 2010 64-bit and Windows 8. More Japanese and English words are included and the total number of words in dictionaries is 6.47 million (http://www.logovista.co.jp).

In Europe, the developments of computer-aided translation systems are noteworthy.

In the Czech Republic, the MemSource Technologies released in January MemSource Editor for translators as a free tool to work with MemSource Cloud and MemSource Server. The Editor is multiplatform and can be currently installed on Windows and Macintosh (http://www.memsource.com). In April, this company released MemSource Cloud 2.0. MemSource Plugin, the former CAT component for Microsoft Word, is replaced by the new MemSource Editor, a standalone translation editor. Other new features include adding comments to segments, version control, translation workflow (only in the Team edition), better quality assurance and segmentation (http://blog.memsource.com). In December, MemSource Technologies released MemSource Cloud 2.8. It now encrypts all communication by default. This release also includes redesigned menu and tools. Based on the data about previous jobs, MemSource can suggest relevant linguistics for translation jobs (http://www.memsource.com).

In France, Wordfast LLC released Wordfast Pro 3.0 in April. Its new features include bilingual review, batch TransCheck, filter 100 per cent matches, split and merge TXML files, reverse source/target and pseudo-translation (http://www.wordfast.com). In June Atril and PowerLing updated Déjà Vu X2. Its new features include an incorporated PDF converter and a CodeZapper Macro (http://www.atril.com).

In Germany, Across Language Server v 5.5 was released in November. New features such as linguistic supply chain management are designed to make project and resources planning more transparent. The new version also supports the translation of display texts in various formats, and allows the protection of the translation units to ensure uniform use (http://www.across.net).

In Hungary, Kilgray Translation Technologies released in July MemoQ 6.0 with new features like predictive typing and several new online workflow concepts such as FirstAccept (assign job to the first translator who accepted it on the online workflow), GroupSourcing, Slicing, and Subvendor group (http://kilgray.com). In December, the company released MemoQ 6.2. Its new features include SDL package support, InDesign support with preview, new quality assurance checks and the ability to work with multiple machine translation engines at the same time (http://kilgray.com).

In Luxembourg, Wordbee in October designed a new business analysis module for its Wordbee translation management system, which provides a new dashboard where over 100 real-time reports are generated for every aspect of the localization process (http://www.wordbee.com).

In Switzerland, STAR Group released Service Pack 6 for Transit ^{NXT} and TermStar ^{NXT}. The improvements of Service Pack 6 of Transit ^{NXT} contain the support of Windows 8 and Windows Server 2012, QuarkXPress 9.0-9.2, InDesign CS6, integrated OpenOffice spell check dictionaries, 10 additional Indian languages (http://www.star-group.net).

In the United Kingdom, XTM International, a developer of XML authoring and translation tools, released in April XTM Suite 6.2. Its updates include a full integration with machine translation system, Asia Online Language Studio and the content management system XTRF. In October, the company released XTM Suite 7.0 and a new XTM Xchange module in XTM Cloud intended to increase the supply chain. Version 7.0 includes project management enhancements, allowing users to group files, assign translators to specific groups or languages, and create different workflows for different languages (http://www.xtm-intl.com).

During this period, the following trends are of note.

1 *The systematic compatibility with Windows and Microsoft Office*
Of the sixty-seven currently available systems on the market, only one does not run on the Windows operation systems. Computer-aided translation systems have to keep up with the advances in Windows and Microsoft Office for the sake of compatibility. Wordfast 5.51j, for example, was released in April 2007, three months after the release of Windows Vista, and Wordfast 5.90v was released in July 2010 to support Microsoft Office Word 2007 and 2010.

2 *The integration of workflow control into CAT systems*
Besides re-using or recycling translations of repetitive texts and text-based terminology, systems developed during this period added functions such as project management, spell check, quality assurance, and content control. Take SDL Trados Studio 2011 as an example. This version, which was released in September 2011, has a spell checking function for a larger number of languages and PerfectMatch 2.0 to track changes of the source documents. Most of the systems on the market can also perform 'context match', which is the identical match with identical surrounding segments in the translation document and in the translation memory.

3 *The availability of networked or online systems*
Because of the fast development of new information technologies, most CAT systems during this period were server-based, web-based and even cloud-based CAT systems, which had a huge storage of data. By the end of 2012, there were fifteen cloud-based CAT systems available on the market for individuals or enterprises, such as Lingotek Collaborative Translation Platform, SDL World Server, and XTM Cloud.

4 *The adoption of new formats in the industry*
Data exchange between different CAT systems has always been a difficult issue to handle as different systems have different formats, such as *dvmdb* for Déjà Vu X, and *tmw* for SDL Trados Translator's Workbench 8.0. These program-specific formats cannot be mutually recognizable, which makes it impossible to share data in the industry. In the past, the Localization Industry Standards Association (LISA) played a significant role in developing and promoting data exchange standards, such as SRX (Segmentation Rules eXchange), TMX (Translation Memory eXchange), TBX (Term-Bese eXchange) and XLIFF (XML Localisation Interchange File Format). (http://en.wikipedia.org/wiki/XLIFF). It can be estimated that the compliance of industry standards is also one of the future directions for better data exchange.

Translation technology on a fast track: a comparison of the developments of computer-aided translation with human translation and machine translation

The speed of the development of translation technology in recent decades can be illustrated through a comparison of computer-aided translation with human translation and machine translation.

The development of human translation

Human translation, in comparison with machine translation and computer-aided translation, has taken a considerably longer time and slower pace to develop. The history of human translation can be traced to 1122 BC when during the Zhou dynasty (1122–255 BC), a foreign

affairs bureau known as *Da xing ren* 大行人 was established to provide interpreting services for government officials to communicate with the twelve non-Han minorities along the borders of the Zhou empire (Chan 2009: 29–30). This is probably the first piece of documentary evidence of official interpreting in the world.

Since then a number of major events have taken place in the world of translation. In 285 BC, there was the first partial translation of the Bible from Hebrew into Greek in the form of the *Septuagint* (Worth 1992: 5–19). In 250 BC, the contribution of Andronicus Livius to translation made him the 'father of translation' (Kelly 1998: 495–504). In 67, Zhu Falan made the first translation of a Buddhist sutra in China (Editorial Committee 1988: 103). In 1141, Robert de Retines produced the first translation of the Koran in Latin (Chan 2009: 47). In 1382, John Wycliffe made the first complete translation of the Bible in English (Worth 1992: 66–70). In 1494, William Tyndale was the first scholar to translate the Bible from the original Hebrew and Greek into English (Delisle and Woodsworth 1995: 33–35). In 1611, the King James Version of the Bible was published (Allen 1969). In 1814, Robert Morrison made the first translation of the Bible into Chinese (Chan 2009: 73). In 1945, simultaneous interpreting was invented at the Nuremberg Trials held in Germany (Gaiba 1998). In 1946, the United Bible Societies was founded in New York (Chan 2009: 117). In 1952, the first conference on machine translation was held at the Massachusetts Institute of Technology (Hutchins 2000: 6, 34–35). In 1953, the Fédération Internationale des Traducteurs (FIT), or International Association of Translators, and the Association Internationale des Interprètes de Conference (AIIC), or the International Association of Conference Interpreters, were both founded in Paris (Haeseryn 1989: 379–84; Phelan 2001). In 1964, with the publication of *Toward a Science of Translating* in which the concept of dynamic equivalent translation was proposed, Eugene A. Nida was referred to as the 'Father of Translation Theory' (Nida 1964). In 1972, James S. Holmes proposed the first framework for translation studies (Holmes 1972/1987: 9–24, 1988: 93–98). In 1978, Even-Zohar proposed the Polysystem Theory (Even-Zohar 1978: 21–27).

A total of some seventeen major events took place during the history of human translation, which may be 3,135 years old. This shows that in terms of the mode of production, human translation has remained unchanged for a very long time.

The development of machine translation

In comparison with human translation, machine translation has advanced enormously since its inception in the 1940s. This can be clearly seen from an analysis of the countries with research and development in machine translation during the last seventy years.

Available information shows that an increasing number of countries have been involved in the research and development of machine translation. This is very much in evidence since the beginning of machine translation in 1947. Actually, long before the Second World War was over and the computer was invented, Georges Artsrouni, a French-Armenian engineer, created a translation machine known as 'Mechanical Brain'. Later in the year, Petr Petrovič Smirnov-Troyanskij (1894–1950), a Russian scholar, was issued a patent in Moscow on 5 September for his construction of a machine which could select and print words while translating from one language into another or into several others at the same time (Chan 2004: 289).

But it was not until the years after the Second World War that the climate was ripe for the development of machine translation. The invention of computers, the rise of information theory, and the advances in cryptology all indicated that machine translation could be a reality. In 1947, the idea of using machines in translating was proposed in March by Warren Weaver (1894–1978), who was at that time the vice president of the Rockefeller Foundation, and

Andrew D. Booth of Birkbeck College of the University of London. They wanted to make use of the newly invented computer to translate natural languages. Historically speaking, their idea was significant in several ways. It was the first application of the newly invented computers to non-numerical tasks, i.e. translation. It was the first application of the computer to natural languages, which was later to be known as computational linguistics. It was also one of the first areas of research in the field of artificial intelligence.

The following year witnessed the rise of information theory and its application to translation studies. The role of this theory has been to help translators recognize the function of concepts such as information load, implicit and explicit information, and redundancy (Shannon and Weaver 1949; Wiener 1954). On 15 July 1948, Warren Weaver, director of the Rockefeller Foundation's natural sciences division, wrote a memorandum for peer review outlining the prospects of machine translation, known in history as 'Weaver's Memorandum', in which he made four proposals to produce translations better than word-for-word translations (Hutchins 2000: 18–20).

The first machine translation system, the Georgetown-IBM system for Russian–English translation, was developed in the United States in June 1952. The system was developed by Leon Dostert and Paul Garvin of Georgetown University and Cuthbert Hurd and Peter Sheridan of IBM Corporation. This system could translate from Russian into English (Hutchins 1986: 70–78).

Russia was the second country to develop machine translation. At the end of 1954, the Steklov Mathematical Institute of the Academy of Sciences began work on machine translation under the directorship of Aleksej Andreevič Ljapunov (1911–1973), a mathematician and computer expert. The first system developed was known as FR-I, which was a direct translation system and was also considered one of the first generation of machine translation systems. The system ran on STRELA, one of the first generation of computers (Hutchins 2000: 197–204).

In the same year, the United Kingdom became the third country to engage in machine translation. A research group on machine translation, Cambridge Language Research Group, led by Margaret Masterman, was set up at Cambridge University, where an experimental system was tried on English-French translation (Wilks 2000: 279–298).

In 1955, Japan was the fourth country to develop machine translation. Kyushu University was the first university in Japan to begin research on machine translation (Nagao 1993: 203–208). This was followed by China, which began research on machine translation with a Russian–Chinese translation algorithm jointly developed by the Institute of Linguistics and the Institute of Computing Technology (Dong 1988: 85–91; Feng 1999: 335–340; Liu 1984: 1–14).

Two years later, Charles University in Czechoslovakia began to work on English–Czech machine translation (http://www.cuni.cz).

These six countries were the forerunners in machine translation. Other countries followed suit. In 1959, France set up the Centre d'Études de la Traduction Automatique (CETA) for machine translation (Chan 2009: 300). In 1960, East Germany had its Working Group for Mathematical and Applied Linguistics and Automatic Translation, while in Mexico, research on machine translation was conducted at the National Autonomous University of Mexico (Universidad Nacional Autonoma de Mexico) (http://www.unam.mx). In 1962, Hungary's Hungarian Academy of Sciences conducted research on machine translation. In 1964 in Bulgaria, the Mathematical Institute of the Bulgarian Academy of Sciences in Sofia set up the section of 'Automatic Translation and Mathematical Linguistics' to conduct work on machine translation (http://www.bas.bg; Hutchins 1986: 205–06). In 1965, the Canadian Research Council set up CETADOL (Centre de Traitement Automatique des Données Linguistiques) to work on an English–French translation system (Hutchins 1986: 224).

But with the publication of the ALPAC Report prepared by the Automatic Language Processing Advisory Committee of the National Academy of Sciences, which concluded with the comment that there was 'no immediate or predictable prospect of useful machine translation', funding for machine translation in the United States was drastically cut and interest in machine translation waned considerably (ALPAC 1966; Warwick 1987: 22–37). Still, sporadic efforts were made in machine translation. An important system was developed in the United States by Peter Toma, previously of Georgetown University, known as Systran, an acronym for System Translation. To this day, this system is still one of the most established and popular systems on the market. In Hong Kong, The Chinese University of Hong Kong set up the Hung On-To Research Laboratory for Machine Translation to conduct research into machine translation and developed a practical machine translation system known as 'The Chinese University Language Translator', abbreviated as CULT (Loh 1975: 143–155, 1976a: 46–50, 1976b: 104–05; Loh, Kong and Hung 1978: 111–120; Loh and Kong 1979: 135–148). In Canada, the TAUM group at Montreal developed a system for translating public weather forecasts known as TAUM-METEO, which became operative in 1977.

In the 1980s, the most important translation system developed was the EUROTRA system, which could translate all the official languages of the European Economic Community (Arnold and Tombe 1987: 1143–1145; Johnson, King and Tombe 1985: 155–169; King 1982; King 1987: 373–391; Lau 1988: 186–191; Maegaard 1988: 61–65; Maegaard and Perschke 1991: 73–82; Somers 1986: 129–177; Way, Crookston and Shelton 1997: 323–374). In 1983, Allen Tucker, Sergei Nirenburg, and others developed at Colgate University an AI-based multilingual machine translation system known as TRANSLATOR to translate four languages, namely English, Japanese, Russian, and Spanish. This was the beginning of knowledge-based machine translation in the United States (http://www.colgate.edu). The following year, Fujitsu produced ATLAS/I and ATLAS/II translation systems for translation between Japanese and English in Japan, while Hitachi and Market Intelligence Centre (QUICK) developed the ATHENE English–Japanese machine translation system (Chan 2009: 223). In 1985, the ArchTran machine translation system for translation between Chinese and English was launched in Taiwan and was one of the first commercialized English–Chinese machine translation systems in the world (Chen, Chang, Wang and Su 1993: 87–98). In the United States, the METAL (Mechanical Translation and Analysis of Languages) system for translation between English and German, supported by the Siemens Company in Munich since 1978 and developed at the University of Texas, Austin, became operative (Deprez, Adriaens, Depoortere and de Braekeleer 1994: 206–212; Lehmann, Bennett and Slocum *et al.* 1981; Lehrberger 1981; Little 1990: 94–107; Liu and Liro 1987: 205–218; Schneider 1992: 583–594; Slocum, Bennett, Bear, Morgan and Root 1987: 319–350; White 1987: 225–240). In China, the TranStar English–Chinese Machine Translation System, the first machine-translation product in China, developed by China National Computer Software and Technology Service Corporation, was commercially available in 1988 (http://www.transtar.com.cn). In Taiwan, the BehaviorTran, an English–Chinese machine translation system, was also launched in the same year.

In the 1990s, Saarbrucken in Germany formed the largest and the most established machine translation group in 1996. The SUSY (Saarbrücker Ubersetzungssystem/The Saarbrücken Machine Translation System) project for German to English and Russian to German machine translation was developed between 1972 and 1986 (rz.uni-sb.de). In 1997, *Dong Fang Kuai Che* 東方快車 (Orient Express), a machine translation system developed by the China Electronic Information Technology Ltd. in China, was commercially available (Chan 2004: 336) while in Taiwan, TransBridge was developed for internet translation from English into Chinese (http://

www.transbridge.com.tw). The first year of the twenty-first century witnessed the development of BULTRA (BULgarian TRAnslator), the first English–Bulgarian machine translation tool, by Pro Langs in Bulgaria (Chan 2004: 339).

What has been presented above shows very clearly that from the beginning of machine translation in 1947 until 1957, six countries were involved in the research and development of machine translation, which included Massachusetts Institute of Technology and Georgetown University in the United States in 1952, Academy of Sciences in Russia and Cambridge University in the United Kingdom in 1954, Kyushu University in Japan in 1955, the Institute of Linguistics in China in 1956, and Charles University in Czechoslovakia in 1957. By 2007, it was found that of the 193 countries in the world, 30 have conducted research on computer or computer-aided translation, 9 actively. This means that around 16 per cent of all the countries in the world have been engaged in machine translation, 30 per cent of which are active in research and development. The 31 countries which have been engaged in the research and development of machine translation are: Belgium, Brazil, Bulgaria, Canada, China, Czechoslovakia, Denmark, Finland, France, Germany, Hungary, India, Italy, Japan, Korea, Macau, Malaysia, Mexico, the Netherlands, Luxemburg, Poland, Russia, Singapore, Slovenia, Spain, Sweden, Switzerland, Taiwan, Tunisia, the United Kingdom, and the United States. Of these, the most active countries are China and Japan in Asia, France, Germany, the Netherlands, the United Kingdom, and Russia in Europe, and Canada and the United States in North America. The huge increase in the number of countries engaged in machine translation and the fast development of systems for different languages and language pairs show that machine translation has advanced by leaps and bounds in the last 65 years.

Conclusion

It should be noted that computer-aided translation has been growing rapidly in all parts of the world in the last 47 years since its inception in 1967. Drastic changes have taken place in the field of translation since the emergence of commercial computer-aided translation systems in the 1980s. In 1988, as mentioned above, we only had the Trados system that was produced in Europe. Now we have more than 100 systems developed in different countries, including Asian countries such as China, Japan, and India, and the northern American countries, Canada and the United States. In the 1980s, very few people had any ideas about computer-aided translation, let alone translation technology. Now, it is estimated that there are around 200,000 computer-aided translators in Europe, and more than 6,000 large corporations in the world handle their language problems with the use of corporate or global management computer-aided translation systems. At the beginning, computer-aided translation systems only had standalone editions. Now, there are over seventeen different types of systems on the market.

According to my research, the number of commercially available computer-aided translation systems from 1984 to 2012 is 86. Several observations on these systems can be made. First, about three computer-aided translation systems have been produced every year during the last 28 years. Second, because of the rapid changes in the market, nineteen computer-aided translation systems failed to survive in the keen competition, and the total number of current commercial systems stands at 67. Third, almost half of the computer-aided translation systems have been developed in Europe, accounting for 49.38 per cent, while 27.16 per cent of them have been produced in America.

All these figures show that translation technology has been on the fast track in the last five decades. It will certainly maintain its momentum for many years to come.

References

Allen, Ward (ed.) (1969) *Translating for King James*, Nashville, TN: Vanderbilt University Press.

ALPAC (Automatic Language Processing Advisory Committee) (1966) *Languages and Machines: Computers in Translation and Linguistics*, A Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC: National Academy of Sciences, National Research Council, 1966.

Arnold, Doug J. and Louis des Tombe (1987) 'Basic Theory and Methodology in EUROTRA', in Sergei Nirenburg (ed.) *Machine Translation: Theoretical and Methodological Issues*, Cambridge: Cambridge University Press, 114–135.

Arthern, Peter J. (1979) 'Machine Translation and Computerized Terminology Systems: A Translator's viewpoint', in Barbara M. Snell (ed.) *Translating and the Computer: Proceedings of a Seminar*, London: North-Holland Publishing Company, 77–108.

Brace, Colin (1992) 'Trados: Smarter Translation Software', *Language Industry Monitor* Issue September–October. Available at: http://www.lim.nl/monitor/trados-1.html.

Brace, Colin (1993) 'TM/2: Tips of the Iceberg', *Language Industry Monitor* Issue May–June. Available at: Retrieved from http://www.mt-archive.

Brace, Colin (1994) 'Bonjour, Eurolang Optimizer', *Language Industry Monitor* Issue March-April. Available at: http://www.lim.nl/monitor/optimizer.html.

Chan, Sin-wai (2004) *A Dictionary of Translation Technology*, Hong Kong: The Chinese University Press.

Chan, Sin-wai (2009) *A Chronology of Translation in China and the West*, Hong Kong: The Chinese University Press.

Chen, Gang (2001) 'A Review on Yaxin CAT2.5', *Chinese Science and Technology Translators Journal* 14(2).

Chen, Shuchuan, Chang Jing-shin, Wang Jong-nae, and Su Keh-yih (1993) 'ArchTran: A Corpus-based Statistics-oriented English-Chinese Machine Translation System', in Sergei Nirenburg (ed.) *Progress in Machine Translation*, Amsterdam: IOP Press, 87–98.

Deprez, F., Greert Adriaens, Bart Depoortere, and Gert de Braekeleer (1994) 'Experiences with METAL at the Belgian Ministry of the Interior', *Meta* 39(1): 206–212.

Delisle, Jean and Judith Woodsworth (eds) (1995) *Translators through History*, Amsterdam and Philadelphia: John Benjamins Publishing Company and UNESCO Publishing.

Dong, Zhendong (1988) 'MT Research in China', in Dan Maxwell, Klaus Schubert, and Toon Witkam (eds) *New Directions in Machine Translation*, Dordrecht, Holland: Foris Publications, 85–91.

Editorial Committee, *A Dictionary of Translators in China* 《中國翻譯家詞典》編寫組 (ed.) (1988) 《中國翻譯家詞典》 (*A Dictionary of Translators in China*), Beijing: China Translation and Publishing Corporation 中國對外翻譯出版公司.

Elita, Natalia and Monica Gavrila (2006) 'Enhancing Translation Memories with Semantic Knowledge', *Proceedings of the 1st Central European Student Conference in Linguistics*, 29-31 May 2006, Budapest, Hungary: 24–26.

Eurolux Computers (1992) 'Trados: Smarter Translation Software', *Language Industry Monitor* 11, September–October. Available at: http://www.lim.nl.

Even-Zohar, Itamar (1978) *Papers in Historical Poetics*, Tel Aviv: The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Feng, Zhiwei 馮志偉 (1999)〈中國的翻譯技術：過去、現在和將來〉(Translation Technology in China: Past, Present, and Future), in Huang Changning 黃昌寧 and Dong Zhendong 董振東 (eds) 《計算器語言學文集》(*Essays on Computational Linguistics*), Beijing: Tsinghua University Press, 335–440.

Gaiba, Francesca (1998) *The Origins of Simultaneous Interpretation: The Nuremberg Trial*, Ottawa: University of Ottawa Press.

Garcia, Ignacio and Vivian Stevenson (2005) 'Trados and the Evolution of Language Tools: The Rise of the De Facto TM Standard – And Its Future with SDL', *Multilingual Computing and Technology* 16(7).

Garcia, Ignacio and Vivian Stevenson (2006) 'Heartsome Translation Suite', *Multingual* 17(1): 77. Available at: http://www.multilingual.com.

German, Kathryn (2009) 'Across: An Exciting New Computer Assisted Translation Tool', *The Northwest Linguist* 9–10.

Gotti, Fabrizio, Philippe Langlais, Elliott Macklovitch, Didier Bourigault, Benoit Robichaud, and Claude Coulombe (2005) '3GTM: A Third-generation Translation Memory', *Proceedings of the 3rd Computational Linguistics in the North-East (CLiNE) Workshop,* Gatineau, Québec, Canada, 26–30.

Haeseryn, Rene (1989) 'The International Federation of Translators (FIT) and its Leading Role in the Translation Movement in the World', in Rene Haeseryn (ed.) *Roundtable Conference FIT-UNESCO: Problems of Translator in Africa*, Belgium: FIT, 379–384.

Hall, Amy (2000) 'SDL Announces Release of SDLX Version 2.0', SDL International. Available at: http://www.sdl.com/en/about-us/press/1999/SDL_Announces_Release_of_SDLX_Version_2_0.asp.

Harmsen, R. (2008) 'Evaluation of DVX'. Available at: http://rudhar.com.

Holmes, James S. (1972, 1987) 'The Name and Nature of Translation Studies', in Gideon Toury (ed.) *Translation across Cultures*, New Delhi: Bahri Publications: Pvt. Ltd., 9–24.

Holmes, James S. (1988) 'The Name and Nature of Translation Studies', in James S. Holmes (ed.) *Translated! Papers on Literary Translation and Translation Studies*, Amsterdam: University of Amsterdam, 93–98.

http://anaphraseus.sourceforge.net.
http://blog.memsource.com.
http://developer.apple.com.
http://en.wikipedia.org/wiki/XLIFF
http://felix-cat.com.
http://new.multilingual.com.
http://terminotix.com/news/newsletter.
http://www.across.net.
http://www.alchemysoftware.ie.
http://www.atasoft.com.
http://www.atril.com.
http://www.bas.bg.
http://www.colgate.edu.
http://www.cuni.cz.
http://www.dreye.com.tw.
http://www.gcys.cn.
http://www.heartsome.net.
http://www.hjtek.com.
http://www.hjtrans.com.
http://www.kilgray.com.
http://www.lisa.org.
http://www.logovista.co.jp.
http://www.maxprograms.com.
http://www.memsource.com.
http://www.metatexis.com.
http://www.multicorpora.com.
http://www.multilizer.com.
http://www.omegat.org.
http://www.opentm2.org.
http://www.otek.com.tw.
http://www.promt.com.
http://www.sdl.com.
http://www.sdlintl.com.
http://www.star-group.net.
http://www.systransoft.com.
http://www.thelanguagedirectory.com/translation/translation_software.
http://www.transbridge.com.tw.
http://www.translationzone.com.
http://www.transn.com.
http://www.transtar.com.cn.
http://www.tratool.com.
http://www.unam.mx.
http://www.urelitetech.com.cn.
http://www.wordbee.com.
http://www.wordfast.com.
http://wordfast.net/champollion.net.

http://www.xtm-intl.com.
http://www.zklr.com.
Hutchins, W. John (1986) *Machine Translation: Past, Present and Future*, Chichester: Ellis Horwood.
Hutchins, W. John (1998) 'The Origins of the Translator's Workstation', *Machine Translation* 13(4): 287–307.
Hutchins, W. John (1999) 'The Development and Use of Machine Translation System and Computer-based Translation Tools', in Chen Zhaoxiong (ed.) *Proceedings of the International Conference on MT and Computer Language Information Processing*, Beijing: Research Center of Computer and Language Engineering, Chinese Academy of Sciences, 1–16.
Hutchins, W. John (2000) *Early Years in Machine Translation*, Amsterdam and Philadelphia: John Benjamins.
Johnson, R.I., Margaret King, and Louis des Tombe (1985) 'Eurotra: A Multilingual System under Development', *Computational Linguistics* 11(2–3): 155–169.
Kavak, Pinar (2009) 'Development of a Translation Memory System for Turkish to English', Unpublished MA dissertation, Boğaziçi University, Turkey.
Kay, Martin (1980) 'The Proper Place of Men and Machines in Language Translation', Research Report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, CA.
Kelly, Louis G. (1998) 'Latin Tradition', in Mona Baker (ed.) *Routledge Encyclopedia of Translation Studies*, London and New York: Routledge, 495–504.
Kilgrary Translation Technologies (2011) 'What's New in MemoQ'. Available at: http://kilgray.com/products/memoq/whatsnew.
King, Margaret (1982) *EUROTRA: An Attempt to Achieve Multilingual MT*, Amsterdam: North-Holland.
King, Margaret (ed.) (1987) *Machine Translation Today: The State of the Art*, Edinburgh: Edinburgh University Press.
Környei, Tibor (2000) 'WordFisher for MS Word: An Alternative to Translation Memory Programs for Freelance Translators?' *Translation Journal* 4(1). Available at: http://accurapid.com/journal/11wf.htm.
Lau, Peter Behrendt (1988) 'Eurotra: Past, Present and Future', in Catriona Picken (ed.) *Translating and the Computer 9: Potential and Practice*, London: The Association for Information Management, 186–91.
Lehmann, Winfred P., Winfield S. Bennett, Jonathan Slocum *et al.* (1981) *The METAL System*, New York: Griffiss Air Force Base.
Lehrberger, John (1981) *The Linguistic Model: General Aspects*, Montreal: TAUM Group, University of Montreal.
LISA (2010) 'IBM and the Localization Industry Standards Association Partner to Deliver Open-Source Enterprise-level Translation Tools'. Available at: http://www.lisa.org/OpenTM2.1557.0.html.
Little, Patrick (1990) 'METAL – Machine Translation in Practice', in Catriona Picken (ed.) *Translation and the Computer 11: Preparing for the Next Decade*, London: The Association for Information Management, 94–107.
Liu, Jocelyn and Joseph Liro (1987) 'The METAL English-to-German System: First Progress Report', *Computers and Translation* 2(4): 205–218.
Liu, Yongquan *et al.* 劉湧泉等 (1984) 《中國的機器翻譯》(*Machine Translation in China*), Shanghai: Knowledge Press.
Locke, Nancy A and Marc-Olivier Giguère (2002) 'MultiTrans 3.0', *MultiLingual Computing and Technology* 13(7): 51.
Locke, William Nash and Andrew Donald Booth (eds) (1955) *Machine Translation of Languages: Fourteen Essays*, Cambridge, MA: MIT Press.
Loh, Shiu-chang (1975) 'Machine-aided Translation from Chinese to English', *United College Journal* 12(13): 143–155.
Loh, Shiu-chang (1976a) 'CULT: Chinese University Language Translator', *American Journal of Computational Linguistics* Microfiche, 46, 46–50.
Loh, Shiu-chang (1976b) 'Translation of Three Chinese Scientific Texts into English by Computer', *ALLC Bulletin* 4(2): 104–05.
Loh, Shiu-chang, Kong Luan, and Hung Hing-sum (1978) 'Machine Translation of Chinese Mathematical Articles', *ALLC Bulltein* 6(2): 111–120.
Loh, Shiu-chang and Kong Luan (1979) 'An Interactive On-line Machine Translation System (Chinese into English)', in Barbara M. Snell (ed.) *Translating and the Computer*, Amsterdam: North-Holland, 135–148.
Maegaard, Bente (1988) 'EUROTRA: The Machine Translation Project of the European Communities', *Literary and Linguistic Computing* 3(2): 61–65.

Maegaard, Bente and Sergei Perschke (1991) 'Eurotra: General Systems Design', *Machine Translation* 6(2): 73–82.
Melby, Alan K. (1978) 'Design and Implementation of a Machine-assisted Translation System', *Proceedings of the 7th International Conference on Computational Linguistics*, 14–18 August 1978, Bergen, Norway.
Melby, Alan K. and Terry C. Warner (1995) *The Possibility of Language: A Discussion of the Nature of Language, with Implications for Human and Machine Translation*, Amsterdam and Philadelphia: John Benjamins.
MultiCorpora Inc. (2011) 'MultiCorpora Launches New Translation Management System'. Available at: http://www.multicorpora.com/news/multicorpora-launches-new-translation-management-system.
MultiLingual (1997) 'CIMOS Releases Arabic to English Translation Software', *MultiLingual* 20 December 1997. Available at: http://multilingual.com/newsDetail.php?id=422
MultiLingual (1998) 'SDL Announces Translation Tools', *MultiLingual* 23 September 1998. Available at: http://multilingual.com/newsDetail.php?id=154.
MultiLingual (1999) 'SDL Announces SDL Workbench and Product Marketing Executive', *MultiLingual* 22 Feburary 1999. Available at: http://multilingual.com/newsDetail.php?id=12.
MultiLingual (2003) 'MultiCorpora R&D Releases MultiTrans 3.5', *MultiLingal* 17 October 2003. Available at: http://multilingual.com/newsDetail.php?id=3219.
MultiLingual (2005a). 'STAR Releases Transit Service Pack 14', *MultiLingual* 15 April 2005. Available at: http://multilingual.com/newsDetail.php?id=4169.
MultiLingual (2005b) 'SIMILIS Version 1.4 Released', *MultiLingual* 27 April 2005. Available at: http://multilingual.com/newsDetail.php?id=4187.
MultiLingual (2005c) 'SDL Releases SDLX 2005', *MultiLingual* 5 May 2005. Available at: http://multilingual.com/newsDetail.php?id=4216.
MultiLingual (2005d) 'MultiCorpora Announces the Release of MultiTrans 4', *MultiLingual* 31 August 2005. Available at: http://multilingual.com/newsDetail.php?id=4425.
MultiLingual (2006a) 'Across Rolls out New Version 3.5', *MultiLingual* 20 November 2006. Available at: http://multilingual.com/newsDetail.php?id=5372.
MultiLingual (2006b) 'Lingotek Announces Beta Launch of Language Search Engine', *MultiLingual* 15 August 2006. Available at: http://multilingual.com/newsDetail.php?id=5168.
MultiLingual (2007) 'Kilgray Releases Version 2.0 of MemoQ', *MultiLingual* 25 January 2007. Available at: http://multilingual.com/newsDetail.php?id=5467.
MultiLingual (2008a) 'Across Language Server 4.0 SP1', *MultiLingual* 21 April 2008. Available at: http://multilingual.com/newsDetail.php?id=6228.
MultiLingual (2008b) 'Fusion One and Fusion Collaborate 3.0', *MultiLingual* 28 November 2008. Available at: http://multilingual.com/newsDetail.php?id=6568.
MultiLingual (2009a) 'Fusion 3.1', *MultiLingual* 19 March 2009. Available at: http://multilingual.com/newsDetail.php?id=6734.
MultiLingual (2009b) 'Across Language Server V.5', *MultiLingual* 13 May 2009. Available at: http://multilingual.com/newsDetail.php?id=6834.
MultiLingual (2009c) 'Lingotek Launches Crowdsourcing Translation Platform', *MultiLingual* 22 October 2009. Available at: http://multilingual.com/newsDetail.php?id=7103.
MultiLingual (2010a) 'SDL Trados Studio', *MultiLingual* 3 March 2010. Available at: http://multilingual.com/newsDetail.php?id=7298.
MultiLingual (2010b) 'Collaborative Translation Platform 5.0', *MultiLingual* 27 July 2010. Available at: http://multilingual.com/newsDetail.php?id=7544.
MultiLingual (2011) 'Déjà Vu X2', *MultiLingual* 24 May 2011. Available at: http://multilingual.com/newsDetail.php?id=933.
Nagao, Makoto (1993) 'Machine Translation: The Japanese Experience', in Sergei Nirenburg (ed.) *Progress in Machine Translation*, Amsterdam: IOS Press, 203–208.
Nida, Eugene A. (1964) *Toward a Science of Translating*, Leiden: E.J. Brill.
Phelan, Mary (2001) *The Interpreter's Resource*, Clevedon: Multilingual Matters Ltd.
Prior, Marc (2003) 'Close Windows, Open Doors', *Translation Journal* 7(1). Available at: http://accurapid.com/journal/23linux.htm.
Schmidt, Axel (2006) 'Integrating Localization into the Software Development Process', *TC World* March 2006.
Schneider, Thomas (1992) 'User Driven Development: METAL as an Integrated Multilingual System', *Meta* 37(4): 583–594.

Shannon, Claude L. and Warren Weaver (1949) *The Mathematical Theory of Communication*, Urbana, IL: University of Illinois Press.

Slocum, Jonathan, Winfield S. Bennett, J. Bear, M. Morgan, and Rebecca Root (1987) 'METAL: The LRC Machine Translation System', in Margaret King (ed.) *Machine Translation Today: The State of the Art*, Edinburgh: Edinburgh University Press, 319–350.

Somers, Harold L. (1986) 'Eurotra Special Issue', *Multilingual* 5(3): 129–177.

Sumita, Eiichiro and Yutaka Tsutsumi (1988) 'A Translation Aid System Using Flexible Text Retrieval Based on Syntax-matching', in *Proceedings of the 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Pittsburgh, Pennsylvania: Carnegie Mellon University. Available online at: http:// www.mt-archive. info/TMI-1988-Sumita.pdf.

Wang, Zheng 王正 (2011) 〈翻譯記憶系統的發展歷程與未來趨勢〉(Translation Memory Systems: A Historical Sketch and Future Trends),《編譯論叢》(*Compilation and Translation Review*) 4(1): 133–160.

Warwick, Susan (1987) 'An Overview of Post-ALPAC Developments', in Margaret King (ed.) *Machine Translation Today: The State of the Art*, Edinburgh: Edinburgh University Press, 22–37.

Wassmer, Thomas (2003) 'SDLX TM Translation Suite 2003', *Translation Journal* 7(3).

Wassmer, Thomas (2004) 'Trados 6.5', *MultiLingual Computing and Technology* 15(1): 61.

Wassmer, Thomas (2007) 'Comparative Review of Four Localization Tools: Déjà Vu, MULTILIZER, MultiTrans and TRANS Suite 2000, and Their Various Capabilities', *MultiLingual Computing and Technology* 14(3): 37–42.

Wassmer, Thomas (2011) 'Dr Tom's Independent Software Reviews'. Available at: http://www.localizationworks.com/DRTOM/Trados/TRADOS.

Way, Andrew, Ian Crookston, and Jane Shelton (1997) 'A Typology of Translation Problems for Eurotra Translation Machines', *Machine Translation* 12(4): 323–374.

White, John S. (1987) 'The Research Environment in the METAL Project', in Sergei Nirenburg (ed.) *Machine Translation: Theoretical and Methodological Issues*, Cambridge: Cambridge University Press, 225–240.

Wiener, Norbert (1954) *The Human Use of Human Beings: Cybernetics and Society*, New York: Houghton Mifflin.

Wilks, Yorick (2000) 'Magaret Masterman', in W. John Hutchins (ed.) *Early Years in Machine Translation*, Amsterdam and Philadelphia: John Benjamins Publishing Company, 279–298.

Worth, Roland H. (1992) *Bible Translations: A History through Source Documents*, Jefferson, NC, and London: McFarland and Company, Inc., Publishers.

Xu, Jie (2001) 'Five Amazing Functions of Dr Eye 2001' (Dr Eye 2001 譯典通 5 大非凡功能),《廣東電腦與電訊》(*Computer and Telecom*) (3).

Yngve, Victor H. (2000) 'Early Research at M.I.T. in Search of Adequate Theory', in W. John Hutchins (ed.) *Early Years in Machine Translation*, Amsterdam and Philadelphia: John Benjamins, 39–72.

Zhang, Zheng 張政 (2006)《計算機翻譯研究》(*Studies on Machine Translation*), Beijing: Tsinghua University Press 清華大學出版社.

2

COMPUTER-AIDED TRANSLATION

Major concepts

Chan Sin-wai

THE CHINESE UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Introduction

When the term computer-aided translation is mentioned, we often associate it with the functions a computer-aided translation system offers, such as toolbars, icons, and hotkeys, the built-in tools we can use, such as online dictionaries, browsers, and the computational hitches we often encounter when working on a computer-aided translation project, such as chaotic codes. What is more important is to see beyond the surface of computer-aided translation to find out the major concepts that shape the development of functions in translation technology.

Concepts, which are relatively stable, govern or affect the way functions are designed and developed, while functions, which are fast-changing, realize the concepts through the tasks they perform. As a major goal of machine translation is to help human translators, a number of functions in computer-aided translation systems have been created to enable machine processing of the source with minimum human intervention. Concepts, moreover, are related to what translators want to achieve in translating. Simply put, translators want to have a controllable (*controllability*) and customizable (*customizability*) system, which is compatible with file formats (*compatibility*) and language requirements, and behaves as well as (*simulativity*) or even better than (*emulativity*) a human translator, to allow them to work together (*collaborativity*) to produce quality translations (*productivity*). We have therefore identified seven major concepts which are important in computer-aided translation: simulativity, emulativity, productivity, compatibility, controllability, customizability, and collaborativity. The order in which concepts are arranged can be memorized more easily by their acronym SEPCCCC.

Simulativity

The first concept of computer-aided translation is simulativity, which is about the way in which a computer-aided translation system models the behaviour of a human translator by means of its functions, such as the use of concordancers in text analysis to model after comprehension on the part of the human translator and the creation of a number of quality assurance tools to follow the way checking is done by a human translator.

There are a number of ways to illustrate man–machine simulativity.

(1) Goal of translation

The first is about the ultimate goal of translation technology. All forms of translation (machine translation, computer-aided translation and human translation) aim at obtaining high-quality translations. In the case of machine translation, the goal of a fully automatic high-quality translation (FAHQT) is to be achieved through the use of a machine translation system without human intervention. In the case of computer-aided translation, the same goal is to be achieved through a computer-aided translation system that simulates the behaviour of a human translator through man–machine interaction.

(2) Translation procedure

A comparison of the procedures of human translation with those of computer-aided translation shows that the latter simulates the former in a number of ways. In manual translation, various translation procedures have been proposed by translation scholars and practitioners, ranging from the two-stage models to eight-stage ones, depending on the text type and purposes of translation. In machine translation and computer-aided translation, the process is known as technology-oriented translation procedure.

(a) Two-stage model

In human translation, the first type of translation procedure is a two-stage one, which includes the stage of source text comprehension and the stage of target text formulation, as shown below:

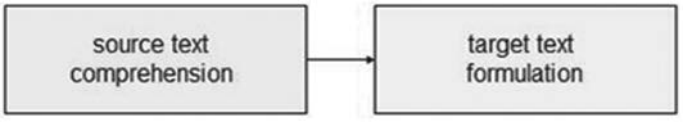


Figure 2.1 A two-stage model for human translation

Figure 2.1 is a model for human translators with the ability of comprehension. As a computer-aided translation system does not have the ability of comprehension, it cannot model after human translation with this two-stage model. It can, however, work on a two-stage translation with the use of its system dictionary, particularly in the case of a language-pair-specific system, as in Figure 2.2:

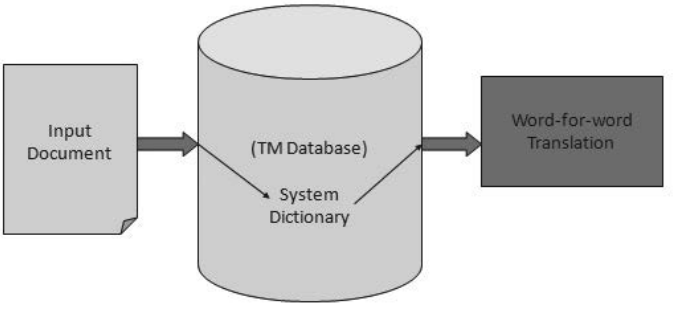


Figure 2.2 A two-stage dictionary-based language-pair-specific model

Another two-stage model of computer-aided translation is a terminology-based system, as shown in Figure 2.3:

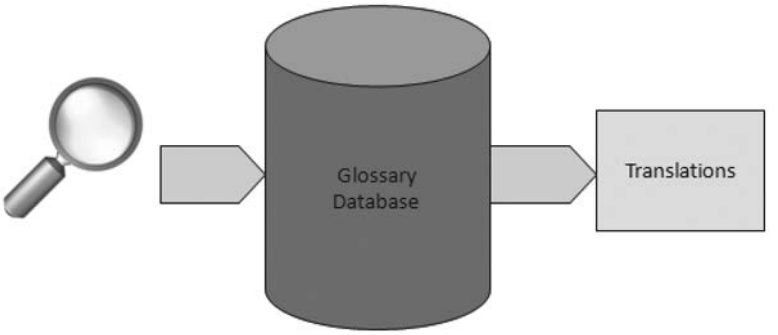


Figure 2.3 A two-stage terminology-based CAT system

(b) Three-stage models

The second type of translation procedure is a three-stage model. This section covers five variations of this model proposed by Eugene Nida and Charles Taber (1969), Wolfram Wilss (1982), Roger Bell (1991), Basil Hatim and Ian Mason, and Jean Delisle (1988) respectively. A three-stage example-based computer-aided translation system is shown to illustrate the simulation of human translation by computer-aided translation.

(i) MODEL BY EUGENE NIDA AND CHARLES TABER

The first model of a three-stage translation procedure involving the three phases of analysis, transfer, and restructuring was proposed by Eugene Nida and Charles Taber ([1969] 1982: 104). They intended to apply elements of Chomsky's transformational grammar to provide Bible translators with some guidelines when they translate ancient source texts into modern target texts, which are drastically different in languages and structures. Nida and Taber describe this three-stage model as a translation procedure in which

> the translator first analyses the message of the source language into its simplest and structurally clearest forms, transfers it at this level, and then restructures it to the level in the receptor language which is most appropriate for the audience which he intends to reach.
>
> (Nida and Taber [1969] 1982: 484)

Analysis is described by these two scholars as 'the set of procedures, including back transformation and componential analysis, which aim at discovering the kernels underlying the source text and the clearest understanding of the meaning, in preparation for the transfer' (Nida and Taber [1969] 1982: 197). Transfer, on the other hand, is described as the second stage 'in which the analysed material is transferred in the mind of the translator from language A to language B' (ibid.: 104). Restructuring is the final stage in which the results of the transfer process are transformed into a 'stylistic form appropriate to the receptor language and to the intended receptors'.

In short, analysis, the first stage, is to analyse the source text, transfer, the second stage, is to transfer the meaning, and restructuring, the final stage, is to produce the target text. Their model is shown in Figure 2.4.

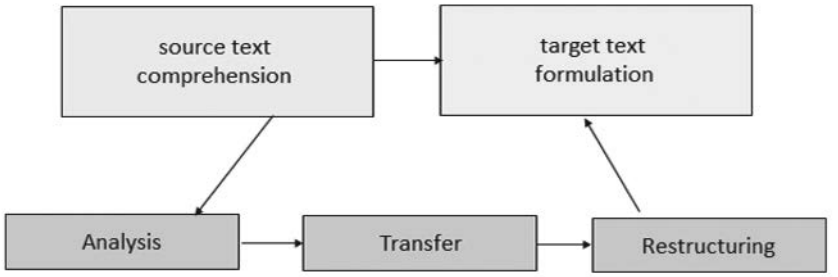


Figure 2.4 Three-stage model by Nida and Taber (1964)

(ii) MODEL BY WOLFRAM WILSS

The second three-stage model was proposed by Wolfram Wilss (1982) who regards translation procedure as a linguistic process of decoding, transfer and encoding. His model is shown in Figure 2.5.

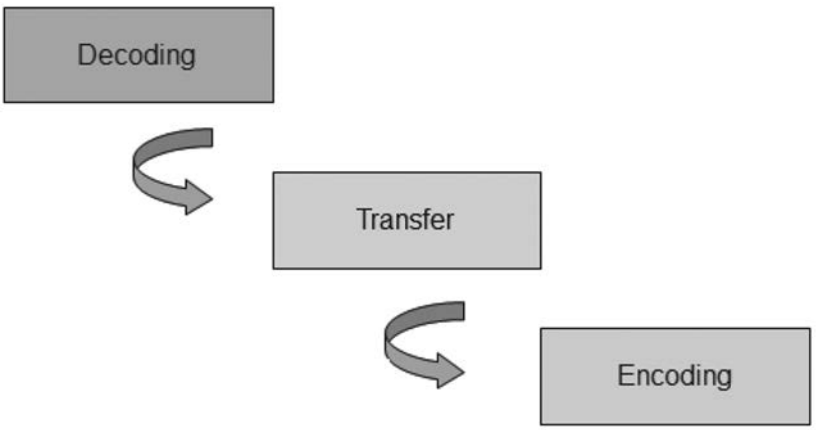


Figure 2.5 Three-stage model by Wolfram Wilss (1982)

(iii) MODEL BY ROGER BELL

Another three-stage model of note is by Roger Bell whose translation procedure framework is divided into three phases: the first phase is source text interpretation and analysis, the second, translation process, and the third, text reformulation (see Figure 2.6). The last phase takes into consideration three factors: writer's intention, reader's expectation, and the target language norms (Bell 1991).

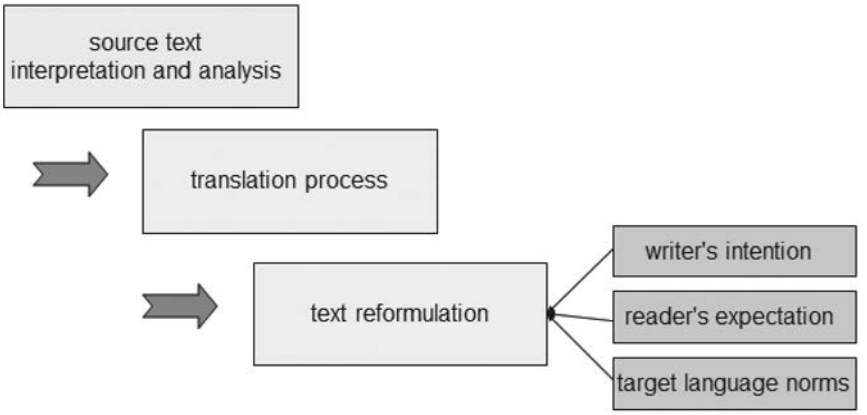


Figure 2.6 Model of Roger Bell

(iv) MODEL BY BASIL HATIM AND IAN MASON

This model, proposed by Basil Hatim and Ian Mason, is a more sophisticated three-stage model, which involves the three steps of source text comprehension, transfer of meaning, and target text assessment (see Figure 2.7). At the source text comprehension level, text parsing, specialized knowledge, and intended meaning are examined. At the meaning transfer stage, consideration is given to the lexical meaning, grammatical meaning, and rhetorical meaning. At the target text assessment level, attention is paid to text readability, target language conventions, and the adequacy of purpose.

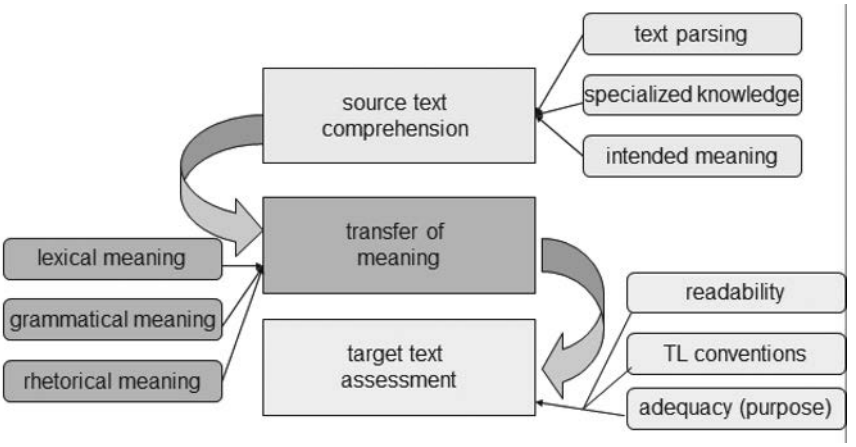


Figure 2.7 A three-stage model by Basil Hatim and Ian Mason

(v) MODEL OF JEAN DELISLE

The fourth model of a three-stage translation procedure was proposed by Jean Delisle (1988: 53–69) (see Figure 2.8). Deslisle believes that there are three stages in the development of a translation equivalence: comprehension, reformulation, and verification: 'comprehension is

based on decoding linguistic signs and grasping meaning, reformulation is a matter of reasoning by analogy and re-wording concepts, and verification involves back-interpreting and choosing a solution' (1988: 53).

Parallel to human translation, a three-stage model in computer-aided translation is the example-based system. The input text goes through the translation memory database and glossary database to generate fuzzy matches and translations of terms before getting the target text. The procedure of an example-based computer-aided translation system is shown in Figure 2.9.

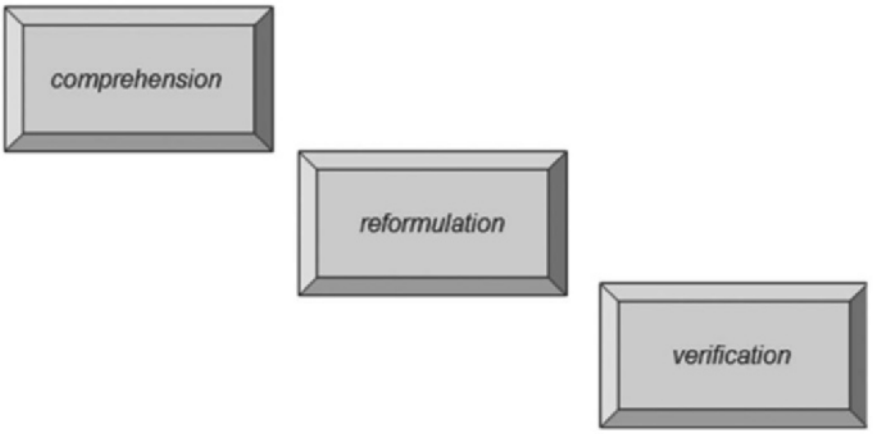


Figure 2.8 A three-stage model of Jean Delisle

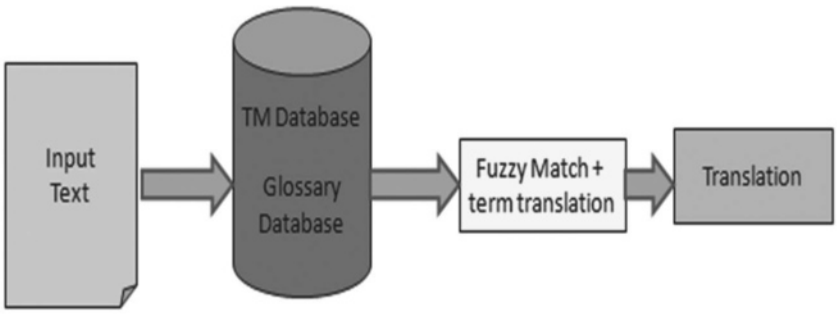


Figure 2.9 Three-stage example-based computer-aided translation model

(c) Four-stage model

The third type of translation procedure is a four-stage one. A typical example is given by George Steiner ([1975] 1992) who believes that the four stages of translation procedure are: knowledge of the author's times, familiarization with author's sphere of sensibility, original text decoding, and target text encoding (see Figure 2.10).

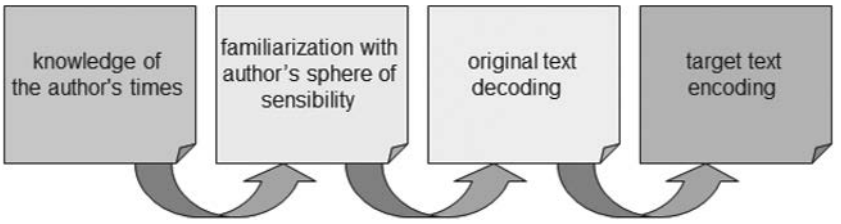


Figure 2.10 Model of George Steiner (1975)

For computer-aided translation, a four-stage model is exemplified by webpage translation provided by Yaxin. The first stage is to input the Chinese webpage, the second stage is to process the webpage with the multilingual maintenance platform, the third stage is to process it with the terminology database, and the final stage is to generate a bilingual webpage. The Yaxin translation procedure is shown in Figure 2.11.

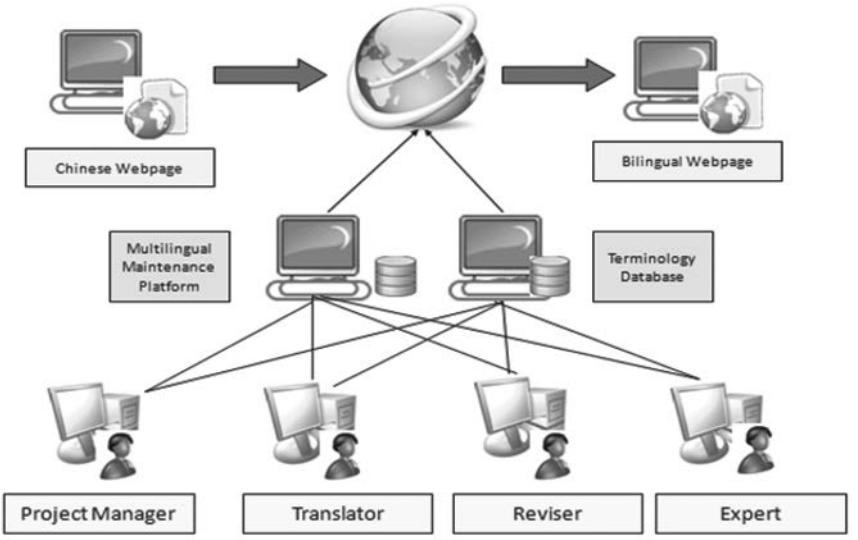


Figure 2.11 Yaxin's four-stage procedure

(d) Five-stage model

The fourth type of translation procedure is a five-stage one, as proposed by Omar Sheikh Al-Shabab (1996: 52) (see Figure 2.12). The first stage is to edit the source text, the second, interpret the source text, the third, interpret it in a new language, the fourth, formulate the translated text, and the fifth, edit the formulation.

In computer-aided translation, a five-stage model is normally practised. At the first stage, the Initiating Stage, tasks such as setting computer specifications, logging in a system, creating a profile, and creating a project file are performed. At the second stage, the Data Preparation Stage, the tasks involve data collection, data creation, and the creation of terminology and translation memory databases. At the third stage, the Data Processing Stage, the tasks include data analysis, the use of system and non-system dictionaries, the use of concordancers, doing

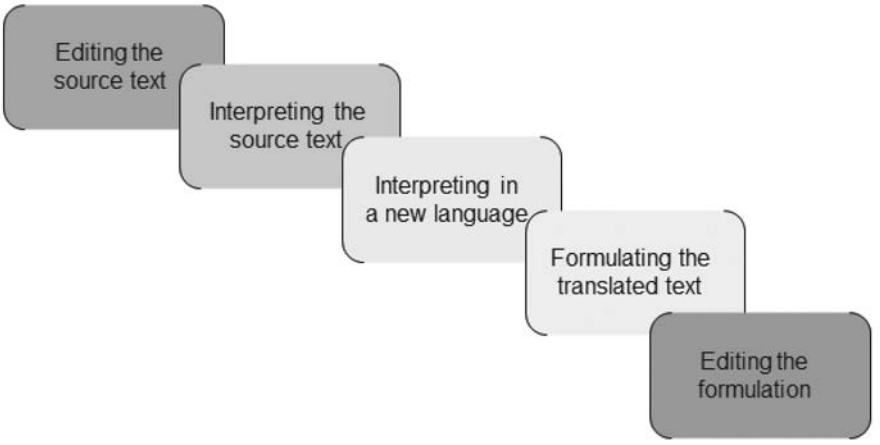


Figure 2.12 Model of Omar Sheikh Al-Shabab

pre-translation, data processing by translation by computer-aided translation systems with human intervention, or by machine translation systems without human intervention, or data processing by localization systems. At the fourth stage, the Data Editing Stage, the work is divided into two types. One type is data editing for computer-aided translation systems, which is about interactive editing, the editing environments, matching, and methods used in computer-aided translation. Another type is data editing for computer translation systems, which is about post-editing and the methods used in human translation. At the last or fifth stage, the Finalizing Stage, the work is mainly updating databases.

The fives stages in computer-aided translation are illustrated in Figure 2.13.

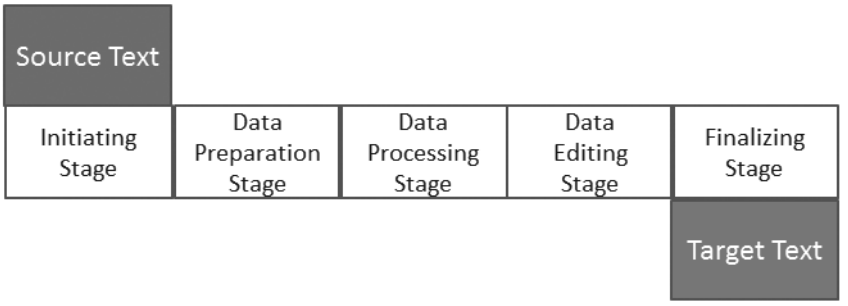


Figure 2.13 Five-stage technology-oriented translation procedure model

It can be seen that though there are both five-stage models in human translation and computer-aided translation and the tasks involved are different, the concept of simulativity is at work at almost all stages.

(e) Eight-stage model

The fifth type of translation procedure is an eight-stage one, as proposed by Robert Bly (1983).

Robert Bly, who is a poet, suggests an eight-stage procedure for the translation of poetry: (a) set down a literal version; (b) find out the meaning of the poem; (c) make it sound like English; (d) make it sound like American; (e) catch the mood of the poem; (f) pay attention to sound; (g) ask a native speaker to go over the version; and (h) make a final draft with some adjustments (see Figure 2.14).

Figure 2.14 Model by Robert Bly (1983)

In computer-aided translation, there is no eight-stage model. But other than the five-stage model, there is also a seven-stage model, which is shown in Figure 2.15.

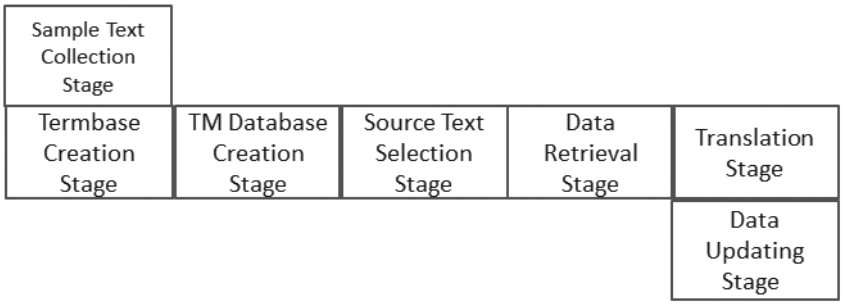


Figure 2.15 Seven-stage computer-aided translation procedure

The seven stages of computer-aided translation go from sample text collection to termbase creation, translation memory database creation, source text selection, data retrieval, source text translation and finally data updating.

All in all, we can say that when compared to human translation, computer-aided translation is simulative, following some of the stages in human translation.

Emulativity

There are obviously some functions which are performable by a computer-aided translation system, but not by a human translation. This is how technology can emulate human translation. Computer-aided translation, with the help of machine translation, simulates human translation, and it also emulates human translation in a number of areas of computer-aided translation, some of which are mentioned below.

Alt-tag translation

This function of machine translation allows the user to understand the meaning of text embedded within images (Joy 2002). The images on a web site are created by IMG tag (inline image graphic tag), and the text that provides an alternative message to viewers who cannot see the graphics is known as ALT-tag, which stands for 'alternative text'. Adding an appropriate ALT-tag to every image within one's web site will make a huge difference to its accessibility. As translators, our concern is the translation of the alternative text, as images are not to be translated anyway.

Chatroom translation

Machine translation has the function to translate the contents of a chatroom, known as 'chat translation' or 'chatroom translation'. Chat translation systems are commercially available for the translation of the contents of the Chatroom on the computer. As a chat is part of conversational discourse, all the theoretical and practical issues relating to conversational discourse can be applied to the study of chat translation. It should be noted that this kind of online jargon and addressivity are drastically different from what we have in other modes of communication.

The function of Chatroom is available in some systems, such as Fluency, as one of the resources. This function has to be purchased and enabled in the Fluency Chat Server to allow clients to be connected to this closed system for internal communications. For standalone version users, the function of Chat will be provided by Fluency Chat Server provided by its company, the Western Standard (Western Standard 2011: 39).

Clipboard translation

This is to copy a text to the clipboard from any Windows application for a machine translation system to translate the clipboard text, and the translated text can then be pasted in the original or any other location. One of the systems that translate clipboards is Atlas.

Conversion between metric and British systems

A function that can be easily handled by machine translation but not so easily by human translation is the conversion of weight, volume, length, or temperature from metric to British or vice versa. Fluency, for example, can do the metric/British conversion the target text box with the converted units.

Currency conversion

Some computer-aided translation systems can do currency conversion. With the use of Currency Converter, a function in Fluency, and access to the Internet to get the currency conversion rates, systems can convert a currency in a country into the local country currency. The number of currencies that can be handled by a system is relatively large. Fluency, for example, supports the conversion of currencies of around 220 countries. The conversion of multiple currencies is also supported.

Email translation

This refers to the translation of emails by a machine translation system (Matsuda and Kumai 1999; Rooke 1985: 105–115). The first online and real-time email translation was made in 1994 by the CompuServe service which provided translation service of emails from and to English and French, German or Spanish. Email translation has since become a very important part of daily communication and most web translation tools have email translators to translate emails. As emails are usually conversational and often written in an informal or even ungrammatical way, they are difficult for mechanical processing (Fais and Ogura 2001; Han, Gates and Levin 2006). One of the systems that translates emails is Atlas.

Foreign language translation

One of the most important purposes of using translation software is to translate a source text the language of which is unfamiliar to the user so as to explain its contents in a language familiar to the user. It is found that a majority of the commercial machine translation systems are for translation among Indo-European languages or major languages with a large number of speakers or users. Software for translation between major languages and minor languages are relatively small in number.

Gist translation

Another area where machine translation differs fundamentally from human translation is gist translation, which refers to a translation output which expresses only a condensed version of the source text message. This type of rough translation is to get some essential information about what is in the text for a user to decide whether to translate it in full or not to serve some specific purposes.

Highlight and translate

This function allows the user to highlight a part of the text and translate it into the designated language. The highlighted text is translated on its own without affecting the rest of the text.

Instant transliteration

This refers to a function of machine translation which can transliterate the words of a text with a certain romanization system. In the case of Chinese, the Hanyu Pinyin Romanization system for simplified characters is used in mainland China, while the Wade-Giles Romanization system for traditional characters is used in Taiwan.

Mouse translation

This is to translate sentences on a web page or on applications by simply clicking the mouse. Systems that provide mouse translation include Atlas.

Online translation

This is the translation of a text by an online machine translation system which is available at all times on demand from users. With the use of online translation service, the functions of information assimilation, message dissemination, language communication, translation entertainment, and language learning can be achieved.

Pre-translation

Machine translation is taken to be pre-translation in two respects. The first is as a kind of preparatory work on the texts to be translated, including the checking of spelling, the compilation of dictionaries, and the adjustment of text format. The second is taken to be a draft translation of the source text which can be further revised by a human translator.

Sentence translation

Unlike human translation which works at the textual level, machine translation is sentential translation. In other words, machine translation is a sentence-by-sentence translation. This type of translation facilitates the work of post-editing and methods which are frequently used in translating sentences in translation practice to produce effective translations can be used to produce good translations from machine translation systems.

Web translation

This refers to the translation of information on a web page from one language into another. Web-translation tools are a type of translation tools which translate information on a web page from one language into another. They serve three functions: (1) as an assimilation tool to transmit information to the user; (2) as a dissemination tool to make messages comprehensible; and (3) as a communication tool to enable communication between people with different language backgrounds.

Productivity

As translation technology is a field of entrepreneurial humanities, productivity is of great importance. Productivity in computer-aided translation is achieved through the use of technology, collective translation, recycling translations, reusing translations, professional competence, profit-seeking, labour-saving, and cost-saving.

Using technology to increase productivity

The use of technology to increase productivity needs no explanation. As early as 1980, when Martin Kay discussed the proper place of men and machines in language translation, he said:

> Translation is a fine and exacting art, but there is much about it that is mechanical and routine and, if this were given over to a machine, the productivity of the translator would not only be magnified but his work would become more rewarding, more exciting, more human.
>
> (Kay 1980: 1)

All computer-aided translation systems aim to increase translation productivity. In terms of the means of production, all translation nowadays is computer-aided translation as virtually no one could translate without using a computer.

Collective translation to increase productivity

Gone are the days when bilingual competence, pen and paper, and printed dictionaries made a translator. Gone are the days when a single translator did a long translation project all by himself. It is true that in the past, translation was mainly done singly and individually. Translation was also done in a leisurely manner. At present, translation is done largely through team work linked by a server-based computer-aided translation system. In other words, translation is done in a collective manner.

Recycling translations to increase productivity

To recycle a translation in computer-aided translation is to use exact matches automatically extracted from a translation memory database. To increase productivity, the practice of recycling translations is followed in computer-aided translation. Networked computer-aided translation systems are used to store centralized translation data, which are created by and distributed among translators. As this is the case, translators do not have to produce their own translations. They can simply draw from and make use of the translations stored in the bilingual database to form their translation of the source text. Translation is therefore produced by selection.

Reusing translations to increase productivity

To reuse a translation in computer-aided translation is to appropriate terms and expressions stored in a term database and translation memory database. It should be noted that while in literary translation, translators produce translations in a creative manner, translators in practical translation reuse and recycle translations as the original texts are often repetitive. In the present age, over 90 per cent of translation work is in the area of practical translation. Computer-aided translation is ideal for the translation of repetitive practical writings. Translators do not have to translate the sentences they have translated before. The more they translate, the less they have to translate. Computer-aided translation therefore reduces the amount a translator needs to translate by eliminating duplicate work. Some systems, such as Across, allow the user to automatically reuse existing translations from the translation memory. It can be seen that 'reduce, reuse, recycle' are the three effective ways of increasing profitability (de Ilarraza, Mayor and Sarasola 2000).

Professional competence to increase productivity

Translators have to work with the help of translation technology. The use of computer-aided translation tools has actually been extended to almost every type of translation work. Computer-aided translation tools are aimed at supporting translators and not at replacing them. They make sure that translation quality is maintained as 'all output is human input'. As far as the use of tools is concerned, professional translation is technological. In the past, translators used only printed dictionaries and references. Nowadays, translators use electronic concordancers, speech technology, online terminology systems, and automatic checkers. Translation is about the use of a workbench or workstation in translation work.

Translation competence or knowledge and skills in languages are not enough today. It is more realistic to talk about professional competence, which includes linguistic competence, cultural competence, translation competence, translator competence, and technological competence. Professional competence is important for translators as it affects their career development. A remark made by Timothy Hunt is worth noting: 'Computers will never replace translators, but translators who use computers will replace translators who don't' (Sofer 2009: 88). What has happened in the field of translation technology shows that Hunt's remark may not be far off the mark. In the 1980s, very few people had any ideas about translation technology or computer-aided translation. Now, SDL alone has more than 180,000 computer-aided translators. The total number of computer-aided translators in the world is likely to be several times higher than the SDL translators.

Profit-seeking to increase productivity

Translation is in part vocational, in part academic. In the training of translators, there are courses on translation skills to foster their professionalism, and there are courses on translation theories to enhance their academic knowledge. But there are very few courses on translation as a business or as an industry. It should be noted that translation in recent decades has increasingly become a field of entrepreneurial humanities as a result of the creation of the project management function in computer-aided translation systems. This means translation is now a field of humanities which is entrepreneurial in nature. Translation as a commercial activity has to increase productivity to make more profits.

Labour-saving to increase productivity

Computer-aided translation systems help to increase productivity and profits through labour-saving, eliminating repetitive translation tasks. Through reusing past translations, an enormous amount of labour is saved. Computer-aided translation tools support translators by freeing them from boring work and letting them concentrate on what they can do best over machines, i.e. handling semantics and pragmatics. Generally, this leads to a broader acceptance by translators. The role of a translator, therefore, has changed drastically in the modern age of digital communication. Rather than simply translating the document, a computer-aided translator has to engage in other types of work, such as authoring, pre-editing, interactive editing, post-editing, term database management, translation memory database management, text alignment and manual alignment verification. It is estimated that with the use of translation technology, the work that was originally borne by six translators can be taken up by just one.

Cost-saving to increase productivity

Computer-aided translation is also cost-saving. It helps to keep the overhead cost down as what has been translated needs not to be translated again. It helps to improve budget planning.

Other issues relating to cost should also be taken into account. First, the actual cost of the tool and its periodic upgrades. Second, the licensing policy of the system, which is about the ease of transferring licences between computers or servers, the incurring of extra charges for client licences, the lending of licences to one's vendors, freelances, and the eligibility for free upgrades. Third, the cost that is required for support, maintenance, or training. Fourth, the affordability of the system for one's translators. Fifth, the user-friendliness of the system to one's computer technicians and translators, which affects the cost of production.

Compatibility

The concept of compatibility in translation technology must be considered in terms of file formats, operating systems, intersystem formats, translation memory databases, terminology databases, and the languages supported by different systems.

Compatibility of file formats

One of the most important concepts in translation technology is the type of data that needs to be processed, which is indicated by its format, being shown by one or several letters at the end of a filename. Filename extensions usually follow a period (dot) and indicate the type of information stored in the file. A look at some of the common file types and their file extensions shows that in translation technology, text translation is but one type of data processing, though it is the most popular one.

There are two major types of formats: general documentation types and software development types.

(I) General documentation types

(1) Text files

All computer-aided translation systems which use Microsoft Word as text editor can process all formats recognized by Microsoft Word. Throughout the development of translation technology, most computer-aided translation systems process text files (*.txt*). For Microsoft Word 2000–2003, text files were saved and stored as *.doc* (document text file/word processing file); for Microsoft Word 2007–2011, documents were saved and stored as *.docx* (Document text file (Microsoft Office 2007)), *.dotx* (Microsoft Word 2007 Document Template). Other types of text files include *.txt* (Text files), *.txml* (WordFast files), and *.rtf* (Rich Text files).

All automatic and interactive translation systems can process text files, provided the text processing system has been installed in the computer before processing begins. Some of the computer-aided translation systems which can only translate text files include: Across, AidTransStudio, Anaphraseus (formerly known as OpenWordfast), AnyMem (*.docx* or higher), Araya, Autshumato Integrated Translation Environment (ITE), CafeTran, Déjà Vu, Esperantilo, Fluency, Fusion, OmegaT, Wordfast, and WordFisher. Computer-aided translation systems which can translate text files as well as other formats include CafeTran, Esperantilo, Felix, Fortis, GlobalSight, Google Translator Toolkit, Heartsome Translation Suite, Huajian IAT, Lingo, Lingotek, MadCap Lingo, MemoQ, MemOrg, MemSource, MetaTexis, MultiTrans,

OmegaT+, Pootle, SDL-Trados, Similis, Snowman, Swordfish, TM-database, Transit, Wordfast, XTM, and Yaxin.

(2) Web-page files

HyperText Markup Language (HTML) is a markup language that web browsers use to interpret and compose text, images and other material into visual or audible web pages. HTML defines the structure and layout of a web page or document by using a variety of tags and attributes. HTML documents are stored as *.asp* (Active Server Pages), *.aspx* (Active Server Page Extended), *.htm* (Hypertext Markup Language), *.html* (Hypertext Markup Language Files), *.php* (originally: Personal Home Page; now: Hypertext Preprocessor), *.jsp* (JavaServer Pages), *.sgml* (Standard Generalized Markup Language File), *.xml* (Extensible Markup Language file), *.xsl* (Extensible Stylesheet Language file) files format, which were available since late 1991. Due to the popularity of web pages, web translation has been an important part of automatic and interactive translation systems. Many systems provide comprehensive support for the localization of HTML-based document types. Web page localization is interchangeable with web translation or web localization.

Systems that handle HTML include Across, AidTransStudio, Alchemy Publisher, Araya, Atlas, CafeTran, CatsCradle, Déjà Vu, Felix, Fluency, Fortis, GlobalSight, Google Translator Toolkit, Heartsome Translation Suite, Huajian IAT, Lingo, Lingotek, LogiTerm, MemoQ, MemOrg, MetaTexis, MultiTrans, Okapi Framework, OmegaT, OmegaT+, Open Language Tools, Pootle, SDL-Trados, Similis, Snowman, Swordfish, TM-database, TransSearch, Transit, Transolution, and XTM.

(3) PDF files

Portable Document Format (PDF) (*.pdf*) is a universally accepted file interchange format developed by Adobe in the 1990s. The software that allows document files to be transferred between different types of computers is Adobe Acrobat. A PDF file can be opened by the document format, which might require editing to make the file look more like the original, or can be converted to an *rtf* file for data processing by a computer-aided translation system.

Systems that can translate Adobe PDF files and save them as Microsoft Word documents include Alchemy Publisher, CafeTran, Lingo, Similis, and Snowman.

(4) Microsoft Office PowerPoint files

Microsoft PowerPoint is a presentation program developed to enable users to create anything from basic slide shows to complex presentations, which are comprised of slides that may contain text, images, and other media. Versions of Microsoft Office PowerPoint include *Microsoft PowerPoint* 2000–2003, *.ppt* (General file extension), *.pps* (PowerPoint Slideshow), *.pot* (PowerPoint template); Microsoft PowerPoint 2007/2011, which are saved as *.pptx* (Microsoft PowerPoint Open XML Document), *.ppsx* (PowerPoint Open XML Slide Show), *.potx* (PowerPoint Open XML Presentation Template), and *.ppsm* (PowerPoint 2007 Macro-enabled Slide Show).

Systems that can handle Powerpoint files include Across, AidTransStudio, Alchemy Publisher, CafeTran, Déjà Vu, Felix, Fluency, Fusion, GlobalSight, Lingotek, LogiTerm, MadCap Lingo, MemoQ, MemSource, MetaTexis, SDL-Trados, Swordfish, TM-database, Transit, Wordfast, XTM, and Yaxin.

(5) Microsoft Excel files
Different versions of Microsoft Excel include Microsoft Excel 2000–2003 *.xls* (spreadsheet), *.xlt* (template); Microsoft Excel 2007: *.xlsx* (Microsoft Excel Open XML Document), *.xltx* (Excel 2007 spreadsheet template), *.xlsm* (Excel 2007 macro-enabled spreadsheet)

The computer-aided translation systems that can translate Excel files include Across, AidTransStudio, Déjà Vu, Felix, GlobalSight, Lingotek, LogiTerm, and MemoQ, MemOrg, MetaTexis, MultiTrans, Snowman, Wordfast, and Yaxin.

(6) Microsoft Access files
One of the computer-aided translation systems which can handle Access with *.accdb* (Access 2007–2010) file extension is Déjà Vu.

(7) Image files
The processing of image data, mainly graphics and pictures, is important in computer-aided translation. The data is stored as *.bmp* (bitmap image file), *.jpg* (Joint Photographic Experts Group), and *.gif* (Graphics Interchange Format). One of the computer-aided translation systems that is capable of handling images is CafeTran.

(8) Subtitle files
One of the most popular subtitle files on the market is *.srt* (SubRip Text). OmegaT is one of the computer-aided systems that supports subtitle files.

(9) Adobe InDesign files
Adobe InDesign is desktop publishing software. It can be translated without the need of any third party software by Alchemy Publisher and AnyMem. For Alchemy Publisher, the *.indd* file must be exported to an *.inx* format before it can be processed. Other computer-aided translation systems that support Adobe Indesign files include Across, Déjà Vu, Fortis, GlobalSight, Heartsome Translation Suite, Okapi Framework, MemoQ, MultiTrans, SDL-Trados, Swordfish, Transit, and XTM.

(10) Adobe FrameMaker Files
Adobe FrameMaker is an authoring and publishing solution for XML. FrameMaker files, *.fm*, *.mif* and *.book*, can be opened directly by a system if it is installed with Adobe FrameMaker.

Computer-aided translation systems that can translate Adobe FrameMaker files include Across, Alchemy Publisher (which requires a PPF created by Adobe FrameMaker before translating it. Alchemy Publisher supports FrameMaker 5.0, 6.0, 7.0, 8.0, 9.0, FrameBuilder 4.0, and FrameMaker + *sgml*), CafeTran, Déjà Vu, Fortis, GlobalSight, Heartsome Translation Suite, Lingo, Lingotek, MadCap Lingo, MemoQ, MetaTexis, MultiTrans, SDL-Trados, Swordfish, Transit, Wordfast, and XTM.

(11) Adobe PageMaker files
Systems that support Adobe PageMaker 6.5 and 7 files include Déjà Vu, GlobalSight, MetaTexis, and Transit.

(12) AutoCAD files
AutoCAD, developed and first released by Autodesk, Inc. in December 1982, is a software application for computer-aided design (CAD) and drafting which supports both 2D and 3D

formats. This software is now used internationally as the most popular drafting tool for a range of industries, most commonly in architecture and engineering.

Computer-aided translation systems that support AutoCad are CafeTran, Transit, and TranslateCAD.

(13) DTP tagged text files

DTP stands for Desktop Publishing. A popular desktop publishing system is QuarkXPress.

Systems that support desktop publishing include Across, CafeTran, Déjà Vu, Fortis, GlobalSight, MetaTexis, MultiTrans, SDL-Trados, and Transit.

(14) Localization files

Localization files include files with the standardized format for localization *.xliff* (XML Localization Interchange File Format) files, *.ttx* (XML font file format) files, and *.po* (Portable Object).

Computer-aided translation systems which process XLIFF files include Across Language Server, Araya, CafeTran, Esperantilo, Fluency, Fortis, GTranslator, Heartsome Translation Suite, MadCap Lingo, Lingotek, MemoQ, Okapi Framework, Open Language Tools, Poedit, Pootle, Swordfish, Transolution, Virtaal, and XTM.

(II) Software development types

(1) Java Properties files

Java Properties files are simple text files that are used in Java applications. The file extension of Java Properties file is *.properties*.

Computer-aided translation systems that support Java Properties File include Déjà Vu, Fortis, Heartsome Translation Suite, Lingotek, Okapi Framework, OmegaT+, Open Language Tools, Pootle, Swordfish, and XTM.

(2) OpenOffice.org/StarOffice

StarOffice of the Star Division was a German company that ran from 1984 to 1999. It was succeeded by OpenOffice.org, an open-sourced version of StarOffice owned by Sun Microsystems (1999–2009) and by Oracle Corporation (2010–2011), which ran from 1999−2011.Currently it is Apache OpenOffice. The format of OpenOffice is *.odf* (Open Document Format).

Computer-aided translation systems which process this type of file include AidTransStudio, Anaphraseus, CafeTran, Déjà Vu, Heartsome Translation Suite, Lingotek, OmegaT, OmegaT+, Open Language Tools, Pootle, Similis, Swordfish, Transolution, and XTM.

(3) Windows resource files

These are simple script files containing startup instructions for an application program, usually a text file containing commands that are compiled into binary files such as *.exe* and *.dll*. File extensions include *.rc* (Record Columnar File), *.resx* (NET XML Resource Template). Computer-aided translation systems that process this type of files include Across, Déjà Vu, Fortis, Lingotek, MetaTexis, and Okapi Framework.

Compatibility of operating systems

One of the most important factors which determined the course of development of computer-aided translation systems is their compatibility with the current operating systems on the market. It is therefore essential to examine the major operating systems running from the beginning of computer-aided translation in 1988 to the present, which include, among others, the Windows of Microsoft and the OS of Macintosh.

Microsoft Operating Systems

In the world of computing, Microsoft Windows has been the dominant operating system. From the 1981 to the 1995, the x86-based MS-DOS (Microsoft Disk Operating System) was the most commonly used system, especially for IBM PC compatible personal computers. Trados's Translator's Workbench II, developed in 1992, is a typical example of a computer-aided translation system working on DOS.

DOS was supplemented by Microsoft Windows 1.0, a 16-bit graphical operating environment, released on 20 November 1985 (Windows 2012). In November 1987, Windows 1.0 was succeeded by Windows 2.0, which existed till 2001. Déjà Vu 1.0, released in 1993, was one of the systems compatible with Windows 2.0. Windows 2.0 was supplemented by Windows 286 and Windows 386.

Then came Windows 3.0, succeeding Windows 2.1x. Windows 3.0, with a graphical environment, is the third major release of Microsoft Windows, and was released on 22 May 1990. With a significantly revamped user interface and technical improvements, Windows 3 became the first widely successful version of Windows and a rival to Apple Macintosh and the Commodore Amiga on the GUI front. It was followed by Windows 3.1x. During its lifespan from 1992–2001, Windows 3.1x introduced various enhancements to the still MS-DOS-based platform, including improved system stability, expanded support for multimedia, TrueType fonts, and workgroup networking. Trados's Translator's Workbench, released in 1994, was a system that was adaptable to Windows 3.1x.

Except for Windows and DOS, OS/2 is also one of the operation systems that support computer-aided translation systems, especially in late 1980s and early 1990s.

Apple Operating Systems

Mac OS (1984–2000) and OS X (2001–) are two series of graphical user interface-based operating systems developed by Apple Inc. for their Macintosh line of computer systems. Mac OS was first introduced in 1984 with the original Macintosh and this series was ended in 2000. OS X, first released in March 2001, is a series of Unix-based graphical interface operating systems. Both series share a general interface design, but have very different internal architectures.

Only one computer-aided translation system, AppleTrans, is designed for OS X. Its initial released was announced in February 2004 and the latest updated version was version 1.2(v38) released in September 2006, which runs on Mac OS X 10.3 or later.

Another computer-aided translation system, Wordfast Classic was released to upgrade its support of the latest text processor running on Mac OS X, such as Wordfast Classic 6.0, which is compatible for MS Word 2011 for Mac.

Other computer-aided translation systems that can run on Mac OS or OS X are cross-platform software, rather than software developed particularly for Mac. Examples are Java-based applications, such as Autshumato, Heartsome, OmegaT, Open Language Tools and

Swordfish. Besides, all cloud-based systems can support Mac OS and OS X, including Wordbee, XTM Cloud, Google Translator's Toolkit, Lingotek Collaborative Translation Platform, MemSource Cloud, and WebWordSystem.

OS/2 is a series of computer operating systems, initially created by Microsoft and IBM, then later developed by IBM exclusively. The name stands for 'Operating System/2'.

Until 1992, the early computer-aided translation systems ran either on MS-DOS or OS/2. For example, IBM Translation Manager/2 (TM/2) was released in 1992 and run on OS/2. ALPS's translation tool also ran on OS/2. But OS/2 had a much smaller market share compared with Windows in early 1990s. Computer-aided translation system developers therefore gradually shifted from OS/2 and MS-DOS to Windows or discontinued the development of OS/2 and MS-DOS compatible computer-aided translation systems. By the end of the 1990s, most computer-aided translation systems mainly ran on Windows, although some developers offered operating-system customization services upon request. OS/2 4.52 was released in December 2001. IBM ended its support to OS/2 on 31 December 2006.

Compatibility of databases

Compatibility of translation memory databases

TMX (Translation Memory eXchange), created in 1998, is widely used as an interchange format between different translation memory formats. TMX files are XML (eXtensible Markup Language) files whose format was originally developed and maintained by OSCAR (Open Standards for Container/Content Allowing Re-use) of the Localization Industry Standards Association. The latest official version of the TMX specification, version 1.4b, was released in 2005. In March 2011 LISA was declared insolvent; as a result its standards were moved under the Creative Commons licence and the standards specification relocated. The technical specification and a sample document of TMX can be found on the website of The Globalization and Localization Association.

TMX has been widely adopted and is supported by more than half of the current computer-aided translation systems on the market. The total number of computer-aided translation systems that can import and export translation memories in TMX format is 54, including Across, Alchemy Publisher, Anaphraseus, AnyMem, Araya, ATLAS, Autshumato, CafeTran, Crowdin, Déjà Vu, EsperantiloTM, Felix, Fluency, Fortis, Fusion, GE-CCT, GlobalSight, Google Translator Toolkit, Heartsome, Huajian IAT, Lingotek, LogiTerm, LongRay CAT, MadCap Lingo, MemoQ, MemSource, MetaTexis, MT2007, MultiTrans, OmegaT, OmegaT+, Open Language Tools, OpenTM2, OpenTMS, PROMT, SDL Trados, Snowball, Snowman, Swordfish, Systran, Text United, The Hongyahu, TM Database, Transit, Translation Workspace, Transwhiz, TraTool, Webwordsystem, Wordbee Translator, Wordfast Classic and Wordfast Pro, XTM, Yaxin CAT, and 翻訳ブレイン (Translation Brain).

Compatibility of terminology databases

Compatibility of terminology databases is best illustrated by TermBase eXchange (TBX), which covers a family of formats for representing the information in a high-end termbase in a neutral intermediate format in a manner compliant with the Terminological Markup Framework (TMF) (Melby 2012: 19–21).

Termbase Exchange is an international standard as well as an industry standard. The industry standard version differs from the ISO standard only by having different title pages. Localization

Industry Standards Association, the host organization for OSCAR that developed Termbase Exchange, was dissolved in February 2011. In September 2011, the European Telecommunications Standards Institute (ETSI) took over maintenance of the OSCAR standards. ETSI has established an interest group for translation/localization standards and a liaison relationship with the International Organization for Standardization (ISO) so that TBX can continue to be published as both an ISO standard and an industry standard.

There are many types of termbases in use, ranging from huge termbases operated by governments, to medium-size termbases maintained by corporations and non-governmental organizations, to smaller termbases maintained by translation service providers and individual translators. The problem addressed by the designers of term exchange was that existing termbases are generally not interoperable. They are based on different data models that use a variety of data categories. And even if the same data category is used for a particular piece of information, the name of the data category and the values allowed for the data category may be different.

Compatibility of rules

SEGMENTATION RULES EXCHANGE

Segmentation Rules eXchange (SRX) is an XML-based standard that was maintained by Localization Industry Standards Association, until it became insolvent in 2011 and then this standard is now maintained by the Globalization and Localization Association (GALA).

Segmentation Rules eXchange provides a common way to describe how to segment text for translation and other language-related processes. It was created when it was realized that translation memory exchange leverage is lower than expected in certain instances due to differences in how tools segment text. Segmentation Rules eXchange is intended to enhance the translation memory exchange so that translation memory data that is exchanged between applications can be used more effectively. Having the segmentation rules that were used when a translation memory was created will increase the leverage that can be achieved when deploying the translation memory data.

Compatibility with the languages supported

As computer-aided translation systems cannot identify languages, language compatibility is therefore an important concept in translation technology. There are a large number of languages and sub-languages in the world, totalling 6,912. But the number of major languages computers can process is relatively small. It is therefore important to know whether the languages that require machine processing are supported by a system or not.

With the aid of unicode, most of the languages in the world are supported in popular computer-aided translation systems. Unicode is a computing industry standard for the consistent encoding, representation and handling of text expressed in most of the world's writing systems.

There are basically two major types of language and sub-language codes. Some systems, such as OmegaT and XTM, use letters for language codes (2 or 3 letters) and language-and-region codes (2+2 letters), which can be selected from a drop-down list. OmegaT follows the ISO 639 Code Tables in preparing its code list. French for example, is coded *fr* with the language- and region code for French (Canada) as *fr-CA*.

The following is a list of languages supported by Wordfast Classics and XTM, two of the nine computer-aided translation systems chosen for analysis in this chapter.

WORDFAST CLASSIC

Wordfast can be used to translate any of the languages supported by Microsoft Word. The number of languages supported by Microsoft is 91, with a number of sub-languages for some major languages.

[*Afro-Asiatic*] Arabic (Algeria), Arabic (Bahrain), Arabic (Egypt), Arabic (Iraq), Arabic (Jordan), Arabic (Kuwait), Arabic (Lebanon), Arabic (Libya), Arabic (Morocco), Arabic (Oman), Arabic (Qatar), Arabic (Saudi Arabia), Arabic (Syria), Arabic (Tunisia), Arabic (U.A.E.), Arabic (Yemen), Hebrew, Maltese

[*Altaic*] Azeri (Cyrillic), Azeri (Latin), Japanese, Korean, Turkish

[*Austro-Asiatic*] Vietnamese

[*Austronesian*] Indonesian, Malay (Brunei Darussalam), Malaysian

[*Basque*] Basque

[*Dravidian*] Kannada, Malayalam, Tamil, Telugu

[*Indo-European*] Afrikaans, Albanian, Armenian, Assamese, Belarusian, Bengali, Bulgarian, Byelorussian, Catalan, Croatian, Czech, Danish, Dutch, Dutch (Belgian), English (Australia), English (Belize), English (Canadian), English (Caribbean), English (Ireland), English (Jamaica), English (New Zealand), English (Philippines), English (South Africa), English (Trinidad), English (U.K.), English (U.S.), English (Zimbabwe), Faroese, Farsi, French (Belgian), French (Cameroon), French (Canadian), French (Congo), French (Cote d'Ivoire), French (Luxembourg), French (Mali), French (Monaco), French (Reunion), French (Senegal), French (West Indies), Frisian (Netherlands), Gaelic (Ireland), Gaelic (Scotland), Galician, German, German (Austria), German (Liechtenstein), German (Luxembourg), Greek, Gujarati, Hindi, Icelandic, Italian, Kashmiri, Konkani, Latvian, Lithuanian, Macedonian (FYRO), Marathi, Nepali, Norwegian (Bokmol), Norwegian (Nynorsk), Oriya, Polish, Portuguese, Portuguese (Brazil), Punjabi, Rhaeto-Romance, Romanian, Romanian (Moldova), Russian, Russian (Moldova), Sanskrit, Serbian (Cyrillic), Serbian (Latin), Sindhi, Slovak, Slovenian, Sorbian, Spanish (Argentina), Spanish (Bolivia), Spanish (Chile), Spanish (Colombia), Spanish (Costa Rica), Spanish (Dominican Republic), Spanish (Ecuador), Spanish (El Salvador), Spanish (Guatemala), Spanish (Honduras), Spanish (Nicaragua), Spanish (Panama), Spanish (Paraguay), Spanish (Peru), Spanish (Puerto Rico), Spanish (Spain), Spanish (Traditional), Spanish (Uruguay), Spanish (Venezuela), Swedish, Swedish (Finland), Swiss (French), Swiss (German), Swiss (Italian), Tajik, Ukrainian, Urdu, Welsh

[*Kartvelian*] Georgian

[*Niger-Congo*] Sesotho, Swahili, Tsonga, Tswana, Venda, Xhosa, Zulu

[*Sino-Tibetan*] Burmese, Chinese, Chinese (Hong Kong SAR), Chinese (Macau SAR), Chinese (Simplified), Chinese (Singapore), Chinese (Traditional), Manipuri, Tibetan

[*Tai-Kadai*] Laothian, Thai

[*Turkic*] Tatar, Turkmen, Uzbek (Cyrillic), Uzbek (Latin)

[*Uralic*] Estonian, Finnish, Hungarian, Sami Lappish

XTM

The languages available in XTM are 157, not including varieties within a single language. These languages are as follows:

[*Afro-Asiatic*] Afar, Amharic, Arabic, Hausa, Hebrew, Maltese, Oromo, Somali, Sudanese Arabic, Syriac, Tigrinya,

[*Altaic*] Azeri, Japanese, Kazakh, Korean, Mongolian, Turkish

[*Austro-Asiatic*] Khmer, Vietnamese

[*Austronesian*] Fijian, Indonesian, Javanese, Malagasy, Malay, Maori, Nauru, Samoan, Tagalog, Tetum, Tonga

[*Aymaran*] Aymara

[*Bantu*] Kikongo

[*Basque*] Basque

[*Constructed Language*] Esperanto, Interlingua, Volapk

[*Dravidian*] Kannada, Malayalam, Tamil, Telugu

[*English Creole*] Bislama

[*Eskimo-Aleut*] Greenlandic, Inuktitut, Inupiak

[*French Creole*] Haitian Creole

[*Hmong-Mien*] Hmong

[*Indo-European*] Afrikaans, Armenian, Assamese, Asturian, Bengali, Bihari, Bosnian, Breton, Bulgarian, Byelorussian, Catalan, Corsican, Croatian, Czech, Danish, Dari, Dhivehi, Dutch, English, Faroese, Flemish, French, Frisian, Galician, German, Greek, Gujarati, Hindi, Icelandic, Irish, Italian, Kashmiri, Konkani, Kurdish, Latin, Latvian, Lithuanian, Macedonian, Marathi, Montenegrin, Nepali, Norwegian, Occitan, Oriya, Pashto, Persian, Polish, Portuguese, Punjabi, Rhaeto-Romance, Romanian, Russian, Sanskrit, Sardinian, Scottish Gaelic, Serbian, Sindhi, Singhalese, Slovak, Slovenian, Sorbian, Spanish, Swedish, Tajik, Ukrainian, Urdu, Welsh, Yiddish

[*Kartvelian*] Georgian

[*Ngbandi-based Creole*] Sango

[*Niger-Congo*] Chichewa, Fula, Igbo, Kinyarwanda, Kirundi, Kiswahili, Lingala, Ndebele, Northern Sotho, Sesotho, Setswana, Shona, Siswati, Tsonga, Tswana, Twi, Wolof, Xhosa, Yoruba, Zulu

[*Northwest Caucasian*] Abkhazian

[*Quechuan*] Quechua

[*Romanian*] Moldavian

[*Sino-Tibetan*] Bhutani, Burmese; Chinese, Tibetan

[*Tai-Kadai*] Laothian, Thai

[*Tupi*] Guarani

[*Turkic*] Bashkir, Kirghiz, Tarar, Turkmen, Uyghur, Uzbek

[*Uralic*] Estonian, Finnish, Hungarian

Several observations can be made from the languages supported by the current eleven systems.

(1) The number of languages supported by language-specific systems is small as they need to be supplied with language-specific dictionaries to function well. Yaxin is best for English–Chinese translation, covering two languages, while most non-language-specific systems support around or above 100 languages.

(2) For the seven systems developed in Europe, the United Kingdom, and the United States, which include Across, Déjà Vu, MemoQ, OmegaT, SDL Trados, Wordfast, and XTM, the Indo-European languages take up around 51.89 per cent, while the proportion of the non-Indo-European languages is 48.11 per cent. Table 2.1 shows the details:

Table 2.1 Statistics of languages supported by 7 CAT systems

Name of the system	*Number of languages supported*	*Number of language families supported*	*Number and percentage of Indo-European languages*	*Number and percentage of non-Indo-European languages*
Across	121	18	61 (50.41%)	60 (49.59%)
Déjà Vu	132	21	66 (50%)	66 (50%)
MemoQ	102	16	54 (52.94%)	48 (47.06%)
OmegaT	90	14	48 (53.33%)	42 (46.67%)
SDL Trados	115	18	62 (53.91%)	53 (46.09%)
Wordfast	91	13	54 (59.34%)	37 (40.66%)
XTM	157	26	68 (43.31%)	89 (56.69%)

Controllability

One of the main differences between human translation and computer-aided translation lies in the degree of control over the source text. In human translation, there is no need, or rather it is not the common practice, to control how and what the author should write. But in computer-aided translation, control over the input text may not be inappropriate as the output of an unedited or uncontrolled source language text is far from satisfactory (Adriaens and Macken 1995: 123–141; Allen and Hogan 2000: 62–71; Arnold *et al.* 1994; Hurst 1997: 59–70; Lehtola, Tenni and Bounsaythip 1998: 16–29; Mitamura 1999: 46–52; Murphy *et al.* 1998; Nyberg *et al.* 2003: 245–281; Ruffino 1985: 157–162).

The concept of controllability is realized in computer-aided translation by the use of controlled language and the method of pre-editing.

Controllability by the use of controlled language

An effective means of achieving controllability in translation technology is controlled language (see Figure 2.16). The idea of controlled language was created, partly at least, as a result of the problems with natural languages which are full of complexities, ambiguities, and robustness (Nyberg *et al.* 2003: 245–281). A strong rationale for controlled language is that a varied source text generates a poor target text, while a controlled source text produces a quality target text. (Bernth 1999). Controlled language is therefore considered necessary (Caeyers 1997: 91–103; Hu 2005: 364–372).

Controlled language, in brief, refers to a type of natural language developed for specific domains with a clearly defined restriction on controlled lexicons, simplified grammars, and style rules to reduce the ambiguity and complexity of a text so as to make it easier to be understood by users and non-native speakers and processed by machine translation systems (Chan 2004: 44; Lux and Dauphin 1996: 193–204).

Control over the three stages of a translation procedure, which include the stage of inputting a source text, the stage of transfer, and the stage of text generation, is generally regarded as a safe guarantee of quality translation. Control of the source text is in the form of controlled authoring, which makes the source text easier for computer processing (Allen 1999; Chan 2004: 44; van der Eijk and van Wees 1998: 65–70; Zydron 2003). The text produced is a 'controlled language text' (Melby 1995: 1). There is also control over the transfer stage. And the output of a machine translation system is known as 'controlled translation' (Carl 2003: 16–24; Gough and Way 2004: 73–81; Rico and Torrejon 2004; Roturier 2004; Torrejón 2002: 107–116), which is alternatively known as a 'controlled target language text' (Chan 2004: 44). In short, a controlled text is easier to be processed by machine translation systems to produce a quality output.

Goals and means of controlled language

Controlled language is used by both humans and computers. The goals of controlled language are to make the source text easier to read and understand. These goals are to be achieved at the lexical and sentential levels.

At the lexical level, controlled language is about the removal of lexical ambiguity and the reduction in homonymy, synonymy, and complexity. This is to be achieved by one-to-one correspondence in the use and translation of words, known as one-word one-meaning. An example is to use only the word 'start' but not similar words such as 'begin', 'commence',

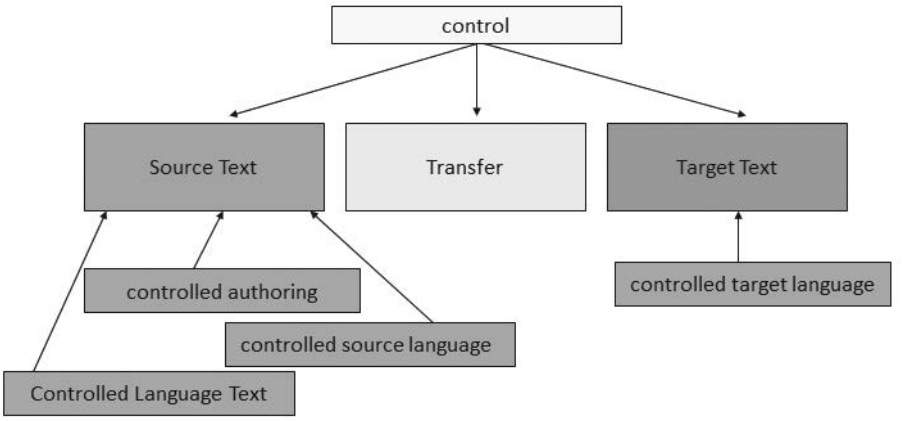


Figure 2.16 Controlled language

'initiate', and 'originate'. The second method is to use the preferred language, such as American English but not British English. The third method is to have a limited basic vocabulary (Bjarnestam 2003; Chen and Wu 1999; Probst and Levin 2002: 157–167; Wasson 2000: 276–281), which can be illustrated by the use of a controlled vocabulary of 3,100 words in aircraft-maintenance documentation at the European Association of Aerospace Industries (AECMA) in 1980 (AECMA 1995).

At the sentential level, controlled language is about the removal of syntactical ambiguity, the simplification of sentence structures, limitations on sentence length, and constraints on voice, tense, and other grammatical units. To do all these, there are a limited number of strictly stipulated writing rules to follow. The European Association of Aerospace Industries had 57 writing rules. Short sentences are preferred over long and complex sentences. And there is also a limit on the number of words in a sentence. For procedural text, there should be no more than twenty words. For descriptive texts, the number is twenty-five. There are also rules governing grammatical well-formedness (Loong 1989: 281–297), restricted syntax, and the use of passive construction in procedural texts. At the suprasentential level, there is a limit of six sentences in a paragraph, the maximum number of clauses in a sentence, and the use of separate sentences for sequential steps in procedural texts.

This means setting limits on the length of a sentence, such as setting the number of words at twenty, using only the active voice, and expressing one instruction or idea by one sentence.

Controlled language checkers

Controlled language cannot be maintained manually; it relies on the use of different kinds of checkers, which are systems to ensure that a text conforms to the rules of a particular controlled language (Fouvry and Balkan 1996: 179–192). There is the automatic rewriting system, which is specially developed for controlled language, rewriting texts automatically into controlled language without changing the meaning of the sentences in the original in order to produce a high-quality machine translation. There is the controlled language checker, which is software that helps an author to determine whether a text conforms to the approved words and writing rules of a particular controlled language.

Checkers can also be divided into two types: in-house controlled language checker and commercial controlled language checker. In-house controlled language checkers include the

PACE (Perkins Approved Clear English) of Perkins Engines Ltd, the Controlled English of Alcatel Telecom, and the Boeing Simplified English Checker of the Boeing Company (Wojcik and Holmback 1996: 22–31). For commercial controlled language checkers, there are a number of popular systems. The LANTmaster Controlled Checker, for example, is a controlled language checker developed by LANT in Belgium. It is based on work done for the METAL (Mechanical Translation and Analysis of Languages) machine translation project. It is also based on the experience of the Simplified English Grammar and Style Checker (SECC) project (Adriaens 1994: 78–88; Adriaens and Macken 1995: 123–141). The MAXit Checker is another controlled language software developed by Smart Communications Incorporation to analyse technical texts written in controlled or simplified English with the use of more than 8,500 grammar rules and artificial intelligence to check the clarity, consistency, simplicity, and global acceptance of the texts. The Carnegie Group also produced the ClearCheck, which performs syntactic parsing to detect such grammatical problems as ambiguity (Andersen 1994: 227).

Advantages and disadvantages of controlled language

The advantages of controlled language translation are numerous, including high readability, better comprehensibility, greater standardization, easier computer processing, greater reusability, increased translatability, improved consistency, improved customer satisfaction, improved competitiveness, greater cost reduction in global product support, and enhanced communication in global management.

There are a number of disadvantages in using controlled language, such as expensive system construction, high maintenance cost, time-consuming authoring, and restrictive checking process.

Controlled language in use

As the advantages of using controlled language outweigh its disadvantages, companies started to use controlled language as early as the 1970s. Examples of business corporations which used controlled languages include Caterpillar Fundamental English (CFE) of the Caterpillar Incorporation in 1975 (Kamprath *et al.* 1998: 51–61; Lockwood 2000: 187–202), Smart Controlled English of the Smart Communications Ltd in 1975, Douglas Aircraft Company in 1979, the European Association of Aerospace Industries (AECMA) in 1980, the KANT Project at the Center for Machine Translation, Carnegie Mellon University in 1989 (Allen 1995; Carbonell *et al.* 1992: 225–235; Mitamura *et al.* 1994: 232–233; Mitamura and Nyberg 1995: 158–172; Mitamura *et al.* 2002: 244–247; Nyberg and Mitamura 1992: 1069–1073; Nyberg *et al.* 1997; Nyberg *et al.* 1998: 1–7; Nyberg and Mitamura 2000: 192–195), the PACE of Perkins Engines Ltd. in 1989, ScaniaSwedish in Sweden in 1995 (Almqvist and Hein 1996: 159–164; Hein 1997), General Motors in 1996, Ericsson English in Sweden in 2000, Nortel Standard English in the United Kingdom in 2002, and Oce Technologies English in Holland in 2002.

Controlled language in computer-aided translation systems

The concept of controlled language is realized in controlled authoring in computer-aided translation systems. Authoring checking tools are used to check and improve the quality of the source text. There is an automatic rewriting system which is usually used as a tool to realize controlled authoring. One of the computer-aided translation systems that performs controlled

authoring is Star Transit. This system provides automatic translation suggestions from the translation memory database from a speedy search engine and it is an open system that can integrate with many authoring systems.

Customizability

Customizability, etymologically speaking, is the ability to be customized. More specifically, it refers to the ability of a computer or computer-aided translation system to adapt itself to the needs of the user. Customizing a general-purpose machine translation system is an effective way to improve MT quality.

Editorial customization

Pre-editing is in essence a process of customization. The customization of machine translation systems, which is a much neglected area, is necessary and essential as most software on the market are for general uses and not for specific domains. Practically, system customization can be taken as part of the work of pre-editing as we pre-edit the words and expressions to facilitate the production of quality translation.

The degree of customization depends on the goals of translation, and the circumstances and the type of text to be translated.

Language customization

It is true that there are many language combinations in computer-aided translation systems to allow the user to choose any pair of source and target languages when creating a project, yet many users only work with a limited set of source and target languages. XTM, a cloud-based system, allows the user to set language combinations through the Data section. In the language combinations section, the project administrator or user can reduce and customize the available languages to be used, set the language combinations for the entire system and set specific language combinations for individual customers (XTM International 2012: 15).

Language customization in XTM, for example, can be conducted on the Customize tab where there are three options for the user to modify and use language combinations. The first option is 'system default language combinations', which is the full set of unmodified language combinations. The second option is 'system defaults with customized language combinations', which is the full set of language combinations in which the user may have customized some parameters. The third option is 'customized language combinations only', which include only the language combinations that the user has customized. It is possible to add or delete the source and target languages in the selected customized option.

Lexicographical customization

Lexicographical customization is best shown in the creation of custom dictionaries for each customer, other than the dictionaries for spell checking. This means that multiple translators working on projects for the same customer will use the same custom dictionary.

Linguistic customization

As far as linguistic customization is concerned, there are basically two levels of customization: lexical customization and syntactical customization.

Lexical customization

Lexical customization is to customize a machine translation system by preparing a customized dictionary, in addition to the system dictionary, before translating. This removes the uncertainties in translating ambiguous words or word combinations. It must be pointed out, however, that the preparation of a customized dictionary is an enormous task, involving a lot of work in database creation, database maintenance, and database management.

Syntactical customization

Syntactical customization, on the other hand, is to add sentences or phrases to the database to translate texts with many repetitions. Syntactical customization is particularly important when there is a change of location for translation consumption. The translation memory databases built up in Hong Kong for the translation of local materials, for example, may not be suitable for the production of translations targeted at non-Hong Kong readers, such as those in mainland China.

Resource customization

Website customization

Some computer-aided translation systems allow the user to create resource profile settings. Each profile in Fluency, for example, has four customized uniform resource locators (URLs) associated with it. URLs are the Internet addresses of information. Each document or file on the Internet has a unique address for its location. Fluency allows the user to have four URLs of one's preference, two perhaps for specialized sites and two general sites.

Machine translation system customization

Some systems are connected to installed machine translation systems the terminology databases of which can be customized for the generation of output, thus achieving terminological consistency in the target text.

Collaborativity

Collaborativity is about continuously working and communicating with all parties relating to a translation project, from the client to the reviewer, in a shared work environment to generate the best benefits of team work. Computer-aided translation is a modern mode of translation production that works best in team translation. In the past and decreasingly at present, individual translation has been the norm of practice. At present and increasingly in the future, team translation is the standard practice.

A number of systems, such as Across and Wordfast, can allow users to interact with each other through the translation memory server and share translation memory assets in real time.

Translation is about management. Translation business operates on projects. Translation technology is about project management, about how work is to be completed by translation teams. With the use of translation technology, the progress of translation work is under control and completed with higher efficiency. The best way to illustrate this point is project collaboration, which allows translators and project managers to easily access and distribute projects and easily monitor their progress.

The work of translation in the present digital era is done almost entirely online with the help of a machine translation or computer-aided translation system. This can be illustrated with SDL-Trados 2014, which is a computer-aided translation system developed by SDL International and generally considered to be the most popular translation memory system on the market.

Figure 2.17 shows the dashboard of SDL-Trados 2014.

Figure 2.17 Dashboard of SDL-Trados 2014

Name	Status	Date Due	Created At	Type	Locat...	Server
translator02.pptx	In Progress	10/04/2014 18:00:00	03/04/2014 14:10:35	Single file project	C:\User...	
SamplePhotoPrinter.doc_en-US...	In Progress	14/04/2014 18:00:00	21/03/2014 15:36:43	Single file project	C:\User...	
Project 24	In Progress	[none]	18/03/2014 15:48:59	Standard Studio project	C:\User...	
Project 23	In Progress	[none]	18/03/2014 15:46:54	Standard Studio project	C:\User...	
Proz April	In Progress	[none]	18/03/2014 15:19:08	Standard Studio project	C:\User...	
SamplePhotoPrinter.doc_en-US...	In Progress	[none]	28/02/2014 16:36:44	Single file project	C:\User...	
2012-08-27-SPECIFICATION-C...	In Progress	[none]	20/02/2014 15:08:04	Single file project	C:\User...	
Project 22	In Progress	[none]	15/01/2014 09:52:40	Standard Studio project	C:\User...	
FLMailerJan09v1.html_en-US_it-I...	In Progress	[none]	07/01/2014 16:28:15	Single file project	C:\User...	
Project 21	In Progress	[none]	07/01/2014 16:04:05	Standard Studio project	C:\User...	
Proz membership Eng.rtf_en-US...	In Progress	[none]	07/01/2014 15:55:11	Single file project	C:\User...	
Project 20	In Progress	[none]	05/12/2013 17:24:49	Standard Studio project	C:\User...	
Proz membership Eng.doc_en-U...	In Progress	[none]	05/12/2013 17:21:22	Single file project	C:\User...	
SamplePhotoPrinter.doc_en-US...	In Progress	[none]	28/11/2013 17:31:13	Single file project	C:\User...	
globe.docx_it-IT_en-GB	In Progress	[none]	25/11/2013 14:20:19	Single file project	C:\User...	
Group Share 2014 PB_en-US_it-...	In Progress	15/11/2013 08:40:22	13/11/2013 08:58:01	Studio package	C:\User...	
Project 19	In Progress	[none]	11/11/2013 23:45:25	Standard Studio project	C:\User...	
HP Printer page.htm_en-US_it-IT	In Progress	[none]	08/11/2013 21:32:10	Single file project	C:\User...	
Project 2	In Progress	[none]	08/11/2013 21:21:13	Standard Studio project	C:\User...	
Project 1	In Progress	[none]	07/11/2013 21:46:06	Standard Studio project	C:\User...	
retest.docx_en-US_it-IT	In Progress	[none]	07/11/2013 19:04:57	Single file project	C:\User...	

Figure 2.18 List of current projects

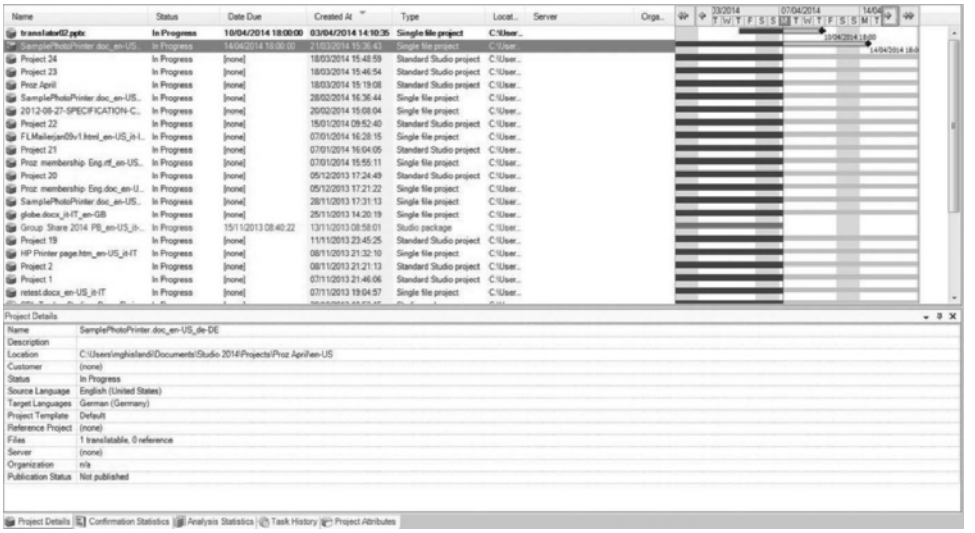


Figure 2.19 Project details

Workflow of a translation project

To start a project, the first stage of the workflow is the creation of a termbase and a translation memory database, as shown in Figure 2.20.

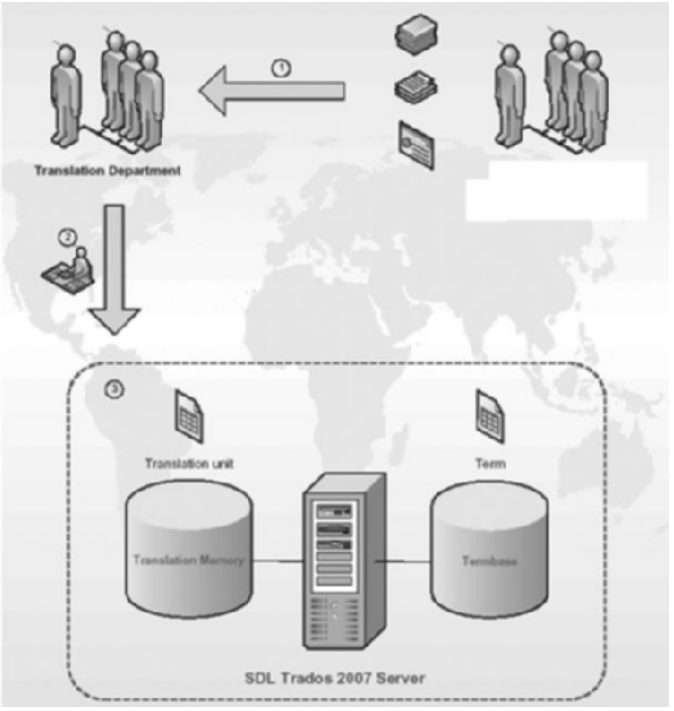


Figure 2.20 Workflow of a translation project: the first stage

In other words, when the Project Manager has any publications, files or web pages to translate, he will send them to the translators of a department or unit, or freelancers for processing. They will create translation units and term databases from these pre-translated documents and save these databases in the SDL-Trados 2014 Server. This is the first stage of the workflow.

After the creation of translation memory and term databases, as shown in Figure 2.21, the Project Manager can then initiate a translation project and monitor its progress with the use of SDL-Trados 2014 (as indicated by ①). He can assign and distribute source files to in-house and / or freelance translators by emails (as indicated by ②). Translators can then do the translation by (i) reusing the translation memories and terms stored in the databases; (ii) adding new words or expressions to the translation memory and term databases (as indicated by ③). When the translation is done, translators send their translated files back to the Project Manager on or before the due date (as indicated by ④). When the Project Manager receives the translated files, he updates the project status, finalizes the project and marks it as 'complete' (as indicated by ⑤).

To make sure that SDL-Trados 2014 has a smooth run, a technical support unit to maintain the SDL-Trados server may be necessary (as indicated by ⑥).

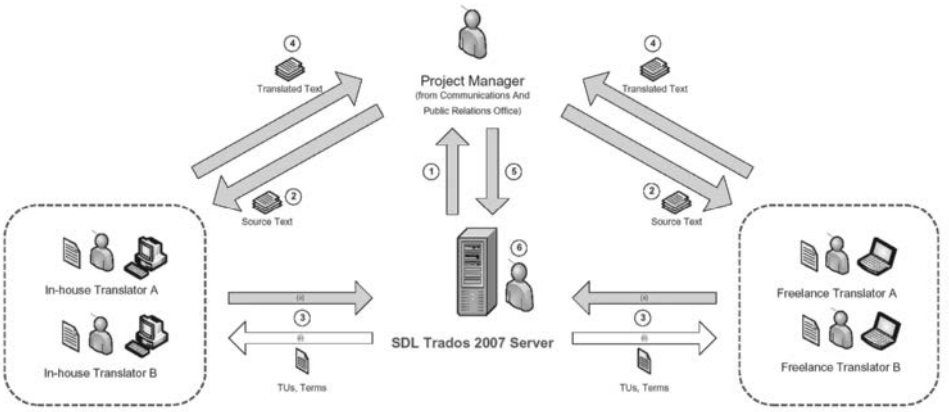


Figure 2.21 Workflow of a translation project: the second stage

A translation team usually consists of the following members.

Project manager

A project manager is a professional in the field of project management. The responsibilities of a project manager include the following:

1 plan, execute, and close projects
(When planning a project, the project manager works on the overall resources and budget of the project. When executing a project, the project manager can add or import customers and subcontract projects.)
2 create clear and attainable project objectives;
3 build the project requirements; and
4 manage cost, time, and scope of projects.

Terminologist

A terminologist is one who manages terms in the terminology database. There are two types of terminologists: (1) customer-specific terminologists who can only access the terminology of one customer; and (2) global experts who can access all the terms in the systems for all customers.

Conclusion

This chapter is possibly the first attempt to analyse the concepts that have governed the growth of functionalities in computer-aided translation systems. As computing science and related disciplines advance, more concepts will be introduced and more functions will be developed accordingly. However, it is believed that most of the concepts discussed in this chapter will last for a long time.

References

Adriaens, Geert (1994) 'Simplified English Grammar and Style Correction in an MT Framework: The LRE SECC Project', in *Translating and the Computer 16*, London: The Association for Information Management, 78–88.

Adriaens, Geert and Lieve Macken (1995) 'Technological Evaluation of a Controlled Language Application: Precision, Recall and Convergence Tests for SECC', in *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation* (*TMI-95*), 5–7 July 1995, University of Leuven, Leuven, Belgium, 123–141.

AECMA (1995) 'A Guide for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language–Issue 1', Brussels, Belgium.

Al-Shabab, Omar Sheikh (1996) Interpretation and the Language of Translation: Creativity and Convention in Translation, London: Janus Publishing Company.

Allen, Jeffrey (1995) *Review of the Caterpillar KANT English-French MT System*, Internal Technical Report, Peoria, IL: Technical Information Department, Caterpillar Inc.

Allen, Jeffrey (1999) 'Adapting the Concept of "Translation Memory" to "Authoring Memory" for a Controlled Language Writing Environment', in *Translating and the Computer 20*, London: The Association for Information Management.

Allen, Jeffrey and Christopher Hogan (2000) 'Toward the Development of a Post-editing Module for Raw Machine Translation Output: A Controlled Language Perspective', in *Proceedings of the 3rd International Workshop on Controlled Language Applications (CLAW 2000)*, 29–30 April 2000, Seattle, WA, 62–71.

Almqvist, Ingrid and Anna Sågvall Hein (1996) 'Defining ScaniaSwedish – Controlled Language for Truck Maintenance', in *Proceedings of the 1st International Workshop on Controlled Language Applications* (*CLAW-96*), Leuven, Belgium, 159–164.

Andersen, Peggy (1994) 'ClearCheck Demonstration', in Proceedings of the 1st Conference of the Association for Machine Translation in the Americas: Technology Partnerships for Crossing the Language Barrier (AMTA-1), 5–8 October 1994, Columbia, MD, 227.

Arnold, Doug J., Lorna Balkan, R. Lee Humphreys, Seity Meijer, and Louisa Sadler (1994) *Machine Translation: An Introductory Guide*, Manchester and Oxford: NCC Blackwell.

Bell, Roger T. (1991) *Translation and Translating: Theory and Practice*, London and New York: Longman.

Bernth, Arendse (1999) *Tools for Improving E-G MT Quality*, Yorktown Heights, NY: IBM T.J. Watson Research Center.

Bjarnestam, Anna (2003) 'Internationalizing a Controlled Vocabulary Based Search Engine for Japanese', in *Proceedings of the Localization World Conference 2003*, 14–16 October 2003, Seattle, WA.

Bly, Robert (1983) *The Eight Stages of Translation*, Boston, MA: Rowan Tree Press.

Caeyers, Herman (1997) 'Machine Translation and Controlled English', in *Proceedings of the 2nd Workshop of the European Association for Machine Translation*: *Language Technology in Your Organization?* 21–22 May 1997, University of Copenhagen, Copenhagen, Denmark, 91–103.

Carbonell, Jaime G., Teruko Mitamura, and Eric H. Nyberg (1992) 'The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics,...)', in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*: *Empiricist vs Rationalist Methods in MT* (*TMI-92*), Montreal, Quebec, Canada, 225–235.

Carl, Michael (2003) 'Data-assisted Controlled Translation', in Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop: Controlled Language Translation (EAMT-CLAW-2003), Dublin City University, Ireland, 16–24.

Chan, Sin-wai (2004) *A Dictionary of Translation Technology*, Hong Kong: The Chinese University Press.

Chen, Kuang-Hua and Chien-Tin Wu (1999) 'Automatically Controlled-vocabulary Indexing for Text Retrieval', in *Proceedings of the International Conference on Research in Computational Linguistics* (*ROCLING-XII*), Taipei, Taiwan.

de Ilarraza, Arantxa Diaz, Aingeru Mayor, and Kepa Sarasola (2000) 'Reusability of Wide-coverage Linguistic Resources in the Construction of Multilingual Technical Documentation', in *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the New Millenium* (*MT-2000*), University of Exeter, England.

Delisle, Jean (1988) *Translation: An Interpretive Approach*, Patricia Logan and Monica Creery (trans.), Ottawa and London: University of Ottawa Press.

Fais, Laurel and Kentaro Ogura (2001) 'Discourse Issues in the Translation of Japanese Email', in *Proceedings of the 5th Pacific Association for Computational Linguistics Conference* (*PACLING-2001*), Fukuoka, Japan.

Fouvry, Frederik and Lorna Balkan (1996) 'Test Suites for Controlled Language Checkers', in *Proceedings of the 1st International Workshop on Controlled Language Applications*, Katholieke Universiteit, Leuven, Belgium, 179–192.

Gough, Nano and Andy Way (2004) 'Example-based Controlled Translation', in Proceedings of the 9th Workshop of the European Association for Machine Translation: Broadening Horizons of Machine Translation and Its Applications, Foundation for International Studies, Malta, 73–81.

Han, Benjamin, Donna Gates, and Lori S. Levin (2006) 'Understanding Temporal Expressions in Emails', in Proceedings of the Human Language Technology Conference – Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-2006), New York.

Hein, Anna Sagvall (1997) 'Scania Swedish-A Basis for Multilingual Translation', in *Translating and the Computer 19*, London: The Association for Information Management.

http://www.alchemysoftware.ie/index.html.

http://www.helicon.co.at/aboutus.html.

http://www.internetworldstats.com/stats.htm.

http://www.lisa.org/Glossary.

http://www2.multilizer.com/company.

http://www.passolo.com.

http://www.schaudin.com.

Hu, Qingping 胡清平 (2005)〈受控語言及其在漢英機器翻譯裏的應用前景〉(Controlled Language and Its Prospective Application in Chinese-English Machine Translation), in Luo Xuanmin 羅選民 (ed.) 《語言認識與翻譯研究》(*Language, Cognition and Translation Studies*), Beijing: Foreign Language Press 外文出版社, 364–372.

Hurst, Matthew F. (1997) 'Parsing for Targeted Errors in Controlled Languages', in Ruslan Mitkov and Nicolas Nicolov (eds) *Recent Advances in Natural Language Processing*, Amsterdam and Philadelphia: John Benjamins, 59–70.

Joy, Lorna (2002) 'Translating Tagged Text – Imperfect Matches and a Good Finished Job', *Translating and the Computer 24*, London: The Association for Information Management.

Kamprath, Christine, Eric Adolphson, Teruko Mitamura, and Eric H. Nyberg (1998) 'Controlled Language Multilingual Document Production: Experience with Caterpillar Technical English', in *Proceedings of the 2nd International Workshop on Controlled Language Applications* (*CLAW-98*), Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 51–61.

Kay, Martin (1980) 'The Proper Place of Men and Machines in Language Translation', Research Report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, CA.

Lehtola, Aarno, Jarno Tenni, and Catherine Bounsaythip (1998) 'Controlled Language—An Introduction', in *Proceedings of the 2nd International Workshop on Controlled Language Applications*, Carnegie Mellon University, Pittsburgh, PA, 16–29.

Lockwood, Rose (2000) 'Machine Translation and Controlled Authoring at Caterpillar', in Robert C. Sprung (ed.) *Translating into Success: Cutting-edge Strategies for Going Multilingual in a Global Age*, Amsterdam and Philadelphia: John Benjamins, 187–202.

Loong, Cheong Tong (1989) 'A Data-driven Control Strategy for Grammar Writing Systems', *Machine Translation* 4(4): 281–297.

Lockwood, Rose (2000) 'Machine Translation and Controlled Authoring at Caterpillar', in Robert C. Sprung (ed.) *Translating into Success: Cutting-edge Strategies for Going Multilingual in a Global Age*, Amsterdam and Philadelphia: John Benjamins, 187–202.

Lux, Veronika and Eva Dauphin (1996) 'Corpus Studies: A Contribution to the Definition of a Controlled Language', in *Proceedings of the 1st International Workshop on Controlled Language Applications* (*CLAW-96*), Leuven, Belgium, 193–204.

Matsuda, Junichi and Hiroyuki Kumai (1999) 'Transfer-based Japanese-Chinese Translation Implemented on an E-mail System', in *Proceedings of MT Summit VII: MT in the Great Translation Era*, 13–17 September 1999, Kent Ridge Digital Labs, Singapore.

Melby, Alan K. (1995) The Possibility of Language: A Discussion of the Nature of Language, with Implications for Human and Machine Translation, Amsterdam and Philadelphia: John Benjamins.

Melby, Alan K. (2012) 'Terminology in the Age of Multilingual Corpora', *The Journal of Specialized Translation* 18: 7–29.

Mitamura, Teruko, Eric H. Nyberg, and Jaime G. Carbonell (1994) 'KANT: Knowledge-based, Accurate Natural Language Translation', in *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas: Technology Partnerships for Crossing the Language Barrier* (*AMTA-1*), 5–8 October 1994, Columbia, MD, 232–233.

Mitamura, Teruko and Eric H. Nyberg (1995) 'Controlled English for Knowledge Based MT: Experience with the KANT System', in *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation* (*TMI-95*), Leuven, Belgium, 158–172.

Mitamura, Teruko (1999) 'Controlled Language for Multilingual Machine Translation', in *Proceedings of MT Summit VII: MT in the Great Translation Era*, Singapore, 46–52.

Mitamura, Teruko, Eric H. Nyberg, Kathy Baker, Peter Cramer, Jeongwoo Ko, David Svoboda, and Michael Duggan (2002) 'The KANTOO MT System: Controlled Language Checker and Lexical Maintenance Tool', in Stephen D. Richardson (ed.) *AMTA-02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, Machine Translation: From Research to Real Users*, 6–12 October 2002, Tiburon, CA, 244–247.

Murphy, Dawn, Jane Mason, and Stuart Sklair (1998) 'Improving Translation at the Source', in *Translating and the Computer 20*, London: The Association for Information Management.

Nida, Eugene A. (1969) 'Science of Translation', *Language* 45(3): 483–498.

Nida, Eugene A. and Charles R. Taber ([1969] 1982) *The Theory and Practice of Translation*, Leiden: E.J. Brill.

Nyberg, Eric H. and Teruko Mitamura (1992) 'The KANT System: Fast, Accurate, High-quality Translation in Practical Domains', in *Proceedings of the 14th International Conference of Computational Linguistics* (*COLING-92*), 23–28 August 1992, Nantes, France, 1069–1073.

Nyberg, Eric H., Teruko Mitamura, and Jaime G. Carbonell (1997) 'The KANT Machine System: From R&D to Initial Deployment', in *LISA Workshop on Intergrating Advanced Translation Technology*, Seattle, WA.

Nyberg, Eric H., Christine Kamprath, and Teruko Mitamura (1998) 'The KANT Translation System: from R&D to Large-Scale Deployment', *LISA Newsletter* 2(1): 1–7.

Nyberg, Eric H. and Teruko Mitamura (2000) 'The KANTOO Machine Translation Environment', in John S. White (ed.) *Envisioning Machine Translation in the Information Future*, Berlin: Springer Verlag, 192–195.

Nyberg, Eric H., Teruko Mitamura, and Willem-Olaf Huijsen (2003) 'Controlled Language for Authoring and Translation', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 245–281.

Probst, Katharina and Lori S. Levin (2002) 'Challenges in Automated Elicitation of a Controlled Bilingual Corpus', in *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation* (*TMI-2002*), Keihanna, Japan, 157–167.

Rico, Celia and Enrique Torrejon (2004) 'Controlled Translation as a New Translation Scenario – Training the Future User', in *Translating and the Computer 26*, London: The Association for Information Management.

Rooke, Robert (1985) 'Electronic Mail', in Catriona Picken (ed.) *Translation and Communication: Translating and the Computer 6*, London: The Association for Information Management, 105–115.

Roturier, Johann (2004) 'Assessing Controlled Language Rules: Can They Improve Performance of Commercial Machine Translation Systems?' *Translating and the Computer 26*, London: The Association for Information Management.

Ruffino, J. Richard (1985) 'The Impact of Controlled English on Machine Translation', in Patricia E. Newman (ed.) *American Translators Association Conference – 1985*, Medford, NJ: Learned Information, Inc., 157–162.

Sofer, Morry (2009) *The Translator's Handbook*, Rockville, MD: Schreiber Publishing.

Steiner, George ([1975] 1992) *After Babel: Aspect of Language and Translation*, 2nd edition, Oxford: Oxford University Press.

Torrejón, Enrique (2002) 'Controlled Translation: A New Teaching Scenario Tailor-made for the Translation Industry', in *Proceedings of the 6th Workshop of the European Association for Machine Translation: Teaching Machine Translation*, Manchester, England, 107–116.

van der Eijk, Pim and Jacqueline van Wees (1998) 'Supporting Controlled Language Authoring', in Proceedings of the 3rd Workshop of the European Association for Machine Translation: Translation Technology: Integration in the Workflow Environment, Geneva, Switzerland, 65–70.

Wasson, Mark (2000) 'Large-scale Controlled Vocabulary Indexing for Named Entities', in *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, DC, USA, 276–281.

Western Standard (2011) *Fluency Translation Suite 2011Fluency User Manual* V2.5.04. Utah: © 2009–2011 Western Standard Translation. Available at: http://www.westernstandard.com/FluencyInstalls/FluencyDocumentation.pdf.

Wilss, Wolfram (1982) *The Science of Translation: Problems and Methods*, Tubingen, Germany: Gunter Narr Verlag.

Windows (2012) "A History of Windows," *Windows*.

Wojcik, Richard H. and Heather Holmback (1996) 'Getting a Controlled Language Off the Ground at Boeing', in *Proceedings of the 1st International Workshop on Controlled Language Applications* (*CLAW-96*), Katholieke Universiteit Leuven, Belgium, 22–31.

XTM International Ltd. (2012) *XTM for CMS Explained*, Bucks: XTM International Ltd. Available at: http://www.xtm-intl.com/files/content/xtm/resources/XTM%20for%20CMS%20Explained%202012-01.pdf.

Zydron, Andrzej (2003) 'Xml: tm – Using XML Technology to Reduce the Cost of Authoring and Translation', in *Translating and the Computer 25*, London: The Association for Information Management.

3

COMPUTER-AIDED TRANSLATION

Systems

Ignacio Garcia

UNIVERSITY OF WESTERN SYDNEY, AUSTRALIA

Introduction

Computer-aided Translation (CAT) systems are software applications created with the specific purpose of facilitating the speed and consistency of human translators, thus reducing the overall costs of translation projects while maintaining the earnings of the contracted translators and an acceptable level of quality. At its core, every CAT system divides a text into 'segments' (normally sentences, as defined by punctuation marks) and searches a bilingual memory for identical (*exact match*) or similar (*fuzzy match*) source and translation segments. Search and recognition of terminology in analogous bilingual glossaries are also standard. The corresponding search results are then offered to the human translator as prompts for adaptation and reuse.

CAT systems were developed from the early 1990s to respond to the increasing need of corporations and institutions to target products and services toward other languages and markets (localization). Sheer volume and tight deadlines (simultaneous shipment) required teams of translators to work concurrently on the same source material. In this context, the ability to reuse vetted translations and to consistently apply the same terminology became vital. Once restricted to technical translation and large localization projects in the nineties, CAT systems have since expanded to cater for most types of translation, and most translators, including non-professionals, can now benefit from them.

This overview of CAT systems includes only those computer applications specifically designed with translation in mind. It does not discuss word processors, spelling and grammar checkers, and other electronic resources which, while certainly of great help to translators, have been developed for a broader user base. Nor does it include applications such as concordancers which, although potentially incorporating features similar to those in a typical CAT system, have been developed for computational linguists.

Amongst the general class of translation-focused computer systems, this will centre only on applications that assist human translators by retrieving human-mediated solutions, not those that can fully provide a machine-generated version in another language. Such Machine Translation (MT) aids will be addressed only in the context of their growing presence as optional adjuncts in modern-day CAT systems.

CAT systems fundamentally enable the reuse of past (human) translation held in so-called translation memory (TM) databases, and the automated application of terminology held in terminology databases. These core functionalities may be supplemented by others such as alignment tools, to create TM databases from previously translated documents, and term extraction tools, to compile searchable term bases from TMs, bilingual glossaries, and other documents. CAT systems may also assist in extracting the translatable text out of heavily tagged files, and in managing complex translation projects with large numbers and types of files, translators and language pairs while ensuring basic linguistic and engineering quality assurance.

CAT Systems have variously been known in both the industry and literature as CAT tools, TM, TM tools (or systems or suites), translator workbenches or workstations, translation support tools, or latterly translation environment tools (TEnTs). Despite describing only one core component, the vernacular term of TM has been widely employed: as a label for a human-mediated process, it certainly stands in attractive and symmetrical opposition to MT. Meanwhile, the CAT acronym has been considered rather too catholic in some quarters, for encompassing strict translation-oriented functionality plus other more generic features (word processing, spell checking etc.).

While there is presently no consensus on an 'official' label, CAT will be used here to designate the suites of tools that translators will commonly encounter in modern workflows. Included within this label will be the so-called localization tools – a specific sub-type which focuses on the translation of software user interfaces (UIs), rather than the 'traditional' user help and technical text. Translation Memory or TM will be used in its actual and literal sense as the database of stored translations.

Historically, CAT system development was somewhat *ad hoc*, with most concerted effort and research going into MT instead. CAT grew organically, in response to the democratization of processing power (personal computers opposed to mainframes) and perceived need, with the pioneer developers being translation agencies, corporate localization departments, and individual translators. Some systems were built for in-house use only, others to be sold.

Hutchins' *Compendium of Translation Software: Directory of Commercial Machine Translation Systems and Computer-aided Translation Support Tools* lists (from 1999 onwards) 'all known systems of machine translation and computer-based translation support tools that are currently available for purchase on the market' (Hutchins 1999–2010: 3). In this *Compendium*, CAT systems are included under the headings of 'Terminology management systems', 'Translation memory systems/components' and 'Translator workstations'. By January 2005, said categories boasted 23, 31 and 9 products respectively (with several overlaps), and although a number have been discontinued and new ones created, the overall figures have not changed much during the past decade. Some *Compendium* entries have left a big footprint in the industry while others do not seem to be used outside the inner circle of its developer.

The essential technology, revolving around sentence-level segmentation, was fully developed by the mid-1990s. The offerings of leading brands would later increase in sophistication, but for over a decade the gains centred more on stability and processing power than any appreciably new ways of extracting extra language-data leverage. We refer to this as the classic period, discussed below in the next section. From 2005 onwards, a more granular approach towards text reuse has emerged; the amount of addressable data expanded, and the potential scenarios for CAT-usage widen. These new trends are explored in the Current CAT Systems section.

Classic CAT systems (1995–2005)

The idea of computers assisting the translation process is directly linked to the development of MT, which began *c.*1949. Documentary references to CAT, as we understand it today, are

already found in the Automatic Language Processing Advisory Committee (ALPAC) report of 1966, which halted the first big wave of MT funding in the United States. In that era of vacuum tube mainframes and punch-cards, the report understandably found that MT (*mechanical translation*, as it was mostly known then) was a more time-consuming and expensive process than the traditional method, then frequently facilitated by dictation to a typist. However, the report did support funding for Computational Linguistics, and in particular for what it called the 'machine-aided human translation' then being implemented by the Federal Armed Forces Translation Agency in Mannheim. A study included in the report (Appendix 12) showed that translators using electronic glossaries could reduce errors by 50 per cent and increase productivity by over 50 per cent (ALPAC 1996: 26, 79–86).

CAT systems grew out of MT developers' frustration at being unable to design a product which could truly assist in producing faster, cheaper and yet still useable translation. While terminology management systems can be traced back to Mannheim, the idea of databasing translations *per se* did not surface until the 1980s. During the typewriter era, translators presumably kept paper copies of their work and simply consulted them when the need arose. The advent of the personal computer allowed document storage as softcopy, which could be queried in a more convenient fashion. Clearly, computers might somehow be used to automate those queries, and that is precisely what Kay ([1980] 1997: 323) and Melby (1983: 174–177) proposed in the early 1980s.

The Translation Support System (TSS) developed by ALPS (Automated Language Processing Systems) in Salt Lake City, Utah, in the mid-1980s is considered the first prototype of a CAT system. It was later re-engineered by INK Netherlands as INK TextTools (Kingscott 1999: 7). Nevertheless, while the required programming was not overly complicated, the conditions were still not ripe for the technology's commercialization.

By the early 1990s this had changed: micro-computers with word processors displaced the typewriter from the translators' desks. Certain business-minded and technologically proficient translators saw a window of opportunity. In 1990, Hummel and Knyphausen (two German entrepreneurs who had founded Trados in 1984 and had already been using TextTools) launched their MultiTerm terminology database, with the first edition of the Translator's Workbench TM tool following in 1992. Also in 1992, IBM Deutschland commercialized its in-house developed Translation Manager 2, while large language service provider STAR AG (also German) launched its own in-house system, Transit, onto the market (Hutchins 1998: 418–419).

Similar products soon entered the arena. Some, such as Déjà Vu, first released in 1993, still retain a profile today; others such as the Eurolang Optimiser, well-funded and marketed at its launch (Brace 1992), were shortly discontinued. Of them all, it was Trados – thanks to successful European Commission tender bids in 1996 and 1997 – that found itself the tool of choice of the main players and, thus, the default industry standard.

By the mid-1990s, translation memory, terminology management, alignment tools, file conversion filters and other features were all present in the more advanced systems. The main components of that technology, which would not change much for over a decade, are described below.

The editor

A CAT system allows human translators to reuse translations from translation memory databases, and apply terminology from terminology databases. The editor is the system front-end that translators use to open a source file for translation, and query the memory and

terminology databases for relevant data. It is also the workspace in which they can write their own translations if no matches are found, and the interface for sending finished sentence pairs to the translation memory and terminology pairs to the term base.

Some classic CAT systems piggy-backed their editor onto third-party word processing software; typically Microsoft Word. Trados and Wordfast were the best known examples during this classic period. Most, however, decided on a proprietary editor. The obvious advantage of using a word-processing package such as Word is that users would already be familiar with its environment. The obvious disadvantage, however, is that if a file could not open normally in Word, then it could not be translated without prior processing in some intermediary application capable of extracting its translatable content. A proprietary editor already embodies such an intermediate step, without relying on Word to display the results.

Whether bolt-on or standalone, a CAT system editor first segments the source file into translation units, enabling the translator to work on them separately and the program to search for matches in the memory. Inside the editor window, the translator sees the active source segment displayed together with a workspace into which the system will import any hits from the memory and/or allow a translation to be written from scratch. The workspace can appear below (vertical presentation) or beside (horizontal or tabular presentation) the currently active segment.

The workflow for classic Trados in both its configurations, as Word macro, and the later proprietary 'Tag Editor' is the model for vertical presentation. The translator opens a segment, translates with assistance from matches if available, then closes this segment and opens the next. Any TM and glossary information relevant to the open segment appeared in a separate window, called Translator's Workbench. The inactive segments visible above and below the open segment provided the translator with co-text. Once the translation was completed and edited, the result was a bilingual ('uncleaned') file requiring 'clean up' into a monolingual target-language file. This model was followed by other systems, most notably Wordfast.

When the source is presented in side-by-side, tabular form, Déjà Vu being the classic example, the translator activates a particular segment by placing the cursor in the corresponding cell; depending on the (user adjustable) search settings, the most relevant database information is imported into the target cell on the right, with additional memory and glossary data presented either in a sidebar or at bottom of screen.

Independently of how the editor presents the translatable text, translators work either in interactive mode or in pre-translation mode. When using their own memories and glossaries they most likely work in interactive mode, with the program sending the relevant information from the databases as each segment is made 'live'. When memories and glossaries are provided by an agency or the end client, the source is first analysed against them and then any relevant entries either sorted and sent to the translators, or directly inserted into the source file in a process known as *pre-translation*. Translators apparently prefer the interactive mode but, during this period, most big projects involved pre-translation (Wallis 2006).

The translation memory

A translation memory or TM, the original coinage attributed to Trados founders Knyphausen and Hummel, is a database that contains past translations, aligned and ready for reuse in matching pairs of source and target units. As we have seen, the basic database unit is called a segment, and is normally demarcated by explicit punctuation – it is therefore commonly a sentence, but can also be a title, caption, or the content of a table cell.

A typical TM entry, sometimes called a translation unit or TU, consists of a source segment linked to its translation, plus relevant metadata (e.g. time/date and author stamp, client name, subject matter, etc.). The TM application also contains the algorithm for retrieving a matching translation if the same or a similar segment arises in a new text.

When the translator opens a segment in the editor window, the program compares it to existing entries in the database:

- If it finds a source segment in the database that precisely coincides with the segment the translator is working on, it retrieves the corresponding target as an exact match (or a 100 per cent match); all the translator need do is check whether it can be reused as-is, or whether some minor adjustments are required for potential differences in context.
- If it finds a databased source segment that is similar to the active one in the editor, it offers the target as a *fuzzy match* together with its degree of similarity, indicated as a percentage and calculated on the Levenshtein distance, i.e. the minimum number of insertions, deletions or substitutions required to make it equal; the translator then assesses whether it can be usefully adapted, or if less effort is required to translate from scratch; usually, only segments above a 70 per cent threshold are offered, since anything less is deemed more distracting than helpful.
- If it fails to find any stored source segment exceeding the pre-set match threshold, no suggestion is offered; this is called a *no match*, and the translator will need to translate that particular segment in the conventional way.

How useful a memory is for a particular project will not only depend on the number of segments in the database (simplistically, the more the better), but also on how related they are to the source material (the closer, the better). Clearly, size and specificity do not always go hand-in-hand.

Accordingly, most CAT tools allow users to create as many translation memories as they wish – thereby allowing individual TMs to be kept segregated for use in specific circumstances (a particular topic, a certain client, etc.), and ensuring internal consistency. It has also been common practice among freelancers to periodically dump the contents of multiple memories into one catch-all TM, known in playful jargon as a 'big mama'.

Clearly, any active TM is progressively enhanced because its number of segments grows as the translator works through a text, with each translated segment sent by default to the database. The more internal repetition, the better, since as the catchcry says 'with TM one need never translate the same sentence twice'. Most reuse is achieved when a product or a service is continually updated with just a few features added or altered – the ideal environment being technical translation (Help files, manuals and documentation), where consistency is crucial and repetition may be regarded stylistically as virtue rather than vice.

There have been some technical variations of strict sentence-based organization for the memories. Star-Transit uses file pairs as reference materials indexed to locate matches. Canadian developers came up with the concept of *bi-texts*, linking the match not to an isolated sentence but to the complete document, thus providing context. LogiTerm (Terminotix) and MultiTrans (MultiCorpora) are the best current examples, with the latter referring this as TextBased (rather than sentence-based) TM. In current systems, however, the lines of differentiation between stress on text or on sentence are blurred, with conventional TM indicating also when an exact match comes from the same context by naming it, depending on the brand, *context*, *101%*, *guaranteed* or *perfect* match, and text-based able to import and work with sentence-based memories. All current systems can import and export memories in Translation Memory

eXchange (TMX) format, an open XML standard created by OSCAR (Open Standards for Container/Content Allowing Re-use), a special interest group of LISA (Localization Industry Standards Association).

The terminology feature

To fully exploit its data-basing potential, every CAT system requires a terminology feature. This can be likened conceptually to the translation memory of reusable segments, but instead functions at term level by managing searchable/retrievable glossaries containing specific pairings of source and target terms plus associated metadata.

Just as the translation memory engine does, the terminology feature monitors the currently active translation segment in the editor against a database – in this case, a bilingual glossary. When it detects a source term match, it prompts with the corresponding target rendering. Most systems also implement some fuzzy terminology recognition to cater for morphological inflections.

As with TMs, bigger is not always better: specificity being equally desirable, a glossary should also relate as much as possible to a given domain, client and project. It is therefore usual practice to compile multiple term bases which can be kept segregated for designated uses (and, of course, periodically dumped into a 'big mama' term bank too).

Term bases come in different guises, depending upon their creators and purposes. The functionalities offered in the freelance and enterprise versions of some CAT systems tend to reflect these needs.

Freelance translators are likely to prefer unadorned bilingual glossaries which they build up manually – typically over many years – by entering source and target term pairings as they go. Entries are normally kept in local computer memory, and can remain somewhat *ad hoc* affairs unless subjected to time-consuming maintenance. A minimal approach offers ease and flexibility for different contexts, with limited (or absent) metadata supplemented by the translator's own knowledge and experience.

By contrast, big corporations can afford dedicated bureaus staffed with trained terminologists to both create and maintain industry-wide multilingual term bases. These will be enriched with synonyms, definitions, examples of usage, and links to pictures and external information to assist any potential users, present or future. For large corporate projects it is also usual practice to construct product-specific glossaries which impose uniform usages for designated key terms, with contracting translators or agencies being obliged to abide by them.

Glossaries are valuable resources, but compiling them more rapidly via database exchanges can be complicated due to the variation in storage formats. It is therefore common to allow export/import to/from intermediate formats such as spreadsheets, simple text files, or even TMX. This invariably entails the loss or corruption of some or even all of the metadata. In the interests of enhanced exchange capability, a Terminology Base eXchange (TBX) open standard was eventually created by OSCAR/LISA. Nowadays most sophisticated systems are TBX compliant.

Despite the emphasis traditionally placed on TMs, experienced users will often contend that it is the terminology feature which affords the greatest assistance. This is understandable if we consider that translation memories work best in cases of incremental changes to repetitive texts, a clearly limited scenario. By contrast, recurrent terminology can appear in any number of situations where consistency is paramount.

Interestingly, terminology features – while demonstrably core components – are not always 'hard-wired' into a given CAT system. Trados is one example, with its MultiTerm tool

presented as a stand-alone application beside the company's translation memory application (historically the Translator's Workbench). Déjà Vu on the other hand, with its proprietary interface, has bundled everything together since inception.

Regardless, with corporations needing to maintain lexical consistency across user interfaces, Help files, documentation, packaging and marketing material, translating without a terminology feature has become inconceivable. Indeed, the imposition of specific vocabulary can be so strict that many CAT systems have incorporated quality assurance (QA) features which raise error flags if translators fail to observe authorised usage from designated term bases.

Translation management

Technical translation and localization invariably involve translating great numbers (perhaps thousands) of files in different formats into many target languages using teams of translators. Modest first-generation systems, such as the original Wordfast, handled files one at a time and catered for freelance translators in client-direct relationships. As globalization pushed volumes and complexities beyond the capacities of individuals and into the sphere of translation bureaus or language service providers (LSPs), CAT systems began to acquire a management dimension.

Instead of the front end being the translation editor, it became a 'project window' for handling multiple files related to a specific undertaking – specifying global parameters (source and target languages, specific translation memories and term bases, segmentation rules) and then importing a number of source files into that project. Each file could then be opened in the editor and translated in the usual way.

These changes also signalled a new era of remuneration. Eventually all commercial systems were able to batch-process incoming files against the available memories, and pre-translate them by populating the target side of the relevant segments with any matches. Effectively, that same analysis process meant quantifying the number and type of matches as well as any internal repetition, and the resulting figures could be used by project managers to calculate translation costs and time. Individual translators working with discrete clients could clearly project-manage and translate alone, and reap any rewards in efficiency themselves. However, for large agencies with demanding clients, the potential savings pointed elsewhere.

Thus by the mid-1990s it was common agency practice for matches to be paid at a fraction of the standard cost per word. Translators were not enthused with these so-called 'Trados discounts' and complained bitterly on the Lantra-L and Yahoo Groups CAT systems users' lists.

As for the files themselves, they could be of varied types. CAT systems would use the relevant filters to extract from those files the translatable text to present to the translator's editor. Translators could then work on text that kept the same appearance, regardless of its native format. Inline formatting (bold, italics, font, colour etc.) would be displayed as read-only tags (typically numbers displayed in colours or curly brackets) while structural formatting (paragraphs, justification, indenting, pagination) would be preserved in a template to be reapplied upon export of the finished translation. The proper filters made it possible to work on numerous file types (desktop publishers, HTML encoders etc.) without purchasing the respective licence or even knowing how to use the creator software.

Keeping abreast of file formats was clearly a challenge for CAT system developers, since fresh converter utilities were needed for each new release or upgrade of supported types. As the information revolution gathered momentum and file types multiplied, macros that sat on third-party software were clearly unwieldy, so proprietary interfaces became standard (witness Trados' shift from Word to Tag Editor).

There were initiatives to normalize the industry so that different CAT systems could talk effectively between each other. The XML Localisation Interchange File Format (XLIFF) was created by the Organization for the Advancement of Structured Information Standards (OASIS) in 2002, to simplify the processes of dealing with formatting within the localization industry. However, individual CAT designers did not embrace XLIFF until the second half of the decade.

By incorporating project management features, CAT systems had facilitated project sharing amongst teams of translators using the same memories and glossaries. Nevertheless, their role was limited to assembling a translation 'kit' with source and database matches. Other in-house or third-party systems (such as LTC Organiser, Project-Open, Projetex, and Beetext Flow) were used to exchange files and financial information between clients, agencies and translators. Workspace by Trados, launched in 2002 as a first attempt at whole-of-project management within a single CAT system, proved too complex and was discontinued in 2006. Web-based systems capable of dealing with these matters in a much simpler and effective fashion started appearing immediately afterwards.

Alignment and term extraction tools

Hitherto the existence of translation memories and term bases has been treated as a given, without much thought as to their creation. Certainly, building them barehanded is easy enough, by sending source and target pairings to the respective database during actual translation. But this is slow, and ignores large amounts of existing matter that has already been translated known variously as parallel corpora, bi-texts or legacy material.

Consider for example the Canadian Parliament's Hansard record, kept bilingually in English and French. If such legacy sources and their translations could be somehow lined up side-by-side (as if already in a translation editor), then they would yield a resource that could be easily exploited by sending them directly into a translation memory. Alignment tools quickly emerged at the beginnings of the classic era, precisely to facilitate this task. The first commercial alignment tool was T Align, later renamed Trados WinAlign, launched in 1992.

In the alignment process parallel documents are paired, segmented and coded appropriately for import into the designated memory database. Segmentation would follow the same rules used in the translation editor, theoretically maximizing reuse by treating translation and alignment in the same way within a given CAT system. The LISA/OSCAR Segmentation Rules eXchange (SRX) open standard was subsequently created to optimize performance across systems.

Performing an alignment is not always straightforward. Punctuation conventions differ between languages, so the segmentation process can frequently chunk a source and its translation differently. An operator must therefore work manually through the alignment file, segment by segment, to ensure exact correspondence. Alignment tools implement some editing and monitoring functions as well so that segments can be split or merged as required and extra or incomplete segments detected, to ensure a perfect 1:1 mapping between the two legacy documents. When determining whether to align apparently attractive bi-texts, one must assess whether the gains achieved through future reuse from the memories will offset the attendant cost in time and effort.

Terminology extraction posed more difficulties. After all, alignments could simply follow punctuation rules; consistently demarcating terms (with their grammatical and morphological inflections, noun and adjectival phrases) was another matter. The corresponding tools thus began appearing towards the end of the classic period, and likewise followed the same well-

worn path from standalones (Xerox Terminology Suite being the best known) to full CAT system integration.

Extraction could be performed on monolingual (usually the source) or bilingual text (usually translation memories) and was only semi-automated. That is, the tool would offer up terminology candidates from the source text, with selection based on frequency of appearance. Since an unfiltered list could be huge, users set limiting parameters such as the maximum number of words a candidate could contain, with a stopword list applied to skip the function words. When term-mining from translation memories, some programs were also capable of proposing translation candidates from the target text. Whatever their respective virtues, term extractors could only propose: everything had to be vetted by a human operator.

Beyond purely statistical methods, some terminology extraction tools eventually implemented specific parsing for a few major European languages. After its acquisition of Trados in 2006, SDL offered users both its SDLX PhraseFinder and Trados MultiTerm Extract. PhraseFinder was reported to work better with those European languages that already had specific algorithms, while MultiTerm Extract seemed superior in other cases (Zetzsche 2010: 34).

Quality assurance

CAT systems are intended to help translators and translation buyers by increasing productivity and maintaining consistency even when teams of translators are involved in the same project. They also contribute significantly to averting errors through automated quality assurance (QA) features that now come as standard in all commercial systems.

CAT QA modules perform linguistic controls by checking terminology usage, spelling and grammar, and confirming that any non-translatable items (e.g. certain proper nouns) are left unaltered. They can also detect if numbers, measurements and currency are correctly rendered according to target language conventions. At the engineering level, they ensure that no target segment is left untranslated, and that the target format tags match the source tags in both type and quantity. With QA checklist conditions met, the document can be confidently exported back to its native format for final proofing and distribution.

The first QA tools (such as QA Distiller, Quintillian, or Error Spy) were developed as third-party standalones. CAT systems engineers soon saw that building in QA made technical and business sense, with Wordfast leading the way.

What is also notable here is the general trend of consolidation, with QA tools following the same evolutionary path as file converters, word count and file analysis applications, alignment tools and terminology extraction software. CAT systems were progressively incorporating additional features, leaving fewer niches where third-party developers could remain commercially viable by designing plug-ins.

Localization tools: a special CAT system sub-type

The classic-era CAT systems described above worked well enough with Help files, manuals and web content in general; they fell notably short when it came to software user interfaces (UIs) with their drop-down menus, dialogue boxes, pop-up help, and error messages. The older class of texts retained a familiar aspect, analogous to a traditional, albeit electronically enhanced, 'book' presentation of sequential paragraphs and pages. The new texts of the global information age operated in a far more piecemeal, visually oriented and random-access fashion, with much of the context coming from their on-screen display. The contrast was simple yet profound: the printable versus the viewable.

Moreover, with heavy computational software (for example, 3D graphics) coded in programming languages, it could be problematic just identifying and extracting the translatable (i.e. displayable) text from actual instructions. Under the circumstances, normal punctuation rules were of no use in chunking, so localizers engineered a new approach centred on 'text strings' rather than segments. They also added a visual dimension – hardly a forte of conventional CAT – to ensure the translated text fitted spatially, without encroaching on other allocated display areas.

These distinctions were significant enough to make localization tools notably different from the CAT systems described above. However, to maintain consistency within the UI and between the UI *per se* and its accompanying Help and documentation, the linguistic resources (glossaries, and later memories too) were shared by both technologies.

The best known localization tools are Passolo (now housed in the SDL stable) and Catalyst (acquired by major US agency TransPerfect). There are also many others, both commercial (Multilizer, Sisulizer, RCWinTrans) and open source (KBabel, PO-Edit). Source material aside, they all operated in much the same way as their conventional CAT brethren, with translation memories, term bases, alignment and term extraction tools, project management and QA.

Eventually, as industry efforts at creating internationalization standards bore fruit, software designers ceased hard-coding translatable text and began placing it in XML-based formats instead. Typical EXE and DLL files give way to Java and .NET, and more and more software (as opposed to text) files could be processed within conventional CAT systems.

Nowadays, the distinctions which engendered localization tools are blurring, and they no longer occupy the field exclusively. There are unlikely to disappear altogether, however, since new software formats will always arise and specialized tools will always address them faster.

CAT systems uptake

The uptake of CAT systems by independent translators was initially slow. Until the late 1990s, the greatest beneficiaries of the leveraging and savings were those with computer power – corporate buyers and language service providers. But CAT ownership conferred an aura of professionalism, and proficient freelancers could access the localization industry (which, as already remarked, could likewise access *them*). In this context, from 2000 most professional associations and training institutions became keen allies in CAT system promotion. The question of adoption became not if but which one – with the dilemma largely hinging on who did the translating and who commissioned it, and epitomized by the legendary Déjà Vu versus Trados rivalry.

Trados had positioned itself well with the corporate sector, and for this reason alone was a pre-requisite for certain jobs. Yet by and large freelancers preferred Déjà Vu, and while today the brand may not be so recognizable, it still boasts a loyal user base.

There were several reasons why Déjà Vu garnered such a loyal following. Freelancers considered it a more user-friendly and generally superior product. All features came bundled together at an accessible and stable price, and the developer (Atril) offered comprehensive – and free – after-sales support. Its influence was such that its basic template can be discerned in other CAT systems today. Trados meanwhile remained a rather unwieldy collection of separate applications that required constant and expensive upgrades. For example, freelancers purchasing Trados 5.5 Freelance got rarefied engineering or management tools such as WorkSpace, T-Windows, and XML Validator, but had to buy the fundamental terminology application MultiTerm separately (Trados 2002). User help within this quite complex scenario also came at a price.

The pros and cons of the two main competing packages, and a degree of ideology, saw passions run high. The Lantra-L translators' discussion list (founded in 1987, the oldest and one of the most active at the time) would frequently reflect this, especially in the famed Trados vs. Déjà Vu 'holy wars', the last being waged in August 2002.

Wordfast, which first appeared in 1999 in its 'classic' guise, proved an agile competitor in this environment. It began as a simple Word macro akin to the early Trados, with which it maintained compatibility. It also came free at a time when alternatives were costly, and began to overtake even Déjà Vu in freelancers' affections. Users readily accepted the small purchase price the developer eventually set in October 2002.

LogiTerm and especially MultiTrans also gained a significant user base during the first years of the century. MetaTexis, WordFisher and TransSuite 2000 had also a small but dedicated base that shows in their users' Yahoo Groups. Completing the panorama were a number of in-house only systems, such Logos' Mneme and Lionbridge's ForeignDesk. However, the tendency amongst most large translation agencies was to either stop developing and buy off-the-shelf (most likely Trados), or launch their own offerings (as SDL did with its SDLX).

There are useful records for assembling a snapshot of relative CAT system acceptance in the classic era. From 1998 onwards, CAT system users began creating discussion lists on Yahoo Groups, and member numbers and traffic on these lists give an idea of respective importance. By June 2003 the most popular CAT products, ranked by their list members, were Wordfast (2205) and Trados (2138), then Déjà Vu (1233) and SDLX (537). Monthly message activity statistics were topped by Déjà Vu (1169), followed by Wordfast (1003), Trados (438), Transit (66) and SDLX (30).

All commercial products were Trados compatible, able to import and export the RTF and TTX files generated by Trados. Windows was the default platform in all cases, with only Wordfast natively supporting Mac.

Not all activity occurred in a commercial context. The Free and Open Source Software (FOSS) community also needed to localize software and translate documentation. That task fell less to conventional professional translators, and more to computer-savvy and multilingual collectives who could design perfectly adequate systems without the burden of commercial imperatives. OmegaT, written in Java and thus platform independent, was and remains the most developed open software system.

Various surveys on freelancer CAT system adoption have been published, amongst them LISA 2002, eColore 2003, and LISA 2004, with the most detailed so far by London's Imperial College in 2006. Its most intriguing finding was perhaps not the degree of adoption (with 82.5 per cent claiming ownership) or satisfaction (a seeming preference for Déjà Vu), but the 16 per cent of recipients who reported buying a system without ever managing to use it (Lagoudaki 2006: 17).

Current CAT systems

Trados was acquired by SDL in 2005, to be ultimately bundled with SDLX and marketed as SDL Trados 2006 and 2007. The release of SDL Trados Studio 2009 saw a shift that finally integrated all functions into a proprietary interface; MultiTerm was now included in the licence, but still installed separately. Curiously, there has been no new alignment tool while SDL has been at the Trados helm: it remains WinAlign, still part of the 2007 package which preserves the old Translator's Workbench and Tag Editor. Holders of current Trados licences (Studio 2011 at time of writing) have access to all prior versions through downloads from SDL's website.

Other significant moves were occurring: Lingotek, launched in 2006, was the first fully web-based system and pioneered the integration of TM with MT. Google released its own web-based Translator Toolkit in 2009, a CAT system pitched for the first time at non-professionals. Déjà Vu along with X2, Transit with NXT and MultiTrans with Prism (latest versions at writing) have all kept a profile. Wordfast moved beyond its original macro (now Wordfast Classic) to Java-coded Wordfast Professional and web-based Wordfast Anywhere.

Translation presupposes a source text, and texts have to be written by someone. Other software developers had looked at this supply side of the content equation and begun creating authoring tools for precisely the same gains of consistency and reuse. Continuing the consolidation pattern we have seen, CAT systems began incorporating them. Across was the first, linking to crossAuthor. The flow is not just one-way: Madcap, the developer of technical writing aid Flare, has moved into the translation sphere with Lingo.

Many other CAT systems saw the light in the last years of the decade and will also gain a mention below, when illustrating new features now supplementing the ones carried out from the classic era. Of them, MemoQ (Kilgray), launched in 2009, seems to have gained considerable freelance following.

The status of CAT systems – their market share, and how they are valued by users – is less clear-cut than it was ten years ago when Yahoo Groups user lists at least afforded some comparative basis. Now developers seek tighter control over how they receive and address feedback. SDL Trados led with its Ideas, where users could propose and vote on features to extend functionality, then with SDL OpenExchange, allowing the more ambitious to develop their own applications. Organizing conferences, as memoQfest does, is another way of both showing and garnering support.

The greatest determining factors throughout the evolution of CAT have been available computer processing power and connectivity. The difference in scope between current CAT systems and those in the 1990s can be better understood within the framework of two trends: cloud computing, where remote (internet) displaced local (hard drive) storage and processing; and Web 2.0, with users playing a more active role in web exchanges.

Cloud computing in particular has made it possible to meld TM with MT, access external databases, and implement more agile translation management systems capable of dealing with a myriad of small changes with little manual supervision. The wiki concept and crowd sourcing (including crowd-based QA) have made it possible to harness armies of translation aficionados to achieve outbound-quality results. Advances in computational linguistics are supplying grammatical knowledge to complement the purely statistical algorithms of the past. Sub-segmental matching is also being attempted. On-screen environments are less cluttered and more visual, with translation editors capable of displaying in-line formatting (fonts, bolding etc.) instead of coded tags. Whereas many editing tasks were ideally left until after re-export to native format, CAT systems now offer advanced aids – including Track Changes – for revisers too. All these emerging enhanced capabilities, which are covered below, appropriately demarcate the close of the classic CAT systems era.

From the hard-drive to the web-browser

Conventional CAT systems of the 1990s installed locally on a hard-drive; some such as Wordfast simply ran as macros within Word. As the technology expanded with computer power, certain functionalities would be accessed over a LAN and eventually on a server. By the middle 2000s, some CAT systems were already making the connectivity leap to software as a service (SaaS).

The move had commenced at the turn of this century with translation memories and term bases. These were valuable resources, and clients wanted to safeguard them on servers. This forced translators to work in 'web-interactive' mode – running their CAT systems locally, but accessing client-designated databases remotely via a login. It did not make all translators happy: it gave them less control over their own memories and glossaries, and made work progress partially dependent on internet connection speed. Language service providers and translation buyers, however, rejoiced. The extended use of Trados-compatible tools instead of Trados had often created engineering hitches through corrupted file exports. Web access to databases gave more control and uniformity.

The next jump came with Logoport. The original version installed locally as a small add-in for Microsoft Word, with the majority of computational tasks (databasing and processing) now performed on the server. Purchased by Lionbridge for in-house use, it has since been developed into the agency's current GeoWorkz Translation Workspace.

The first fully-online system arrived in the form of Lingotek, launched in 2006. Other web-based systems soon followed: first Google Translator Toolkit and Wordfast Anywhere, then Crowd.in, Text United, Wordbee and XTM Cloud, plus open source GlobalSight (Welocalize) and Boltran. Traditional hard drive-based products also boast web-based alternatives, including SDL Trados (WorldServer) and Across.

The advantages of web-based systems are obvious. Where teams of translators are involved, a segment just entered by one can be almost instantly reused by all. Database maintenance becomes centralized and straightforward. Management tasks can also be simplified and automated – most convenient in an era with short content lifecycles, where periodic updates have given way to streaming changes.

Translators themselves have been less enthused, even though browser-based systems neatly circumvent tool obsolescence and upgrade dilemmas (Muegge 2012: 17–21). Among Wordfast adherents, for example, the paid Classic version is still preferred over its online counterpart, the free Wordfast Anywhere. Internet connectivity requirements alone do not seem to adequately explain this, since most professional translators already rely on continuous broadband for consulting glossaries, dictionaries and corpora. As countries and companies invest in broadband infrastructure, response lagtimes seem less problematic too. Freelancer resistance thus presumably centres on the very *raison d'être* of web-based systems: remote administration and resource control.

Moving to the browser has not favoured standardization and interoperability ideals either. With TMX having already been universally adopted and most systems being XLIFF compliant to some extent, retreating to isolated log-in access has hobbled further advances in cross-system communicability. A new open standard, the Language Interoperability Portfolio (Linport), is being developed to address this. Yet as TAUS has noted, the translation industry still is a long way behind the interoperability achieved in other industries such as banking or travel (Van der Meer 2011).

Integrating machine translation

Research into machine translation began in the mid-twentieth century. Terminology management and translation memory happened to be an offshoot of research into full automation. The lack of computational firepower stalled MT progress for a time, but it was renewed as processing capabilities expanded. Sophisticated and continually evolving MT can be accessed now on demand through a web browser.

Although conventional rule-based machine translation (RBMT) is still holding its ground, there is a growing emphasis on statistical machine translation (SMT) for which, with appropriate

bilingual and monolingual data, it is easier to create new language-pair engines and customize existing ones for specific domains. What is more, if source texts are written consistently with MT in mind (see 'authoring tools' above) output can be significantly improved again. Under these conditions, even free on-line MT engines such as Google Translate and Microsoft Bing Translator, with *light* (or even no) post-editing may suffice, especially when *gisting* is more important than stylistic correctness.

Post-editing, the manual 'cleaning up' of raw MT output, once as marginal as MT itself, has gradually developed its own principles, procedures, training, and practitioners. For some modern localization projects, enterprises may even prefer customized MT engines and trained professional post-editors. As an Autodesk experiment conducted in 2010 showed, under appropriate conditions MT post-editing also 'allows translators to substantially increase their productivity' (Plitt and Masselott 2010: 15).

Attempts at augmenting CAT with automation began in the 1990s, but the available desktop MT was not really powerful or agile enough, trickling out as discrete builds on CD-ROM. As remarked above, Lingotek in 2006 was the first to launch a web-based CAT integrated with a mainframe powered MT; SDL Trados soon followed suit, and then all the others. With machines now producing useable first drafts, there are potential gains in pipelining MT-generated output to translators via their CAT editor. The payoff is twofold: enterprises can do so in a familiar environment (their chosen CAT system), whilst leveraging from legacy data (their translation memories and terminology databases).

The integration of TM with MT gives CAT users the choice of continuing working the traditional way (accepting or repairing *exact* matches, repairing or rejecting the *fuzzy* ones and translating from the source the *no matches*) or to populate those *no matches* with MT solutions for treatment akin to conventional fuzzy matches: modify if deemed helpful enough, or discard and translate from scratch.

While the process may seem straightforward, the desired gains in time and quality are not. As noted before, fixing fuzzy matches below a certain threshold (usually 70 per cent) is not viable; similarly, MT solutions should at least be of gisting quality to be anything other than a hindrance. This places translation managers at a decisional crossroad: trial and error is wasteful, so how does one *predict* the suitability of a text before MT processing?

Unfortunately, while the utility of MT and post-editing for a given task clearly depends on the engine's raw output quality, as yet there is no clear way of quantifying it. Standard methods such as the BLEU score (Papineni *et al.* 2002: 311–318) measure MT match quality against a reference translation, and thus cannot help to exactly predict performance on a previously untranslated sentence. Non-referenced methods, such as those based on confidence estimations (Specia 2011: 73–80), still require finetuning.

The next generation of CAT systems will foreseeably ascribe segments another layer of metadata to indicate whether the translation derives from MT (and if so which), and the steps and time employed achieving it. With the powerful analytic tools currently emerging, we might shortly anticipate evidence-based decisions regarding the language pairs, domains, engines, post-editors, and specific jobs for which MT integration into CAT localization workflow makes true business sense.

Massive external databases

Traditionally, when users first bought a CAT system, it came with empty databases. Unless purchasers were somehow granted external memories and glossaries (from clients, say) everything had to built up from zero. Nowadays that is not the only option, and from day one

it is possible to access data in quantities that dwarf any translator's – or for that matter, entire company's – lifetime output.

Interestingly, this situation has come about partly through SMT, which began its development using published bilingual corpora – the translation memories (minus the metadata) of the European Union. The highly useable translations achieved with SMT were a spur to further improvement, not just in the algorithms but in data quality and quantity as well. Since optimal results for any given task depend on feeding the SMT engine domain-specific information, the greater the volume one has, the better, and the translation memories created since the 1990s using CAT systems were obvious and attractive candidates.

Accordingly, corporations and major language service providers began compiling their entire TM stock too. But ambitions did not cease there, and initiatives have emerged to pool *all* available data in such a way that it can be sorted by language, client and subject matter. The most notable include the TAUS Data Association (TDA, promoted by the Translation Automation Users Society TAUS), MyMemory (Translated.com) and Linguee.com.

Now, these same massive translation memories that have been assembled to empower SMT can also significantly assist human translation. Free on-line access allows translators to tackle problematic sentences and phrases by querying the database, just as they would with the concordance feature in their own CAT systems and memories. The only hitch is working within a separate application, and transferring results across: what would be truly useful is the ability to access such data without ever needing to leave the CAT editor window. It would enable translators to query worldwide repositories of translation solutions and import any *exact* and *fuzzy* matches directly.

Wordfast was the first to provide a practical implementation with its Very Large Translation Memory (VLTM); it was closely followed by the Global, shared TM of the Google Translator Toolkit. Other CAT systems have already begun incorporating links to online public translation memories: MultiTrans has enabled access to TDA and MyMemory since 2010, and SDL Trados Studio and memoQ had MyMemory functionality soon afterwards.

Now that memories and glossaries are increasingly accessed online, it is conceivable that even the most highly resourced corporate players might also see a benefit to increasing their reach through open participation, albeit quarantining sensitive areas from public use. Commercial secrecy, ownership, prior invested value, and copyright are clearly counterbalancing issues, and the trade-off between going public and staying private is exercising the industry's best minds. Yet recent initiatives (e.g. TAUS) would indicate that the strain of coping with sheer translation volume and demand is pushing irrevocably toward a world of open and massive database access.

Sub-segmental reuse

Translation memory helps particularly with internal repetition and updates, and also when applied to a source created for the same client and within the same industry. Other than that, a match for the average size sentence is a coincidence. Most repetition happens below the sentence level, with the stock expressions and conventional phraseology that make up a significant part of writing. This posed a niggling problem, since it was entirely possible for sentences which did not return fuzzy matches to contain shorter *perfect* matches that were going begging.

Research and experience showed that low-value matches (usually under 70 per cent) overburdened translators, so most tools were set to ignore anything under a certain threshold. True, the concordancing tool can be used to conduct a search, but this is inefficient (and

random) since it relies on the translator's first identifying the need to do so, and it takes additional time. It would be much better if the computer could find and offer these phrase-level (or 'sub-segmental') matches all by itself – automated concordancing, so to speak.

Potential methods have been explored for years (Simard and Langlais 2001: 335–339), but have proven elusive to achieve. The early leader in this field was Déjà Vu with its Assemble feature, which offered portions that had been entered into the term base, the lexicon or the memory when no matches were available. Some translators loved it; others found it distracting (Garcia 2003).

It is only recently that all major developers have engaged with the task, usually combining indexing with predictive typing, suggestions popping up as the translator types the first letters. Each developer has its own implementation and jargon for sub-segmental matching: MultiTrans and Lingotek, following TAUS, call it Advanced Leveraging; memoQ refers to Longest Substring Concordance; Star-Transit has Dual Fuzzy, and Déjà Vu X2 has DeepMiner. Predictive typing is variously described as AutoSuggest, AutoComplete, AutoWrite etc.

A study sponsored by TAUS in 2007 reported that sub-segmental matching (or advanced leveraging in TAUS-speak), increased reuse by an average of 30 per cent over conventional reuse at sentence level only.

As discovered with the original Déjà Vu Assemble, what is a help to some is a distraction to others, so the right balance is needed between what (and how many) suggestions to offer. Once that is attained, one can only speculate on the potential and gains of elevating sub-segmental match queries from internal databases to massive external ones.

CAT systems acquire linguistic knowledge

In the classic era, it was MT applications that were language specific, with each pair having its own special algorithms; CAT systems were the opposite, coming as empty vessels that could apply the same databasing principles to whatever language combination the user chose. First generation CAT systems worked by seeking purely statistical match-ups between new segments and stored ones; as translation aids they could be powerful, but not 'smart'.

The term extraction tool Xerox Terminology Suite was a pioneer in introducing language-specific knowledge within a CAT environment. Now discontinued, its technology resurfaced in the second half of the decade in the Similis system (Lingua et Machina). Advertised as a 'second-generation translation memory', Similis boasts enhanced alignment, term extraction, and sub-segmental matching for the seven European Union languages supported by its linguistic analysis function.

Canada-based Terminotix has also stood out for its ability to mix linguistics with statistics, to the extent that its alignments yield output which for some purposes is deemed useful enough without manual verification. Here an interesting business reversal has occurred. As already noted, CAT system designers have progressively integrated third-party standalones (file converters, QA, alignment, term extraction), ultimately displacing their pioneers. But now that there is so much demand for SMT *bi-texts*, quick and accurate alignments have become more relevant than ever. In this climate, Terminotix has bucked the established trend by unbundling the alignment tool from its LogiTerm system and marketing it separately as Align Factory.

Apart from alignment, term extraction is another area where tracking advances in computational linguistics can pay dividends. Following the Xerox Terminology Suite model, SDL, Terminotix and MultiCorpora have also created systems with strong language specific term extraction components. Early in the past decade term extraction was considered a luxury,

marketed by only the leading brands at a premium price. By decade's end, all newcomers (Fluency, Fortis, Snowball, Wordbee, XTM) were including it within their standard offerings.

Now at least where the major European languages are concerned, the classic 'tabula rasa' CAT paradigm no longer stands, and although building algorithms for specific language pairs remains demanding and expensive, more CAT language specialization will assuredly follow.

Upgrades to the translator's editor

Microsoft Word-based TM editors (such as Trados Workbench and Wordfast) had one great blessing: translators could operate within a familiar environment (Word) whilst remaining oblivious to the underlying coding that made the file display. Early proprietary interfaces could handle other file types, but could become uselessly cluttered with in-line formatting tags (displayed as icons in Tag Editor, paint-brushed sections in SDLX, or a numeric code in curly brackets).

If for some reason the file had not been properly optimized at the source (e.g., text pasted in from a PDF, OCR output with uneven fonts etc.), the number of tags could explode and negate any productivity benefits entirely. If a tag were missing, an otherwise completed translation could not be exported to native format – a harrowing experience in a deadline-driven industry. Tags were seemingly the bane of a translator's existence. The visual presentation was a major point of differentiation between conventional CAT systems and localization tools. That situation has changed somewhat, with many proprietary editors edging closer to a seamless 'what-you-see-is-what-you-get' view.

Conventional CAT has not particularly facilitated the post-draft editing stage either. A decade ago, the best available option was probably in Déjà Vu, which could export source and target (plus metadata) to a table in Word for editing, then import it back for finalization (TM update, export to native format).

In word processing, Track Changes has been one effective way to present alterations in a document for another to approve. It is only at the time of writing that this feature is being developed for CAT systems, having emerged almost simultaneously in SDL Trados and MemoQ.

Where to from here?

A decade ago CAT systems were aimed at the professional translator working on technical text, and tended to be expensive and cumbersome. The potential user base is now much broader, and costs are falling. Several suites are even free, such as OmegaT, Virtaal, GlobalSight and other open source tools, but also the Google Translation Toolkit and Wordfast Anywhere. Many at least have a free satellite version, so that while the project creator needs a licence, the person performing the translation does not: Across, Lingotek memoQ, MemSource, Similis, Snowball, Text United, Wordbee and others.

One sticking point for potential purchasers was the often hefty up-front licence fee, and then feeling 'locked in' by one's investment. Web-based applications (Madcap Lingo, Snowball, Text United, Wordbee) have skirted this obstacle by adopting a subscription approach, charged monthly or on the volume of words translated. This allows users to both shop and move around.

Modern CAT systems now assist with most types of translation, and suit even the casual translator engaged in sporadic work. Some translation buyers might prefer to have projects done by bilingual users or employees, in the belief that subject matter expertise will offset a

possible lack of linguistic training. Another compensating factor is sheer numbers: if there are enough people engaged in a task, results can be constantly monitored and if necessary corrected or repaired. This is often referred to as crowdsourcing. For example, Facebook had its user base translate its site into various languages voluntarily. All CAT systems allow for translators to work in teams, but some – like Crowd.in, Lingotek or Translation WorkSpace – have been developed specifically with mass collaboration in mind.

A decade ago, CAT systems came with empty memory and terminology databases. Now, MultiTrans, SDL Trados Studio and memoQ can directly access massive databases for matches and concordancing; Logiterm can access Termium and other major term banks. In the past, CAT systems aimed at boosting productivity by reusing *exact* and *fuzzy* matches and applying terminology. Nowadays, they can also assist with non-match segments by populating with MT and post-editing or, if preferred, enhancing manual translation with predictive typing and sub-segmental matching from existing databases.

As for typing *per se*, history is being revisited with a modern twist. In the typewriter era, speed could be increased by having expert translators dictate to expert typists. With the help of speech recognition software, dictation has returned for major supported languages at least.

Translators have been using stand-alone speech recognition applications in translation editor environments over the last few years. However, running heavy programs concurrently (say Trados and Dragon NaturallySpeaking) can strain computer resources. Aliado.SAT (Speech Aided Translation) is the first system that is purpose-built to package TM (and MT) with speech recognition.

Translators who are also skilled interpreters might perhaps achieve more from 'sight translating' than from MT post-editing or assembling sub-segmental strings or predictive typing. The possibilities seem suggestive and attractive. Unfortunately, there are still no empirical studies to describe how basic variables (text type, translator skill profile) can be matched against different approaches (MT plus post-editing, sub-segmental matching, speech recognition, or combinations thereof) to achieve optimal results.

Given all this technological ferment, one might wonder how professional translation software will appear by the end of the present decade. Technology optimists seem to think that MT post-editing will be the answer in most situations, making the translator-focused systems of today redundant. Pessimists worry even now that continuous reuse of matches from internal memory to editor window, from memory to massive databases and STM engines, and then back to the editor, will make language itself fuzzier; they advocate avoidance of the technology altogether except for very narrow domains.

Considering recent advances, and how computing in general and CAT systems in particular have evolved, any prediction is risky. Change is hardly expected to slacken, so attempting to envision state-of-the-art in 2020 would be guesswork at best. What is virtually certain is that by then, the systems of today will look as outdated as DOS-based software looks now.

While it is tempting to peer into possible futures, it is also important not to lose track of the past. That is not easy when change is propelling us dizzyingly and distractingly forward. But if we wish to fully understand what CAT systems have achieved in their first twenty years, we need to comprehensively document their evolution before it recedes too far from view.

Further reading and relevant resources

With the Hutchings *Compendium* now discontinued, the TAUS Tracker web page may soon become the best information repository for products under active development. Just released, it contained only 27 entries at the time of writing (even major names such as Déjà Vu or

Lingotek have not made its list yet). ProZ's CAT Tool comparison – successor to its popular 'CAT Fight' feature that was shelved some years ago – also proposes to help freelance translators make informed decisions by compiling all relevant information con CAT systems in one place.

ProZ, the major professional networking site for translators, includes also 'CAT Tools Support' technical forums and group buy schemes. There are also user bases on Yahoo Groups, some of which (Déjà Vu, Wordfast, the old Trados) are still quite active; these CAT Tool Support forums allow for a good appraisal of how translators engage with these products.

The first initiative to use the web to systematically compare features of CAT systems was Jost Zetzsche's TranslatorsTraining.com. Zetzsche is also the author of *The Tool Kit* newsletter, now rebranded *The Tool Box*, which has been an important source of information and education on CAT systems (which he calls TEnTs, or 'translation environment tools'). Zetzsche has also authored and regularly updated the electronic book *A Translator's Tool Box for the 21st Century: A Computer Primer for Translators*, now in its tenth edition.

Of the several hard copy industry journals available in the nineties (*Language Industry Monitor*, *Language International*, *Multilingual Computing and Technology* and others), only *Multilingual* remains, and continues offering reviews of new products (and new versions of established ones) as well as general comments on the state of the technology. Reviews and comments can also be found in digital periodicals such as *Translation Journal*, *ClientSide News*, or *TCWorld*; they can be found also in newsletters published by translators' professional organizations (*The ATA Chronicle*, *ITI Bulletin*), and academic journals such as *Machine Translation* and *Journal of Specialised Translation*.

Articles taken from these and other sources may be searched from within the Machine Translation Archives, a repository of articles also compiled by Hutchings. Most items related to CAT systems will be found in the 'Methodologies, techniques, applications, uses' section under 'Aids and tools for translators', and also on 'Systems and project names'.

References

ALPAC (Automatic Language Processing Advisory Committee) (1966) *Language and Machines: Computers in Translation and Linguistics*, A Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC: National Research Council.

Brace, Colin (1992, March–April) 'Bonjour, Eurolang Optimiser', *Language Industry Monitor*. Available at: http://www.lim.nl/monitor/optimizer.html.

Garcia, Ignacio (2003) 'Standard Bearers: TM Brand Profiles at Lantra-L', *Translation Journal* 7(4).

Hutchins, W. John (1998) 'Twenty Years of Translating and the Computer', *Translating and the Computer 20*. London: The Association for Information Management.

Hutchins, W. John (1999–2010) Compendium of Translation Software: Directory of Commercial Machine Translation Systems and Computer-aided Translation Support Tools. Available at: http://www.hutchinsweb.me.uk/Compendium.htm.

Kay, Martin (1980/1997) 'The Proper Place of Men and Machines in Language Translation', *Machine Translation* 12(1–2): 3–23.

Kingscott, Geoffrey (1999, November) 'New Strategic Direction for Trados International', *Journal for Language and Documentation* 6(11). Available at: http://www.crux.be/English/IJLD/trados.pdf.

Lagoudaki, Elina (2006) *Translation Memories Survey*, Imperial College London. Available at: http://www3.imperial.ac.uk/portal/pls/portallive/docs/1/7294521.PDF.

Melby, Alan K. (1983) 'Computer Assisted Translation Systems: The Standard Design and a Multi-level Design', in *Proceedings of the ACL-NRL Conference on Applied Natural Language Processing*, Santa Monica, CA, USA, 174–177.

Muegge, Uwe (2012) 'The Silent Revolution: Cloud-based Translation Management Systems', *TC World* 7(7): 17–21.

Papineni, Kishore A., Salim Roukos, Todd Ward, and Zhu Wei-Jing (2002) 'BLEU: A Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002,* 7–12 July 2002, University of Pennsylvania, PA, 311–318.

Plitt, Mirko and François Masselot (2010) 'A Productivity Test of Statistical Machine Translation: Post-editing in a Typical Localisation Context', *The Prague Bulletin of Mathematical Linguistics* 93: 7–16.

Simard, Michel and Philippe Langlais (2001) 'Sub-sentential Exploitation of Translation Memories', in *Proceedings of the MT Summit VIII: Machine Translation in the Information Age*, Santiago de Compostela, Spain, 335–339.

Specia, Lucia (2011) 'Exploiting Objective Annotations for Measuring Translation Post-editing Effort', in *Proceedings of the 15th Conference of the European Association for Machine Translation* (*EAMT 2011*), Leuven, Belgium, 73–80.

TAUS (Translation Automation User Society) (2007) *Advanced Leveraging: A TAUS Report*. Available at http://www.translationautomation.com/technology-reviews/advanced-leveraging.html.

Trados (2002) Trados 5.5 Getting Started Guide, Dublin, Ireland: Trados.

van der Meer, Jaap (2011) *Lack of Interoperability Costs the Translation Industry a Fortune: A TAUS Report*. Available at: http://www.translationautomation.com/reports/lack-of-interoperability-costs-the-translation-industry-a-fortune.

Wallis, Julian (2006) 'Interactive Translation vs. Pre-translation in the Context of Translation Memory Systems: Investigating the Effects of Translation Method on Productivity, Quality and Translator Satisfaction', unpublished MA Thesis in Translation Studies, Ottawa, Canada: University of Ottawa.

Zetzsche, Jost (2004–) *The Tool Box Newsletter*, Winchester Bay, OR: International Writers' Group.

Zetzsche, Jost (2010) 'Get Those Things Out of There!' *The ATA Chronicle* 34–35, March.

Zetzsche, Jost (2012) *A Translator's Tool Box for the 21st Century: A Computer Primer for Translators* (version 10), Winchester Bay, OR: International Writers' Group.

4

COMPUTER-AIDED TRANSLATION

Translator training

Lynne Bowker

UNIVERSITY OF OTTAWA, CANADA

Over the past 75 years, computer technology has evolved considerably and has become increasingly prevalent in most areas of society. The translation profession is no exception, and this field has witnessed changes in the way that translation work is approached and carried out as a result of the increasing availability and integration of computer-based tools and resources. Indeed, translation technologies have become so firmly embedded in the translation profession that it now seems unthinkable for a translator to approach the task of translating without the use of some kind of computer tool.

If the task of translation itself has been affected by the use of computers, so too has the way in which translators are trained, which must now necessarily include training in the use of technologies. The need to integrate training in the use of computer-aided translation (CAT) tools into translator education programs has introduced a host of challenges, raising questions such as which types of tools are relevant for translators, what do translators need to learn about technologies, who should be responsible for teaching translators about computer aids, and when should technologies be introduced into the curriculum. The answers to such questions are not always clear cut, and they may be influenced by practical considerations that differ from one educational institution to the next. Nevertheless, over the past two decades, translation technologies have staked claim to a place in the translation curriculum, and as translator trainers continue to grapple with the challenges associated with technology training, some possible solutions and best practices are beginning to emerge.

Why do translators need to learn about translation technologies?

There is a longstanding debate about whether translation constitutes an art, a craft, or a science. Indeed some purists take the attitude that true translation is something best learned in the absence of technology. However, the reality of the twenty-first century is such that the vast majority of practising translators need to be able to leverage the possibilities offered by computer tools in order to remain competitive and to meet the evolving demands of the marketplace. Indeed in 2011, a survey of employers in the European translation industry was conducted in the context of the European-Union funded OPTIMALE project (Optimising Professional

Translator Training in a Multilingual Europe) (Toudic 2012). This survey revealed that the ability to use translation memory systems is considered essential or important by over three-quarters of the 538 employers who responded to this question (2012: 9). Similarly, 74 per cent of the 526 respondents viewed more general information technology skills, such as the ability to process files in and convert files to different formats, to be essential or important (2012: 9). Moreover, a quarter of the 526 respondents considered it essential or important for translators to be able to work with machine translation systems, which may include pre- or post-editing (2012: 10). Meanwhile, the ability to undertake software and website localization is also considered an essential or important skill by one-third of the 526 employers who responded to this question (2012: 10). Finally, 69 per cent of the 539 respondents indicated that the ability to extract and manage terminology was an essential or important skill for translators to possess (2012: 8).

Fuelled by a host of societal, political, economic, and technological trends, the demand for translation as a means of cutting through language barriers has grown exponentially in recent decades. Among these trends, we have witnessed:

- the shift to an information society with a knowledge-based economy;
- the creation and expansion of political and economic unions and agreements (e.g. the European Union, the North American Free Trade Agreement);
- the development of new and increasingly sophisticated products (e.g. digital cameras, smart phones, medical equipment), which often require extensive accompanying documentation;
- the globalization of commerce and the rise of e-commerce; and
- the growth of the World Wide Web coupled with the desire for localized content.

In the face of such trends, the volume of text to be translated worldwide has increased significantly, and the demand for translation far outstrips the supply of qualified translators. Indeed, as language professionals belonging to the baby boom generation have begun to retire, the shortage of qualified translators has been exacerbated.

On top of the increased volume of text to be translated and the relative shortage of qualified workers, deadlines for producing translations are getting ever shorter as organizations struggle to provide multiple language versions of the same document or product at the same time. Taken in combination, these trends are putting translators around the world under enormous pressure.

For both the translators who are faced with the prospect of processing higher volumes of text in seemingly ever shorter turnaround times – and for their employers – translation technologies present an attractive option for helping them to increase productivity and throughput. However, CAT tools cannot merely be assimilated into the translator's workflow without any effort.

Indeed, there have been several reports indicating that a considerable number of translators do not seem to be sufficiently well trained in the use of CAT tools. For example, according to Wheatley, who presents the results of a translation memory survey conducted as part of the European Union-funded eCoLoRe project, 34 per cent of respondents found it difficult to learn how to use translation technologies, 38 per cent felt that they would benefit from additional training, and 25 per cent felt unconfident with regard to their technological skills (Wheatley 2003: 4). Similarly, Lagoudaki, who conducted an international survey on the use of translation memory systems by language professionals, observed that 16 per cent of respondents who had already invested in such tools found it challenging to learn how to use

them properly, while an additional 4 per cent reported having a lack of time or energy to identify a suitable tool and learn how to use it (Lagoudaki 2006: 17).

It is possible that some of the translators who are currently working received their education at a time before technologies had risen to such prominence in the profession and when instruction in their use may not have been a standard component of the university curriculum. However, as the popularity of CAT tools has increased, most translator education institutes have taken steps to incorporate some form of technology instruction into their programs. Nevertheless, there seems to be some evidence that calls into question the effectiveness of the technology education that is currently being offered as part of translation programs. For example, authors such as Jaatinen and Immonen (2004: 37) and Samson (2005: 104) have reported that employers and clients frequently complain that translators – even recent graduates – are not necessarily proficient users of CAT tools. Meanwhile, as several researchers have noted, the use of technologies is contributing to changes in the nature of translation work (Melby 2006; Garcia 2010a/b; Pym 2011a). This in turn means that integrating CAT tools into a translator education program can require a fundamental shift in how we view – and therefore how we teach – translation. Clearly, then, there is room for contemplating and adjusting the way that the use of translation technologies is taught as part of a translator education program, and a number of new initiatives are indeed being developed and implemented by universities across the globe.

For instance, a major European Union translator education initiative – the European Master's in Translation (EMT) – was launched in 2009 with a view to improving the quality of translator training and to ensuring that the next generation of professional translators will be able to meet the needs of the twenty-first century marketplace. Currently held by 54 different university programs across Europe, the EMT is a quality label for translator training courses at the master's level which can be given to higher education programs that meet commonly accepted quality standards for translator training. The EMT promotes quality translator training and helps translators to keep up with the requirements of our knowledge society. As reported by Gambier (2009), the EMT expert group identified a number of technology-related competences that are considered important for professional translators and for which adequate training must be provided as part of translator training programs. These competences include being able to effectively use search engines, corpus analysis tools and term extractors for information mining; knowing how to produce and prepare a translation in different file formats and for different technical media; knowing how to use a range of CAT tools; understanding the possibilities and limits of machine translation; and being able to learn and adapt to new and emerging tools (Gambier 2009: 6–7). An EMT spin-off project known as QUALETRA (Quality in Legal Translation) seeks to do the same for eight European university programs that specifically train legal translators.

Meanwhile, in Canada, the Collection of Electronic Resources in Translation Technologies (CERTT) project (Bowker and Marshman 2010; Marshman and Bowker 2012) and the Translation Ecosystem project (Mihalache 2012) are examples of how translator trainers are addressing technology-related education needs. Both CERTT and the Translation Ecosystem are available through LinguisTech, a translation technology learning platform developed by the Language Technologies Research Centre. CERTT and the Translation Ecosystem aim to provide translation students across Canada with translation technology knowledge and know-how that will enable them not only to master the tools, but also to develop the strategic and reflexive skills needed for adopting best practices and for making informed decisions with regard to tool selection and use.

Which types of tools are relevant for translators?

Before considering some possible approaches to technology-related training, and the accompanying challenges that they present, let us begin with a brief survey of some of the different types of technologies that may appear in a translation curriculum. Note that the goal here is not to describe specific features of these tools, nor to explain how they work – such descriptions can be found elsewhere in this volume – but rather to provide a general idea of different categories of tools to show the range of technologies available to translators and to provide a very general indication of how these might fit into the translation curriculum.

The range of general office software, in particular word processors or text editors, but also spreadsheets and desktop publishing programs, are among those tools used most regularly by translators in their daily work. In addition, translators regularly find themselves needing to use general tools such as file conversion programs or file compression software. In the past, when the presence of computers in our everyday lives was far less ubiquitous, and before the plethora of specialized and sophisticated CAT tools had arrived on the market, training in the use of these more basic tools was sometimes integrated into the translator training curriculum. Nowadays, as computers have become increasingly prevalent, translation students arriving in the translation classrooms in the early twenty-first century are undoubtedly far more computer-savvy than were their counterparts in preceding decades. Nevertheless, while these students might be comfortable with the basic functions of such programs, there may still be considerable room for them to develop into 'power users' who can optimize the functionality of a word processor or other type of office software. However, since the translation curriculum must make room for intensive learning with regard to the more sophisticated CAT tools now in existence, the training provided for the more basic tools must come in other forms. For example, translator education institutes may keep on hand a series of tutorials and exercises covering the more advanced features of office tools and encourage students to do these independently so that they might be better prepared for the program. Similarly, students may organize or participate in peer-to-peer training sessions, where they share tips and tricks that they have acquired for making better use of general software. Some educational institutions may offer computer training workshops or seminars organized through a central student services or computer services unit. Finally, many professional associations offer workshops or training sessions on a range of tools – from the general to the more specialized – and students may be directed towards these offerings. This will not only provide an opportunity to learn about a tool, but it also helps students to develop a professional network and instills the important notion that lifelong learning – particularly in relation to technology – is essential for continued success in the translation profession.

Electronic resources, such as term banks or dictionaries, as well as resources such as pre-constructed online corpora and associated processing tools such as concordancers, may find a place in a documentation course on a translator education program. Again, while the use of pre-existing resources might seem relatively intuitive, an early and gradual introduction to technologies allows students to build a solid foundation. As Dillon and Fraser (2006: 76) and Lagoudaki (2006: 16) note, translators with strong basic computer skills seem better positioned to graduate to using more specialized tools. In addition to teaching how to search pre-constructed corpora, documentation courses may also provide translator trainees with an opportunity to learn how to design and compile their own 'do-it-yourself' corpora. This requires learning how to critically evaluate texts to decide whether they can usefully be included in a corpus. It also entails learning how to interpret the output of a concordance, taking into account the limitations of the corpus.

The use of terminology management systems, which allow students to build and query their own termbases, will undoubtedly be incorporated into a terminology course on a translator training program. These tools can be used in standalone mode, and students learn how to design term record templates, record terminological information, and search for solutions within the termbase. Term extractors, which seek to automatically identify potential terms in an electronic corpus, may also be introduced in a terminology course.

Translation memory systems, which are used to interrogate an associated database of previously translated texts along with their corresponding source texts, are typically introduced in a course dedicated to translation technologies. These tools are normally at the heart of a larger tool suite, sometimes referred to as a Translation Environment Tool (TEnT) or translator's workstation. As part of this larger suite, they may interact with other tools, such as word processors, terminology management systems, concordancers, and machine translation systems; however, in many cases, students learn about these different modules independently. From a didactic viewpoint, this makes sense, as students can more manageably digest information about the underlying concepts and the operation of the individual components, but as we will see in an upcoming section, this approach does not facilitate an understanding of the way the different tools interact, nor of the ways in which users must interact with the tools in order to optimize their performance.

Once considered to be a tool for replacing translators, machine translation (MT) systems are now more widely accepted as a sort of CAT tool, which requires some interaction with a professional translator, such as in the form of pre- or post-editing. In the past, MT systems were frequently left off the curriculum of translator training programs altogether, or given only cursory attention, on the grounds that such tools were not used by practising professionals. Increasingly, however, translators are being asked to work with MT systems in some form. For example, an MT system may be integrated with a TM system in order to generate possible solutions when none are found in the TM database. It is therefore becoming increasingly relevant to include at least an introduction to MT as part of the regular translation curriculum. Additional specialized courses may be added to provide training in pre-editing or writing for translation, as well as in post-editing, where translators learn techniques for revising the output of an MT system.

Localization tools are tools that allow translators to take the content of a website or a software package and adapt it for use by speakers of another language. Localization tools include functions that allow translators to extract text strings from the software code, and to reinsert them back into the code once they have been translated. In addition, localization tools make it easy to adapt other elements of software or websites, such as shortcut keys, colours, or currency or date formats, so that they are more appropriate to users from another culture. Localization tools may be introduced as part of a core course in translation technologies, but they are more likely to be examined in more detail as part of a specialized elective course dedicated to localization, if such a specialization is offered as part of a given translator training program.

Voice recognition tools, which allow users to dictate a text directly into a word processor or text editor, have not yet become commonplace on the translator training curriculum. This is in part because, until recently, the technology did not produce accurate enough results to make it worthwhile for translators to adopt a dictation approach to working. This is changing however, and it is likely that voice recognition tools will feature more prominently in the translation curriculum in coming years.

Translation workflow tools are another type of tool that has not yet taken a firm hold in the translation curriculum. These tools are designed to help manage translation projects where

there are multiple team members who must share resources and work with the same texts. They also have features that facilitate interactions between clients and translators, such as a means of placing an order, or sending an estimate or an invoice. While these tools do not typically get addressed in core translation technology courses, they may be introduced in courses that deal with professional issues or courses that focus on project management, if such electives are part of the translation program in question.

A similarly specialized set of tools are those used for dubbing, subtitling or audiovisual translation. Normally, these tools are not introduced as part of the regular curriculum, but may be included in a course on audiovisual translation if such a course is part of the program.

Meanwhile, new tools and technologies continue to emerge, such as wiki-based collaborative authoring or collaborative terminology management platforms, designed with a view to facilitating work by translators who collaborate online. Undoubtedly, these too will soon find a place in the translation curriculum.

While this list is not exhaustive, it has nonetheless provided a general idea of the broad range of technologies that might be introduced in a translator training program. It has also suggested, in a general way, where these tools might currently be found in the curriculum of many programs.

Which specific tools should be included in the curriculum?

The previous section introduced an extensive range of types of CAT tools that trainee translators are likely to encounter as part of their studies. As the number of tools available in each category continues to rise, translator trainers are faced with the dilemma of having to select specific products to represent these different categories of tools in order to give students an opportunity to get some practical hands-on experience. A host of factors may influence the decision about which tools to select, and as part of that decision process, translation technology trainers such as O'Brien and Kenny (2001: 22) have pointed out that one of the main challenges is the so-called 'skills versus knowledge' debate. In other words, should a university course attempt to train translators how to use the leading TM tools on the market (e.g. to increase their chances of employment)? Or should it aim to impart knowledge of the technology in a more generic way in order to equip students with the ability to evaluate and to learn to use such tools themselves? Most trainers seem to lean towards the latter approach, recognizing that tools that are popular today may well be out of fashion tomorrow.

The good news in this scenario is that, if the goal is to teach students the underlying principles which are common across a particular category of tool, then the decision about which particular tool is chosen becomes less important. However, in order to allow trainees to develop critical evaluation skills, an educational institution should not focus on a single tool for training purposes. Moreover, there is a good argument to be made that having a minimum of two to three tools available for observation makes it much easier to distinguish the basic features of the tool type in general from the quirks or options of an individual product.

Undoubtedly, a fundamental understanding of the underlying concepts and principles, as can be obtained from studying a small number of tools, is essential. However, ideally, any training program should be flexible enough to adapt to evolving commercial needs. In the case of translation technology, there would seem to be at least two further arguments to be made in favour of exposing the students to an even wider selection of tools during the course of their training. First, as noted above, there are a plethora of tools available on the market today, and even if a translator is in a position where he or she is able to work with only a single tool, it will first be necessary to select this tool. Deciding which tool can best meet the needs at hand

is a task that can be facilitated through a comparative evaluation. Therefore, if translators are going to find themselves needing to conduct such comparative evaluations, they will be better equipped to do so if they have previously been given the opportunity to gain such experience by evaluating and comparing a selection of tools as part of their training.

In addition, the reality of today's market would seem to be that translators typically need to be comfortable using multiple tools. A 2006 survey of 874 translation memory users revealed that most use multiple tools, with three to four being the average (Lagoudaki 2006: 23). If students will be faced with the need to use more than one tool of a particular type in the workplace, then they will surely benefit from having the chance to learn and experiment with several as part of their studies. In addition, the more exposure they have to a variety of tools, the less likely they are to be naïve users once they enter the workforce (Dillon and Fraser 2006: 75).

Another consideration is the complexity of the tools selected. If more than one tool of a certain type is to be learned, it may make sense to begin with one that integrates fewer 'extra' features. As several trainers logically point out, translators with strong basic computer skills can more easily graduate to using complex software (Biau Gil 2006: 93; Dillon and Fraser 2006: 76; Lagoudaki 2006: 16).

In certain cases, this simplicity is in fact an advantage; many newer users may be less intimidated by a 'core' tool package containing only the main functions that they are likely to use than by a product that includes numerous additional programs whose uses may be more or less clear (e.g., functions intended to assist in managing complex workflow, dealing with heavily coded documents, and other similar tasks). Moreover, the volume of accompanying documentation for these programs is also likely to be more manageable for a new user, such as a student, when the product itself is more targeted to specific, translation-centred functions.

Nevertheless, once users become more comfortable with the use of such tools, or once they enter the workforce and find themselves working in specific contexts or for clients that require more advanced functions, they may eventually regret the absence of some of these more 'peripheral' tools, or the necessity of adding another tool to their repertoires in order to have access to them. However, as noted above, evidence from the literature (e.g. Lommel and Ray 2004; Lagoudaki 2006) would seem to suggest that it is rarely enough for a translator to be comfortable using only one tool. The general consensus seems to be that every tool has its strengths and weaknesses, and the choice of which one to use depends on the job at hand. Still, the fact remains that if multiple tools must be learned, there is a certain logic to learning the most straightforward tool first and working up to a more complex system.

If the cost of purchasing multiple tools in a single category is prohibitive for a translator training institute, one option may be to try to incorporate the use of demo versions of these tools. Most commercial products do create and distribute demo versions with a view to allowing potential clients to have an opportunity to test and evaluate the tool before committing to it. However, these demo versions are often restricted in some way (e.g. time-limited versions, limited functionality), which may hinder their usefulness as a teaching tool. Depending on the way in which the functionality is limited, it may be more or less feasible for a demo version to be usefully incorporated into a training program.

It may be more attractive for an educational institution to turn instead to freely available open source products, and to incorporate these into training programs. In this way, students can be introduced to a wider range of products and can have the opportunity to learn multiple tools and to comparatively evaluate them. Open source tools have the added advantage that they can be installed by students on their own computers and used outside the requirements of specific courses. This may encourage students to begin using tools more extensively and to allow them to start building up resources – such as translation memory and terminology

databases – early on, before they even get started on their career. As noted by Fulford (2001: 228), one obstacle that hinders established translators from adopting TM technology is the difficulty of transferring legacy translations (i.e. those created outside a TM environment) into a TM database. Encouraging students to get into the habit of using a TM early on will hopefully mitigate this problem. Moreover, even if translators end up switching from using a free TM system during their student days to using a commercial product after graduation or later in their career, or if they end up using multiple tools, it is becoming increasingly easy to transfer TM databases and termbases between different systems – including between free systems and commercial products – without a great loss of time or investment of effort.

What do translators need to learn about translation technologies?

Once translator trainers have acquired a selection of tools for translation students to work with, an important question becomes *what* do students need to learn about these tools. CAT tools can be extremely sophisticated, each incorporating a variety of features and functions that work in a slightly different manner, or are referred to by a different proprietary term, or are accessed through a different style of interface. It is clear, therefore, that any training must involve learning the particular steps required to operate a given tool. In other words, trainers certainly need to provide students with step-by-step instructions for using this tool. However, as noted above, while it is clearly important for translator training institutes to turn out graduates whose overall skill set is in line with the needs of the market, this market is somewhat volatile. Therefore, technology training cannot be set up solely to address the latest trends but must take a more balanced approach that includes providing students with transferable skills, such as the ability to engage in critical analysis and problem solving.

In addition to providing a 'how to' manual, instructors must also seek to provide a framework that goes beyond merely describing a tool's features or explaining *how* it functions. In other words, to prepare translators to become effective users of translation technologies, trainers need to provide opportunities for students to learn not only *how* but also *when* and *why* to use a given tool. For instance, for each category of tool that is being learned, students should be given a series of tasks and questions for reflection that will encourage them, as tool users, to reflect on *why* it might be helpful to adopt a given tool as well as to consider what a tool can and cannot do, and the positive and negative effects that tool use may have on the translation process and product in different situations.

It is clear that translation technology cannot be taught or understood in a vacuum, so translator training programs must include practical experience with tools in order to support theoretical understanding. This practical experience may in turn stand students in good stead as they reach the job market. However, in many cases, the pertinence of hands-on training for future work will depend on the ultimate employment of translation graduates.

Surveys of technology use have highlighted some variations in the use of tools – in particular, translation memories (TMs) – in different user groups. Surveys conducted by the *Association of Translators and Interpreters of Ontario (ATIO)* of independent (ATIO 2005) and salaried (ATIO 2007) translators showed a substantial difference in responses, with 44 per cent of salaried translators reporting using TM tools, but only 27 per cent of independent translators indicating TM use. In her survey, Lagoudaki (2006: 19) observes that the vast majority of the freelance respondents who used TM tools did so by choice, with much smaller proportions required to by the translation agencies they worked for and even fewer by their clients. In contrast, considerably more of the company employees were required to use these tools by the translation agencies they worked for. Citing Lommel and Ray's survey (2004), Lagoudaki (2006: 15) also notes that companies are

more likely to be open to TM use than individual users, given their potential for cost-savings and productivity gains. It is nevertheless difficult to generalize about TM use by freelancers and how it differs from use by companies: reported levels of TM use for freelancers responding to surveys range from 27 per cent (ATIO 2005: n.p.) or 28 per cent (Fulford and Granell-Zafra 2005: n.p.) to 81 per cent (Lagoudaki 2006: 15), depending in part on the context of the study.

Moreover, those who work as freelancers will likely be best served by experience with different kinds of tools and functions – and in fact, may ultimately need almost a different technological skill set – from those who go on to work for large corporations, or in the public sector. Clearly, the scope of projects undertaken and the complexity of workflow (including among other factors the size and structure of a documentation/translation team and the volume of translation carried out) will play a large role in the selection of a TM system. Thus, the student who goes on to work in a freelance environment may benefit most from experience with central TM functions, while those who ultimately work with translation agencies or in larger documentation and translation environments may need to become more familiar with project and TM management functions that freelancers are more rarely called upon to use.

In addition, while those working in larger organizations may have easily available technical support for many applications including TM systems and the management of TMs themselves, freelancers generally need to manage their own technological environments independently. For these users, the challenges of installing, updating, managing and using more complex programs may outweigh the advantages of the additional features they offer.

Thus in a single translator training program it is extremely difficult to predict and meet the specific needs of all future translators. This is particularly true as during their training, many students may not yet have a clear idea of the type of job that they will eventually have. Moreover, many will likely work in multiple contexts (either consecutively or simultaneously).

Where in the curriculum should translators learn about translation technologies?

Teaching and learning translation technologies are not straightforward tasks. As alluded to above, one factor that may hinder students from developing a well-rounded understanding of technologies may be the fact that, in many cases, the opportunity to learn about translation technologies may be restricted to a 'core' course dedicated solely to this subject. Such core courses are certainly valuable, and in fact are essential to understanding the underlying principles of how tools work and how they may be useful. Moreover, core courses provide opportunities for comparative evaluation of different tools as well as in-depth exploration of a fuller range of the features offered by the tools. In short, these courses provide an occasion for students to think about technology. However, in these courses, tools are often examined in isolation rather than as part of an integrated translation environment or interactive tool suite. For example, the features of a terminology management system may be explored in some depth by working with this tool in stand-alone mode; however, the practices needed to optimize the tool for use in conjunction with a translation memory system might not be adequately addressed. Because these 'core' courses do not always provide students with sufficient opportunity to use the tools in the context of an actual translation project, it means they may not be thinking specifically about how tool performance is affected by – or can affect – the translation process and product, and how they themselves can best interact with tools to achieve optimal results.

In many contemporary translator training programs, a main drawback to the way that translation technologies are taught can thus be summarized as a lack of integration on two

levels. First, as noted above, tools are primarily viewed in isolation rather than as part of an integrated translation environment. This approach, which introduces students to the basic functions of the tools, is necessary as a first stage of teaching and learning where knowledge and instructional content are broken down for easy digestion. However, it does not allow students to appreciate fully how the performance and use of these tools fits into the bigger picture of translation practice. Therefore, the next stage of learning requires evaluating the task in its natural wider context, which many include as part of a larger interactive tool suite.

Second, on many translator training programs, the tools are only seen and used in 'core' courses – i.e. courses with a specific focus on technology – rather than being integrated across a range of applied courses in the translator training program. The resulting gap between theory and practice does not provide students with an accurate picture of how they are likely to work – and in fact may be expected or required to work – in many professional contexts. To truly learn how tools fit into the translation process, technology-related tasks must be contextualized rather than severed from realistic experience.

Another challenge that may arise in 'core' courses is that students work with various language pairs and directions. Whereas practical translation classes tend to focus on a specific language pair, core technology courses often bring together a mixed language group, which often requires technology trainers to provide source texts or research questions in a lingua franca, and some students must work in the 'wrong' direction, which is not an authentic experience for them. Moreover, the trainer cannot usually provide in-depth assessments or feedback since he or she is not usually an expert in all the language directions used by the students.

This lack of integration is not usually a result of trainers' unwillingness or failure to recognize the importance of technologies. Indeed, a number of researchers have suggested that integrating technologies more fully across the translator training curriculum could benefit students (and their eventual employers) (e.g. Clark *et al.* 2002; Samson 2005; Jaatinen and Jääskeläinen 2006; Kenny 2007; Bowker and Marshman 2010). However, many challenges are involved in achieving this goal. The question then remains: how can translator trainees' needs be met effectively in a university context?

Situated learning promotes the use of tools as aids in practical translation courses as well as in core technology courses and offers a chance for reflection on the role and impact of translation technologies in the bigger picture. Active and situated learning strategies are increasingly being adopted in numerous facets of translator education (e.g., Biau Gil 2006; Gouadec 2003; Jaatinen and Immonen 2004; Kenny 2007; Kiraly 2000; Shuttleworth 2002). Using this approach, learning takes place in an environment that simulates as much as possible an authentic workplace setting. In the case of translation technology education, this means embedding tool use in practical translation courses, rather than contributing to the siloization of tools by restricting their use to the 'core' technology courses.

Under such realistic conditions, students work and build knowledge and skills in a collaborative fashion, thus taking on the role of active learners, rather than passive receivers of potentially abstract and decontextualized knowledge, which may appear divorced from real-world requirements or practices. The challenge for translator educators is to establish a framework that will support the embedding of technologies into – and especially across – translator education programs.

Another effective way to introduce situated learning is through internships or work placements. An increasing number of translator education programs are incorporating such opportunities into their programs. One example is the AGORA project, which is a European Master's in Translation (EMT) spin-off project to assess the feasibility of cross-border placements and internships for EMT translation students.

When should translators learn about translation technologies?

As discussed by Kelly (2005: 113), decisions about sequencing the different elements of a translator training program (e.g. theory, practice, language skills) have long been debated. Moreover, the simple time pressures of trying to prepare students to translate professionally with a limited number of course hours mean that choices of content at each stage must be carefully weighed to maximize results. With regard to technology, there is no consensus on when tools should be introduced. On the one hand, students will benefit from the opportunity to practice realistic work habits by using such tools early and often, but on the other hand, they need a certain amount of translation experience to avoid becoming naive users of technology.

Dillon and Fraser (2006: 69), for example, suggest that inexperienced translators do not have the breadth or depth of knowledge needed to allow them to properly evaluate the advantages or disadvantages of using a given tool. Meanwhile, Bowker (2005: 19) observes that novice translators sometimes exhibit 'blind faith' in technologies because they lack the confidence or experience required to critically evaluate the tools' output.

Of course, it is worth noting that translators use many different kinds of tools, ranging from the relatively straightforward word processors and term banks to the more sophisticated corpus processing tools, translation memories and beyond. Common sense suggests that it should be possible to introduce more general tools earlier in the translator training process, while reserving some of the more complex tools for later integration. With an early and gradual introduction to technologies, students will build a solid foundation. As Dillon and Fraser (2006: 76) and Lagoudaki (2006: 16) note, translators with strong basic computer skills seem better positioned to graduate to using more specialized tools.

Although there may be no straightforward answer to *when* tools should be introduced, simply not introducing them is not a reasonable solution. Rather, observations such as those above seem to reinforce the notion that more and better training in technology use for translation is needed. The translation classroom offers an unparalleled venue for students to observe when using technology has helped them to find good solutions, when it has not, and why. Such discussions can help to develop students' judgment not only about technologies, but also about translation strategies in general.

Additional challenges to integrating technologies into translator training

In addition to the challenges discussed above, numerous other obstacles can hinder the successful integration of technologies into a translation program. Some are practical in nature, such as lack of access to appropriate hardware and software. Hopefully, such issues will become decreasingly problematic as prices for computer-related products continue to drop and partnership agreements between universities and tools vendors become more commonplace, and as open source tools become fully developed. However, a number of thorny issues remain, and are less straightforward to address, though some tentative solutions are proposed below.

Managing expectations by encouraging critical reflection

One significant challenge presented by translation technologies is the management of expectations with regard to what these tools can and cannot do. In their enthusiasm to reap the aggressively marketed potential benefits associated with computer-aided translation tool use, some less experienced translators and technology users may have expectations of tools that go far beyond the capacities of today's technologies, and certainly beyond their intended

uses. As observed by Dillon and Fraser (2006: 75), inexperienced translators seem to have 'an uncharacteristically positive view of TM [translation memory]', coupled with 'a higher level of ignorance of the limitations of TM'. This type of attitude may result either in disillusionment with tools when these expectations are not met or – even more seriously – in uncritical reliance on the output of these tools. While gains in income, productivity, and quality are often reported (e.g., Gauthier 2012; Vallianatou 2005), numerous authors (e.g., Bédard 2000; Topping 2000; Bowker 2005, 2006; Pym 2011b) have identified problems that can arise from uncritical use of tools, such as inappropriately recycling the contents of a translation memory or applying one tool in a context that more properly calls for the use of another. In the context of translation technologies, those translators who have developed keen critical reflection skills will be the ones best placed to determine where the benefits and pitfalls lie in relation to tool use.

Experienced translator educators, who have witnessed some unfortunate results arising from uncritical attitudes, may be understandably reluctant to use technologies in class or to introduce them into translator education before students have acquired enough experience to be critical of tools' output. Without a coherent structure to guide the implementation of technologies in a translator education program, and in the absence of reflection by both educators and students about the contexts and ways in which tools may (and, equally importantly, should not) be used, it is difficult to bring these two extremes together in a way that allows tools to be implemented to their – and more importantly their users' – best advantage.

To mitigate this situation, translator trainers might consider presenting the practical instructions for the use of tools in a framework that accents both background knowledge and critical thinking. For example, rather than simply adopting the tutorials provided by the tool developers, it could be useful for translator trainers to augment these by including as an introduction to each tutorial or exercise the essential information for understanding what the tool is designed to do, how it can be useful to a translator, and how it compares to some other tools or approaches. This type of information aims to help users determine whether a tool is of particular interest and to lead them to consider whether it may meet their personal needs. The tutorials and exercises could be accompanied by a series of questions for reflection on key points about the tools (e.g., user reactions to a tool and its use, comparisons to other similar tools he or she may have used, advantages and disadvantages compared to a manual approach and/or to using other tools, situations in which the tool might be useful). This encourages users to consider the tool at a relatively high level based on both background knowledge and practical experience, rather than simply on the basis of whether they were able to accomplish a specific required task with it. This approach goes beyond that used by many others in the field, as few of the tutorials and resources that are provided by developers encourage evaluation and comparison of tools according to users' specific needs. However, these are among the most important aspects of instruction in translation technologies, and they will go a long way toward helping educators and users to manage expectations surrounding technology use in the field of translation.

Expanding and centralizing resources for increased accessibility

Another problem facing trainers is the lack of easy access to authentic complementary resources (e.g. relevant exercises, sample termbases, corpora, bitexts, sample source texts suitable for technological processing) required to introduce these tools to students and to work with them in a realistic way (eCoLoTrain 2006: 20). Tools such as terminology management systems and translation memory systems are 'empty' when first installed, and users must create relevant

term records to build up the termbases and TM databases. In a similar vein, term extractors are designed to operate on corpora, but to get realistic and usable results, these corpora must be well-designed and reasonably large. Designing and compiling these resources and accompanying exercises can be time-consuming and labour-intensive. For an educator who already has a very full schedule, this additional workload can act as a deterrent, preventing him or her from effectively introducing computerized translation tools to students and to working with them in the noncore courses (Shih 2006: 17; eCoLoTrain 2006: 20; Marshman and Bowker 2012: 79).

A related challenge that may hinder the successful integration of tools into a wider range of translation courses is a lack of centralization and management of technology-related resources. All too often, tutorials, exercises, and resources developed for use with particular tools are dispersed among the various educators who create and use them, and thus are not known to or available for use by others. Storing and organizing various types and versions of documents relating to different tools, as well as coordinating their use in different courses, also pose challenges for instructors. Moreover, when educators leave an institution, the resources and expertise they have developed are not always passed on to others and may be lost. As a result of these obstacles, work may be duplicated unnecessarily and overall coverage of tools and their functions may be uneven, with the risk that some elements may be covered repeatedly in different courses while others are neglected altogether, leaving students both frustrated and ill prepared with regard to tool use.

By pooling resources such as corpora, termbases or TM databases, and storing them in a central and easily accessible location (e.g. using a course management system or shared directory), trainers could have access to a wider range of materials to provide a more authentic and situated learning experience for their students.

Training the trainers

It is widely recognized (e.g. Arrouart 2003: 478; Bowker 2003: 74; Jaatinen and Jääskeläinen 2006: 84; Kenny 2007: 203) that a key question when contemplating a more integrated approach to technology training is whether the trainers who teach other subjects (e.g. terminology, specialized translation) are comfortable using the relevant tools. As it has been less than twenty years since technology has really begun to permeate a wide range of translation activities, numerous instructors likely received their own training before technologies were incorporated into the translation curriculum in any significant way. While they are almost inevitably aware of the increasingly important role of technology in the field, they will not necessarily be familiar with the finer details of the tools, and/or may not have considered how such tools could be used in teaching.

It is tempting to dismiss this as a generational issue that will be resolved as senior trainers retire and are replaced by colleagues who are familiar with technologies. However, as Samson points out (2005: 103), the problem is a long-term one because tools are evolving rapidly, and instructors who specialize in other areas of translation may not have the time or inclination to keep up with the latest technologies.

Nevertheless, in spite of the fact that they may not currently integrate translation tools into their teaching, many trainers are acutely aware of the benefits that such integration could bring. The results of the eCoLoTrain (2006: 21–22) survey, which set out to uncover the perceptions and requirements of translator trainers with regard to translation technologies, show that the majority of the 86 respondents feel that it is extremely or very important to teach both general (76.75 per cent) and specialized (70.93 per cent) technology skills as part of a translator training program.

However, while the eCoLoTrain (2006) survey participants support the inclusion of technology in translator training programs, most feel that they themselves would need further training to become highly proficient users (particularly of specialized software), and especially to be able to teach others. Barriers to translation technology uptake cited (eCoLoTrain 2006: 20) include 'my own computer skills are not good enough to teach others with the computer' (18.18 per cent) and 'do not know about software tools' (10.91 per cent). Similarly, a survey conducted by Kelly (2008: 117–18) asking Spanish translator educators to evaluate their own knowledge in important areas of the discipline identified technologies as one of the key areas in which training was required. However, encouragingly, among those trainers who are not yet familiar with relevant tools, there is an obvious interest in learning. For instance, in answer to the question 'Do you know how to use terminology management software?', 48.8 per cent of respondents said 'yes', 9.3 per cent responded 'no' and the remaining 41.9 per cent replied 'no, but I would like to learn' (eCoLoTrain 2006: 15). It would therefore seem that a resource for helping such educators to learn more about translation technologies would be welcome and useful. One such effort currently underway is the Collection of Electronic Resources in Translation Technologies (CERTT) project, based at the University of Ottawa in Canada. The general idea is to provide a point for 'one-stop shopping' – a single centralized and relatively comprehensive resource that both instructors and students can access to find everything they need to begin using, or to facilitate or increase their use of, translation technologies as part of their academic experience (Bowker and Marshman 2010; Marshman and Bowker 2012).

Addressing the needs of a wide range of student learners

Once translator trainers have developed necessary resources and knowledge and feel ready to integrate these more fully into their teaching, they then face another challenge. As confirmed by Clark *et al.* (2002: 65–66) and Arrouart (2003: 478–479), among others, one of the most constant and greatest challenges in teaching technologies is that students arrive in translation classes with varying degrees of technological competence. Course groups may include students with an advanced grasp of and significant experience with tools alongside those who have little experience and who may even be intimidated by information technology. Thus trainers must walk a metaphorical tightrope trying to ensure that the more technologically savvy students are not bored, while the more technologically challenged ones are not frustrated or overwhelmed. Moreover, the latter may experience considerable difficulties in courses that involve technologies, perhaps particularly when these are not the main focus of a course, but rather tools intended to facilitate learning, discussion and practice. Clearly, when students find using tools the greatest challenge in these courses, the effect is quite the opposite. In such a case, it is important for the trainer to ensure that tools-related difficulties do not become the focus of the learning situation, overshadowing the larger objective of learning to translate.

On a related note, it is clear that individuals learn differently and thus have different training preferences (Kelly 2005: 47). When learning about translation technologies, some prefer a classroom setting, others want to do exercises independently, and still others favour using documentation that explains tools and their uses (Wheatley 2003: 5). Initial comfort and confidence levels may also influence the effectiveness of different learning strategies. Biau Gil (2006: 93–95) notes that users who initially have better general computer skills seem better able to learn to use new tools independently. Accommodating these varied learning styles and needs requires a flexible approach.

Concluding remarks

Early users of CAT tools primarily had to come to grips with understanding and applying them as quickly as possible, often without any formal instruction. Meanwhile, translator trainers had to contend with developing expertise, designing methodologies, and preparing curricula and resources required to teach these tools. Even today, there are numerous challenges to be faced by educators as they seek to comprehensively organize, categorize and contextualize the bewildering and evolving array of tools and technologies available to translation professionals. The diverse approaches to overcoming these challenges reflect the diverse backgrounds and circumstances of the trainers, as well as the state of flux that characterizes technology development in general.

The integration of new tools into courses will always require preparation and effort on the part of trainers. In addition, technical difficulties can certainly not be completely avoided. However, it is much better for students to begin to come to grips with new technologies and their associated challenges during their studies, rather than waiting until they enter the high-volume, high-stress environment of today's professional workplace. Students who begin their careers with established and tested translation practices that work for them are better prepared to continue these good practices in their professional life. Moreover, the literature highlights the fact that students who have already developed basic skills are more likely to be able to adapt easily to new tools and situations (e.g. to a new CAT tool used by an employer or client). By allowing students to become more familiar and more comfortable with CAT tools gradually and by giving them access to a range of tools throughout their program of studies, educators should be able to help students to significantly improve comfort levels with technologies and knowledge of the field, and also develop better and more realistic translation practices throughout their training.

Finally, by encouraging a fuller integration of technologies in the academic life of students and trainers, we hope the translator training program will better reflect current practice in the translation field today, including the necessary integration of effective CAT tool use into the translator's day-to-day work. As pointed out by Kiraly (2000: 13), there is a difference between helping students to develop 'translation competence', which gives them the skills to produce an acceptable target text in one language on the basis of a text written in another, and aiding them in the acquisition of 'translator competence', which also involves assisting them with the development of a host of other skills, including proficiency in new technologies. In the words of Kiraly (2000: 13–14):

> Translator competence does not primarily refer to knowing *the* correct translation for words, sentences or even texts. It does entail being able to use tools and information to create communicatively successful texts that are accepted as good translations within the community concerned. … With the changes in the translation profession in mind, it is time to reconsider the viability of conventional approaches for educating translators, which date back almost half a century, when the translation profession was something altogether different from what it is today.

References

Arrouart, Catherine (2003) 'Les mémoires de traduction et la formation universitaire: quelques pistes de réflexion', *Meta* 48(3): 476–479.

Association of Translators and Interpreters of Ontario (ATIO) (2005) *Results of the 2005 Survey of Independent Translators*. Available at: http://www.atio.on.ca/info/ind_survey/survey05_intro.html.

ATIO Salaried Translators Committee (2007) 'Do You Know Maria? Results of the 2007 Survey of Salaried Translators', *InformATIO* 36(2): 6–7. Available at: http://www.atio.on.ca/Membership/Sal_Survey/Sal_Tran_Srvy_Rslts.asp.

Bédard, Claude (2000) 'Mémoire de traduction cherche traducteur de phrases….', *Traduire* 186: 41–49.

Biau Gil, José Ramón (2006) 'Teaching Electronic Tools for Translators Online', in Anthony Pym, Alexander Perekrestenko, and Bram Starink (eds) *Translation Technology and Its Teaching (with Much Mention of Localization)*, Tarragona, Spain: Intercultural Studies Group, Universitat Rovira i Virgili, 89–96. Available at: http://isg.urv.es/library/papers/Biau_Teaching.pdf.

Bowker, Lynne (2003) 'Teaching Translation Technology: Towards an Integrated Approach', *Tradução e Comunicação* 12: 65–79.

Bowker, Lynne (2005) 'Productivity vs Quality? A Pilot Study on the Impact of Translation Memory Systems', *Localisation Focus* 4(1): 13–20.

Bowker, Lynne (2006) 'Translation Memory and "Text"', in Lynne Bowker (ed.) *Lexicography, Terminology and Translation: Text-based Studies in Honour of Ingrid Meyer*, Ottawa: University of Ottawa Press, 175–187.

Bowker, Lynne and Elizabeth Marshman (2010) 'Toward a Model of Active and Situated Learning in the Teaching of Computer-aided Translation: Introducing the CERTT Project', *Journal of Translation Studies* 13(1–2): 199–226.

Clark, Robert, Andrew Rothwell, and Mark Shuttleworth (2002) 'Integrating Language Technology into a Postgraduate Translation Programme', in Belinda Maia, Johann Haller, and Margherita Ulrych (eds) *Training the Language Services Provider for the New Millenium*, Porto: Faculdade de Letras da Universidade do Porto, 63–70.

Dillon, Sarah and Janet Fraser (2006) 'Translators and TM: An Investigation of Translators' Perceptions of Translation Memory Adoption', *Machine Translation* 20(2): 67–79.

eCoLoTrain (2006) 'Translator Training Survey – Results'. Avalable at: http://www.iti.org.uk/uploadedFiles/surveys/eColoTrain-Results%20April%202006%20graphic.pdf.

Fulford, Heather (2001) 'Exploring Terms and Their Linguistic Environment in Text: A Domain-independent Approach to Automated Term Extraction', *Terminology* 7(2): 259–279.

Fulford, Heather and Joaquin Granell-Zafra (2005) 'Translation and Technology: A Study of UK Freelance Translators', *The Journal of Specialised Translation* 4: 2–17. Available at: http://portal.acm.org/citation.cfm?id=1236526&jmp=cit&coll=&dl=.

Gambier, Yves (2009) 'European Master's in Translation: Competences for Professional Translators, Experts in Mulilingual and Multimedia Communication'. Available at: http://ec.europa.eu/dgs/translation/programmes/emt/key_documents/emt_competences_translators_en.pdf.

Garcia, Ignacio (2010a) 'Is Machine Translation Ready Yet?' *Target* 22(1): 7–21.

Garcia, Ignacio (2010b) 'The Proper Place of Professionals (and Non-professionals and Machines) in Web Translation', *Revista Tradumática: Traducció i Technologies de la Informació i la Comunicació* 8. Available at: http://www.fti.uab.es/tradumatica/revista/num8/articles/02/02central.htm.

Gauthier, François (2012) '2012 Survey on Rates and Salaries', K. Montin (trans.), Montréal: Ordre des traducteurs, terminologies et interprètes agréés du Québec (OTTIAQ). Available at: http://www.ottiaq.org/gestion/upload/publications/survey-results-2012.pdf.

Gouadec, Daniel (2003) 'Position Paper: Notes on Translator Training', in Anthony Pym, Carmina Fallada, José Ramón Biau, and Jill Orenstein (eds) *Innovation and E-learning in Translator Training*, Tarragona: Universitat Rovira i Virgili, 11–19.

Jaatinen, Hannu and Jarkko Immonen (2004) 'Finnish University Meets Needs of Translation Industry', *Multilingual Computing and Technology* 15(4): 37–40.

Jaatinen, Hannu and Riitta Jääskeläinen (2006) 'Introducing IT in Translator Training: Experiences from the COLC Project', in Anthony Pym, Aexander Perekrestenko, and Bram Starink (eds) *Translation Technology and Its Teaching*, Tarragona: Intercultural Studies Group, Universitat Rovira i Virgili, 83-88. Available at: http://isg.urv.es/library/papers/JaatinenJaaskelainen_IntroducingIT.pdf.

Kelly, Dorothy (2005) *A Handbook for Translator Trainers*, Manchester: St. Jerome Publishing.

Kelly, Dorothy (2008) 'Training the Trainers: Towards a Description of Translator Trainer Competence and Training Needs Analysis', *TTR* 21(1): 99–125.

Kenny, Dorothy (2007) 'Translation Memories and Parallel Corpora: Challenges for the Translation Trainer', in Dorothy Kenny and Kyongjoo Ryou (eds) *Across Boundaries: International Perspectives on Translation Studies*, Newcastle upon Tyne: Cambridge Scholars Publishing, 192–208.

Kiraly, Don (2000) A Social Constructivist Approach to Translator Education, Manchester: St. Jerome Publishing.

Lagoudaki, Elina (2006) 'Translation Memory Systems: Enlightening Users' Perspective'. Available at: http://www3.imperial.ac.uk/pls/portallive/docs/1/7307707.PDF.

Lommel, Arle and Rebecca Ray (2004) 'The LISA 2004 Translation Memory Survey', Localization Industry Standards Association (LISA). Available at: http://www.lisa.org/Translation-Memory-S.518.0.html.

Marshman, Elizabeth and Lynne Bowker (2012) 'Translation technologies as seen through the eyes of educators and students: Harmonizing views with the help of a centralized teaching and learning resource', in Séverine Hubscher-Davidson and Michal Borodo (eds) *Global Trends in Translator and Interpreter Training: Mediation and Culture*, London/New York: Continuum, 69–95.

Melby, Alan K. (2006) 'MT + TM + QA: The Future Is Ours', *Revista Tradumática: Traducció i Technologies de la Informació i la Comunicació* 4. Available at: http://www.fti.uab.es/tradumatica/revista/num4/articles/04/04.pdf.

Mihalache, Iulia (2012) 'Competency-based Approach to Translator's Training: The Example of LinguisTech', in Paul Lam (ed.) *Proceedings of the 7th International Conference on e-Learning (ICEL 2012)*, Reading, UK: Academic Publishing International, 311–320.

O'Brien, Sharon and Dorothy Kenny (2001) 'In Dublin's Fair City: Teaching Translation Technology at Dublin City University', *Language International* 13(5): 20–23.

Pym, Anthony (2011a) 'Democratizing Translation Technologies: The Role of Humanistic Research', Paper presented at the Luspio Translation Automation Conference, April 5. Available at: http://usuaris.tinet.cat/apym/on-line/research_methods/2011_rome_formatted.pdf.

Pym, Anthony (2011b) 'What Technology Does to Translating', *Translation and Interpreting* 3(1): 1–9. Available at: http://www.trans-int.org/index.php/transint/issue/current.

Samson, Richard (2005) 'Computer Assisted Translation', in Martha Tennent (ed.) *Training for the New Millennium: Pedagogies for Translation and Interpreting*, Amsterdam and Philadelphia: John Benjamins, 101–126.

Shih, Chung-ling (2006) 'Computer-aided Translation Teaching of the Passive Construction', in *Proceedings of the International Conference on Computer-aided Translation: Theory and Practice*. Available at: http://traserver.tra.cuhk.edu.hk/cattap/papers/S06-Computer-Aided%20Translation%20Teaching%20of%20the%20Passive%20Construction%20by%20Prof%20Shih.pdf.

Shuttleworth, Mark (2002) 'Combining MT and CAT on a Technology-oriented Translation Masters', *Teaching Machine Translation: Proceedings of the 6th EAMT Workshop*, 14–15 November 2002, Manchester, UK. Available at: http://www.mt-archive.info/EAMT-2002-Shuttleworth.pdf.

Topping, Suzanne (2000) 'Sharing Translation Database Information: Considerations for Developing an Ethical and Viable Exchange of Data', *Multilingual Computing and Technology* 11(5): 59–61.

Toudic, Daniel (2012) 'The OPTIMALE Employer Survey and Consultation: Synthesis Report'. Available at: http://www.translator-training.eu/attachments/article/52/WP4_Synthesis_report.pdf

Vallianatou, Fotini (2005) 'CAT Tools and Productivity: Tracking Words and Hours', *Translation Journal* 9(4). Available at: http://translationjournal.net/journal/34CAT.htm.

Wheatley, Alan (2003) 'eContent Localization Resources for Translator Training: A Major Breakthrough for Translator Training'. Available at: http://www.iti.org.uk/uploadedFiles/surveys/eCoLoRe%20results.pdf.

5
MACHINE TRANSLATION
General

Liu Qun

DUBLIN CITY UNIVERSITY, IRELAND

Zhang Xiaojun

DUBLIN CITY UNIVERSITY, IRELAND

Definition: Machine Translation (MT)

Machine translation (MT) is a sub-field of computational linguistics (CL) or natural language processing (NLP) that investigates the use of software to translate text or speech from one natural language to another. The core of MT itself is the automation of the full translation process, which is different with the related terms such as machine-aided human translation (MAHT), human-aided machine translation (HAMT) and computer-aided translation (CAT).

History

1950s and 1960s: Pioneers

The idea of MT may be traced back to the seventeenth century (Hutchins and Somers 1992: 5). In 1629, René Descartes proposed the idea of a universal language to share one symbol in different tongues. The possibilities of using computers to translate natural languages were proposed as early as 1947 by Warren Weaver of the Rockefeller Foundation and Andrew D. Booth, a British scientist. In the next two years, Weaver was urged by his colleagues to specify his ideas. In 1949, he wrote a memorandum entitled 'Translation' which concludes his four proposals and assumptions on mechanical translation:

1 the problem of multiple meaning might be tackled by examinations of immediate contexts;
2 there may be logical features common in all language;
3 the cryptographic methods concerned with the basic statistical properties of communication can be applied in mechanical translation; and
4 there may be linguistic universals.

These proposals were practised and realized by successors completely or partially and each proposal was regarded as one of the important approaches in the later MT studies.

Yehoshua Bar-Hillel was one of the successors and practitioners. He was appointed as the first full-time researcher in MT by Massachusetts Institute of Technology (MIT) in 1951. One year later, he convened the first MT conference at MIT to outline the future researches on MT clearly. The MIT conference in 1952 brought together those who had contact with MT and who might have a future interest. Two years later, on the 8 January 1954, the first public demonstration of an MT system was reported by US newspapers. In the report, the demonstration system used only 250 words and 6 grammar rules to translate 49 Russian sentences into English and it achieved success. It was the result of a joint project by IBM staff and members of the Institute of Linguistics at Georgetown University. The impressive IBM-Georgetown demonstration attracted a great deal of attention, and stimulated the large-scale funding of MT research in the USA and in the world. It marked the beginning of MT as a reality from the idea of the use of computer to translate proposed by Weaver seven years earlier.

After the 1954 demonstration, MT study became a multimillion dollar affair in the United States. The decade from 1956 to 1966 filled with high expectations for MT. The emergence of electronic brains created all kinds of expectations in people and institutions and some researchers hoped to solve the problem of MT early on. Major funding went into the field more and more. With sufficient funding, various methods were tried in MT researches. By the mid-1960s, MT research groups had been established in many countries throughout the world, including most European countries (such as Germany, France, Hungary, Czechoslovakia, Bulgaria, and Belgium), China, Mexico and Japan, as well as the United States. Unfortunately, most of them set out to pursue a mistaken and unattainable goal of MT research which is called 'fully automatic high quality (FAHQ) translation'.

Mid-1960s: The ALPAC Report

Optimism was dominant in MT researches in 1950s because of the rising expectations. Developments in formal linguistics such as syntax seemed to promise a great improvement in translation quality. However, disillusion caused by 'semantic barriers' grew as the complexity of the linguistic problems became more and more apparent and researchers saw no straightforward solutions.

In 1959, Bar-Hillel proposed his second survey report on MT research, which questioned the goals and expectations of the whole field of MT research. In 1960, he revised the report and published it in the journal *Advances in Computers* where he was highly critical of any MT group that declared FAHQ translation its long-term aim. He suggested that MT should adopt less ambitious goals but more cost-effective use of man-machine interaction. But the report was read only within MT circles and its impact went relatively unnoticed. The validity of his argument was not seen until the release of the ALPAC report six years later.

In 1964, the National Academy of Sciences of the United States formed the Automatic Language Processing Advisory Committee (ALPAC) to examine the prospects of MT research. The famous report released two years later showed, among other things, that MT output was not cheaper or faster than full human translation with the conclusion that 'there is no immediate or predictable prospect of useful machine translation'. The committee suggested that funding should rather go into basic linguistic research and the development of methods to improve human translation for there was no advantage in using MT systems. Funding for MT in the United States stopped almost entirely as a consequence. While the ALPAC report may have unfairly considered only the goal of high-quality translation, it shows the dangers of over-promising the capabilities

of MT systems. The ALPAC report was widely regarded as narrow, biased and shortsighted. It also misjudged the economics of the computer-based translation industry.

Fortunately, the negative impact of the ALPAC report did not stop research in other countries. MT researches stepped into a quiet decade after the release of the ALPAC report in 1966.

1970s: Revival

In the United States, the main activity had concentrated on English translation of Russian scientific and technical materials. With the worsening of relationship between the United States and Soviet Union in the Cold War, the requirement of translation between English and Russian was changed. However, in Canada and Europe, translation needs were quite different. To the Canadian government, English–French translation was required to meet the demand of its bicultural policy. In the European countries, there were growing demands for translations from and into all European Community languages. Therefore, the focus of MT activity switched from the United States to Canada and to Europe in 1970s.

At Montreal in Canada, a project named TAUM (Traduction Automatique à l'Université de Montréal) generated two MT systems: Q-system and Météo system. Q-system formalized a computational metalanguage for manipulating linguistic strings and trees in natural language processing. Météo system translated weather forecasts from English to French for the whole of Canada. Its throughput started at 7,500 words a day and reached more than 80,000 words a day or close to 30 million words a year until it was replaced by its successor in 2001. Météo has been successfully doing this job since 1976, which was designed for the restricted vocabulary and limited syntax of meteorological reports.

Between 1960 and 1971, the group at Grenoble University in France established by Bernard Vauquois developed an interlingua system, CETA, for translating Russian into French. An interlingua, namely 'pivot language', was used at the grammatical level (Slocum 1985: 1–17).

Russian MT research was greatly affected by Igor Mel'čuk's meaning-text model (Kahane 2003: 546–569), which was an ambitious interlingua system combining a stratificational dependency approach with a strong emphasis on the lexicographic aspects of an interlingua.

The United States MT researches were also revived from the mid-1970s. The Linguistics Research Center (LRC) at the University of Texas adopted a similar model of 'pivot language' in its METAL system. Another famous system, Systran, was founded in 1968. The Russian–English system has been used by the US Air Force since 1970. A French–English version was bought by the European Commission in 1976, and thereafter systems for more European language pairs were developed. Xerox Corporation has used Systran to translate technical manuals since 1978. There is another long-established system named Logos, which was an English–Vietnamese system for translating aircraft manuals during the 1970s. The Logos system was based on a direct translation approach. The Pan American Health Organization in Washington has successfully developed and widely used two systems for translating Spanish to English and back since the 1970s.

1980s and 1990s: Diversity

Research from the 1980s had three main strands: (1) transfer systems; (2) new kinds of interlingua systems; and (3) corpus-based MT research.

The Grenoble group began development of the second generation linguistic-based transfer system Ariane in the 1980s. Different to the previous pivot language tree, the Ariane system could

incorporate different levels and types of representation on single labelled tree structures. Ariane did not become an operational system and ceased in the late 1980s, but became part of the Eurolang project in the 1990s and the Grenoble team has continued MT research in this project.

The Mu system was developed at the University of Kyoto under Makoto Nagao in the 1980s. The features of the Mu system were case grammar analysis, dependency tree presentation and GRADE, a programming environment for grammar writing (Nagao and Tsujii 1986: 97–103). Since 1986, the Mu system has been converted into an operational system in practice.

From the mid-1970s to the late 1980s, the group at Saarbrücken (Germany) developed SUSY. It was a highly modular multilingual transfer system focusing on the in-depth treatment of inflected languages such as Russian and German (Maas 1977: 582–592). The Eurotra project of the European Community in the 1980s aimed to construct an advanced multilingual transfer system for translation among all the Community languages. It was designed to combine lexical, logical syntactic information and semantic information in different level interfaces. In that period, Eurotra was regarded as the best linguistics-based design (de Roeck 1981).

There was a general revival of interest in interlingua systems during the latter half of the 1980s. Some groups used knowledge-based methods from research on artificial intelligence (AI). DLT system was a leader of development in this field in the 1990s (Witkam 1988: 756–759). It was a multilingual interactive system operating over computer networks with each terminal as a translating machine from and into one language only. Its interlingua was a modified form of Esperanto. Another interlingua project was Rosetta (Landsbergn 1982: 175–181), which explored the use of Montague grammar in interlingual representations. After the Mu system, MT researches in Japan showed a wide variety of approaches. The PIVOT system from NEC was a typical interlingua system which is now available commercially (Muraki 1987: 113–115). The LUTE project of NTT was a knowledge-based experiment (Nomura *et al.* 1985: 621–626). Investigations of AI methods in MT research began in the mid-1970s, focusing on preference semantics and semantic templates. Later, Roger Schank at Yale University developed expert systems of text understanding (Schank and Abelson 1977). The KANT prototype system of the Carnegie-Mellon team was designed as 'meaning-oriented MT in an interlingua paradigm' for translating computer manuals for English and Japanese in both directions (Nyberg III and Mitamura 1992: 1069–1073). The core of the system is the interlingual representation form of networks of propositions.

Rule-based machine translation (RBMT) research continued in both transfer and interlingua systems after 1990. The CAT2 system at Saarbrücken (Sharp 1988) and PaTrans transfer system in Denmark (Orsnes *et al.* 1996: 1115–1118) were two fruits of the Eurotra project. The Eurolang project developed ten language pairs between English, French, German, Italian and Spanish and produced a translator's workstation, Optimizer (Seite *et al.* 1992: 1289–1293). Another remarkable European MT project, LMT, stood for 'logic programming MT' which implemented translation in Prolog in combination with the lexical approach. In 1994, LMT programs were sold to IBM to be modules of TranslationManager/2. CATALYST at the Carnegie-Mellon University, ULTRA at the New Mexico State University, UNITRAN at the University of California were all rule-based domain-adopted interlingual systems in the 1990s.

Since the end of the 1980s, the dominance of linguistic rules-based approaches was broken by the appearance of new corpus-based methods and strategies. First, an IBM research group purely based on statistical methods developed MT system, which carved out the way to statistical machine translation (SMT). Second, a Japanese group tried to discover methods to leverage translation examples, namely example-based machine translation (EBMT).

In 1988, encouraged by its success in speech recognition, an IBM group in the Candide project began to look for the application of statistics in MT. Their research was based on the

vast French and English texts of Canadian parliamentary debate reports to align phrases, word groups and individual words and calculate the probabilities that any segment in source language corresponds to the segment in aligned target language. Fortunately, the results were acceptable. To improve the results, the IBM group proposed to introduce more sophisticated statistical methods in the 1990s (Vogel *et al.* 1996: 836–841; Och and Ney 2003: 19–51).

The example-based experiments were begun at the end of the 1980s even though Makoto Nagao had proposed this idea first in 1984. An underlying assumption was that translation often involves the finding or recalling of analogous examples. The example-based approach extracted and selected the equivalent aligned translation segments from a parallel corpus as the examples in translation process (Nagao 1984: 173–180).

The availability of large corpora had encouraged the research on parallel computation, neural networks and connectionism. MT researches ran into a statistical era in the new century.

2000s afterwards: New trends

SMT systems are currently being developed in a large number of academic and commercial systems. IBM pioneered the research on SMT, and in particular, proposed the word-based SMT methods based on the IBM Model 1-5 (Brown *et al.* 1993: 263–313) which also provide the theoretical basis to word alignment which is fundamental to all other SMT methods. The modern statistical phrase-based models are rooted in works by Franz Och and his colleagues (Och and Weber 1998: 985–989; Och *et al.* 1999: 20–28; Och 2002; Och and Ney 2004: 417–449) and based on later works by Philipp Koehn and his co-operators (Koehn *et al.* 2003: 48–54). Various syntax-based models are proposed and researched to utilize the syntax information in translation. Typical syntax-based models include inverse transduction grammar (Wu 1995: 1328–1335), the hierarchical phrase-based model (Chiang 2005: 265–270, 2007), string-to-tree model (Galley *et al.* 2004: 273–280; Galley *et al.* 2006: 961–968) and tree-to-string model (Liu *et al.* 2006: 609–616; Huang *et al.* 2006: 66–73).

Open source tools are broadly used in the SMT research community. GIZA++ is an implementation of the IBM word-based models and Hidden Markov Model (HMM), and it is commonly used nowadays for word alignment. The code was developed at a Johns Hopkins University summer workshop and later refined by Och and Ney (Al-Onaizan et al. 1999; Och and Ney 2000: 440–447). Moses is an implementation of a phrase-based decoder, including training. It was developed at the University of Edinburgh and enhanced during a Johns Hopkins University summer workshop also (Koehn *et al.* 2007: 17–180).

MT evaluation is another highlight of MT research in the new century. Manual assessment and automatic evaluation are compulsory. The BLEU (bilingual evaluation understudy) scoring tool is most commonly used to evaluate machine translation performance (Papineni et al. 2002: 311–318). Recently, some other metrics such as a metric for evaluation of translation with explicit ordering (METEOR) (Banerjee and Lavie 2005: 65-72; Lavie and Agarwal 2007: 228–231) and translation edit rate (TER) (Snover et al. 2006: 223–231) have also gained popularity. The NIST (National Institute of Standards and Technology) evaluation is the oldest and most prestigious evaluation campaign. The IWSLT (International Workshop on Spoken Language Translation) evaluation campaign has a stronger focus on speech translation. The WMT (Workshop on Machine Translation) evaluation campaign targets translation between European languages. These campaigns also generate standard test sets, manual evaluation data, and reference performance numbers (Koehn 2010).

Integration with other technologies such as speech recognition can put MT into a global intelligent content processing circle. With the vast needs of multimodal and multilingual

communication via various social media and on-line communities, a diversity of MT applications such as speech-to-speech translation, computer-aided translation, photo translation and web translation have developed rapidly in recent years.

Approaches

Vauquois' Triangle

MT approaches can be categorized by the depth of intermediary representations which are used in the translation process: direct, transfer and interlingua, which are often depicted by the Vauquois' Triangle (Vauquois 1968: 254–260):

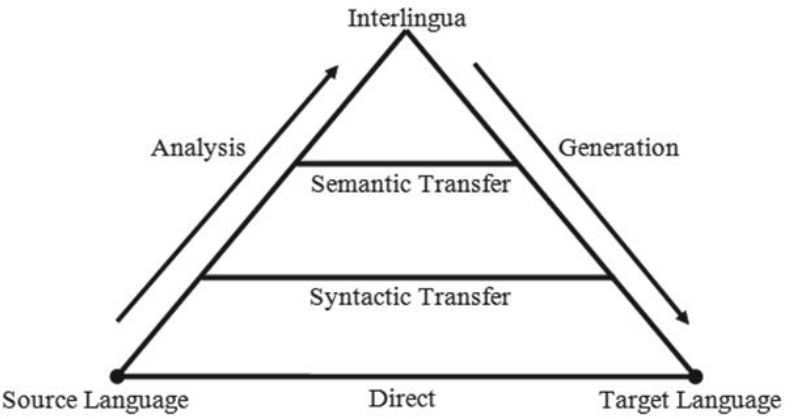


Figure 5.1 Vauquois' Triangle

In this triangle, shown in Figure 5.1, transfer approach is subcategorized to syntactic transfer approach and semantic transfer approach.

Direct

In the direct approach, the target language sentence is translated directly from the source word sequence, possibly with analysis in certain individual linguistic phenomena. No general syntactic or semantic analysis is conducted and the most (even only) important resource for this approach is a dictionary. That is the reason why it is also called dictionary-based approach.

Transfer

In the transfer approach, the intermediary structures are defined for both source and target languages. The translation process is divided into three phases: analysis from the source language sentence to the source intermediary structure; transfer from the source intermediary structure to the target intermediary structure; generation from the target intermediary structure to the target sentence.

Independent analysis means the characteristics of the target language are not considered in the analysis phrase, which will lead to a source intermediary structure which can be used for MT into various target languages. Independent generation means the characteristics of source

language are not considered in the generation phrase, that is to say, the target intermediary structure can be used for MT from various source languages.

In the syntactic transfer approach the transfer process mainly occurs at the syntactic level. Thus the system may have components of morphological analysis, syntactic analysis, syntactic transfer, syntactic generation and morphological generation.

In the semantic transfer approach the transfer process mainly occurs at the semantic level. Accordingly the system may have components of morphological analysis, syntactic analysis, semantic analysis, semantic transfer, semantic generation, syntactic generation and morphological generation.

Interlingua

In the interlingua approach, a universal representation is defined for all the source and target languages. The translation process only contains two phrases: analysis and generation. The interlingua approach is regarded as an appropriate method for multilingual MT because it dramatically reduces the number of components compared with what is needed in direct approach or transfer approach. An interlingual can be a structured representation such as a logic expression, a semantic network, or a knowledge representation and so on, or an artificial or natural language representation. It is also called a knowledge-based approach (Nirenburg 1989: 5–24; Carbonell *et al.* 1978) when a knowledge representation is used as an interlingua. An interlingua is also called a pivot language, a metalanguage or a bridge language. Practices in some large-scale projects showed that an interlingua approach using a human-defined presentation may encounter uncontrollable complexity when many languages are involved (Nagao 1989; Patel-Schneider 1989: 319–351). Recent web-based translation services such as Google Translate, Bing Translator, etc. usually adopt English as a pivot language to support MT between tens of other languages.

Another group of terms used for categorizing MT approaches include rule-based, example-based, statistical and hybrid MT.

Rule-based

An RBMT system uses rules to direct the process of MT. In an RBMT system, rules are encoded by experts with their linguistic insights. Although the automatically extracted rules are also used in some EBMT or SMT approaches, they are not an RBMT approach because they do not use human-encoded rules. In an RBMT system, rules may be used in all the components. Table 5.1 lists the components and corresponding rules used in a typical RBMT system adopting a semantic transfer approach.

Not all these components and rules are necessary for a practical RBMT system. Some of them may be omitted or merged, depending on the specific language pairs or the algorithms used. Also it is possible to have additional components or rule bases for a specific purpose.

A rule-based paradigm provides a mechanism for linguists and computer scientists working together to develop an MT system: computer scientists can focus on algorithm design and implementation, while linguistic experts can focus on the construction of rule bases and lexicons. The development of an RBMT system is time-consuming and labor-intensive and that of a commercial RBMT system may take several years. Human-encoded rules suffer from the low coverage of linguistic phenomena and the conflicts between rules which lead to unsatisfied translation quality when facing large-scale real-life texts.

Table 5.1 Rules used in an RBMT system adopting a semantic transfer approach

Analysis	Morphological analysis	Source morphological rules
	Syntactic analysis (parsing)	Source grammar
	Semantic analysis	Source semantic rules
Transfer	Lexical transfer	Bilingual lexicon
	Syntactic transfer	Syntactic mapping rules
	Semantic transfer	Semantic mapping rules
Generation	Semantic generation	Target semantic rules
	Syntactic generation	Target grammar
	Morphological generation	Target morphological rules

Example-based

In an EBMT system, a sentence is translated by analogy (Nagao 1984: 173–180). A number of existing translation pairs of source and target sentences are used as examples, which is also called parallel corpus. When a new source sentence is to be translated, the examples are retrieved to find similar ones in the source side to match. Then the target sentence is generated by imitating the translation of the matched examples. Because the hit rate for long sentences is very low, usually the examples and the source sentence are broken down into small fragments. Word alignment is necessary between the source and target examples, so that the correspondent target part can be located when only a part of the source example is matched. The typical process of EBMT includes the following phases:

1 *Example retrieval*: Indexing on examples should be built for fast retrieval on large numbers of examples against the input source sentence. Fuzzy match or partial match should be supported in example retrieval.
2 *Source sentence decomposing*: The input source sentence is decomposed into fragments to match example fragments or lexicon entries.
3 *Fragment translation*: Each matched source fragment is translated according the word alignment between source and target examples.
4 *Target sentence recombination*: The translation of the source fragments is assembled into the target sentence.

It is very common to introduce syntactic parser in EBMT approach (Sato and Nagao 1990: 247–252). In such cases an example fragment may be a sub-tree of a syntactic structure instead of a word sequence.

EBMT provides a technique to improve the translation quality by increasing the size of corpus only, and to obtain natural output sentences without deep analysis on the source side. EBMT may result in a high-quality translation while high similarity examples are found. On the contrary, when there is no example found with high similarity, the translation quality may be very low.

Memory-based MT is also called the translation memory (TM) approach, which is broadly used in CAT. TM is a sentence-aligned parallel corpus which is usually accumulated by the user him/herself. When a new source sentence is to be translated, the TM is searched and if there are one or more sentences matched with higher similarity than a certain threshold, the aligned target sentence of the most similar source sentence in the translation memory will be

output without any modification. The memory-based approach provides reference translation for every source sentence, and it is necessary for the output to be post-edited if the source sentence is not matched with 100 per cent similarity or the post-editor does not satisfy the output translation at all. TM approach is regarded as a special case of EBMT.

Statistical

The basic idea of SMT (Brown *et al.* 1990: 79–85; Brown *et al.* 1993: 263–313; Koehn and Knight 2009) is to mathematically model the probability of a target sentence being the translation of a given source sentence $P(T \mid S)$, which is called a translation model. Once the translation model is defined and implemented, the problem of translating a source sentence into a target sentence is converted to searching a target sentence with the highest translation probability $\hat{T} = \arg\max_T P(T \mid S)$. Such a searching process is decoding. For a given translation model, its parameters can be obtained from a given parallel in a training process.

A translation model is usually decomposed to several specific models under a certain framework to model the translation probability in different aspects.

An early framework is the source channel model (Brown *et al.* 1990: 79–85). In this model, first a target sentence is generated by an information source described by a language model $P(T)$, then the target sentence is transmitted through an information channel described by a reverse translation model $P(S \mid T)$, and finally output the source sentence. With the source sentence observed, decoding can be viewed as a search process of finding the optimal target sentence $\hat{T} = \arg\max_S P(T)P(S \mid T)$. The main contribution of the source channel model is introducing the language model to SMT. While the translation model ensures the source and target sentences with the same meaning, the language model guarantees the target sentence is fluent.

A more general framework is the log-linear model (Och and Ney 2002: 295–302). In this model, the translation probability is defined as a log-linear combination based on a set of feature functions:

$$P(T \mid S) = \frac{\exp(\sum_i \lambda_i h_i(S,T))}{\sum_{T'} \exp(\sum_i \lambda_i h_i(S,T'))}$$

where $h_i(S,T)$ is an arbitrary real function defined on source sentence S and target sentence T, and the denominator is a constant for normalization. Thus the decoding can be viewed as a search process to find a target sentence with the highest translation probability $\hat{T} = \arg\max_T \sum_i \lambda_i h_i(S,T)$. A log-linear framework takes the source channel framework as a special case, where the only two features are log $P(T)$ and log $P(S \mid T)$, and the parameters are both equal to 1. In a log-linear model, the parameters λ_i can be obtained in the process of tuning against a specific target function on a held-out development data without overlap with the training data. Features usually include one or more language models, translation models, reordering models, the length of the output sentence, lexicon features, etc. The log-linear model provides the possibility of incorporating any useful features in MT, and balancing the effectiveness of all the features by tuning the parameters discriminatively.

Language models and translation models are the most important models in SMT. The most commonly used language model is an n-gram model. Translation models can be classified into several different formalisms depending on the language units used: word-based models, phrase-based models and syntax-based models.

Word-based models calculate sentence translation probability based on word-to-word translation tables. The IBM Model 1-5 (Brown *et al.* 1993: 263–313) is a typical word-based translation model with increasing complexity. The IBM Model 1 only considers the word-to-word translation probabilities while the later models introduce more sophisticated factors, such as distortion probability which characterizes the word reordering and the fertilization probability which depicts one-to-many mappings between words. IBM models can be trained on a sentence-aligned corpus using an expectation-maximization (EM) algorithm which results in the model parameters as well as the word alignment on the training corpus. HMM is an improved version of IBM Model 2 which models the distortion probability as a Hidden Markov Chain (Vogel *et al.* 1996: 836–841).

Phrase-based models (Och 2002; Koehn *et al.* 2003) are built based on phrase tables which record phrase-to-phrase translation probabilities. Phrase-based models can capture the local context while translating a word, thus outperforming word-based models significantly. However, phrase-based models fail to capture long-distance dependency.

Syntax-based models are built based on synchronized grammars. The rule tables are used to record the probabilities of synchronized syntax rules. A translation rule consists of a source rule, a target rule, and a correspondence between variables in source and target rules. There are many formalisms for syntax-based models depending on the characters of syntax information utilized: some of them, such as stochastic bracketing inverse transduction grammar (Wu 1995: 1328–1335) and hierarchical phrase-based model (Chiang 2005: 263–270), do not use linguistic syntax labels; others use linguistic syntax labels in source side (e.g. tree-to-string model) (Liu *et al.* 2006: 609–616; Huang *et al.* 2006: 66–73), in target side (e.g. string-to-tree model) (Galley *et al.* 2004: 273–280; Galley *et al.* 2006: 961–968) or in both sides (e.g. tree-to-tree model). Syntax-based models can capture long-distance dependencies between words and perform better than phrase-based models on language pairs with very different syntax structures.

Training of language models are on monolingual corpus while that of translation models are on parallel corpus. Word-based models (IBM Model 1-5 and HMM) can be trained directly from a sentence-aligned corpus using EM algorithms and the model parameters and the word alignments will be generated at the same time (Brown *et al.* 1993: 263–313). The training processes of phrase-based models and syntax-based models are based on the word-alignments generated by word-based models, which are also called phrase extraction (Koehn *et al.* 2003: 48–54) and rule extraction (Galley *et al.* 2006: 941–968) respectively.

Decoding means to search an optimal target sentence from the space of all possible target sentences for a given sentence. Stack search algorithm is the most commonly used algorithm for decoding in SMT, where partial translation candidates are grouped in different stacks and a threshold is used for each stack to prune low possibility candidates. The decoding algorithm is closely related to the utilized translation model. For word-based and phrase-based models, the decoding usually runs in a left-to-right style, which means the words in the target sentence are generated from left to right, while for syntax-based models it goes in a bottom-up style, which means that small pieces of target sentences are generated first and large pieces are merged from these small pieces.

Tuning means to train the parameters for the log-linear model in a development data set, which is usually in the same domain as the test data are in. The parameters are tuned to obtain a best score with regard to certain automatic evaluation metrics in the development data. Minimal error rate training (MERT) (Och 2003: 160–167) algorithm is the most commonly used tuning algorithm. Other tuning algorithms are also proposed to improve the performance when a large number of features are used.

Hybrid

Because all the above MT approaches have their advantages and shortcomings, many hybrid MT approaches are proposed to integrate the advantages of different approaches.

The system recommendation approach takes all the outputs of different systems and recommends the best one.

The system combination (Rosti *et al.* 2007: 228–235) approach takes one or more outputs from each system and merges these results in a word, phrase or sentence level.

Pipelined approaches adopt one system as the main system and another system for monolingual pre-processing or post-processing. Typical pipelined hybrid approaches include statistical post-editing for RBMT (Dugast *et al.* 2007: 220–223) and rule-based pre-reordering for SMT (Xia and McCord 2004: 508–514).

Mixture approaches adopt one approach for the main system while using other approaches in one or more components. For example, RBMT may adopt a statistical word segmentation or parsing, while SMT usually utilizes human-encoded rules to translate certain types of name entities such as time, date, numerical expressions and names of persons, locations or organizations.

Almost all the practical MT systems adopt hybrid approaches to a certain extent.

Quality evaluation and estimation

MT evaluation refers to assessing an MT system from various aspects, of which translation quality is its main concern.

Human evaluation

Human evaluation is the most reliable method for MT quality evaluation. It can be divided into:

1 *Scoring-based human evaluation*: Human evaluators are asked to score each system output sentence by sentence, and the the average score on all the sentences and evaluators is the final score of a system. The most common metrics for human scoring is adequacy and fluency. Adequacy reflects how much meaning of the source sentence is conveyed in the target sentence, while fluency measures to what degree the target sentence is smooth idiomatically and grammatically.
2 *Ranking-based human evaluation*: Human evaluators are asked to rank the results of the same sentence given by part or all of the systems. An overall ranking of all systems is finally generated by synthesizing all the human assessments.
3 *Post-edit-based human evaluation*: Human post-editors are asked to post-edit all the results given by every sentence.
4 Finally a *human translation edit rate* (HTER) (Snover *et al.* 2006: 223–231) is calculated for each system based on the system results and post-edited correspondences.

Automatic evaluation

Human evaluation is expensive and time-consuming and thus unsuitable for frequent use during research and development. Various automatic evaluation methods are proposed. Word error rate (WER) is defined based on the Levenshtein distance between the system output and the reference translation. Position-independent error rate (PER) calculates the word error rate

by treating the sentence as a bag of words and ignoring the word order. TER considers the shift operation in addition to the insertion, deletion and substitution operations used in WER. BLEU (Papineni *et al.* 2002: 311–318) computes the n-gram precision rather than words error rate against multiple references. METEOR (Banerjee and Lavie 2005: 65–72) takes further considerations of stemming and synonym for evaluation. Diagnostic metrics (Yu 1993: 117–26; Zhou *et al.* 2008: 1121–8) calculate correctness on a number of linguistic checkpoints predefined and distributed in the test sentences rather than the score of each sentence, which provide a better understanding of the systems from a linguistic point of view.

Automatic translation quality evaluation plays an important role in SMT research since it provides the target function for parameter tuning. However, the correlations between current automatic evaluation metrics and human translation are not satisfactory.

Automatic quality estimation

Because the current MT quality is not stable, many users hope to know the translation quality before they use it. Automatic quality estimation technologies are developed for this purpose (Specia *et al.* 2009: 28–35). Quality estimation can be done on sentence level or word level. Usually a statistical classifier is trained to predict the translation quality for each sentence or word.

Application

Although MT does not reach the so-called ideal FAHQ, it finds its way to be applied in many cases with acceptable quality.

The most popular MT application may be the on-line translation services provided by search engines such as Google Translate and Microsoft Bing Translator. Such products support translation between tens of languages and provide application programming interfaces (APIs) for other applications.

Another type of application for MT is the integration with CAT tools. MT is used for post-editing by profession translators and brings significant improvement on their work efficiency.

MT is also used in on-line or off-line human interaction situations by integration with tools such as instant messengers and emails.

The combination of MT with other technologies also produces a diversity of applications, for example, speech translations, snapshot translations and cross-lingual information retrieval.

References

Al-Onaizan, Yaser, Jan Curin, Micahel Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah Smith, and David Yarowsky (1999) *Statistical Machine Translation*, Technical Report, John Hopkins University Summer Workshop (WS 99) on Language Engineering, Center for Language and Speech Processing, Baltimore, MD.

Banerjee, Satanjeev and Alon Lavie (2005) 'METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements', in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 29 June 2005, University of Michigan, Ann Arbor, MI, 65–72.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (1990) 'A Statistical Approach to Machine Translation', *Computational Linguistics* 16(2): 79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) 'The Mathematics of Statistical Machine Translation', *Computational Linguistics* 19(2): 263–313.

Carbonell, Jaime G., Richard E. Cullinford, and Anatole V. Gershman (1978) *Knowledge-based Machine Translation*, Technical Report, Department of Computer Science, Yale University, New Haven, CT.

Chiang, David (2005) 'A Hierarchical Phrase-based Model for Statistical Machine Translation', in *Proceedings of the 43rd Annual Meeting of the Association of Computational Linguistics (ACL)*, 25–30 June 2005, University of Michigan, Ann Arbor, MI, 263–270.

Chiang, David (2007) 'Hierarchical Phrase-based Translation', *Computational Linguistics* 33(2): 201–228.

de Roeck, Anne (1981) 'Anatomy of Eurotra: A Multilingual Machine Translation System', in *Actes du Congrès international informatique et sciences humaines*, Liège: Université de Liège, 298–303.

Dugast, Loïc, Jean Senellart, and Philipp Koehn (2007) 'Statistical Post-editing on Systran's Rule-based Translation System', in *Proceedings of the 2nd Workshop on Statistical Machine Translation*, 23 June 2007, Prague, Czech Republic. Stroudsburg, PA: Association of Computational Linguistics, 220–223.

Galley, Michel, Mark Hopkins, Kevin Knight, and Daniel Marcu (2004) 'What's in a Translation Rule?' in Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, 2–7 May 2004, Boston, MA, 273–280.

Galley, Michel, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer (2006) 'Scalable Inference and Training of Context-rich Syntactic Translation Models', in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 20 July 2006, Sydney, 961–968.

Huang, Liang, Kevin Knight, and Aravind Joshi (2006) 'Statistical Syntax-directed Translation with Extended Domain of Locality', in *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation*, 8–12 August 2006, Cambridge, MA, 66–73.

Hutchins, W. John and Harold L. Somers (1992) *An Introduction to Machine Translation*, London: Academic Press.

Kahane, Sylvain (2003) 'The Meaning-text Theory', in Vilmos Agel, Ludwig M. Eichinger, Hans Werner Eroms, Peter Hellwig, Hans Jürgen Heringer, and Henning Lobin (eds) *Dependency and Valency: An International Handbook of Contemporary Research*, Berlin and New York: Walter de Gruyter, 546–569.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu (2003) 'Statistical Phrase-based Translation', in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Stroudsburg, PA, 48–54.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007) 'Moses: Open Source Toolkit for Statistical Machine Translation', in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 23–30 June 2007, Prague, Czech Republic, 177–180.

Koehn, Philipp and Kevin Knight (2009) 'U.S. Patent No. 7,624,005', Washington, DC: U.S. Patent and Trademark Office.

Koehn, Philipp (2010) *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Landsbergn, Jan (1982) 'Machine Translation Based on Logically Isomorphic Montague Grammars', in Jan Horecky (ed.) *Proceedings of the 9th International Conference on Computational Linguistics*, 5–10 July 1982, Prague/Amsterdam/New York/Oxford: North-Holland Publishing Company, 175–181.

Lavie, Alon and Abhaya Agarwal (2007) 'Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments', in *Proceedings of the 2nd Workshop on Statistical Machine Translation*, 23 June 2007, Prague, Czech Republic, 228–231.

Liu, Yang, Qun Liu, and Shouxun Lin (2006) 'Tree-to-string Alignment Template for Statistical Machine Translation', in Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 20 July 2006, Sydney, Australia, 609–616.

Maas, Heinz-Dieter (1977) 'The Saarbrücken Automatic Translation System (SUSY)', in *Overcoming the Language Barrier*, 3-6 May 1977, Munich: Verlag Dokumentation, 585–592.

Muraki, Kazunori (1987) 'PIVOT: Two-phase Machine Translation System', in *Proceedings of the Machine Translation Summit*, 17–19 September 1987, Kanagawa, Japan, Tokyo: Ohmsha Ltd., 113–115.

Nagao, Makoto (1984) 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle', in Alick Elithorn and Ranan Banerji (eds) *Artificial and Human Intelligence*, New York: Elsevier North-Holland Inc., 173–180.

Nagao, Makoto and Jun-ichi Tsujii (1986) 'The Transfer Phase of the Mu Machine Translation System', in *Proceedings of COLING '86: 11th Conference on Computational Linguistics*, 25–29 August 1986, University of Bonn, Germany, 97–103.

Nagao, Makoto (1989) *Machine Translation: How Far Can It Go?* Oxford: Oxford University Press.
Nirenburg, Sergei (1989) 'Knowledge-based Machine Translation', *Machine Translation* 4(1): 5–24.
Nomura, Hirosato, Shozo Naito, Yasuhiro Katagiri, and Akira Shimazu (1985) 'Experimental Machine Translation Systems: LUTE', in *Proceedings of the 2nd Joint European-Japanese Workshop on Machine Translation*, December 1985, Geneva, Switzerland, 621–626.
Nyberg III, Eric H. and Teruko Mitamura (1992) 'The KANT System: Fast, Accurate, High-quality Translation in Practical Domains', in *COLING '92 Proceedings of the 14th Conference on Computational Linguistics*, 23–28 August 1992, Nantes, France, 1069–1073.
Och, Franz Josef (2002) 'Statistical Machine Translation: From Single-word Models to Alignment Templates', PhD thesis, RWTH Aachen, Germany.
Och, Franz Josef (2003) 'Minimum Error Rate Training in Statistical Machine Translation', in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 7–12 July 2003, Sapporo Convention Center, Sapporo, Japan, 160–167.
Och, Franz Josef and Hans Weber (1998) 'Improving Statistical Natural Language Translation with Categories and Rules', in Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics and 17th International Conference on Computational Linguistics, 10–14 August 1998, University of Montreal, Quebec, Canada, 985–989.
Och, Franz Josef and Hermann Ney (2000) 'Improved Statistical Alignment Models', in *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics*, 4–9 August 2000, Morristown, NJ, 440–447.
Och, Franz Josef and Hermann Ney (2002) 'Discriminative Training and Maximum Entropy Models for Statistical Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 7–12 July 2002, University of Pennsylvania, Philadelphia, PA, 295–302.
Och, Franz Josef and Hermann Ney (2003) 'A Systematic Comparison of Various Statistical Alignment Models', *Computational Linguistics* 29(1): 19–51.
Och, Franz Josef and Hermann Ney (2004) 'The Alignment Template Approach to Statistical Machine Translation', *Computational Linguistics* 30(4): 417–449.
Och, Franz Josef, Christoph Tillman, and Hermann Ney (1999) 'Improved Aligned Models for Statistical Machine Translation', in Pascale Fung and Joe Zhou (eds) *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 21–22 June 1999, University of Maryland, 20–28.
Orsnes, Bjarne, Bradley Music, and Bente Maegaard (1996) 'PaTrans – A Patent Translation System', in *Proceedings of the 16th Conference on Computational Linguistics*, Copenhagen, Denmark, 1115–1118.
Papineni, Kishore A., Salim Roukos, Todd Ward, and Zhu Wei-Jing (2002) 'BLEU: A Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002,* 7–12 July 2002, University of Pennsylvania, Philadelphia, PA, 311–318.
Patel-Schneider, Peter F. (1989) 'A Four-valued Semantics for Terminological Logics', *Artificial Intelligence* 38(3): 319–351.
Rosti, Antti-Veikko I., Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr (2007) 'Combining Outputs from Multiple Machine Translation Systems', in *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, 23–25 April 2007, Rochester, NY, 228–235.
Sato, Satoshi and Makoto Nagao (1990) 'Toward Memory-based Translation', in Hans Karlgren (ed.) *Proceedings of the 13th Conference on Computational linguistics*, Stroudsburg, PA, 247–252.
Schank, Roger C. and Robert P. Abelson (1977) *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, Hillsdale, NJ: Lawrence Erlbaum Associates.
Seite, Bernard, Daniel Bachut, D. Maret, and Brigitte Roudaud (1992) 'Presentation of Eurolang Project', in *COLING '92 Proceedings of the 14th Conference on Computational Linguistics*, 23–28 August 1992, Nantes, France, 1289–1293.
Sharp, Randall (1988) 'CAT2 — Implementing a Formalism for Multi-lingual MT', in Proceedings of the 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, 3–6 June 1988, Pittsburgh, PA.
Slocum, Jonathan (1985) 'A Survey of Machine Translation : Its History, Current Status, and Future Prospects', *Computational Linguistics* 11(1): 1–17.
Snover, Matthew, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006) 'A Study of Translation Edit Rate with Targeted Human Annotation', in *Proceedings of Association for Machine Translation in the Americas*, 8–12 August 2006, Cambridge, MA, 223–231.

Specia, Lucia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini (2009) 'Estimating the Sentence-level Quality of Machine Translation Systems', in *Proceedings of 13th Annual Conference of the European Association for Machine Translation*, May 2009, Barcelona, Spain, 28–35.

Vauquois, Bernard (1968) 'A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation', in Arthur J.H. Morrell (ed.) *Information Processing 68, Proceedings of the IFIP (Internal Federation for Information Processing) Congress 1968*, 5–10 August 1968, Edinburgh, UK, 254–260.

Vogel, Stephen, Hermann Ney, and Christoph Tillmann (1996) 'HMM-based Word Alignment in Statistical Translation', in *Proceedings of the 16th Conference on Computational Linguistics*, 5–9 August 1996, Center for Sprogteknologi, Copenhagen, Denmark, 836–841.

Witkam, Toon (1988) 'DLT: An Industrial R and D Project for Multilingual MT', in *Proceedings of the 12th Conference on Computational Linguistics*, Budapest, Hungary, 756–759.

Wu, Dekai (1995) 'Stochastic Inversion Transduction, Grammars, with Application to Segmentation, Bracketing, and Alignment of Parallel Corpora', in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, San Francisco, CA, 1328–1335.

Xia, Fei and Michael McCord (2004) 'Improving a Statistical MT System with Automatically Learned Rewrite Patterns', in *Proceedings of the 20th International Conference on Computational Linguistics*, 23–27 August 2004, University of Geneva, Switzerland, 508–514.

Yu, Shiwen (1993) 'Automatic evaluation of output quality for machine translation systems', *Machine Translation* 8(1–2): 117–126.

Zhou, Ming, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang and Tiejun Zhao (2008). 'Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points', in *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008), August 2008. Manchester, UK, 1121–1128.

6
MACHINE TRANSLATION
History of research and applications

W. John Hutchins

FORMERLY UNIVERSITY OF EAST ANGLIA, THE UNITED KINGDOM

From 1949 to 1970

Within a few years of the first appearance of the 'electronic calculators' research had begun on using computers as aids for translating natural languages. The major stimulus was a memorandum in July 1949 by Warren Weaver, who put forward possible lines of research. One was a statistical approach expressed as a dictum that 'When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode".' Weaver referred also to war-time success in code-breaking, from developments by Shannon in information theory and from speculations about universal principles underlying natural languages (Weaver 1949: 15–33). Within a few years research had begun at many US universities, and in 1954 the first public demonstration of the feasibility of translation by computer was given, in a collaboration of IBM and Georgetown University (Hutchins 2004: 102–114). Although using a very restricted vocabulary and grammar it was sufficiently impressive to stimulate massive funding of what became known since that date as 'machine translation' (MT).

This first decade saw the beginnings of the three basic approaches to MT. The first was the 'direct translation' model, where programming rules were developed for translation specifically from one source language (SL) into one particular target language (TL) with a minimal amount of analysis and syntactic reorganization. The second approach was the 'interlingua' model, based on abstract language-neutral representations (codes or symbols independent of both SL and TL), where translation would then be in two stages, from SL to the interlingua and from interlingua to TL. The third approach was less ambitious: the 'transfer approach', where conversion was through a transfer stage from abstract (i.e. disambiguated) representations of SL texts to equivalent TL representations; in this case, translation comprised three stages: analysis, transfer, and generation (or synthesis). (For a general historical survey of MT see Hutchins 1986.)

At the University of Washington Erwin Reifler led a team on German–English and Russian–English translation, which later led to the IBM system developed by Gilbert King on a special memory device (the 'photoscopic disk') developed for the US Air Force and in operation from 1958. The largest MT group in the US was at Georgetown University, which did not continue with the method used in the 1954 experiment but based its system on rules derived from traditional grammars. There were three levels of analysis: morphological

(including identification of idioms), syntagmatic (agreement of nouns and adjectives, government of verbs, modification of adjectives, etc.), and syntactic (subjects and predicates, clause relationships, etc.). Much of the linguistic research for the Russian–English system was undertaken by Michael Zarechnak; the program was based on work by Petr Toma (later designer of Systran) and by Antony Brown (his SLC program for French–English). In this form it was successfully demonstrated in 1961 and 1962, and as a result Russian–English systems were installed at Euratom in Ispra (Italy) in 1963 and at the Oak Ridge National Laboratory of the US Atomic Energy Commission in 1964.

Anthony Oettinger at Harvard University adopted a gradualist approach. From 1954 to 1960 his group concentrated on the compilation of a massive Russian–English dictionary, to serve as an aid for translators (a forerunner of the now common computer-based dictionary aids), to produce crude word-for-word translations for scientists familiar with the subject, and as the basis for more advanced experimental work. From 1959 research turned to a method of syntactic analysis originally developed at the National Bureau of Standards under Ida Rhodes. This 'predictive syntactic analyzer' sought to identify permissible sequences of grammatical categories (nouns, verbs, adjectives, etc.) and to predict the probabilities of the following categories. Multiple parsings were generated to examine all possible predictions, but the results were often unsatisfactory, and by 1965 the Harvard group had effectively ceased MT research.

Research at the Massachusetts Institute of Technology, started by Bar-Hillel in 1951, was directed by Victor Yngve from 1953 until its end in 1965. Whereas other groups saw syntax as an adjunct to lexicographic transfer, as a means of resolving ambiguities and rearranging TL output, Yngve placed syntax at the centre: translation was a three-stage process: a SL grammar analysed input sentences as phrase structure representations, a 'structure transfer routine' converted them into equivalent TL phrase structures, and the TL grammar rules produced output text. An important contribution of MIT was the development of the first string-handling programming language (COMIT). Eventually the limitations of the 'syntactic transfer' approach became obvious, and in 1964 Yngve acknowledged that MT research had come up against 'the semantic barrier ... and that we will only have adequate mechanical translations when the machine can "understand" what it is translating' (Yngve 1964: 279).

There were other US groups at the University of Texas led by Winfried Lehmann, and at the University of California led by Sydney Lamb (who developed his 'stratificational' model of language), both linguistics-based models. There were, however, no American groups taking the interlingua approach. This was the focus of projects elsewhere. At the Cambridge Language Research Unit, Margaret Masterman and her colleagues adopted two basic lines of research: the development of a prototype interlingua producing crude 'pidgin' (essentially word-for-word) translations, and the development of tools for improving and refining MT output, primarily by means of the rich semantic networks of a thesaurus (conceived as lattices of interlocking meanings). At Milan, Silvio Ceccato concentrated on the development of an interlingua based on conceptual analysis of words (species, genus, activity type, physical properties, etc.) and their possible correlations with other words in texts.

In the Soviet Union research was as vigorous as in the United States and showed a similar mix of empirical and basic theoretical approaches. At the Institute of Precision Mechanics the research under D.Y. Panov on English–Russian translation was on lines similar to that at Georgetown, but with less practical success – primarily from lack of adequate computer facilities. More basic research was undertaken at the Steklov Mathematical Institute by Aleksej A. Ljapunov, Olga S. Kulagina and Igor A. Mel'čuk (of the Institute of Linguistics) – the latter working on an interlingua approach that led eventually to his 'meaning-text' model. This combined a stratificational dependency approach (six strata: phonetic, phonemic, morphemic,

surface syntactic, deep syntactic, semantic) with a strong emphasis on lexicographic aspects of an interlingua. Fifty universal 'lexical functions' were identified at the deep syntactic stratum covering paradigmatic relations (e.g. synonyms, antonyms, verbs and their corresponding agentive nouns, etc.) and a great variety of syntagmatic relations (e.g. inceptive verbs associated with given nouns, *conference: open, war: break out*; idiomatic causatives, *compile: dictionary, lay: foundations*, etc.). Interlingua investigations were consonant with the multilingual needs of the Soviet Union and were undertaken at a number of other centres. The principal one was at Leningrad State University, where a team under Nikolaj Andreev conceived an interlingua not as an abstract intermediary representation but as an artificial language complete in itself with its own morphology and syntax, and having only those features statistically most common to a large number of languages.

By the mid-1960s MT research groups had been established in many countries throughout the world, including most European countries (Hungary, Czechoslovakia, Bulgaria, Belgium, Germany, France, etc.), China, Mexico, and Japan. Many of these were short-lived; an exception was the project which begun in 1960 at Grenoble University (see next section).

In the 1950s optimism had been high; developments in computing and in formal linguistics, particularly in the area of syntax, seemed to promise great improvement in quality. There were many predictions of imminent breakthroughs and of fully automatic systems operating within a few years. However, disillusion grew as the complexity of the linguistic problems became more and more apparent. In a review of MT progress, Bar-Hillel (1960: 91–163) criticized the prevailing assumption that the goal of MT research should be the creation of fully automatic high quality translation (FAHQT) systems producing results indistinguishable from those of human translators. He argued that it was not merely unrealistic, given the current state of linguistic knowledge and computer systems, but impossible in principle. He demonstrated his argument with the word *pen*. It can have at least two meanings (a container for animals or children, and a writing implement). In the sentence *The box was in the pen* we know that only the first meaning is plausible; the second meaning is excluded by our knowledge of the normal sizes of (writing) pens and boxes. Bar-Hillel contended that no computer program could conceivably deal with such 'real world' knowledge without recourse to a vast encyclopedic store.

By 1964, the US government sponsors had become increasingly concerned at the lack of progress; they set up the Automatic Language Processing Advisory Committee (ALPAC), which concluded in its report (ALPAC 1966) that MT was slower, less accurate and twice as expensive as human translation and that 'there is no immediate or predictable prospect of useful machine translation'. It saw no need in the United States for further investment in MT research; instead it recommended the development of machine aids for translators, such as automatic dictionaries, and continued support in basic research in computational linguistics.

The ALPAC report brought a virtual end to MT research in the United States for over a decade and it had great impact elsewhere in the Soviet Union and in Europe. However, MT research did continue in Canada, in France and in Germany. Within a few years Petr Toma, one of the members of the Georgetown University project, had developed Systran for operational use by the USAF (1970) and by NASA (in 1974/5), and shortly afterwards Systran was installed by the Commission of the European Communities for translating from English into French (1976) and later between other Community languages.

Throughout this period, research on MT became an 'umbrella' for much contemporary work in structural and formal linguistics (particularly in the Soviet Union), semiotics, logical semantics, mathematical linguistics, quantitative linguistics, and nearly all of what would now be called computational linguistics and language engineering (terms already in use since the early 1960s). Initially, there were also close ties with cybernetics and information theory. In

general, throughout the early period, work on MT (both theoretical and practical) was seen to be of wide relevance in many fields concerned with the application of computers to 'intellectual' tasks; this was true in particular for the research on 'interlingua' aspects of MT, regarded as significant for the development of 'information languages' to be used in document retrieval systems.

From 1970 to 1989

Research did not stop completely, however, after ALPAC. Even in the United States groups continued for a few more years, at the University of Texas and at Wayne State University. But there was a change of direction. Where 'first generation' research of the pre-ALPAC period (1956–1966) had been dominated by mainly 'direct translation' approaches, the 'second generation' post-ALPAC was to be dominated by 'indirect' models, both interlingua and transfer based.

In the 1960s in the US and the Soviet Union MT activity had concentrated on Russian–English and English–Russian translation of scientific and technical documents for a relatively small number of potential users, most of whom were prepared to overlook mistakes of terminology, grammar and style in order to be able to read something which they would have otherwise not known about. Since the mid-1970s the demand for MT has come from quite different sources with different needs and different languages. The administrative and commercial demands of multilingual communities and multinational trade have stimulated the demand for translation in Europe, Canada and Japan beyond the capacity of the traditional translation services. The demand is now for cost-effective machine-aided translation systems which can deal with commercial and technical documentation in the principal languages of international commerce.

At Montreal, research began in 1970 on a syntactic transfer system for English–French translation. The TAUM project (Traduction Automatique de l'Université de Montréal) had two major achievements: first, the Q-system formalism for manipulating linguistic strings and trees (later developed as the Prolog programming language), and secondly, the Météo system for translating weather forecasts. Designed specifically for the restricted vocabulary and limited syntax of meteorological reports, Météo has been successfully operating since 1976 (since 1984 in a new version). The TAUM group attempted to repeat this success in another field, that of aviation manuals, but failed to overcome the problems of complex noun compounds and phrases, and the project ended in 1981.

A similar fate met the ITS system at Brigham Young University. This was a transfer-based interactive multilingual system based on Eldon G. Lytle's junction grammar. The aim was a commercial system but an internal evaluation in 1979 – a decade after the project had begun – concluded that the system had become too complex, and recommended the development of practical computer aids for translators (cf. ALPS, below).

Throughout the 1980s research on more advanced methods and techniques continued. For most of the decade, the dominant strategy was that of 'indirect' translation via intermediary representations, sometimes interlingual in nature, involving semantic as well as morphological and syntactic analysis and sometimes non-linguistic 'knowledge bases'. There was an increasing emphasis on devising systems for particular subject domains and for particular specific purposes, for monolingual users as well as bilingual users (translators), and for interactive operation rather than batch processing.

The most notable research projects were the GETA-Ariane system at Grenoble, SUSY and ASCOF at Saarbrücken, Mu at Kyoto, DLT at Utrecht, Rosetta at Eindhoven, the

knowledge-based MT project at Carnegie-Mellon University (Pittsburgh), and two ambitious international multilingual projects: Eurotra, supported by the European Communities, involving teams in each member country; and the Japanese CICC project with participants in China, Indonesia and Thailand.

Between 1960 and 1971 the group established by Bernard Vauquois at Grenoble University developed an interlingua system for translating Russian mathematics and physics texts into French. The 'pivot language' of CETA (Centre d'Etudes pour la Traduction Automatique) was a formalism for representing the logical properties of syntactic relationships. It was not a pure interlingua as it did not provide interlingual expressions for lexical items; these were translated by a bilingual transfer mechanism. Syntactic analysis produced first a phrase-structure (context-free) representation, then added dependency relations, and finally a 'pivot language' representation in terms of predicates and arguments. After substitution of TL lexemes (French), the 'pivot language' tree was converted first into a dependency representation and then into a phrase structure for generating French sentences. A similar model was adopted by the group at Texas during the 1970s in its METAL system: sentences were analysed into 'normal forms', semantic propositional dependency structures with no interlingual lexical elements.

By the mid-1970s, the future of the interlingua approach was in doubt. The main problems identified were attributed to the rigidity of the levels of analysis (failure at any one stage meant failure to produce any output at all), the inefficiency of parsers (too many partial analyses which had to be 'filtered' out), and in particular loss of information about surface forms of the SL input which might have been used to guide the selection of TL forms and the construction of acceptable TL sentence structures.

After the disappointing results of its interlingua system, the Grenoble group (GETA, Groupe d'Etudes pour la Traduction Automatique) began development of its influential Ariane system. Regarded as the paradigm of the 'second generation' linguistics-based transfer systems, Ariane influenced projects throughout the world in the 1980s. Of particular note were its flexibility and modularity, its algorithms for manipulating tree representations, and its conception of static and dynamic grammars. However, like many experimental MT systems, Ariane did not become an operational system, and active research on the system ceased in the late 1980s.

Similar in conception to the GETA-Ariane design was the Mu system developed at the University of Kyoto under Makoto Nagao. Prominent features of Mu were the use of case grammar analysis and dependency tree representations, and the development of a programming environment for grammar writing (GRADE). Another experimental system was developed at Saarbrücken (Germany), a multilingual transfer system SUSY (Saarbrücker Übersetzungssystem), displaying a heterogeneity of techniques: phrase structure rules, transformational rules, case grammar and valency frames, dependency grammar, the use of statistical data, etc.

The best known project of the 1980s was the Eurotra project of the European Communities. Its aim was the construction of an advanced multilingual transfer system for translation among all the Community languages – on the assumption that the 'direct translation' approach of the Communities' Systran system was inherently limited. Like GETA-Ariane and SUSY the design combined lexical, logico-syntactic and semantic information in multilevel interfaces at a high degree of abstractness. No direct use of extra-linguistic knowledge bases or of inference mechanisms was made, and no facilities for human assistance or intervention during translation processes were to be incorporated. A major defect was the failure to tackle problems of the lexicon, both theoretically and practically; by the end of the 1980s no operational system was in prospect and the project ended.

During the latter half of the 1980s there was a general revival of interest in interlingua systems, motivated in part by contemporary research in artificial intelligence and cognitive

linguistics. The DLT (Distributed Language Translation) system at the BSO software company in Utrecht (the Netherlands), under the direction of Toon Witkam, was intended as a multilingual interactive system operating over computer networks, where each terminal was to be a translating machine from and into one language only. Texts were to be transmitted between terminals in an intermediary language, a modified form of Esperanto. A second interlingua project in the Netherlands was the Rosetta project at Philips (Eindhoven) directed by Jan Landsbergen. The aim was to explore the use of Montague grammar in interlingual representations, and as a secondary goal, the exploration of the reversibility of grammars, i.e. grammatical rules and transformations that could work in both directions between languages.

In the latter half of the 1980s Japan witnessed a substantial increase in MT research activity. Most of the computer companies (Fujitsu, Toshiba, Hitachi, etc.) began to invest large sums into an area which government and industry saw as fundamental to the coming 'fifth generation' of the information society. The research, initially greatly influenced by the Mu project at Kyoto University, showed a wide variety of approaches. While transfer systems predominated there were also interlingua systems, e.g. the PIVOT system at NEC and the Japanese funded multilingual multinational project, from the mid-1980s to the mid-1990s, already mentioned above.

As in the previous decade, many research projects were established in the 1980s outside North America, Western Europe, and Japan – in Korea (sometimes in collaborative projects with Japanese and American groups), in Taiwan (e.g. the ArchTran system), in mainland China at a number of institutions, and in Southeast Asia, particularly in Malaysia.

There was also an increase in activity in the Soviet Union. From 1976 most research was concentrated at the All-Union Centre for Translation in Moscow. Systems for English–Russian (AMPAR) and German–Russian translation (NERPA) were developed based on the direct approach, but there was also work under the direction of Yurij Apres'jan based on Mel'čuk's 'meaning-text' model – Mel'čuk himself had been obliged to leave the Soviet Union in 1977. This led to the advanced transfer systems FRAP (for French–Russian), and ETAP (for English–Russian). Apart from this group, however, most activity in the Soviet Union focused on the production of relatively low-level operational systems, often involving the use of statistical analyses – where the influence of the 'Speech Statistics' group under Raimund Piotrowski (Leningrad State University) has been particularly significant for the development of many later commercial MT systems in Russia.

During the 1980s, many researchers believed that the most likely means for improving MT quality would come from natural language processing research within the context of artificial intelligence (AI). Investigations of AI methods in MT began in the mid-1970s with Yorick Wilks' work on 'preference semantics' and 'semantic templates'. A number of projects applied knowledge-based approaches – some in Japan (e.g. the LUTE project at NTT, and the ETL research for the Japanese multilingual project), others in Europe (e.g. at Saarbrücken and Stuttgart), and many in North America. The most important group was at Carnegie-Mellon University in Pittsburgh under Jaime Carbonell and Sergei Nirenburg, which experimented with a number of knowledge-based MT systems (Goodman and Nirenburg 1991).

The 1980s witnessed the emergence of a variety of operational MT systems. First there were a number of mainframe systems. Best known is Systran, operating in many pairs of languages; others were: Logos for German–English translation and for English–French in Canada; the internally developed systems for Spanish–English and English–Spanish translation at the Pan American Health Organization; systems developed by the Smart Corporation for large organizations in North America; the Metal system from Siemens for German–English translation; and major systems for English–Japanese and Japanese–English translation came from Japanese computer companies, Fujitsu, Hitachi and Toshiba.

The wide availability of microcomputers and of text-processing software led to a commercial market for cheaper MT systems, exploited in North America and Europe by companies such as ALPS, Weidner, Linguistic Products, Tovna and Globalink, and by many Japanese companies, e.g. Sharp, NEC, Oki, Mitsubishi, Sanyo. Other microcomputer-based systems came from China, Taiwan, Korea, Bolivia, Eastern and Central Europe, e.g. PROMT from Russia.

Finally, not least, there was the beginning of systems offering some kind of translation for spoken language. These were the phrase-book and PC-based systems which included the option of voice output from written text – it seems that Globalink in 1995 was the earliest. But automatic speech synthesis of text-to-text translation is not at all the same as genuine 'speech-to-speech translation'. Research on speech translation did not start until the late 1980s (see below).

Applications of MT up to 2000: translation tools

Until the middle of the 1990s there were just two basic ways in which machine translation systems were used. The first was the traditional large-scale system mounted on mainframe computers in large companies. The purpose was to use MT in order to produce publishable translations. The output of MT systems was thus revised (post-edited) by human translators or editors familiar with both source and target languages. Revision for MT differs from the revision of traditionally produced translations; the computer program is regular and consistent with terminology, unlike the human translator, but typically it contains grammatical and stylistic errors which no human translator would commit. Hence, there was opposition from translators (particularly those with the task of post-editing) but the advantages of fast and consistent output have made large-scale MT cost-effective. In order to improve the quality of the raw MT output many large companies included methods of 'controlling' the input language (by restricting vocabulary and syntactic structures) in order to minimize problems of disambiguation and alternative interpretations of structure and thus improve the quality. Companies such as the Xerox Corporation used the Systran systems with a 'controlled language' from the late 1970s (Elliston 1978: 149–158) for the translation of English language documents into Scandinavian languages. Many companies followed their example, and the Smart Corporation specializes to this day in setting up 'controlled language' MT systems for large companies in North America. In a few cases, it was possible to develop systems specifically for the particular 'sublanguage' of the texts to be translated, as in the Météo system mentioned above. Indeed, nearly all systems operating in large organizations are in some way 'adapted' to the subject areas they operate in: earth moving machines, job applications, health reports, patents, police data, and many more.

Personal Computers became widely marketed since the early 1980s and software for translation became available soon afterwards: ALPS (later Alpnet) in 1983, Weidner in 1984 (later acquired by the Japanese company Bravis). They were followed from the mid-1980s onwards by many companies marketing PCs – including most of the Japanese manufacturers of PCs – and covering an increasingly wider range of language pairs and on an increasingly wide range of operating systems. Since the mid-1990s a huge range of translation software has been available (Hutchins 2003: 161–174).

What has always been uncertain is how purchasers have been using these PC systems. In the case of large-scale (mainframe) 'enterprise' systems it is clear that MT is used to produce drafts which are then edited by bilingual personnel. This may also be the case for PC systems, i.e. it may be that they have been and are used to create 'drafts' which are edited to a higher quality.

On the other hand, it seems more likely that users want just to get some idea of the contents (the basic 'message') of foreign texts and are not concerned about the quality of translations. This usage is generally referred to as 'assimilation' (in contrast to the use for publishable translations: 'dissemination'). We know (anecdotally) that some users of PC systems have trusted them too much and have sent 'raw' (unedited) MT translations as if they were as good as human translations.

The same comments apply to the marketing since the early 1990s of hand-held translation devices or 'pocket translators'. Many, such as the Ectaco range of special devices, are in effect computerized versions of the familiar phrase-book or pocket dictionary, and are clearly marketed primarily to the tourist and business traveller. The small dictionary sizes are obviously limited. Although sold in large numbers, there is no indication of how successful in actual use they may be. Recently, since the end of the 1990s they have been largely replaced by online MT services (see below).

Mainframe, client-server and PC systems are overwhelmingly 'general purpose' systems, i.e. they are built to deal with texts in any subject domain. Of course, 'enterprise' systems (particularly controlled language systems) are over time focused on particular subject areas, and adaptation to new areas is offered by most large MT systems (such as Systran). A few PC-based systems are available for texts in specific subject areas, e.g. medical texts and patents (the English/Japanese Transer systems). On the whole, however, PC systems deal with specific subjects by the provision of subject glossaries. For some systems the range of dictionaries is very wide, embracing most engineering topics, computer science, business and marketing, law, sports, cookery, music, etc.

Few translators have been happy with fully automatic translation. In particular they do not want to be post-editors of poor quality output. They prefer dedicated computer-based aids, in particular since the early 1990s the availability of 'translation memories'. An early advocate of translation aids was Martin Kay (1980), who criticized the current approaches to MT as technology-driven rather than user-driven. He argued that the real need was assistance in translation tasks. These aids include facilities for multilingual word-processing, for creating in-house glossaries and termbanks, for receiving and sending texts over telecommunication networks, for accessing remote sources of information, for publishing quality documents, and for using interactive or batch MT systems when appropriate. Above all, translators need access to previous translations in 'translation memories', i.e. bilingual corpora of aligned sentences and text segments. Translators can find examples of existing translations of text which match or are similar to those in hand. Not only is consistency improved and quality maintained, but sections of repetitive texts are not translated again unnecessarily. Ideas for translation memories date back to proposals by Arthern (1979) and Kay (1980), but it was not until the early 1990s that they came onto the market with systems from Trados, SDL, Atril, Champollion, etc. Systems which integrate a variety of aids are known as translators' workstations or workbenches and have been commercially available from a number of vendors (Trados, STAR, IBM). (For a historical survey see Hutchins 1998: 287–307.)

A special application of MT since the early 1990s has been the localization of software products. (For a survey see Esselink 2003: 67–86). Software producers seek to market versions of their systems in other languages, simultaneously or very closely following the launch of the version in the original language (usually English), and so localization has become a necessity in the global markets of today. The repetitive nature of the documentation (e.g. software manuals), changing little from one product to another and from one edition to the next, made the use of translation memories and the development of 'controlled' terminologies for MT systems particularly attractive. But, localization involves more than just translation of texts. It

means the adaptation of products (and their documentation) to particular cultural conditions, ranging from the correct expression of dates (day-month-year vs. month-day-year), times (12-hour vs. 24-hour), address conventions and abbreviations, to the reformatting (re-paragraphing) and re-arranging of complete texts to suit expectations of recipients.

Corpus-based MT research – 1989 to the present

The dominant framework of MT research until the end of the 1980s was based on essentially linguistic rules of various kinds: rules for syntactic analysis, lexical rules, rules for lexical transfer, rules for syntactic generation, rules for morphology, etc. The rule-based approach was most obvious in the dominant transfer systems of the 1980s (Ariane, Metal, SUSY, Mu and Eurotra), but it was also the basis of all the various interlingua systems – both those which were essentially linguistics-oriented (DLT and Rosetta), and those which were knowledge-based (KANT). Rule-based methods continued into the 1990s: the CAT2 system (a by-product of Eurotra) at Saarbrücken, the Catalyst project at Carnegie-Mellon University (a domain-specific knowledge-based system) for the Caterpillar company, a project at the University of Maryland based on the linguistic theory of 'principles and parameters', and Pangloss, an ARPA-funded research project at Carnegie-Mellon, Southern California, and New Mexico State University.

Since 1989, however, the dominance of the rule-based approach has been broken by the emergence of new methods and strategies which are now loosely called 'corpus-based' methods. The most dramatic development was the revival of a purely statistics-based approach to MT in the Candide project at IBM, first reported in 1988 (Brown *et al.* 1988, 1990: 79–85), and developed to its definitive form in 1993 (Brown *et al.* 1993: 263–311). Statistical methods were common in the earliest period of MT research (such as the distributional analysis of texts at the RAND Corporation), but the results had been generally disappointing. With the success of newer stochastic techniques in speech recognition, the IBM team at Yorktown Heights began to look again at their application to MT. The distinctive feature of Candide was that statistical methods were used as the sole means of analysis and generation; no linguistic rules were applied. The researchers at IBM acknowledged that their approach was in effect a return to the statistical approach suggested by Warren Weaver (1949). The system was tested on the large corpus of French and English texts contained in the reports of Canadian parliamentary debates (the Canadian Hansard). What surprised most researchers (particularly those involved in rule-based approaches) was that the results were so acceptable: almost half the phrases translated either matched exactly the translations in the corpus, or expressed the same sense in slightly different words, or offered other equally legitimate translations.

Stages of translation in statistical machine translation (SMT) systems are: first, alignment of bilingual corpora (i.e. texts in original language and texts in target language, or texts in comparable corpora which are not directly alignable), either by word or phrase; then, frequency matching of input words against words in the corpus, extraction of most probable equivalents in the target language ('decoding'); reordering of the output according to most common word sequences using a 'language model', a monolingual corpus providing word frequencies of the TL; and finally production of the output in the target language. In broad terms the process was in effect a revival of the 'direct translation' approach of some MT pioneers (see the quote from Weaver above), but refined of course by sophisticated statistical techniques.

Since this time, statistical machine translation (SMT) has become the major focus of most MT research groups, based primarily on the IBM model, but with many subsequent refinements (Ney 2005). The original emphasis on word correlations between source and target languages has been replaced by correlations between 'phrases' (i.e. sequences of words, not necessarily

'traditional' noun phrases, verb phrases or prepositional phrases), by the inclusion of morphological and syntactic information, and by the use of dictionary and thesaurus resources. Subsequent refinements have been the inclusion of structural information (usually dependency relations) in hierarchical trees similar to some earlier rule-based systems. For transfer from source to target, SMT systems incorporate string-to-string (or phrase-to-string) transfer relations based on the bilingual corpora, and the output is revised (corrected) via frequency information from monolingual corpora ('language models'). The SMT approach has been applied to an ever widening range of language pairs. The main centres for SMT research are the universities of Aachen, Edinburgh, and Southern California, and they have been recently joined by the Google Corporation. There are a number of ambitious SMT projects. Within Europe and funded by the European Union is the Euromatrix project involving many European researchers in an 'open' network under the general leadership of the Edinburgh centre. The project began in 2006 (Koehn 2007) with the aim of developing SMT systems between all the languages of the European Union. Some language pairs already exist, many in different versions, particularly between languages such as English, French, German, Spanish. A major effort of the project has been the development of SMT for 'minor' languages not previously found in MT systems, such as Estonian, Latvian, Slovenian, Macedonian, etc. The project does not exclude rule-based methods when appropriate (i.e. as hybrid systems – see below); and given the complexity of translation and the range of types of languages it is presumed that multiple approaches will be essential. (An insightful summary of achievements in SMT systems for translation of European languages is found in Koehn *et al.* 2009: 65–72.) Apart from the Euromatrix project, groups active in Europe include researchers at many German and Spanish universities, researchers at the Charles University Prague, who have made fundamental contributions to the SMT of morphologically rich languages (Czech and others), and researchers in the Baltic countries.

The second major 'corpus-based' approach – benefiting likewise from improved rapid access to large databanks of text corpora – was what is known as the 'example-based' (or 'memory-based') approach (Carl and Way 2003). Although first proposed in 1981 by Makoto Nagao (1984: 173–180), it was only towards the end of the 1980s that experiments began, initially in some Japanese groups and during the DLT project mentioned above. The underlying hypothesis of example-based machine translation (EBMT) is that translation by humans often involves the finding or recalling of analogous examples, i.e. how a particular expression or some similar phrase has been translated before. The EBMT approach is founded on processes of extracting and selecting equivalent phrases or word groups from a databank of parallel bilingual texts, which have been aligned either by statistical methods (similar perhaps to those used in SMT) or by more traditional rule-based methods. For calculating matches, some research groups use semantic methods, e.g. a semantic network or a hierarchy (thesaurus) of domain terms; other groups use statistical information about lexical frequencies in the target language. A major problem is the re-combination of selected target language examples (generally short phrases) in order to produce fluent and grammatical output. Nevertheless, the main advantage of the approach (in comparison with rule-based approaches) is that since the texts have been extracted from databanks of actual translations produced by professional translators there is an assurance that the results should be idiomatic. Unlike SMT, there is little agreement on what might be a 'typical' EBMT model (cf. Turcato and Popowich 2003: 59–81), and most research is devoted to example-based methods which might be applicable to any MT system (rule-based or statistical).

Although SMT is now the dominant framework for MT research, it is recognized that the two corpus-based approaches are converging in many respects: SMT systems are making more

use of phrase-based alignments and of linguistic data, and EBMT systems are making wider use of statistical analysis techniques.

Increasingly, resources for MT (components, algorithms, corpora, etc.) are widely available as 'open source' materials. For SMT well known examples are: *GIZA++* for alignment, and the *Moses* basic translation engine. For rule-based MT there is the *Apertium* system from Spain which has been the basis of MT systems for Spanish, Portuguese, Galician, Catalan, Welsh, Swedish, Danish, Slovenian, etc.

Many researchers believe that the future for MT lies in the development of hybrid systems combining the best of the statistical and rule-based approaches. In the meantime, however, until a viable framework for hybrid MT appears, experiments are being made with multi-engine systems and with adopting statistical techniques with rule-based (and example-based) systems. The multi-engine approach involves the translation of a given text by two or more different MT architectures (SMT and RBMT, for example) and the integration or combination of outputs for the selection of the 'best' output – for which statistical techniques can be used (in what are called 'combination systems'). An example of appending statistical techniques to rule-based MT is 'statistical post-editing'. i.e. the submission of the output of an RBMT system to a 'language model' of the kind found in SMT systems.

Evaluation

Evaluations of MT systems date back to the earliest years of research: Miller and Beebe-Center (1956: 73–80) were the first; Henisz-Dostert evaluated the Georgetown Russian–English system (Henisz-Dostert 1967: 57–91) and John Carroll (1966: 55–66) did the study that influenced the negative conclusions of ALPAC – all were based on human judgments of comprehensibility, fluency, fidelity, etc. and all were evaluations of Russian–English systems. In the years from 1970 to 1990 the European Commission undertook in-depth evaluations of the Systran English–French and English–Italian systems before they were adopted (van Slype 1979: 59–81). In the 1990s there were numerous workshops dedicated specifically to the problems of evaluating MT, e.g. Falkedal 1991, Vasconcellos 1992, and the workshops attached to many MT conferences. The methodologies developed by Japan Electronic Industry Development Association (Nomura and Isahara 1992: 11–12) and those designed for the evaluation of ARPA (later DARPA) supported projects were particularly influential (ARPA 1994), and MT evaluation proved to have significant implications for evaluation in other areas of computational linguistics and other applications of natural language processing. Initially, most measures of MT quality were performed by human assessments of such factors as comprehensibility, intelligibility, fluency, accuracy and appropriateness – for such evaluation methods the research group at ISSCO has been particularly important – e.g. King *et al.* (2003: 224–231). However, human evaluation is expensive in time and effort and so efforts have been made, particularly since 2000, to develop automatic (or semi-automatic) methods.

One important consequence of the development of the statistics-based MT models (SMT, above) has in fact been the application of statistical analysis to the automatic evaluation of MT systems. The first metric was BLEU from the IBM group, followed later by the NIST (National Institute for Standards and Techniques); for BLEU see Papineni *et al.* (2002: 311–318); for NIST see Doddington (2002: 138–145). Both have been applied by (D)ARPA in its evaluations of MT projects supported by US research funds.

BLEU and NIST (and other subsequently developed metrics such as METEOR) are based on the availability of human produced translations (called 'reference texts'). The output from an MT system is compared with one or more 'reference texts'; MT texts which are identical

or very close to the 'reference' in terms of word sequences score highly, MT texts which differ greatly either in individual word occurrences or in word sequences score poorly. The metrics tend to rank rule-based systems lower than SMT systems even though the former are often more acceptable to human readers. Nevertheless, current automatic evaluation is undeniably valuable for monitoring whether a particular system (SMT or EBMT) has or has not improved over time. Many researchers are currently seeking metrics which produce results more closely matching human judgments; or indeed, metrics based directly on collaborative human evaluations from 'crowd sourcing' (e.g. using *Mechanical Turk*, as in Callison-Burch 2009: 286–295).

A consequence of the change from rule-based approaches to statistics-based methods has been that MT researchers do not any longer need to have considerable knowledge of the source and target languages of their systems; they can rely upon metrics based on human produced 'reference texts' to suggest improvements; furthermore, the use of statistics-based methods means that researchers can produce systems much more quickly than with the previous laborious rule-based methods.

Speech translation since 1990

Reports of the speech translation research in Japan appeared from 1988 onwards (e.g., the research at ATR, by Tomita *et al.* 1988: 57–77). Reports of the JANUS system at Carnegie-Mellon came in 1993 (Woszczyna *et al.* 1993: 195–200) and in the same year news of the Verbmobil project based in Germany (Wahlster 1993: 127–135) and of the SLT project in the SRI group in Cambridge (Rayner *et al.* 1993: 217–222). The NESPOLE research project came in 2001 (Lavie *et al.* 2001).

The research in speech translation is faced with numerous problems, not just variability of voice input but also the nature of spoken language. By contrast with written language, spoken language is colloquial, elliptical, context-dependent, interpersonal, and frequently in the form of dialogues. MT has focused primarily on well-formed, technical and scientific language and has tended to neglect informal modes of communication. Speech translation therefore represents a radical departure from traditional MT. Some of the problems of spoken language translation may be reduced by restricting communication to relatively narrow domains. Business communication was the focus of the government funded research at a number of German universities (the Verbmobil project), where the aim was the development of a system for three-way negotiation between English, German and Japanese (Wahlster 2000). The focus of the ATR research in Japan has been telephone communication between English and Japanese primarily in the area of booking hotel accommodation and registration for conferences. The potentialities of speech translation in the area of health-communication are obvious. Communication may be from doctor to patient or interactive, or may be via a screen displaying possible 'health' conditions. Examples are the MedSLT project from SRI where voice input locates potential phrases and the translation is output by speech synthesis (Rayner and Bouillon 2002: 69–76), and the interactive multimodal assistance provided by the Converser system (Seligman and Dillinger 2006). A somewhat similar 'phrasebook' approach is found in the DIPLOMAT system from Carnegie-Mellon (Frederking *et al.* 1997: 261–262). The system was developed for the US Army for communication from English to Serbo-Croat, Haitian Creole and Korean: spoken input is matched against fixed phrases in the database and translations of the phrases are output by a speech synthesizer. Nearly all the systems were somewhat inflexible and limited in range – the weakest point continues to be speech recognition.

One of the most obvious applications of speech translation is the assistance of tourists in foreign countries. In most cases, translation is restricted to 'standard' phrases extracted from corpora of dialogues and interactions in tourist situations, although, in recent years, researchers have turned to systems capable of dealing with 'spontaneous speech'. Despite the amount of research in an apparently highly restricted domain it is clear that commercially viable products are still some way in the future.

Usage and applications since 1990

Since the early 1990 the use of unrevised MT output has grown greatly, such that now it may well be true that 'raw' unedited MT is the principal form in which people encounter translation from any source.

For the general public, the main source of translation since the mid-1990s has been the availability of free MT services on the Internet (Gaspari and Hutchins 2007: 199–206). Initially, online MT services in the early 1990s were not free. In 1988 Systran in France offered a subscription to its translation software using the French postal services Minitel network. At about the same time, Fujitsu made its Atlas EnglishóJapanese and Japanese–English systems available through the online service Niftyserve. Then in 1992 CompuServe launched its MT service (based on the Intergraph DP/Translator), initially restricted to selected forums, but which proved highly popular, and in 1994 Globalink offered an online subscription service – texts were submitted online and translations returned by email. A similar service was provided by Systran Express. However, it was the launch of AltaVista's Babelfish free MT service in 1997 (based on the various Systran MT systems) that attracted the greatest publicity. Not only was it free but results were (virtually) immediate. Within the next few years, the Babelfish service was joined by FreeTranslation (using the Intergraph system), Gist-in-Time, ProMT, PARS, and many others; in most cases, these were online versions of already existing PC-based (or mainframe) systems. The great attraction of these services was (and is) that they are free to users – it is evidently the expectation of the developers is that free online use will lead either to sales of PC translation software, although the evidence for this has not been shown, or to the use of fee-based 'valued-added' post-editing services offered by providers such as FreeTranslation. While online MT has undoubtedly raised the profile of MT for the general public, there have, of course, been drawbacks.

To most users the idea of automatic translation was something completely new – many users 'tested' the services by inputting sentences containing idiomatic phrases, ambiguous words and complex structures, and even proverbs and deliberately opaque sayings, and not surprisingly the results were unsatisfactory. A favourite method of 'evaluation' was back translation: i.e. translation, into another language and then back into the original language (Somers 2007: 209–233). Not surprisingly, users discovered that MT suffered from many limitations – all well-known to company users and to purchasers of PC software. Numerous commentators have enjoyed finding fault with online MT and, by implication, with MT itself. On the other hand, there is no doubt that the less knowledge users have of the language of the original texts the more value they attach to the MT output; and some users must have found that online MT enabled them to read texts that they would have previously had to pass over.

Largely unknown by the general public is the use of MT systems by the intelligence services. The languages of most interest are, for obvious reasons, Arabic, Chinese, Persian (Farsi). The older demand for translation from Russian (see above) has almost disappeared. The need is for the translation of huge volumes of text. The coming of statistical machine translation has answered this need to a great extent: SMT systems are based on large corpora, often

concentrating on specific topics (politics, economics, etc.), and the systems can be delivered quickly. As a sideline we may mention one intriguing application of SMT methods to the decipherment of ancient languages (Ravi and Knight 2011: 12–21) – reviving the cryptographic speculations of Weaver in 1949 (see the above section).

Collaboration in the acquisition of lexical resources dates from the beginning of MT research (e.g. the Harvard Russian–English dictionary was used by the MT project at the National Physical Laboratory). A notable effort in the late 1980s was the Electronic Dictionary Research project, which was supported by several Japanese computer manufacturing companies. The need grew with the coming of corpus-based systems (see above). Since the latter part of the 1990s large lexical resources have been collected and made available in the United States through the Linguistic Data Consortium and in Europe through the European Language Resources Association (ELRA), which in 1998 inaugurated its major biennial series of conferences devoted to the topic – the Language Resources and Evaluation Conferences (LREC). The Internet itself is now a source for lexical data, such as Wikipedia. One of the earliest examples of 'mining' bilingual texts from the World Wide Web was described by Resnick (1999: 527–534).

The languages most often in demand and available commercially are those from and to English. The most frequently used pairs (for online MT services and apparently for PC systems) are English/Spanish and English/Japanese. These are followed by (in no particular order) English/French, English/German, English/Italian, English/Chinese, English/Korean, and French/German. Other European languages such as Catalan, Czech, Polish, Bulgarian, Romanian, Latvian, Lithuanian, Estonian, and Finnish, were more rarely found in the commercial PC market or online until the last decade. Until the middle of the 1990s, Arabic/English, Arabic/French and Chinese/English were also rare, but this situation has now changed for obvious political reasons. Other Asian languages have been relatively neglected: Malay, Indonesian, Thai, Vietnamese and even major languages of India: Hindu, Urdu, Bengali, Punjabi, Tamil, etc. And African languages (except Arabic dialects) are virtually invisible. In terms of population these are not 'minor' languages – many are among the world's most spoken languages. The reason for neglect is a combination of low commercial viability and lack of language resources (whether for rule-based lexicons and grammars or for statistical MT corpora). There is often no word-processing software (indeed some languages lack scripts), no spellcheckers (sometime languages lack standard spelling conventions), no dictionaries (monolingual or bilingual), indeed a general lack of language resources (e.g. corpora of translations) and of qualified and experienced researchers. (For an overview see Somers 2003a: 87–103.)

Summary

Machine translation has come a long way from its tentative and speculative beginnings in the 1950s. We can see three stages of development, each spanning two decades. The first 20 years include the pioneering period (1949–1966) when numerous different approaches were investigated: dictionary-based word-for-word systems, experiments with interlinguas, syntax-based systems with multiple levels of analysis, and the first operational systems (IBM and Georgetown). The period ended with the influential ALPAC report of 1966. The next two decades (1967–1989) saw the development of linguistic rule-based systems, mainly in the framework of transfer grammars, and experiments with sophisticated interlingua and artificial intelligence systems; in the same decade there was increased application of MT for commercial users, including the use of controlled languages and sublanguages, and applications such as

localization; and there was also the first computer-based translation aids. The third period, since the early 1990s, has seen the domination of corpus-based approaches, translation memories, example-based MT, and in particular statistical MT; but there has also been much greater attention to evaluation methods; lastly, applications and usages of MT have widened markedly, most significantly by the access to and use of MT and resources over the Internet.

References

ALPAC (1966) *Languages and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC: National Academy of Sciences, National Research Council, 1966.

ARPA (1994) A*RPA Workshop on Machine Translation*, 17–18 March 1994, Sheraton Premier Hotel at Tyson's Corner, Vienna, Austria.

Arthern, Peter J. (1979) 'Machine Translation and Computerized Terminology Systems: A Translator's Viewpoint', in Barbara M. Snell (ed.) *Translating and the Computer: Proceedings of a Seminar*, London, 14 November 1978, Amsterdam: North-Holland, 77–108.

Bar-Hillel, Yehoshua (1960) 'The Present Status of Automatic Translation of Languages', *Advances in Computers* 1: 91–163.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, Robert L. Mercer, and Paul S. Roossin (1988) 'A Statistical Approach to French / English Translation', in *Proceedings of the 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages,* 12–14 June 1988, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (1990) 'A Statistical Approach to Machine Translation', *Computational Linguistics* 16(2): 79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational Linguistics* 19(2): 263–311.

Callison-Burch, Chris (2009) 'Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk', in *EMNLP-2009: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 6–7 August 2009, Singapore, 286–295.

Carl, Michael and Andy Way (eds) (2003) *Recent Advances in Example-based Machine Translation*, Dordrecht: Kluwer Academic Publishers.

Carroll, John B. (1966) 'An Experiment in Evaluating the Quality of Translations', *Mechanical Translation and Computational Linguistics* 9(3–4): 55–66.

Doddington, George (2002) 'Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics', in *HLT 2002: Human Language Technology Conference: Proceedings of the 2nd International Conference on Human Language Technology Research*, 24–27 March 2002, San Diego, CA, 138–145.

Elliston, John S.G. (1978) 'Computer-aided Translation: A Business Viewpoint', in Barbara M. Snell (ed.) *Translating and the Computer: Proceedings of a Seminar*, 14 November 1978, London, Amsterdam: North-Holland, 149–158.

Esselink, Bert (2003) 'Localisation and Translation', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 67–86.

Falkedal, Kirsten (ed.) (1991) *Proceedings of the Evaluators' Forum*, 21–24 April 1991, Les Rasses, Vaud, Switzerland.

Frederking, Robert E., Ralf D. Brown, and Christopher Hogan (1997) 'The DIPLOMAT Rapid-deployment Speech MT System', in Virginia Teller and Beth Sundeim (eds) *Proceedings of the MT Summit VI: Machine Translation: Past, Present, Future*, 29 October – 1 November 1997, San Diego, CA, 261–262.

Gaspari, Federico and W. John Hutchins (2007) 'Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects', in *Proceeding of the MT Summit XI*, 10–14 September 2007, Copenhagen, 199–206.

Goodman, Kenneth and Sergei Nirenburg (eds) (1991) *The KBMT Project: A Case Study in Knowledge-based Machine Translation*, San Mateo, CA: Morgan Kaufmann.

Goutte, Cyril, Nicola Cancedda, Marc Dymetman, and George Foster (eds) (2009) *Learning Machine Translation*, Cambridge, MA: MIT Press.

Henisz-Dostert, Bozena (1967) 'Experimental Machine Translation', in William M. Austin (ed.) *Papers in Linguistics in Honor of Léon Dostert*, The Hague: Mouton, 57–91.

Hutchins, W. John (1986) *Machine Translation: Past, Present, Future*, Chichester: Ellis Horwood and New York: Halsted Press.

Hutchins, W. John (1998) 'The Origins of the Translator's Workstation', *Machine Translation* 13(4): 287–307.

Hutchins, W. John (2003) 'Commercial Systems: The State of the Art', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 161–174.

Hutchins, W. John (2004) 'The Georgetown-IBM Experiment Demonstrated in January 1954', in Robert E. Frederking and Kathryn B. Taylor (eds) *Proceedings of Machine Translation: From Real Users to Research: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004*, 28 September – 2 October 2004, Washington, DC,/Berlin: Springer Verlag, 102–114.

Kay, Martin (1980) 'The Proper Place of Men and Machines in Language Translation', Research Report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, CA.

King, Margaret, Andrei Popescu-Belis, and Eduard Hovy (2003) 'FEMTI: Creating and Using a Framework for MT Evaluation', in *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches*, 23–27 September 2003, New Orleans, LA, USA, 224–231.

Koehn, Philipp (2007) 'EuroMatrix – Machine Translation for all European Languages', Invited Talk at *MT Summit XI*, 10–14 September 2007, Copenhagen, Denmark.

Koehn, Philipp (2009) *Statistical Machine Translation*, Cambridge: Cambridge University Press.

Koehn, Philipp, Alexandra Birch, and Ralf Steinberger (2009) '462 Machine Translation Systems for Europe', in *MT Summit XII: Proceedings of the 12th Machine Translation Summit*, 26–30 August 2009, Ottawa, Ontario, Canada, 65–72.

Lavie, Alon, Chad Langley, Alex Waibel, Fabio Pianesi, Gianni Lazzari, Paolo Coletti, Loredana Taddei, and Franco Balducci (2001) 'Architecture and Design Considerations in NESPOLE!: A Speech Translation System for E-commerce Applications', in *HLT-2001: Proceedings of the 1st International Conference on Human Language Technology Research*, 18–21 March 2001, San Diego, CA.

Miller, George A. and J.G. Beebe-Center (1956) 'Some Psychological Methods for Evaluating the Quality of Translations', *Mechanical Translation* 3(3): 73–80.

Nagao, Makoto (1984) 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle', in Alick Elithorn and Ranan Banerji (eds) *Artificial and Human Intelligence*, Amsterdam: North-Holland, 173–180.

Ney, Hermann (2005) 'One Decade of Statistical Machine Translation', in *Proceedings of the MT Summit X: The 10th Machine Translation Summit,* 12–16 September 2005, Phuket, Thailand, i-12-i-17.

Nomura, Hirosato and Hitoshi Isahara (1992) 'The JEIDA Methodology and Survey', in Muriel Vasconcellos (ed.) *MT Evaluation: Basis for Future Directions: Proceedings of a Workshop Sponsored by the National Science Foundation*, 2–3 November 1992, San Diego, CA, 11–12.

Papineni, Kishore, Salim Roukos, Todd Ward, and Zhu Wei-Jing (2002) 'BLEU: A Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002,* 7–12 July 2002, University of Pennsylvania, Philadelphia, PA, 311–318.

Ravi, Sujith and Kevin Knight (2011) 'Deciphering Foreign Language', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 19–24 June 2011, Portland, OR, 12–21.

Rayner, Manny, Hiyan Alshawi, Ivan Bretan, David Carter, Vassilios Digalakis, Björn Gambäck, Jaan Kaja, Jussi Karlgren, Bertil Lyberg, Steve Pulman, Patti Price, and Christer Samuelsson (1993) 'A Speech to Speech Translation System Built from Standard Components', in *HLT '93: Proceedings of the Workshop on Human Language Technology*, 21–24 March 1993, Plainsboro, NJ, 217–222.

Rayner, Manny and Pierrette Bouillon (2002) 'Flexible Speech to Speech Phrasebook Translator', in *Proceedings of the ACL-2002 Workshop on Speech-to-speech Translation*, 11 July 2002, Philadelphia, PA, 69–76.

Resnick, Philip (1999) 'Mining the Web for Bilingual Text', in *ACL-1999: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 20–26 June 1999, University of Maryland, College Park, MD, 527–534.

Seligman, Mark and Mike Dillinger (2006) 'Usability Issues in an Interactive Speech-to-speech Translation System for Healthcare', in *HLT-NAACL 2006: Proceedings of the Workshop on Medical Speech Translation*, 9 June 2006, New York, 1–8.

Somers, Harold L. (2003a) 'Translation Technologies and Minority Languages', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 87–103.

Somers, Harold L. (2007) 'Machine Translation and the World Wide Web', in Ahmad Kurshid, Chistopher Brewster, and Mark Stevenson (eds) *Words and Intelligence II: Essays in Honor of Yorick Wilks*, Dordrecht: Springer Verlag, 209–233.

Tomita, Masaru, Marion Kee, Hiroaki Saito, Teruko Mitamura, and Hideto Tomabechi (1988) 'Towards a Speech-to-speech Translation System', *Interface: Journal of Applied Linguistics* 3(1): 57–77.

Turcato, Davide and Fred Popowich (2003) 'What Is Example-based Machine Translation?' in Michael Carl and Andy Way (eds) *Recent Advances in Example-based Machine Translation*, Dordrecht: Kluwer Academic Publishers, 59–81.

van Slype, Georges (1979) *Critical Study of Methods for Evaluating the Quality of Machine Translation*, Final Report BR19142, Brussels: Bureau Marcel van Dijk [for] European Commission.

Vasconcellos, Muriel (ed.) (1992) 'MT Evaluation: Basis for Future Directions', in *Proceedings of a Workshop Sponsored by the National Science Foundation*, 2–3 November 1992, San Diego, CA.

Wahlster, Wolfgang (1993) 'Verbmobil: Translation of Face-to-face Dialogs', in *Proceedings of the MT Summit IV: International Cooperation for Global Communication*, 20–22 July 1993, Kobe, Japan, 127–135.

Wahlster, Wolfgang (ed.) (2000) *Verbmobil: Foundations of Speech-to-speech Translation*, Berlin: Springer Verlag.

Weaver, Warren (1949) 'Translation'. Reprinted in William N. Locke and Andrew D. Booth (eds) *Machine Translation of Languages: Fourteen Essays*, Cambridge, MA: Technology Press of the Massachusetts Institute of Technology, 15–33.

Woszczyna, Monika, Noah Coccaro, Andreas Eisele, Alon Lavie, Arthur E. McNair, Thomas Polzin, Ivica Rogina, Carolyn P. Rose, Tilo Sloboda, Masaru Tomita, Junya Tsutsumi, Naomi Aoki-Waibel, Alex H. Waibel, and Wayne Ward (1993) 'Recent Advances in JANUS: A Speech Translation System', in *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*: *MT in the Next Generation* (*TMI-93*), 14–16 July 1993, Kyoto, Japan, 195–200.

Yngve, Victor H. (1964) 'Implications of Mechanical Translation Research', in *Proceedings of the American Philosophical Society*, 108(4): 275–281.

7

EXAMPLE-BASED MACHINE TRANSLATION

Billy Wong Tak-ming

THE OPEN UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Jonathan J. Webster

CITY UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Introduction

Machine translation (MT) is the mechanization and automation of the process of translating from one natural language into another. Translation is a task which needs to tackle the 'semantic barriers' between languages using real world encyclopedic knowledge, and requires a full understanding of natural language. Accordingly different approaches have been proposed for addressing the challenges involved in automating this task. At present the major approaches include rule-based machine translation (RBMT) which heavily relies on linguistic analysis and representation at various linguistic levels, and example-based machine translation (EBMT) and statistical machine translation (SMT), both of which follow a more general corpus-based approach and make use of parallel corpora as a primary resource.

This chapter presents an overview of the EBMT technology. In brief, EBMT involves extracting knowledge from existing translations (examples) in order to facilitate translation of new utterances. A comprehensive review of EBMT can be found in Somers (2003: 3–57) and the latest developments in Way (2010: 177–208).

After reviewing the history of EBMT and the controversies it has generated over the past decades, we will examine the major issues related to examples, including example acquisition, granularity, size, representation and management. The fundamental stages of translation for an EBMT system will be discussed with attention to the various methodologies and techniques belonging to each stage. Finally the suitability of EBMT will be discussed, showing the types of translation that are deemed suitable for EBMT, and how EBMT interoperates with other MT approaches.

Origin

The idea of using existing translation data as the main resource for MT is most notably attributed to Nagao (1984: 173–180). Around the same time, there were other attempts at similarly exploiting parallel data as an aid of human translation. Kay (1976, 1997: 3–23) for example introduced the concept of translation memory (TM) which has become an important feature in many computer-aided translation (CAT) systems. TM can be understood as a

'restricted form of EBMT' (Kit *et al.* 2002: 57–78) in the sense that both involve storing and retrieving previous translation examples; nevertheless in EBMT the translation output is produced by the system while in TM this is left to human effort. Arthern (1978: 77–108) on the other hand proposes 'a programme which would enable the word processor to "remember" whether any part of a new text typed into it had already been translated, and to fetch this part together with the translation'. Similarly, Melby (1995) and Warner mention the ALPS system, one of the earliest commercial MT systems which dates back to the 1970s, and incorporated what they called a 'Repetition Processing' tool.

Conceptually, Nagao's EBMT attempts to mimic human cognitive behavior in translating as well as language learning:

> Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases … then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference.
>
> (Nagao 1984: 175)

Nagao (1992) further notes:

> Language learners do not learn much about a grammar of a language… They just learn what is given, that is, a lot of example sentences, and use them in their own sentence compositions.
>
> (ibid.: 82)

Accordingly, there are three main components of EBMT: (1) matching source fragments against the examples, (2) identifying the corresponding translation fragments, and then (3) recombining them to give the target output.

A major advantage of EBMT over RBMT is its ability to handle extra-grammatical sentences, which though linguistically correct cannot be accounted for in the grammar of the system. EBMT also avoids the intractable complexity of rule management which can make it difficult to trace the cause of failure, or to predict the domino effect of the addition or deletion of a rule. EBMT addresses such inadequacies by incorporating the 'learning' concept for handling the translation of expressions without structural correspondence in another language (Nagao 2007: 153–158), and also by extending the example base simply by adding examples to cover various kinds of language use.

Definition

EBMT offers a high flexibility in the use of examples and implementation of each of the three components (matching, alignment and recombination), leading to systems with, for instance, rule-based matching or statistical example recombination. The underlying principle for EBMT, according to Kit *at al.* (2002: 57–78), is to 'remember everything translated in the past and use everything available to facilitate the translation of the next utterance' where 'the knowledge seems to have no overt formal representation or any encoding scheme. Instead … in a way as straightforwardly as text couplings: a piece of text in one language matches a piece of text in another language.'

EBMT implies the application of examples – as the main source of system knowledge – at run-time, as opposed to a pre-trained model where bilingual data are only used for training in advance but not consulted during translation. Examples can be pre-processed and represented in the forms of string (sentence or phrase), template, tree structure or/and other annotated representations appropriate for the matching and alignment processes.

Examples

Acquisition

As it relates to the source of system knowledge, example acquisition is critical to the success of EBMT. Examples are typically acquired from translation documents, including parallel corpora and multilingual webpages, as well as from TM databases. Multilingual texts from sources such as the European and Hong Kong parliaments constitute high-quality data. The Europarl corpus (Koehn 2005: 79–86) for example covers twenty language pairings. The BLIS (The Bilingual Laws Information System of Hong Kong) corpus (Kit *et al.* 2003: 286–292, 2004: 29–51, 2005: 71–78) provides comprehensive documentation of the laws of Hong Kong in Chinese–English bilingual versions aligned at the clause level, with 10 million English words and 18 million Chinese characters. Legal texts like this kind are known to be more precise and less ambiguous than most other types of text. In the past decade the growing number of web-based documents represents another major source of parallel texts (Resnik 1998: 72–82; Ma and Liberman 1999: 13–17; Kit and Ng 2007: 526–529).

Possible sources of examples include not only such highly parallel bitexts, which though increasingly available still remain limited in volume, language and register coverage, especially for certain language pairs. Efforts have also been made to collect comparable non-parallel texts such as multilingual news feeds from news agencies. They are not exactly parallel but convey overlapping information in different languages; hence some sentences/paragraphs/texts can be regarded as meaning equivalent. Shimohata *et al.* (2003: 73–80) describe such as 'shar[ing] the main meaning with the input sentence despite lacking some unimportant information. It does not contain information additional to that in the input sentence.' In order to facilitate the development of an 'Example-based Rough Translation' system, a method is proposed to retrieve such meaning-equivalent sentences from non-parallel corpora using lexical and grammatical features such as content words, modality and tense. Munteanu and Marcu (2005: 477–504), on the other hand, proposes to accomplish the same purpose by means of machine learning strategy.

Apart from gathering available resources, new bitexts can also be 'created' by using MT system to translate monolingual texts into target languages. Gough *et al.* (2002: 74–83) reports on experiments in which they first decomposed sentences into phrases and then translated them with MT systems. The resulting parallel phrases could then be used as examples for an EBMT system. The output quality is proved better than that from translating the whole input sentence via online MT systems.

Granularity and size

In principle, an example can be as simple as a pair of translated texts in two languages, of any size at any linguistic level: word, phrase, clause, sentence and even paragraph. Thus, a bilingual dictionary can be viewed as a 'restricted example base', i.e. translation aligned at the word level. More flexibly, an example can simply be a pair of text chunks of an arbitrary length, not necessarily matching a linguistically meaningful structure or constituent.

In practice, because sentence boundaries are relatively easier to identify than those of finer linguistic constituents, the most common 'grain-size' for examples is the sentence. Example sentences, however, have to be decomposed into smaller chunks in the process of matching and recombination. There is usually a trade-off between granularity and recall of examples: the larger the example, the lower the probability to reuse. On the other hand, the smaller the example is, the greater is the probability of ambiguity. For examples at the word level, it is not surprising to find many source words with multiple possible target counterparts. An optimal balance may be better achieved at the sub-sentential level. Cranias *et al.* (1994: 100–104) state that 'the potential of EBMT lies in the exploitation of fragments of text smaller than sentences'.

Another important consideration is the size of the example base. In general the translation quality of an EBMT system improves as the example base is enlarged. However there may be a ceiling on the number of examples after which further addition of examples will not further improve the quality and may even degrade the system performance. The speed of computation also depends on the number of examples: the more examples, the longer will be the processing time required at run time.

Representation and management

A number of representation schemas are proposed for storing examples. The simplest representation is in the form of text string pairs aligned at various granularity levels without additional information. Giza++ (Och and Ney 2003: 19–51) is the most popular choice for implementing word level alignment. On the other hand Way and Gough (2003: 421–457) and Gough and Way (2004b: 95–104) discuss an approach based on bilingual phrasal pairs, i.e. 'marker lexicon'. Their approach follows the Marker Hypothesis (Green 1979: 481–496), which assumes that every natural language has its own closed set of lexemes and morphemes for marking the boundary of syntactic structure. Alternatively, Kit *et al.*'s (2003: 286–292, 2004: 29–51) lexical-based clause alignment approach achieves a high alignment accuracy via reliance on basic lexical resources.

Examples may also be annotated with various kinds of information. Similar to conventional RBMT systems, early attempts at EBMT stored examples as syntactic tree structures following constituency grammar. This offers the advantage of clear boundary definition, ensuring that example fragments are well-formed constituents. Later works such as Al-Adhaileh and Kong (1999: 244–249) and Aramaki *et al.* (2001: 27–32, 2005: 219–226) employed dependency structures linking lexical heads and their dependents in a linguistic expression. Planas and Furuse (1999: 331–339) presents a multi-level lattice representation combining typographic, orthographic, lexical, syntactic and other information. Forcada (2002) represents sub-sentential bitexts as a finite-state transducer. In their Data-Oriented Translation model, Way (2001: 66–80, 2003: 443–472) and Hearne and Way (2003: 165–172) use linked phrase-structure trees augmented with semantic information. In Microsoft's MT system reported in Richardson *et al.* (2001: 293–298) and Brockett *et al.* (2002: 1–7), a graph structure 'Logical Form' is used for describing labeled dependencies among content words, with information about word order and local morphosyntactic variation neutralized. Liu *et al.*'s (2005: 25–32) 'Tree String Correspondence' structure has only a parse tree in the source language, together with the target string and the correspondences between the leaf nodes of the source tree and the target substrings.

A unique approach to EBMT which does without a parallel corpus is reported in Markantonatou *et al.* (2005: 91–98) and Vandeghinste *et al.* (2005: 135–142). Their example base consists only of a bilingual dictionary and monolingual corpora in the target language. In

the translation process, a source text is first translated word-for-word into the target language using the dictionary. The monolingual corpora are then used to help determine a suitable translation in case of multiple possibilities, and to guide a correctly ordered recombination of target words. This approach is claimed to be suitable for language pairs without a sufficiently large parallel corpus available.

Webster *et al.* (2002: 79–91) links EBMT with Semantic Web technology, and demonstrates how a flat example base can be developed into a machine-understandable knowledge base. Examples of statutory laws of Hong Kong in Chinese–English parallel version are enriched with metadata describing their hierarchical structures and inter-relationships in Resource Description Framework (RDF) format, thus significantly improving example management and sub-sentential alignment.

In some systems, similar examples are combined and generalized as templates in order to reduce the size of the example base and improve example retrieval performance. Equivalence classes such as 'person's name', 'date', 'city's name' and linguistic information like gender and number that appear in examples with the same structure are replaced with variables. For example, the expression '*John Miller flew to Frankfurt on December 3rd*' can be represented as '<PERSON-M> *flew to* <CITY> *on* <DATE>' which can easily be matched with another sentence '*Dr Howard Johnson flew to Ithaca on 7 April 1997*' (Somers 2003: 3–57). To a certain extent such example templates can be viewed as 'a special case of translation rules' (Maruyama and Watanabe 1992: 173–184) in RBMT. In general the recall rate of example retrieval can be improved by this approach, but possibly with precision trade-off. Instances of studies of example templates include Malavazos *et al.* (2000), Brown (2000: 125–131) and McTait (2001: 22–34).

Examples need to be pre-processed before being put to use, and be properly managed. For instance, Zhang *et al.* (2001: 247–252) discuss the pre-processing tasks of English–Chinese bilingual corpora for EBMT, including Chinese word segmentation, English phrase bracketing, and term tokenization. They show that a pre-processed corpus improves the quality of language resources acquired from the corpus: the average length of Chinese and English terms was increased by around 60 percent and 10 percent respectively, and the coverage of bilingual dictionary by 30 percent.

When the size of example base is scaled up, there is the issue of example redundancy. Explained in Somers (2003: 3–57), overlapping examples (source side) may mutually reinforce each other or be in conflict, depending on the consistency of translations (target side). Whether such redundancy needs to be constrained depends on the application of examples: a prerequisite for systems relying on frequency for tasks such as similarity measurement in example matching, or a problem to be solved where this is not the case.

Stages

Matching

The first task of EBMT is to retrieve examples which closely match the source sentence. This process relies on a measure of text similarity, and is one of the most studied areas in EBMT. Text similarity measurement is a task common in various applications of natural language processing with many measures available. It is also closely related to how examples are represented and stored, and accordingly can be performed on string pairs or annotated structures. In order to better utilize available syntactic and semantic information, it may be further facilitated by language resources like thesauri and a part-of-speech tagger.

When examples are stored as string pairs at the sentence level, they may first need to be decomposed into fragments to improve example retrieval. In Gough *et al.* (2002: 74–83) and Gough and Way (2004b: 95–104), example sentences are split into phrasal lexicons with the aid of a closed set of specific words and morphemes to 'mark' the boundary of phrases. Kit *et al.* (2002: 57–78) uses a multi-gram model to select the best sentence decomposition with the highest occurring frequencies in an example base. Roh *et al.* (2003: 323–329) discusses two types of segmentation for sentences: 'chunks' that include proper nouns, time adverbs and lexically fixed expressions, and 'partitions' that are selected by syntactic clues such as punctuation, conjunctions, relatives and main verbs.

The similarity measure for example matching can be as simple as a character-based one. Two string segments are compared for the number of characters required for modification, whether in terms of addition, deletion or substitution, until the two are identical. This is known as edit-distance, which has been widely applied in other applications like spell-checking, translation memory and speech processing. It offers the advantages of simplicity and language independence, and avoids the need to pre-process the input sentence and examples. Nirenburg *et al.* (1993: 47–57) extends the basic character-based edit-distance measure to account for necessary keystrokes in editing operations (e.g. deletion = 3 strokes, substitution = 3 strokes). Somers (2003: 3–57) notes that in languages like Japanese certain characters are more discriminatory than others, thus the matching process may only focus on these key characters.

Nagao (1984: 173–180) employs word-based matching as the similarity measure. A thesaurus is used for identifying word similarity on the basis of meaning or usage. Matches are then permitted for synonyms and near-synonyms in the example sentences. An early method of this kind was reported on Sumita and Iida (1991: 185–192), where similarity between two words is measured by their distance in a hierarchically structured thesaurus. In Doi *et al.* (2005: 51–58) this method is integrated with an edit-distance measure. Highlighting an efficiency problem in example retrieval, they note that real-time processing for translation is hard to achieve, especially if an input sentence has to be matched against all examples individually using a large example base. Accordingly they propose the adoption of multiple strategies including search space division, word graphs and the A* search algorithm (Nilsson 1971) to improve retrieval efficiency. In Aramaki *et al.* (2003: 57–64), example similarity is measured based on different weights assigned to content and function words in an input string that are matched with an example, together with their shared meaning as defined in a dictionary.

The availability of annotated examples with linguistic information allows the implementation of similarity measures with multiple features. In the multi-engine Pangloss system (Nirenburg *et al.* 1994: 78–87), the matching process combines several variously weighted requirements including exact matches, number of word insertions or deletions, word-order differences, morphological variants and parts-of-speech. Chatterjee (2001) discusses the evaluation of sentence similarity at various linguistic levels, i.e. syntactic, semantic and pragmatic, all of which need to be considered in the case of dissimilar language pairs where source and target sentences with the same meaning may vary in their surface structures. A linear similarity evaluation model is then proposed which supports a combination of multiple individually weighted linguistic features.

For certain languages the word-based matching process requires pre-processing of both the input sentences and examples in advance. This may include tokenization and word segmentation for languages without clear word boundaries like Chinese and Japanese, and lemmatization for morphologically rich languages such as Arabic.

When examples are stored as structured objects, the process of example retrieval entails more complex tree-matching. Typically it may involve parsing an input sentence into the same

representation schema as examples, searching the annotated example base for best matched examples, and measuring similarity of structured representations. Liu *et al.* (2005: 25–32) presents a measure of syntactic tree similarity accounting for all the nodes and meaning of headwords in the trees. Aramaki *et al.* (2005: 219–226) proposes a tree matching model, whose parameters include the size of tree fragments, their translation probability, and context similarity of examples, which is defined as the similarity of the surrounding phrases of a translation example and an input phrase.

Recombination

After a set of translation examples are matched against an input sentence, the most difficult step in the EBMT process is to retrieve their counterpart fragments from the example base and then combine them into a proper target sentence. The problem is twofold, as described by Somers (2003: 3–57): (1) identifying which portion of an associated translation example corresponds to which portion of the source text, and (2) recombining these portions in an appropriate manner. The first is partially solved when the retrieved examples are already decomposed from sentences into finer fragments, either at the beginning when they are stored or at the matching stage. However, in case more than one example is retrieved, or multiple translations are available for a source fragment, there arises the question of how to decide which alternative is better.

Furthermore, the recombination of translation fragments is not an independent process, but closely related to the representation of examples. How examples are stored determines what information will be available for performing recombination. In addition, as the final stage of EBMT, the performance of recombination is to a large extent affected by the output quality from the previous stages. Errors occurring at the matching stage or earlier are a kind of noise which interferes with recombination. McTait (2001: 22–34) shows how tagging errors resulting from applying part-of-speech analysis to the matching of examples unexpectedly lower both the recall of example retrieval and accuracy of translation output. Further complications occur when examples retrieved do not fully cover the input sentence in question.

The most critical point in recombination is to adjust the fragment order to form a readable, at best grammatical, sentence in the target language. Since each language has its own syntax to govern how sentential structures are formed, it will not work if the translation fragments are simply sequenced in the same order as in the source sentence. However, this is the approach of some EBMT systems such as that reported in Way and Gough (2003: 421–457). In Doi and Sumita (2003: 104–110) it is claimed that such a simple approach is suitable for speech translation, since sentences in a dialog usually do not have complicated structures, and many long sentences can be split into mutually independent portions.

With reference to a text-structured example base, Kit *et al.* (2002: 57–78) suggests that it is preferable to use the probabilistic approach for recombination. Taking an empirical case-based knowledge engineering approach to MT, they give an example of a tri-gram language model, and point out some other considerations such as insertion of function words for better readability. Techniques in SMT have also been used in the hybrid EBMT-SMT models of Groves and Way (2005a: 301–323; 2005b: 183–190), which uses Pharaoh (Koehn 2004: 115–124), a decoder for selecting a translation fragment order in the highest probability; and the MaTrEx system (Du *et al.* 2009: 95–99) which uses another decoder called Moses (Koehn *et al.* 2007: 177–180).

For EBMT systems using examples in syntactic tree structures, where the correspondence between source and target fragments is labeled explicitly, recombination is then a task of tree

unification. For instance, in Sato (1995: 31–49), possible word-dependency structures of translation are first generated based on the retrieved examples; the best translation candidate is then selected by a number of criteria such as the size and source-target context of examples. In Watanabe (1995: 269–291) where examples are represented as graphs, the recombination involves a kind of graph unification, which they refer to as a 'gluing process'. In Aramaki *et al.* (2005: 219–226), the translation examples stored in a dependency structure are first combined, with the source dependency relation preserved in the target structure, and then output with the aid of a probabilistic language model to determine the word-order.

Other systems without annotated example bases may be equipped with information about probable word alignment from dictionaries or other resources that can facilitate the recombination process (e.g. Kaji *et al.* 1992: 672–678; Matsumoto *et al.* 1993: 23–30). Some systems like Franz *et al.* (2000: 1031–1035), Richardson *et al.* (2001: 293–298), and Brockett *et al.* (2002: 1–7) rely on rule-based generation engines supplied with linguistic knowledge of target languages. Alternatively, in Nederhof (2001: 25–32) and Forcada (2002), recombination is carried out via a finite state transition network (FSTN), according to which translation generation becomes akin to giving a 'guided tour' from the source node to the target node in the FSTN.

A well-known problem in recombination, namely 'boundary friction' (Nirenburg *et al.* 1993: 47–57; Collins 1998), occurs when translation fragments from various examples need to be combined into a target sentence. Grammatical problems often occur because words with different syntactic functions cannot appear next to each other. This is especially true for certain highly inflected languages like German. One solution is to smooth the recombined translation, by adjusting the morphological features of certain words in the translation, or inserting some additional function words, based on a grammar or probabilistic model of the target language. Another proposal from Somers *et al.* (1994) is to attach each fragment with 'hooks' indicating the possible contexts of the fragment in a corpus, i.e. the words and parts-of-speech which can occur before and after. Fragments which can be connected together are shown in this way. Brown *et al.* (2003: 24–31) puts forth the idea of translation-fragment overlap. They find that examples with overlapping fragments are more likely to be combined into valid translations if there are sentences in an example base that also share these overlapping fragments. Based on their study of the occurrence frequencies of combined fragments from the Internet, Gough *et al.* (2002: 74–83) finds that valid word combinations usually have much higher occurrence frequencies than invalid ones.

Suitability

Sublanguage translation

EBMT is usually deemed suitable for sublanguage translation, largely due to its reliance on text corpora, most of which belongs to specific domains. In other words, EBMT systems are optimized to texts in the same domain as their examples. The contribution of an example domain to improving EBMT translation quality is illustrated in Doi *et al.* (2003: 16–18). A system with more 'in-domain examples' (domain of input sentences matches with that of examples) performs better than those with either only out-of-domain examples or fewer in-domain examples. One negative finding in Denoual (2005: 35–42) however shows that given the same number of examples, homogeneous data (in-domain) yield neither better nor worse EBMT quality than heterogeneous data (mixed domain). Even though the usefulness of in-domain examples is not yet completely clear, EBMT has been widely adopted in different

specific areas in recent years, including translation of sign language (Morrissey *et al.* 2007: 329–336), DVD subtitles (Flanagan 2009: 85–92) and idioms (Anastasiou 2010).

A series of works in Gough and Way (2003: 133–140; 2004a: 73–81) and Way and Gough (2005: 1–36) further substantiates the suitability of EBMT for domain-specific translation. Their system is based on examples written in controlled language (CL), a subset of natural language whose grammar and lexicon are restricted in order to minimize ambiguity and complexity. In return, their system outperforms both RBMT and SMT systems in evaluation, largely due to the fact that RBMT suffers from greater complexity in fine-tuning its system to support CL, while SMT requires extremely large volumes of training text which is hard to come by for CL. An EBMT system, on the other hand, can be developed with smaller amounts of training examples.

Interoperation with other MT paradigms

While there may be few 'pure' EBMT systems (Lepage and Denoual 2005: 81–90; Somers *et al.* 2009: 53–60), example-based method has been widely integrated with other MT approaches to provide complementary advantages. Mention of interoperability of EBMT occurs as early as in Furuse and Iida (1992: 139–150) which notes that 'an example-based framework never contradicts other frameworks such as a rule-based and a statistically-based framework, nor is it difficult to integrate it with them'.

The following discussion reviews how EBMT were combined with rule-based and statistical systems.

With RBMT

Since its initial proposal, EBMT has long been applied as a solution for what is too difficult to resolve in RBMT. This includes notably special linguistic expressions such as Japanese adnominal particle constructions (Sumita *et al.* 1990: 203–211; Sumita and Iida 1991: 185–192), 'parameterizable fixed phrases' in economics news stories (Katoh and Aizawa 1994: 28–33), compound nouns and noun phrases which are syntactically and semantically idiosyncratic (Yamabana *et al.* 1997: 977–982), all of which can be handled by simply collecting translation examples to cover them.

The strengths and weaknesses of the two approaches are more thoroughly analysed in Carl *et al.* (2000: 223–257). In general, applying EBMT to longer translation units (phrasal) ensures better translation quality, but lower coverage of the types of source texts that can be reliably translated. In contrast, applying RBMT to shorter translation units (lexical) enables a higher coverage but inferior output quality to EBMT. An MT system architecture is accordingly proposed to integrate the two approaches. Source chunks in an input sentence are first matched by an example-based module against an example base. The unmatched parts and the reordering of translated chunks are then handled by a rule-based module.

With SMT

Statistical methods have been widely used in EBMT, including example retrieval (Doi *et al.* 2005: 51–58), as well as matching and recombination (Liu *et al.* 2005: 25–32). Example-based methods can also be utilized to assist SMT. Marcu (2001: 386–393) built a translation memory from a bilingual corpus with statistical methods to train SMT models. Langlais and Simard (2002: 104–113) integrated bilingual terminologies into an SMT system, leading to improved

translation performance. Watanabe and Sumita (2003: 410–417) designed an example-based SMT decoder that can retrieve translation examples from a parallel corpus whose source part is similar to the input sentence, and modify the target part of the example to produce translation output. Based on a series of experiments carried out by Groves and Way (2005a: 301–323, 2005b: 183–190, 2006: 115–124) involving the addition of example chunks to an SMT system, they note that, 'while there is an obvious convergence between both paradigmatic variants, more gains are to be had from combining their relative strengths in novel hybrid systems'.

Summary

In general, the main idea of EBMT is simple: from all that was translated in the past, use whatever is available to facilitate the translation of the next utterance. How translation data (examples) are stored and applied (in matching and recombination) can vary, as long as the examples are used in run-time. Being empirical in nature, the example base represents real language use, covers the constructions which really occur, and is relatively easy to extend – simply adding more examples. This is particularly true for examples stored as bilingual string pairs, which have benefited, and will continue to benefit from the massively growing number of documents on the web. The problem of developing MT systems for language pairs suffering from a scarcity of resources has been to some extent resolved.

The current status of EBMT also reveals, however, the possible constraints of this approach. So far there is no publicly available MT system purely based on the example-based approach, except a few on a limited scale for research purposes. Most systems exhibit a certain degree of hybridity with other MT approaches. (On the other hand, example-based method has also been widely applied in other MT paradigms.) From a practical perspective, it is perhaps more pragmatic to focus on how the strength of example-based method can be adequately utilized, together with other available methods, to tackle various kinds of translation problems.

References

Al-Adhaileh, Mosleh Hmoud, and Enya Kong Tang (1999) 'Example-based Machine Translation Based on the Synchronous SSTC Annotation Schema', in *Proceedings of the MT Summit VII: MT in the Great Translation Era*, 13–17 September 1999, Kent Ridge Digital Labs, Singapore, 244–249.

Anastasiou, Dimitra (2010) *Idiom Treatment Experiments in Machine Translation*, Newcastle upon Tyne: Cambridge Scholars Publishing.

Aramaki, Eiji, Sadao Kurohashi, Satoshi Sato, and Hideo Watanabe (2001) 'Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation', in *Proceedings of the Machine Translation Summit VIII: Machine Translation in the Information Age*, 18–22 September 2001, Santiago de Compostela, Spain, 27–32.

Aramaki, Eiji, Sadao Kurohashi, Hideki Kashioka, and Hideki Tanaka (2003) 'Word Selection for EBMT Based on Monolingual Similarity and Translation Confidence', in *HLT-NAACL-PARALLEL '03: Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, 27 May – 1 June 2003, Edmonton, Canada, 57–64.

Aramaki, Eiji, Sadao Kurohashi, Hideki Kashioka, and Naoto Kato (2005) 'Probabilistic Model for Example-based Machine Translation', in *Proceedings of Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 219–226.

Arthern, Peter J. (1978) 'Machine Translation and Computerized Terminology Systems: A Translator's Viewpoint', in *Proceedings of Translating and the Computer*, London: ASLIB, 77–108.

Brockett, Chris, Takako Aikawa, Anthony Aue, Arul Menezes, Chris Quirk and Hisami Suzuki (2002) 'English-Japanese Example-based Machine Translation Using Abstract Linguistic Representations', in *COLING-02 Workshop: Machine Translation in Asia*, 24 August – 1 September 2002, Taipei, Taiwan, 1–7.

Brown, Ralf D. (2000) 'Automated Generalization of Translation Examples', in *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, 29 July – 6 August 2000, Saarbrücken, Germany, 125–131.

Brown, Ralf D., Rebecca Hutchinson, Paul N. Bennett, Jaime G. Carbonell, and Peter Jansen (2003) 'Reducing Boundary Friction Using Translation-fragment Overlap', in *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches*, 23–27 September 2003, New Orleans, LA, 24–31.

Carl, Michael, Catherine Pease, Lonid L. Iomdin, and Oliver Streiter (2000) 'Towards a Dynamic Linkage of Example-based and Rule-based Machine Translation', *Machine Translation* 15(3): 223–257.

Chatterjee, Niladri (2001) 'A Statistical Approach for Similarity Measurement between Sentences for EBMT', in *Proceedings of the Symposium on Translation Support Systems (STRANS)*, 15–17 February 2001, Kanpur, India.

Collins, Brona (1998) 'Example-based Machine Translation: An Adaptation Guided Retrieval Approach', PhD thesis, Trinity College, Dublin.

Cranias, Lambros, Harris Papageorgiou, and Stelios Piperidis (1994) 'A Matching Technique in Example-based Machine Translation', in *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, 5–9 August 1994, Kyoto, Japan, 100–104.

Denoual, Etienne (2005) 'The Influence of Example-data Homogeneity on EBMT Quality', in *Proceedings of the Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 35–42.

Doi, Takao and Eiichiro Sumita (2003) 'Input Sentence Splitting and Translating', in *HLT-NAACL-PARALLEL '03: Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, 27 May – 1 June 2003, Edmonton, Canada, 104–110.

Doi, Takao, Eiichiro Sumita, and Hirofumi Yamamoto (2003) 'Adaptation Using Out-of-domain Corpus within EBMT', in *Proceedings of the HLT-NAACL: Conference Combining Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics*, 27 May – 1 June 2003, Edmonton, Canada, 16–18.

Doi, Takao, Hirofumi Yamamoto, and Eiichiro Sumita (2005) 'Graph-based Retrieval for Example-based Machine Translation Using Edit Distance', in *Proceedings of Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 51–58.

Du, Jinhua, He Yifan, Sergio Penkale, and Andy Way (2009) 'MaTrEx: The DCU MT System for WMT 2009', in *Proceedings of the EACL: The 4th Workshop on Statistical Machine Translation*, 30–31 March 2009, Athens, Greece, 95–99.

Flanagan, Marian (2009) 'Using Example-based Machine Translation to Translate DVD Subtitles', in *Proceedings of the 3rd International Workshop on Example-based Machine Translation*, 12–13 November 2009, Dublin City University, Dublin, Ireland, 85–92.

Forcada, Mikel L. (2002) 'Using Multilingual Content on the Web to Build Fast Finite-state Direct Translation Systems', in *The 2nd ELSNET TMI Workshop on MT Roadmap*, Keihanna, Japan.

Franz, Alexander, Keiko Horiguchi, Duan Lei, Doris Ecker, Eugene Koontz, and Kazami Uchida (2000) 'An Integrated Architecture for Example-based Machine Translation', in *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, 29 July – 6 August 2000, Saarbrücken, Germany, 1031–1035.

Furuse, Osamu and Hitoshi Iida (1992) 'An Example-based Method for Transfer-driven Machine Translation', in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT (TMI-92)*, 25–27 June 1992, Montreal, Canada, 139–150.

Gough, Nano and Andy Way (2003) 'Controlled Generation in Example-based Machine Translation', in *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches*, 23–27 September 2003, New Orleans, LA, 133–140.

Gough, Nano and Andy Way (2004a) 'Example-based Controlled Translation', in *Proceedings of the 9th Conference of the European Association for Machine Translation (EAMT-04)*, 26–27 April 2004, University of Malta, Valetta, Malta, 73–81.

Gough, Nano and Andy Way (2004b) 'Robust Large-scale EBMT with Marker-based Segmentation,' in *Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, 4–6 October 2004, Baltimore, MD, 95–104.

Gough, Nano, Andy Way, and Mary Hearne (2002) 'Example-based Machine Translation via the Web', in Stephen D. Richardson (ed.) *AMTA-02: Proceedings of the 5th Conference of the Association for Machine*

Translation in the Americas, Machine Translation: From Research to Real Users, 6–12 October 2002, Tiburon, CA, 74–83.

Green, Thomas R.G. (1979) 'The Necessity of Syntax Markers: Two Experiments with Artificial Languages', *Journal of Verbal Learning and Behavior* 18(4): 481–496.

Groves, Declan and Andy Way (2005a) 'Hybrid Data-driven Models of Machine Translation', *Machine Translation* 19(3–4): 301–323.

Groves, Declan and Andy Way (2005b) 'Hybrid Example-based SMT: The Best of Both Worlds?' in *HLT-NAACL-PARALLEL '03: Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, 27 May – 1 June 2003, Edmonton, Canada, 183–190.

Groves, Declan and Andy Way (2006) 'Hybridity in MT: Experiments on the Europarl Corpus', in *Proceedings of the 11th Conference of the European Association for Machine Translation (EAMT-06)*, 19–20 June 2006, University of Oslo, Norway, 115–124.

Hearne, Mary and Andy Way (2003) 'Seeing the Wood for the Trees: Data-oriented Translation', in *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches*, 23–27 September 2003, New Orleans, LA, 165–172.

Kaji, Hiroyuki, Yuuko Kida, and Yasutsugu Morimoto (1992) 'Learning Translation Templates from Bilingual Text', in *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, 23–28 August 1992, Nantes, France, 672–678.

Katoh, Naoto and Teruaki Aizawa (1994) 'Machine Translation of Sentences with Fixed Expressions', in *Proceedings of the 4th Conference on Applied Natural Language Processing*, 13–15 October 1994, Stuttgart/ San Francisco: Morgan Kaufmann, 28–33.

Kay, Martin (1976) 'The Proper Place of Men and Machines in Language Translation', in *Proceedings of the Foreign Broadcast Information Service Seminar on Machine Translation*. (Reprinted in *Machine Translation* (1997) 12(1–2): 3–23.)

Kay, Martin (1997) 'The Proper Place of Men and Machines in Language Translation', *Machine Translation*, 12, 3–23. First appeared in the *Proceedings of the Foreign Broadcast Information Service Seminar on Machine Translation*, 1976.

Kit, Chunyu and Jessica Y.H. Ng (2007) 'An Intelligent Web Agent to Mine Bilingual Parallel Pages via Automatic Discovery of URL Pairing Patterns', in *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops: Workshop on Agents and Data Mining Interaction (ADMI-07)*, 2–5 November 2007, Silicon Valley, CA, 526–529.

Kit, Chunyu, Pan Haihua, and Jonathan J. Webster (2002) 'Example-based Machine Translation: A New Paradigm', in Chan Sin-wai (ed.) *Translation and Information Technology*, Hong Kong: The Chinese University Press, 57–78.

Kit, Chungyu, Liu Xiaoyue, Sin King Kui, and Jonathan J. Webster (2005) 'Harvesting the Bitexts of the Laws of Hong Kong from the Web', in *Proceedings of the 5th Workshop on Asian Language Resources (ALR-05)*, 14 October 2005, Jeju Island, Korea, 71–78.

Kit, Chunyu, Jonathan J. Webster, Sin King Kui, Pan Haihua, and Li Heng (2003) 'Clause Alignment for Bilingual HK Legal Texts with Available Lexical Resources', in *Proceedings of the 20th International Conference on the Computer Processing of Oriental Languages (ICCPOL-03)*, 3–6 August 2003, Shenyang, China., 286–292.

Kit, Chunyu, Jonathan J. Webster, Sin King Kui, Pan Haihua, and Li Heng (2004) 'Clause Alignment for Bilingual HK Legal Texts: A Lexical-based Approach', *International Journal of Corpus Linguistics* 9(1): 29–51.

Koehn, Philipp (2004) 'Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation Models', in *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-04)*, 29 September – 2 October 2004, Georgetown University, Washington, DC/Berlin: Springer, 115–124.

Koehn, Philipp (2005) 'Europarl: A Parallel Corpus for Statistical Machine Translation', in *Proceedings of the Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 79–86.

Koehn, Philipp, Hiu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowen, Shen Wade, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst (2007) 'Moses: Open Source Toolkit for Statistical Machine Translation', in *ACL 2007: Proceedings of the Interactive Poster and Demonstration Sessions*, 25–27 June 2007, Prague, Czech Republic, 177–180.

Langlais, Philippe and Michel Simard (2002) 'Merging Example-based and Statistical Machine Translation: An Experiment', in Stephen D. Richardson (ed.) *AMTA-02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, 6–12 October 2002, Tiburon, CA, 104–113.

Lepage, Yves and Etienne Denoual (2005) 'The "Purest" EBMT System Ever Built: No Variables, No Templates, No Training, Examples, Just Examples, Only Examples', in *Proceedings of Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 81–90.

Liu, Zhanzi, Wang Haifeng, and Wu Hua (2005) 'Example-based Machine Translation Based on TSC and Statistical Generation', in *Proceedings of the Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 25–32.

Ma, Xiaoyi and Mark Y. Liberman (1999) 'Bits: A Method for Bilingual Text Search over the Web,' in *Proceedings of the Machine Translation Summit VII: MT in the Great Translation Era*, 13–17 September 1999, Kent Ridge Digital Labs, Singapore, 13–17.

Malavazos, Christos, Stelios Piperidis, and George Carayannis (2000) 'Towards Memory and Template-based Translation Synthesis', in *Proceedings of MT2000: Machine Translation and Multilingual Applications in the New Millenium*, 1–8 January 2000, Exeter, UK.

Marcu, Daniel (2001) 'Towards a Unified Approach to Memory- and Statistical-based Machine Translation', in *Association for Computational Linguistics: 39th Annual Meeting and 10th Conference of the European Chapter: Workshop Proceedings: Data-driven Machine Translation*, 6–11 July 2001, Toulouse, France, 386–393.

Markantonatou, Stella, Sokratis Sofianopoulos, Vassiliki Spilioti, Yiorgos Tambouratzis, Marina Vassiliou, Olga Yannoutsou, and Nikos Ioannou (2005) 'Monolingual Corpus-based MT Using Chunks', in *Proceedings of the Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 91–98.

Maruyama, Hiroshi and Hideo Watanabe (1992) 'Tree Cover Search Algorithm for Example-based Translation', in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT (TMI-92)*, 25–27 June 1992, Montréal, Quebec, Canada, 173–184.

Matsumoto, Yuji, Hiroyuki Ishimoto, and Takehito Utsuro (1993) 'Structural Matching of Parallel Texts', in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*, 22–26 June 1993, Columbus, OH, 23–30.

McTait, Kevin (2001) 'Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns', in *Proceedings of the Machine Translation Summit VIII: Machine Translation in the Information Age*, 18–22 September 2001, Santiago de Compostela, Spain, 22–34.

Melby, Alan K. and Terry Warner (1995) *The Possibility of Language: A Discussion of the Nature of Language*, Amsterdam and Philadelphia: John Benjamins.

Morrissey, Sara, Andy Way, Daniel Stein, Jan Bungeroth, and Hermann Ney (2007) 'Combining Data-driven MT Systems for Improved Sign Language Translation', in *Proceedings of the Machine Translation Summit XI*, 10–14 September 2007, Copenhagen Business School, Copenhagen, Denmark, 329–336.

Munteanu, Dragos and Daniel Marcu (2005) 'Improving Machine Translation Performance by Exploiting Comparable Corpora', *Computational Linguistics* 31(4): 477–504.

Nagao, Makoto (1984) 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle', in Alick Elithorn and Ranan Banerji (eds) *Artificial and Human Intelligence*, Amsterdam: North Holland, 173–180.

Nagao, Makoto (1992) 'Some Rationales and Methodologies for Example-based Approach', in Sofia Ananaidou (ed.) *Proceedings of the International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP)*, 30–31 July 1992, University of Manchester, UK, Manchester: Centre for Computational Linguistics, University of Manchester Institute of Science and Technology, 82–94.

Nagao, Makoto (2007) 'An Amorphous Object Must Be Cut by a Blunt Tool', in Khurshid Ahmad, Christopher Brewster, and Mark Stevenson (eds) *Words and Intelligence II: Essays in Honor of Yorick Wilks*, Dordrecht: Springer, 153–158.

Nederhof, Mark-Jan (2001) 'Approximating Context-free by Rational Transduction for Example-based MT', in *Association for Computational Linguistics: 39th Annual Meeting and 10th Conference of the European Chapter: Workshop Proceedings: Data-driven Machine Translation*, 6–11 July 2001, Toulouse, France, 25–32.

Nilsson, Nils J. (1971) *Problem-solving Methods in Artificial Intelligence*, New York: McGraw-Hill.

Nirenburg, Sergei, Constantine Domashnev, and Dean J. Grannes (1993) 'Two Approaches to Matching in Example-based Translation', in *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation: MT in the Next Generation (TMI-93)*, 14–16 July 1993, Kyoto, Japan, 47–57.

Nirenburg, Sergei, Stephen Beale, and Constantine Domashnev (1994) 'A Full-text Experiment in Example-based Machine Translation', in Daniel Jones (ed.) *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, 14–16 September 1994, the University of Manchester Institute of Science and Technology, Manchester, 78–87.

Och, Franz and Hermann Ney (2003) 'A Systematic Comparison of Various Statistical Alignment Models', *Computational Linguistics* 29(1): 19–51.

Planas, Emmanuel and Osamu Furuse (1999) 'Formalizing Translation Memories', in *Proceedings of the Machine Translation Summit VII: MT in the Great Translation Era*, 13–17 September 1999, Kent Ridge Digital Labs, Singapore, 331–339.

Resnik, Philip (1998) 'Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text', in *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas: Machine Translation and the Information Soup (AMTA-98)*, 28–31 October 1998, Langhorne, PA, 72–82.

Richardson, Stephen D., William B. Dolan, Arul Menezes, and Jessie Pinkham (2001) 'Achieving Commercial-quality Translation with Example-based Methods', in *Proceedings of the Machine Translation Summit VIII: Machine Translation in the Information Age*, 18–22 September 2001, Santiago de Compostela, Spain, 293–298.

Roh, Yoon-Hyung, Munpyo Hong, Choi Sung-Kwon, Lee Ki-Young, and Park Sang-Kyu (2003) 'For the Proper Treatment of Long Sentences in a Sentence Pattern-based English-Korean MT System', in *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches*, 23–27 September 2003, New Orleans, LA, 323–329.

Sato, Satoshi (1995) 'MBT2: A Method for Combining Fragments of Examples in Example-based Machine Translation', *Artificial Intelligence* 75(1): 31–49.

Shimohata, Mitsuo, Eiichiro Sumita, and Yuji Matsumoto (2003) 'Retrieving Meaning-equivalent Sentences for Example-based Rough Translation', in *HLT-NAACL-PARALLEL '03: Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, 27 May – 1 June 2003, Edmonton, Canada, 73–80.

Somers, Harold L. (2003) 'An Overview of EBMT', in Michael Carl and Andy Way (eds) *Recent Advances in Example-based Machine Translation*, Dordrecht: Kluwer Academic Publishers, 3–57.

Somers, Harold L., Ian McLean, and Danny Jones (1994) 'Experiments in Multilingual Example-based Generation', in Marta Gatius and Horacio Rodríguez (eds) *Proceedings of the 3rd Conference on the Cognitive Science of Natural Language Processing (CSNLP-94)*, Dublin.

Somers, Harold L., Sandipan Dandapat, and Sudip Kumar Naskar (2009) 'A Review of EBMT Using Proportional Analogies', in *Proceedings of the 3rd International Workshop on Example-based Machine Translation*, 12–13 November 2009, Dublin City University, Dublin, Ireland, 53–60.

Sumita, Eiichiro and Hitoshi Iida (1991) 'Experiments and Prospects of Example-based Machine Translation', in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, Berkeley, CA, 185–192.

Sumita, Eiichiro, Hitoshi Iida, and Hideo Kohyama (1990) 'Translating with Examples: A New Approach to Machine Translation', in Gary A. Coen (ed.) *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-90)*, 11–13 June 1990, University of Texas at Austin, TX, 203–211.

Vandeghinste, Vincent, Peter Dirix, and Ineke Schuurman (2005) 'Example-based Translation without Parallel Corpora: First Experiments on a Prototype', in *Proceedings of the Machine Translation Summit X: The 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 135–142.

Watanabe, Hideo (1995) 'A Model of a Bi-directional Transfer Mechanism Using Rule Combinations', *Machine Translation* 10(4): 269–291.

Watanabe, Taro and Eiichiro Sumita (2003) 'Example-based Decoding for Statistical Machine Translation', in *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches*, 23–27 September 2003, New Orleans, LA, 410–417.

Way, Andy (2001) 'Translating with Examples', in *Proceedings of the Machine Translation Summit VIII: Workshop on Example-based Machine Translation*, 18–22 September 2001, Santiago de Compostela, Spain, 66–80.

Way, Andy (2003) 'Translating with Examples: The LFG-DOT Models of Translation', in Michael Carl and Andy Way (eds) *Recent Advances in Example-based Machine Translation*, Dordrecht: Kluwer Academic Publishers, 443–472.

Way, Andy (2010) 'Panning for EBMT Gold, or "Remembering Not to Forget"', *Machine Translation* 24(3–4): 177–208.

Way, Andy and Nano Gough (2003) 'wEBMT: Developing and Validating an Example-based Machine Translation System Using the World Wide Web', *Computational Linguistics* 29(3): 421–457.

Way, Andy and Nano Gough (2005) 'Controlled Translation in an Example-based Environment', *Machine Translation* 19(1): 1–36.

Webster, Jonathan J., Sin King Kui, and Hu Qinan (2002) 'The Application of Semantic Web Technology for Example-based Machine Translation (EBMT)', in Chan Sin-wai (ed.) *Translation and Information Technology*, Hong Kong: The Chinese University Press, 79–91.

Yamabana, Kiyoshi, Shin-ichiro Kamei, Kazunori Muraki, Shinichi Doi, Shinko Tamura, and Kenji Satoh (1997) 'A Hybrid Approach to Interactive Machine Translation – Integrating Rule-based, Corpus-based, and Example-based Method', in *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, 23–29 August 1997, Nagoya, Japan, 977–982.

Zhang, Ying, Ralf Brown, Robert Frederking, and Alon Lavie (2001) 'Pre-processing of Bilingual Corpora for Mandarin-English EBMT', in *Proceedings of the Machine Translation Summit VIII: Machine Translation in the Information Age*, 18–22 September 2001, Santiago de Compostela, Spain, 247–252.

8

OPEN-SOURCE MACHINE TRANSLATION TECHNOLOGY

Mikel L. Forcada

UNIVERSITAT D'ALACANT, SPAIN

Free/open-source software

Free software,[1] as defined by the Free Software Foundation, is software that (a) may be freely executed for any purpose, (b) may be freely examined to see how it works and may be freely modified to adapt it to a new need or application (for that, source code, that is, the text of the program as the programmer sees and edits it,[2] must be available, hence the alternative name *open-source*; see below), (c) may be freely redistributed to anyone, and (d) may be freely improved and released to the public so that the whole community of users benefits (source code must be available for this too). These freedoms are regulated by a license, which is packaged with each piece of free software. The Open Source Initiative establishes an alternative definition[3] of *open-source software* which is roughly equivalent, as the most important free licenses are also open-source. The joint term free/open-source will be used in this chapter. Free/open-source licenses are legally based on copyright law; therefore, the freedoms they grant are protected by law in almost every country in the world.

Note that in English, the word *free* is ambiguous: free software advocates explain that it means 'free as in *freedom*, not as in *free beer*'; some use the word *libre* as an alternative to avoid the second sense.[4] Examples of free/open-source software include the LibreOffice and OpenOffice.org word-processing alternatives to Microsoft Word, the Firefox web browser, or the GNU/Linux family of operating systems.

Note also that, in contrast, the term *freeware* is the usual name for software that is distributed free of charge (free as in 'free beer'), but does not necessarily grant the four freedoms explained above (for instance, there may be no way to access the source code).

For most users ('end users' in the jargon of the software industry), the difference between freeware and free/open-source software may not be directly relevant, unless they are particularly conscious about knowing what the software is actually doing in their computer (if the source is available, a programmer can determine, for instance, if the program is respecting the user's private information stored in the computer). If the user is happy about the program as is, freeware is probably fine (three typical examples of freeware would be the Acrobat Reader, used to read PDF files, the Opera web browser, or the Adobe Flash Player, used to view multimedia content). However, if the user wants to get involved in efforts to improve the program, having access to the source in free/open-source software makes it possible to recruit people with the necessary programming skills to do it. Free/open-source software is frequently developed collaboratively by communities of experts share their improvements (in free/open-

source projects), constantly improve the code, and periodically release new versions of the software.

An optional property that some free/open-source software licenses may have is called *copyleft*, which is obviously a pun on the word *copyright*. When the free/open-source license for a piece of software has copyleft, it means that derivatives (modifications) of this software can only be distributed using exactly the same free/open-source license, and therefore may not be distributed as non-free/closed-source software. *Copylefted* licenses provide a way to encourage modifications to be contributed back to the community, so that a shared pool of software, a *commons*,[5] is created. The most popular free/open-source license for software, the GNU General Public License or GPL for short,[6] is a typical *copylefted* license. There are also some popular non-copylefted licenses such as the Apache License,[7] the MIT License,[8] or the 3-clause BSD license.[9]

Machine translation software: licensing

Machine translation (MT) software is special in the way it strongly depends on data.

One can distinguish two main kinds of MT: *Rule-based* (or *knowledge-based*) *machine translation* (RBMT), which uses linguistic and translation knowledge, created and conveniently encoded by experts, to perform the translations, and *corpus-based* (or *data-driven*) *machine translation* (CBMT) which automatically *learns* information learned from (usually very large) corpora of *parallel text*, where all sentences come with their translations. Of course, there is a wide range of hybrid approaches between these two extremes.

On the one hand, rule-based machine translation depends on *linguistic data* such as morphological dictionaries (which specify, for instance, that *went* is the past-tense of *go*), bilingual dictionaries (specifying, for instance, that a French translation of *go* is *aller*), grammars that describe, for instance, the use of the verb *do* in English interrogative and negative clauses, and *structural transfer* rule files describing how structures are transformed from one language to another, for instance to turn English genitive constructions such as *The teacher's library* into the Spanish construction *La biblioteca del profesor*.

On the other hand, corpus-based machine translation (such as *statistical machine translation*, for example, Koehn 2010) depends, as said above, on the availability of data; in most cases, of sentence-aligned parallel text where, for instance, *Machine translation software is special* comes with its French translation *Les logiciels de traduction automatique sont spéciaux*.

The choice of rule-based versus corpus-based MT depends on the actual language setting and the actual translation tasks that will be tackled. For some language pairs, it may be quite hard to obtain and prepare the amounts of sentence-aligned parallel text (of the order of millions of words) required to get reasonable results in 'pure' corpus-based machine translation such as statistical machine translation (SMT), so hard that it might be much easier to encode the available expertise into the language data files (dictionaries, grammar rule files, etc.) needed to build a rule-based machine translation system. For other language pairs, translated texts may be readily available in sufficient amounts and it may be much easier to filter and sentence-align those texts and then train a statistical machine translation system.

In either case, one may clearly distinguish three components: first, an *engine*, the program that performs the translation (also called *decoder* in statistical machine translation); second, the *data* (either linguistic data or parallel corpora) needed for that particular language pair, and, third, optionally, *tools* to maintain these data and turn them into a format which is suitable for the engine to use (for instance, in statistical machine translation, parallel text is used to *learn* a statistical translation table containing sub-sentential translation units – *phrase pairs* in statistical

MT parlance – such as '*machine translation* = *traduction automatique*' and probability information associated to each such pair).

Commercial machine translation

Most commercial machine translation systems are rule-based (although machine translation systems with a strong corpus-based component have started to appear[10]). Most RBMT systems (not all of them, see "Knowledge- or Rule-based Free/Open-source Machine Translation Systems" below) have engines with proprietary technologies which are not completely disclosed (indeed, most companies view their proprietary technologies as their main competitive advantage). Linguistic data are not easily modifiable by end users either; in most cases, one can only add new words or user glossaries to the system's dictionaries, and perhaps some simple rules, but it is not possible to build complete data for a new language pair and use it with the engine.

Free/open-source machine translation

On the one hand, for a rule-based machine translation system to be free/open-source, source code for the engine and tools should be distributed as well as the "source code" (expert-editable form) of linguistic data (dictionaries, translation rules, etc.) for the intended pairs. It is more likely for users of the free/open-source machine translation to change the linguistic data than to modify the machine translation engine; for the improved linguistic data to be used with the engine, tools to maintain them should also be distributed under free/open-source licenses. On the other hand, for a corpus-based machine translation system such as a statistical machine translation system, source code both for the programs that learn the statistical translation models from parallel text as well as for the engines (decoders) that use these translation models to generate the most likely translations of new sentences should be distributed along with data such as the necessary sentence-aligned parallel texts.[11]

Machine translation that is neither commercial nor free/open-source

The previous sections have dealt with commercial machine translation and free/open-source machine translation. However, the correct dichotomy would be between free/open-source MT versus 'non-free/closed-source' MT; indeed, there are a number of systems that do not clearly fit in the categories considered in the last two sections.

For example, there are MT systems on the web that may be freely used (with a varying range of restrictions); some are demonstration or reduced versions of commercial systems, whereas some other freely-available systems are not even commercial.[12] The best examples of web-based MT systems which are not free/open-source but may be freely used for short texts and web pages are Google Translate[13] and Microsoft's Bing Translator:[14] both are basically corpus-based systems. Finally, another possibility would be for the MT engine and tools not to be free/open-source (even using proprietary technologies) but just to be simply freely or commercially available and fully documented, with linguistic data being distributed openly (open-source linguistic data), but there is no such system available.

Types of free/open-source machine translation systems and users

Distributing machine translation systems under free/open-source licenses gives their target users full access to machine-translation technologies. As the target users of free/open-source machine translation systems are very varied, one may find many different types of systems. Note that we are referring here to systems that may be downloaded and installed, either to be used offline or to set up a web service.

There are systems which may easily be installed in a few clicks and directly used (some call these 'zero-install' systems). For instance, Apertium-Caffeine,[15] part of the Apertium platform (see below) is a small package that runs immediately after download in any computer where a Java run-time environment has been installed and prompts the user for the language pairs they would like to have available. It translates plain text as soon as it is typed or pasted in the input window and it is aimed at casual users needing a quick translation for short texts when an Internet connection is not available. A related program, Apertium-Android,[16] offers similar 'offline' translation functionalities for devices running the Android operating system.

There are also free/open-source machine translation systems aimed at professional translators using computer-aided translation (CAT) environments. For instance, for those using the (also free/open-source) OmegaT CAT system,[17] Apertium-OmegaT,[18] also part of the Apertium project, is available for easy installation as an extension (or *plug-in*). It allows translators to get a quick offline translation for segments in their job for which the system cannot find a match in their translation memories.

There are complete fully fledged free/open-source machine translation systems whose installation and usage requires a certain degree of technical expertise. Among the rule-based machine translation systems, OpenLogos, Apertium, and Matxin (see "Knowledge- or rule-based free/open-source machine translation systems" below) are designed to be installed on computers running the GNU/Linux operating system (although they may also be installed on other operating systems). They offer support for different kinds of text formats, and are aimed at heavy usage by more than one user, possibly remotely through a web interface; they are also intended to serve as platforms for research and development. Among corpus-based machine translation systems, with similar target users and expected usages, the prime example would be the statistical machine translation system Moses (see description below); installation (and training on suitable corpora) is a bit more challenging,[19] even if efforts have been made to simplify the process.

Free/open-source machine translation in business

Free/open-source machine translation may be used to engage in basically the same kinds of business as non-free/closed-source machine translation, with differences that will be described immediately.

1 Some companies sell services around a particular machine translation system, such as installing, configuring or customizing the system for a particular user (for instance, to deal with their particular terminology, document formats, document style, etc.); their business in the free/open-source setting changes significantly. Companies marketing such services for a non-free/closed-source system may do so by means of a specific (often exclusive) agreement with the company producing the system, that provides them with tools that other companies would not have access to; this reduces competition. In the case of free/open-source software, the same kinds of services could be offered by any company, as the

software is usually available to them as third parties under a general free/open-source license (and that software includes usually the necessary tools to customize or adapt the system). Access to the source code of the system adds flexibility to the kind of services they can market. But using free/open-source may expose service companies to strong competition by other companies having exactly the same access rights to the software: therefore, the deeper the knowledge of the system, the sharper the competitive edge of the company (with machine translation developers having the sharpest edge of all if they decide to market these services). Note that the usage of free/open-source licenses shifts a good part of the business from a license-centered model to a service-centered model, which happens to be less vulnerable, for instance, to loss of revenue due to unlicensed distribution of copies.

2 The business of companies who would otherwise sell machine translation software licenses is probably the one that changes most with free/open-source machine translation. On the one hand, if the company has developed the system and therefore owns the right to distribute it, it may want to release a reduced-functionality or "basic" version under a free/open-source license while marketing end-user licenses for a fully functional or "premium" system for a fee, as it is done with some non-free/closed-source software. On the other hand, if the company has not developed the system but the license is not a "copylefted" license requiring the distribution of source when distributing a modified system (see "Free/open-source software" above), it may produce a non-free/closed-source derivative of an existing free/open-source system and sell licenses to use them.

3 A third group of companies offer value-added web-based machine translation (translation of documents in specific formats, translated chat rooms, online computer-aided translation systems, etc.). They are currently under strong competition from the main web-based machine translation companies, namely Google and Microsoft; therefore, they have to add value to web-based machine translation. In the non-free/closed-source setting, web-based translation companies have to buy a special license, different from end-user licenses, from machine translation manufacturers or their resellers; in the free/open-source setting, no fees are involved, and, as in the second group, the additional flexibility provided by full access to the source allows these companies to offer innovative services that would be more difficult to develop with non-free/closed-source software; they could develop and deploy these services themselves or hire one of the companies in the third group, who are competing to offer services based on the same free/open-source machine translation system.

4 The fourth group comprises professional translators and translation agencies. Instead of investing in licenses for non-free/closed-source systems, they could either install and use the free/open-source system "as is" and save in license fees; they could additionally hire one of the service companies in the second group above to customize it for their needs. In the case of a free/open-source system, more than one company could actually offer the service: better prices could arise from the competition; in turn, professional translators and translation agencies could offer more competitive translation prices to their customers.

Note that free/open-source machine translation systems often have active (and enthusiastic) user and developer communities that may make technical support available to any of the above businesses. Businesses themselves may choose to engage in the activity of these communities to ensure a better technical support in the future. These interactions may result in improved services or products.

Free/open-source machine translation in research

Machine translation research is undoubtedly very active in view of its relevance in an increasingly globalized world. Many researchers and developers have adopted free/open-source licensing models to carry out and disseminate their research. A clear indicator of this is the fact that over the last decade, many free/open-source MT systems and resources have been released (see "A survey of free/open-source machine translation technologies" below).[20] Series of specialized conferences and workshops such as Machine Translation Marathons[21] or International Workshops on Free Rule Based Machine Translation (FreeRBMT),[22] have been devoted to free/open-source machine translation.

The benefits of using free/open-source MT systems for research are varied. On the one hand, free/open-source development radically guarantees the reproducibility of any experiments, a key point in the advance of any scientific field, and lowers the bar for other researchers to engage in research in that field (Pedersen 2008). On the other hand, it makes it easier for end users to benefit earlier from the latest advances in the field; in particular, as the systems may be freely used, businesses and industries become direct and early beneficiaries of the research that has gone into building them, and the technology advances reach its end users much faster. Finally, as Pedersen (2008) also notes, the fact that the software is distributed and probably pooled makes it possible for the software to survive changes in the staff of research groups or the mere passage of time.

In fact, the existence of actively developed free/open-source machine translation platforms has contributed to the consolidation of what are now considered standard or state-of-the-art approaches to machine translation: consider, for instance, statistical machine translation, where the combination of the free/open-source packages GIZA++ (Och and Ney 2003) and Moses (Koehn *et al.* 2007; see below for details) has become the de facto "baseline" system, both in research and in industry. Some of the free/open-source MT systems featured in "A survey of free/open-source machine translation technologies" below are also designed to be platforms on their own; as a result, there has never been a wider option for researchers starting to do research in this field.

These benefits may be used as arguments to encourage public administrations to preferentially fund machine translation research projects whose developments are to be released under free/open-source licenses as they encourage fast transfer of technology to all interested parties in society; this is for instance, the point of Streiter *et al.* (2006). It certainly makes complete sense for publicly funded machine translation technologies to be freely and openly available to the society that directly or indirectly supports those public institutions with their taxes. Indeed, there is a tendency for research projects that explicitly commit to the free/open-source development to receive more public funding: for instance, the European Union's seventh Framework Programme explicitly lists as a sought outcome of the research that it will fund in the field of Information and Communication Technologies in 2013 (European Commission 2012, pp. 50 and 54) that "a European open-source MT system becomes the most widely adopted worldwide".

A survey of free/open-source machine translation technologies

Knowledge- or rule-based free/open-source machine translation systems

Among the existing knowledge- or rule-based free/open-source machine translation systems, Apertium will be described in detail and Matxin and OpenLogos in less detail; other systems will also be mentioned at the end of this section.

Apertium

Apertium[23] (Forcada *et al*. 2011) is a platform for rule-based MT – with an active community of hundreds of developers around it – which can be used to build machine translation systems for many language pairs, initiated in 2004 as part of a project funded by the Spanish Ministry of Industry, Tourism and Commerce. The design of Apertium is simple, as it was initially aimed at translating between closely related languages such as Spanish and Portuguese to produce draft translations to be post-edited. The core idea is that of building on top of the intuitive notion of word-for-word translation by tackling, using the minimum amount of linguistic analysis possible, the main problems encountered by such a crude approximation: solving the ambiguity of certain lexical items (such as *books* which can be a noun in *the books* or a verb in *He books*), identifying multi-word lexical items that need to be translated as a whole (such as *machine translation* → *traduction automatique*, and not → *traduction de machine*), ensuring locally the right word order (*the blue lamp* → *la lampe bleue*, *Peter's telephone* → *le téléphone de Peter*), or agreement (*the blue lamps* → *les lampes bleues*), etc. Such a simple formulation of the translation strategy makes it easy for people to help in the development of linguistic data for Apertium language pairs, or even to start new language pairs.

Even if this design was not initially aimed at less related language pairs, there are Apertium translators also for these pairs, not with the objective of producing a draft needing a reasonable amount of post-editing (which would be impossible with such a simple design) but rather to provide readers with the gist or general idea of a text written in a language that would otherwise be impenetrable to them. For instance, someone who does not know any Basque will not be able to make any sense of the sentence *Nire amaren adiskideak loreak erosi ditu*. However, the Apertium Basque–English translator (currently under development) produces the approximate translation *My mother's friend the flowers he has bought* which is quite intelligible for an English speaker even if rather hard to post-edit into *My mother's friend has bought the flowers*.

Apertium as a platform provides a language-independent engine, a well-defined, XML-based format to encode linguistic data, and the tools needed to manage these data and to turn them into the format used by the engine. The Apertium engine is a modular pipeline that processes the text by incrementally transforming it as follows:

1 Any formatting information (fonts, font weights, etc.) is detected and hidden so that work concentrates on the actual text.
2 All words are morphologically analysed. For instance, the word *books* above could be analysed as *book*, noun, plural, or *book*, verb, present tense, 3rd person singular. Multi-word lexical units such as *machine translation* are also detected and marked in this step.
3 For words such as *books*, one of the morphological analyses is chosen according to context (this is commonly called *part-of-speech tagging*). This may be done in one or two steps. The first step is optional and uses *constraints* or rules such as 'the word *the* cannot be followed by a personal form of a verb'. The second step uses statistical information obtained from texts about the distribution of word classes of neighbouring words to decide, for instance, that *like* is more likely to be a preposition than a verb in *run like the wind*.
4 Morphologically analysed words are looked up in a bilingual dictionary and translations are attached to them: *book*, noun, plural → *livre*, noun, masculine, plural (default translation); *cahier*, noun, masculine, plural (alternative translation). Multi-word lexical units are translated as a whole: *machine translation*, noun, singular → *traduction automatique*, noun, feminine, singular.

5 In recent Apertium versions, one of these translations is chosen using lexical selection rules which may be hand-written or learned from a bilingual corpus of sentences and their translations. If this module is not available, the default translation is chosen.
6 One or more modules apply (in cascade) the *structural transfer rules* for the language pair to ensure the correct word order, agreement, etc. and produce an adequate translation. For instance, a rule detects the English sequence article–adjective–noun and translates it into French as article–noun–adjective while making sure that the adjective receives the gender of the noun in French, to ensure, for instance that *the blue lamp* is translated into *la lampe bleue*. Another rule may be used to decide that the translation of the English preposition *in* should be *à* before a place name, so that *in the house* → *dans la maison* but *in Paris* → *à Paris*.
7 Structural transfer rules do not deliver target-language words in their final form, but rather in their 'morphologically analysed' form. A module is needed to turn e.g. *bleu*, adjective, feminine, singular into *bleue*.
8 The last linguistic processor takes care of some inter-word phenomena such as *la* + *élection* → *l'élection*, *de* + *égalité* → *d'égalité*, *viendra* + *il ?* → *viendra-t-il?* etc.
9 Formatting information hidden in the first step is placed back into the appropriate positions of the text.

This particular design may be classified as a transfer architecture (Hutchins and Somers 1992: 75); in particular, a shallow-transfer architecture; it may also be seen as what Arnold *et al.* call a *transformer* architecture (Arnold *et al.* 1994: sec. 4.2).

To perform some of these operations, Apertium uses source- and target-language dictionaries describing their morphology, bilingual dictionaries, disambiguation rules, structural transfer (structure transformation) rules, etc. All this information is encoded in clearly specified formats based on XML and grouped in language-pair packages that can be installed on demand. The modularity of the engine reflects on the internal modularity of language-pair packages: for instance, the Spanish–Portuguese and the Spanish–French language pair use essentially the same Spanish morphological dictionary.

At the time of writing this, 35 stable[24] language pairs are available from Apertium. Many of them include small languages which are not supported by any other machine translation systems such as Breton, Asturian, or Sámi. The free/open-source setting in Apertium, which uses the GNU General Public License for all its packages, makes it particularly attractive for minor-language experts to contribute their expertise in the creation of resources, as it is guaranteed that these will be available to the whole language community.

Apertium is one of the most-installed rule-based free/open-source machine translation systems. On the one hand, it has been included in some major GNU/Linux operating system distributions such as Debian and Ubuntu. On the other hand, it is so fast[25] and frugal that it may be installed in devices running the Android operating system, used as a plug-in in the most popular free/open-source computer-aided translation toolkit, OmegaT, or installed in a single click on any Java-enabled computer, regardless of the operating system (see "Types of free/open-source machine translation systems and users" above for details).

Matxin

Matxin[26] (Mayor *et al.* 2011) was born in the same project as Apertium, as the first publicly available machine translation system for Basque, a less-resourced language, and shares some components with it (for instance, monolingual and bilingual dictionaries and the code for

processing them). Matxin translates from Spanish to Basque (and from English to Basque). The authors (Mayor *et al.* 2011) declare that they designed this system for assimilation purposes (to be useful for Basque readers to understand Spanish text) although it has also been used by volunteers in marathon-like events to populate the Basque Wikipedia.[27] Unlike Apertium (see above), Matxin works by performing a deep morphological and syntactical analysis of the source sentence, which delivers a parse tree (a mixed dependency/constituency tree).[28] This is needed because of the stark morpho-syntactic divergence between English or Spanish on the one hand and Basque on the other hand. The source parse tree, encoded as an XML structure, is transformed into a target-language parse tree which is then used to generate the Basque sentence. Transformations use rich linguistic knowledge: for instance, verb sub-categorization frames—describing the arguments taken by each verb together with their case—are used to inform the choice of the correct translation of prepositions and verb chunks, and the order of words in the target language is determined independently of that of words in the source language. Matxin has also been used as a platform to perform machine translation research, for instance on hybrid rule-based/statistical machine translation (España-Bonet *et al.* 2011), and its extension to language pairs not involving Basque has also been explored (Mayor and Tyers 2009). The free/open-source version of Matxin (which has a more reduced vocabulary than the one that may be used on the web) is distributed under the GNU General Public License and may be installed by moderately expert users on computers running the GNU/Linux operating system.

OpenLogos

OpenLogos[29] is the free/open-source version (released by the German Research Center for Artificial Intelligence DFKI)[30] of a historical commercial machine translation system, Logos, which was developed by the Logos Corporation from 1970 to 2000. OpenLogos translates from German and English into French, Italian, Spanish and Portuguese. It uses an incremental or cascaded pipeline structure, and a particular internal symbolic representation called SAL ('semantico-syntactic abstract representation'). The designers of OpenLogos (Barreiro *et al.* 2011) claim that this endows the system with the ability to deal with ambiguity and other particularly hard related cognitive problems in ways which differ from those of other rule-based systems, and which, according to the authors, are inspired in neural computation (a subfield of artificial intelligence). The authors argue that OpenLogos is unique in the way it applies rules to the input stream, and that this makes its customization to application-specific needs very easy with the tools provided with the system. OpenLogos is also distributed under the GNU General Public License and may only be installed by expert users on the GNU/Linux operating system. Public development in the project site[31] appears to have ceased around 2011.

Research systems

There are a few other free/open-source rule-based systems, but they are experimental and are not widely distributed or packaged to be easily installed. An example of these research systems has been recently described by Bond *et al.* (2011). Their system tackles the problem of translation as one of meaning preservation, by using precise grammars for both the analysis of the source language and the generation of the target language, an explicit semantic representation of language meaning and statistical methods to select the best translation among those possible. Using only free/open-source components, they describe the building of a Japanese–English

system with slightly better human evaluation results than a Moses statistical machine translation system trained on a suitable corpus.

Data-driven or corpus-based free/open-source machine translation systems

Due to their nature, data-driven or corpus-based free/open-source machine translation systems usually require the existence of sentence-aligned parallel corpora containing translations of good quality, related to the texts one aims to translate, and usually in very large amounts, substantially larger than the usual translation memories used in computer-aided translation. This means that *training* is a necessary step before one starts to translate.

The following paragraphs describe the most famous free/open-source corpus-based system, the statistical machine translation system Moses, in detail, and then goes on to briefly describe other systems.

Moses

Moses (Haddow 2012) is a statistical machine translation system; it is probably the most widely used and installed free/open-source machine translation system. Hieu Hoang, then a PhD student in the University of Edinburgh, started it as a successor to a freely available but not free/open-source research system called Pharaoh[32] (Koehn 2004). The first version of Moses was made available in 2005, and one can say it became an instant success, partly because the license used (the GNU Lesser General Public License,[33] a partially-copylefted license, unlike the GPL which is *fully copylefted*), was perceived as being more adequate for commercial usage. Moses development has successfully attracted funding from the European Commission.[34]

As Pharaoh, Moses provides a *decoder*, that is, a program that performs the translation, but also a series of training tools to process the sentence-aligned parallel texts (parallel corpus) and extract the necessary information (the translation table) for the decoder.[35] Moses is what in statistical machine translation jargon is called a *phrase*-based system: new translations are performed by breaking down the source sentences into smaller units called *phrases* (sequences of words, not necessarily *phrases* in the linguistic sense) and assembling the available translations of these smaller units, stored as *phrase* pairs in the translation table, into the *most likely* translation, according to the probabilistic information stored with the *phrase* pairs as well as a probabilistic model of the target language. Recent versions of Moses can also learn advanced ('hierarchical') models of translation equivalence, in which phrase pairs are embedded. Examples of phrase pairs would be *basically depends on* (*depend fondamentalement de*) or *statistical machine translation* (*traduction automatique statistique*).

Moses is not a monolithic system, and integrates external components:

- Training relies on existing tools such as *word aligners*, which are trained on the parallel corpus to establish translation links between the words in the source sentence and those in the target sentence, and then used to extract the *phrase* pairs. The preferred word aligner in Moses is GIZA++ (Och and Ney 2003), which is also free/open-source (under the GNU General Public License). Moses used to rely also on external software to train and use probabilistic models of the target language but now it has one of them fully integrated as part of the main *decoder*.
- Models trained with Moses depend on a few parameters. One can run a Moses-trained system with preset values of these parameters, but one usually sets a small part of the training corpus apart (a *development* corpus) and uses it to automatically *tune* it so that

translation performance on that corpus is maximized. This means that Moses has to rely on software that computes *automatic evaluation measures* which compare the output of the system with the *reference* translation given in the reference corpus. Some of this software (for example, the one that computes BLEU, one of the most widely used such measures) comes bundled with Moses, but one can easily integrate other evaluation measures perceived as giving a better indication of translation quality.

- Training can avail of the linguistic information provided by morphological analysers or part-of-speech taggers (such as the ones available from the Apertium system above), by using 'factored models' that *factor* in that linguistic information to help in the probabilistic estimation. Moses can also use source-language and target-language parsers to help train *hierarchical phrase*-based models in which *phrase* pairs may contain *variables* that represent a set of smaller *phrase* pairs. For instance the phrase (X *depends on*, *depend* X *de*) has a variable *X* that could be instantiated by the phrase pairs (*basically, fondamentalement*) or (*radically*, *radicalement*) to obtain translations such as (*basically depends on*, *depend fondamentalement de*) or (*radically depends on*, *depend radicalement de*).

Moses has become the *de facto* baseline system in machine translation research: performance of new systems is always compared with that of a Moses-trained statistical machine translation system. This may be due to the free/open-source nature of Moses, which allows unrestricted usages and therefore could be seen not only as a readily-available research platform but also as a path towards industrial or commercial applications.

Moses may be installed in GNU/Linux-based and in Windows-based systems, and recently successful use of the *decoder* in Android-based devices has been reported. The standard distribution of Moses requires certain expertise to install; however, there are initiatives like *Moses for Mere Mortals*[36] or even commercial installers[37] to make installation easier. Web services such as LetsMT!,[38] KantanMT,[39] or SmartMATE[40] offer Moses training and translation online. Many companies offer translation services powered by Moses-trained systems.

Other systems

Joshua

Joshua[41] (Li *et al.* 2009) is a free/open-source statistical machine translation *decoder* that implements *hierarchical* machine translation models (also implemented in Moses, see above for details). One important feature of Joshua is that it is written in Java, which makes it possible to install in most operating systems through the available Java support. Installing Joshua requires a certain level of expertise, and one has to install other packages in advance, as it only provides a decoder. Like Moses, Joshua is distributed under the GNU Lesser General Public License.

Cunei

Cunei[42] (Phillips 2011) is a free/open-source corpus-based machine translation platform which combines the traditional example-based MT and statistical MT paradigms, allowing the integration of additional contextual information in the translation process. Like Moses, Cunei translates by segmenting the new source sentence in all possible contiguous sub-segments ('phrases') and looking for examples where they appear; in contrast to Moses, it then takes advantage of the context (document type, genre, alignment probability, etc.) in which the examples were found by using the distance between the new sentence and each of the actual

matching examples in the example base; in this way Cunei builds features that model the relevance of the 'phrase' and that are tuned on a development set as usual in statistical MT. Cunei can be considered a research system requiring a certain level of expertise to install it and train it; unfortunately, its designer Aaron Phillips has given up maintaining it as of July 2012. Cunei is distributed using the MIT License, a non-copylefted free/open-source license (see 'Free/open-source software' above).

CMU-EBMT

CMU-EBMT[43] (Carnegie-Mellon University Example-Based Machine Translation, Brown 2011) may be called a classical example-based MT system; that is, one capable of learning a lexicon, performing word and phrase alignment, and indexing and looking up a corpus. CMU-EBMT, however, also uses some typical statistical MT techniques such as target-language modelling, decoding, and parameter tuning. Like Cunei above, installation and usage requires a certain level of expertise. CMU-EBMT is distributed under the GNU General Public License, and has not released any new version since May 2011.

Marie

Marie[44] (Crego *et al.* 2005) provides a statistical machine translation decoder which is an alternative to the phrase-based decoder used in Moses: it uses *bilingual language modelling*, that is, it models translation as the search for the best sequence or chain of sub-sentential translation units, using an explicit statistical model to model sub-chains of length *N* called *N*-grams. Marie is distributed under the GNU General Public License but has not released any versions since 2005.

Great

Great (Gonzalez and Casacuberta 2011) also resorts to *bilingual language modelling* like Marie above, but it uses general probabilistic finite-state models instead of N-grams. The authors report competitive results with standard *phrase*-based models like Moses (see above) which were obtained faster and using a smaller amount of memory for the statistical translation model. A recent version, iGreat,[45] enhanced to be used in interactive machine translation environments, is distributed under the GNU General Public license.

PBMBMT

The Tilburg University phrase-based memory-based machine translation system[46] (PBMBMT), implements a kind of example-based machine translation in which the translation of new words and *phrases* in the sentence is modelled as the search for the *k* best equivalents in the given context using a classifier based on the 'nearest neighbor' strategy, using a suitable distance or dissimilarity measure. The system is available under the GNU General Public License and is suited primarily for research purposes.

OpenMaTrEx

OpenMaTrEx[47] (Dandapat *et al.* 2010) is a basically a wrapper around Moses that uses alternative example-based methods (based on the idea of *marker words*) to extract linguistically motivated

phrase pairs, which can be added to the Moses phrase tables. OpenMaTrEx is licensed under the GNU General Public License. The last release was in May 2011.

Challenges for free/open-source MT technology

Here are two main challenges faced by free/open-source MT technology:

User-friendliness: Available free/open-source machine translation technologies are diverse and cover the two main paradigms: rule-based and corpus-based. However, most of the systems require a fair level of expertise, on the one hand, to install and set-up (which involves training in the case of corpus-based machine translation), and on the other hand, to execute (for instance, most of them assume a command-line interface, i.e., offer no graphical interface). Their integration in the usual professional environment (e.g. to be used from computer-aided translation software in conjunction with translation memories) has barely started. This lack of end-user friendliness is surely delaying the adoption of systems that have obtained substantial public and private funding, and is one of the main challenges faced.

Unification: The multiplicity of systems, each one having different requirements, installation procedures, and user interfaces is another major hindrance to users who would like to switch technologies when choosing tasks to obtain the best possible results. Integrating all the free/open-source systems in a single platform with unified installation procedures, standardized application-oriented interfaces, and user-friendly interfaces is another big challenge, which should in principle be easier to tackle as free/open-source software is involved, but remains basically open.

Notes

1 The reader will find a definition at http://www.gnu.org/philosophy/free-sw.html.
2 Access to the source is not necessary for users to run the program.
3 See http://www.opensource.org/docs/definition.php. The concept of 'open-source' is more of an operational, business-friendly, concept that avoids the political overtones usually associated to the position of the Free Software Foundation.
4 Which gives rise to a common acronym: FLOSS, free/libre/open-source software.
5 The notion of software commons draws from the existing meaning of *commons* as 'a piece of land subject to common use'; its current usage refers to a body of related free/open-source software which is shared and developed by an online community, and also to systems that host that commons to allow this sharing and development, such as SourceForge (http://www.sourceforge.net) or GitHub (http://www.github.com).
6 http://www.gnu.org/licenses/gpl.html.
7 http://www.apache.org/licenses/LICENSE-2.0.
8 http://opensource.org/licenses/MIT.
9 http://opensource.org/licenses/BSD-3-Clause.
10 AutomaticTrans (http://www.eng.automatictrans.es), SDL Language Weaver (http://www.sdl.com/products/sdl-enterprise-language-server/).
11 This last requirement may sound strange to some but is actually the SMT analog of distributing linguistic data for a RBMT system.
12 This is the case, for example, of three non-commercial but freely available machine translation systems between Spanish and Catalan: interNOSTRUM (www.internostrum.com), which has thousands of daily users, and two less-known but powerful systems called SisHiTra (http://sishitra.iti.upv.es. González *et al.* 2006) and N-II (http://www.n-ii.org).
13 http://translate.google.com.
14 http://www.bing.com/translator.
15 http://wiki.apertium.org/wiki/Apertium-Caffeine.
16 http://wiki.apertium.org/wiki/Apertium_On_Mobile.
17 http://www.omegat.org.

18 http://wiki.apertium.org/wiki/Apertium-OmegaT.
19 http://www.statmt.org/moses/?n=Development.GetStarted.
20 For a list of free/open-source machine translation software, see http://www.fosmt.org.
21 The last Machine Translation Marathon, the eighth one, was held in September 2013: http://ufal.mff.cuni.cz/mtm13/.
22 The last FreeRBMT was held in Sweden in 2012: https://www.chalmers.se/hosted/freerbmt12-en.
23 http://www.apertium.org.
24 'Stable' does not imply any claim about the quality of the translations; it is rather a development concept referring to the fact that those language-pair packages do not contain any internal errors or inconsistencies, and do not produce any problems when used with the Apertium engine. Forcada *et al.* (2011) report evaluation results for selected language pairs.
25 With speeds in the range of tens of thousands of words a second in regular desktop computers.
26 http://matxin.sourceforge.net.
27 http://eu.wikipedia.org/wiki/Wikiproiektu:OpenMT2_eta_Euskal_Wikipedia.
28 The source-language syntactical parser in Matxin was written specifically for this system and is currently part of the (also free/open-source) Freeling language analysis toolbox (http://nlp.lsi.upc.edu/freeling; Padró and Stanilovsky 2012).
29 There is another free/open-source project called OpenLogos which has no relation to this machine translation system.
30 http://sourceforge.net/apps/mediawiki/openlogos-mt.
31 http://www.sourceforge.net/projects/openlogos-mt.
32 http://www.isi.edu/licensed-sw/pharaoh.
33 http://www.gnu.org/copyleft/lesser.html.
34 Projects EuroMatrix, EuroMatrix+ and MosesCore.
35 As well as many other tools, for instance, to tune the statistical models for maximum performance in a held-out portion of the training set.
36 http://code.google.com/p/moses-for-mere-mortals.
37 http://www.precisiontranslationtools.com.
38 http://www.letsmt.eu.
39 http://www.kantanmt.com.
40 https://www.smartmate.co.
41 http://joshua-decoder.org.
42 http://www.cunei.org.
43 http://sourceforge.net/projects/cmu-ebmt.
44 http://www.talp.upc.edu/index.php/technology/tools/machine-translation-tools/75-marie.
45 http://sourceforge.net/projects/igreat.
46 http://ilk.uvt.nl/mbmt/pbmbmt.
47 http://www.openmatrex.org.

References

Arnold, Doug J., Lorna Balkan, R. Lee Humphreys, Seity Meijer, and Louisa Sadler (1994) *Machine Translation: An Introductory Guide*, Manchester and Oxford: NCC Blackwell. Available at: http://www.essex.ac.uk/linguistics/external/clmt/MTbook.

Barreiro, Anabela, Bernard Scott, Walter Kasper, and Bernd Kiefer (2011) 'OpenLogos Machine Translation: Philosophy, Model, Resources and Customization', *Machine Translation* 25(2): 107–126.

Bond, Francis, Stephen Oepen, Eric Nichols, Dan Flickinger, Erik Velldal, and Petter Haugereid (2011) 'Deep Open-source Machine Translation', *Machine Translation* 25(2): 87–105.

Brown, Ralf D. (2011) 'The CMU-EBMT Machine Translation System', *Machine Translation* 25(2): 179–195.

Crego, Josep M., José B. Mariño, and Adrià de Gispert (2005) 'An Ngram-based Statistical Machine Translation Decoder', in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005)*, 3193–3196.

Dandapat, Sandipan, Mikel L. Forcada, Declan Groves, Sergio Penkale, John Tinsley, and Andy Way (2010) 'OpenMaTrEx: A Free/Open-source Marker-driven Example-based Machine Translation System', in Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir (eds) *Advances in Natural*

Language Processing: 7th International Conference on NLP, IceTAL 2010, 16–18 August 2010, Reykjavík, Iceland, 121–126.

España-Bonet, Cristina, Gorka Labaka, Arantza Diaz de Ilarraza, Lluis Màrquez, and Kepa Sarasola (2011) 'Hybrid Machine Translation Guided by a Rule-based System', in *Proceedings of the 13th Machine Translation Summit*, 19–23 September 2011, Xiamen, China, 554–561.

European Commission (2012) *ICT – Information and Communication Technologies: Work Programme 2013*, Luxembourg: EU Publications Office.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers (2011) 'Apertium: A Free/Open-source Platform for Rule-based Machine Translation', *Machine Translation* 25(2): 127–144.

González, Jorge and Francisco Casacuberta (2011) 'GREAT: Open Source Software for Statistical Machine Translation', *Machine Translation* 25(2): 145–160.

Haddow, B. (coord. 2012) *Moses Core Deliverable D.1.1: Moses Specification*. Available online: http://www.statmt.org/moses/manual/Moses-Specification.pdf.

Hutchins, W. John and Harold L. Somers (1992) *An Introduction to Machine Translation*, London: Academic Press. Available online at: http://www.hutchinsweb.me.uk/IntroMT-TOC.htm.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Cristine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007) 'Moses: Open Source Toolkit for Statistical Machine Translation', in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, June 2007, Prague, Czech Republic, 177–180.

Koehn, Philipp (2004) 'Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation', in Robert E. Frederking and Kathryn B. Taylors (eds) *Machine Translation: From Real Users to Research*, Berlin: Springer Verlag, 115–124.

Koehn, Philipp (2010) *Statistical Machine Translation*, Cambridge: Cambridge University Press.

Li, Zhifei, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, JonathanWeese, and Omar F. Zaidan (2009) 'Joshua: An Open Source Toolkit for Parsing-based Machine Translation', in *Proceedings of 4th Workshop on Statistical Machine Translation (StatMT '09)*, 135–139.

Mayor, Aingeru and Francis M. Tyers (2009) 'Matxin: Moving towards Language Independence', in Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, and Francis M. Tyers (eds) *Proceedings of the 1st International Workshop on Free/Open-Source Rule-based Machine Translation*, 2–3 November 2009, Alacant, Spain, 11–17.

Mayor, Aingeru, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola (2011) 'Matxin: An Open-source Rule-based Machine Translation System for Basque', *Machine Translation* 25(1): 53–82.

Och, Franz Josef and Hermann Ney (2000) 'Improved Statistical Alignment Models', in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, 1–8 October 2000, Hong Kong, China, 440–447.

Och, Franz Josef and Hermann Ney (2003) 'A Systematic Comparison of Various Statistical Alignment Models', *Computational Linguistics* 29(1): 19–51.

Padró, Lluís and Evgeny Stanilovsky (2012) 'FreeLing 3.0: Towards Wider Multilinguality', in Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds) *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 23–25 May 2012, Istanbul, Turkey, 2473–2479.

Pedersen, Ted (2008) 'Empiricism Is Not a Matter of Faith', *Computational Linguistics* 34(3): 465–470.

Phillips, Aaron B. (2011) 'Cunei: Open-Source Machine Translation with Relevance-based Models of Each Translation Instance', *Machine Translation* 25(2): 161–177.

Streiter, Oliver, Kevin P. Scannell, and Mathias Stuflesser (2006) 'Implementing NLP Projects for Non-central Languages: Instructions for Funding Bodies, Strategies for Developers', *Machine Translation* 20(4): 267–289.

9

PRAGMATICS-BASED MACHINE TRANSLATION

David Farwell

FORMERLY OF THE CATALONIA POLYTECHNIC UNIVERSITY, SPAIN

Stephen Helmreich

FORMERLY OF NEW MEXICO STATE UNIVERSITY, THE UNITED STATES

Introduction

There are three basic computational strategies for fully automatic Machine Translation (MT).

1 The algorithm, using a stochastic model induced from large amounts of extant data, substitutes source language surface forms (strings) for target language surface forms.
2 Alternatively, using linguistic rule bases, it converts source language surface forms into some sort of abstract representation (possibly syntactic, ideally semantic), which it then manipulates in such a way as to make it suitable for generating a target language equivalent.
3 Or, using a knowledge base, context modeling and inferencing engine, it converts the text into a representation of what the author intended to communicate and then uses that representation as the target interpretation for the translation generated.

The first of these strategies is referred to as the *direct method* (and its current instantiations as *statistical* or *example-based* MT) while the second is referred to as the *transfer method* (or, in some cases, where the representation is in fact semantic, the *interlingual method*). The third strategy alone may be referred to as Pragmatics-Based MT (PBMT) since it is in fact attempting to model the communicative intent associated with a text as opposed to its linguistic form or content *per se*.

Pragmatics-based MT, then, relies on:

- the use of context to interpret source language utterances and to produce target language equivalents;
- a context of the utterance which crucially includes nested beliefs environments which are constructed and modified through ascription during processing;
- a discourse context consisting of a knowledge base which is accessed for constructing or modifying the utterance context;
- context-sensitive (non-monotonic) inferencing within the context to resolve ambiguities during interpretation or to select expressions during production.

Within pragmatics-based approaches to machine translation, it is important to note that there are three distinct levels of analysis: that of identifying the intended information content of an utterance (the locutionary level); that of identifying the underlying communicative intent of an utterance (the illocutionary level); and that of identifying the intended effect of the utterance (the perlocutionary level) (Austin 1975; Searle 1969). In particular, the illocutionary level includes more than simply the type of speech act. It also includes the intended implications of the information content as well.

These levels may be illustrated in the following exchange between the characters Coulomb and Constance, in a scene from the film *Jesus of Montreal* (1989). Coulomb, an actor and playwright, has been living in the apartment of his good friend and collaborator Constance while revising the script of a Passion play. He normally is out all day researching the life of Christ at a local library or organizing the staging of the play itself. But on this day he returns early to the apartment. After entering, he takes off his coat and tries unsuccessfully to hang it up as he proceeds to the back of the apartment where there is a bookshelf with various books of interest. The coat, falling from the hook, knocks something over, making a bit of noise. Shortly thereafter Constance emerges from her bedroom in a robe. Coulomb, thinking he may have found her in a compromising situation asks, whispering, if she would like him to leave. She shakes her head while turning away, cracks open the bedroom door and says in a calm voice, *Ben, écoutes, sors* (Ok. Listen. Come on out), *On va pas jouer une scène de Feydeau* (We are not acting out a scene from Feydeau).

This last line in particular is relevant to the discussion because its English subtitle is 'This isn't a bedroom farce' which is not only a possible translation but a very good one. This is because, while at the locutionary level 'We are not acting out a scene from Feydeau' is perfectly appropriate, at the illocutionary level and therefore at the perlocutionary level, it is quite likely to fail. The reasoning is as follows. Constance wishes to let the person in the bedroom (who is a priest as it turns out and thus might well fear being caught in a compromising situation) know that there will be no negative consequences in his being discovered. So, since Constance knows that Feydeau was a turn of the (twentieth) century playwright who was famous for writing bedroom farces and Constance also believes that the person in the bedroom knows who Feydeau was, in saying 'We are not acting out a scene from Feydeau' she is in fact telling the priest that the situation in which they find themselves is not one in which he needs to conceal his identity, i.e., a scene in a bedroom farce, a scene typical of a play by Feydeau. In saying this to the priest, Constance is trying to allay his fears of coming out.

What is interesting here in regard to the subtitling is that the translator feels that the English-speaking audience of the film will likely not know who Feydeau is or that he was famous for writing bedroom farces. Therefore, they will not understand the illocutionary intent or intended perlocutionary effect of what Constance said. So instead of translating on the basis of the locutionary level, the subtitle is based on the illocutionary level which presumably the English-speaking audience would understand. Were it the case that the meaning at the illocutionary level were equally opaque, i.e., the audience is unlikely to know what a bedroom farce is, the translator might have chosen to base the subtitle directly on the intended perlocutionary effect, perhaps something like 'Do not worry about being discovered, it is safe to come out'.

In the remainder of this chapter,

1 a pragmatics-based approach to machine translation is put forward;
2 the theoretical framework for developing such an approach is presented;

3 the requirements for a computational platform on which to implement such a system are outlined;
4 various advantages to the approach are discussed; and
5 some serious difficulties with the approach are addressed.

The objective is, on the one hand, to provide an introduction to pragmatics-based machine translation and, on the other, to demonstrate the importance of pragmatics for high-quality MT.

Motivation

Theoretical motivation

The theoretical motivation for pursuing pragmatics-based approaches to machine translation derives from linguistic phenomena which can only be accounted for by positing an extra-linguistic context. These include such processes as resolving reference, recovering ellipted information, interpreting metonymy and metaphor and resolving lexical, semantic and syntactic ambiguity. These are, of course, mainly problematical as source language phenomena and therefore prerequisite for arriving at an accurate interpretation, but they also may come into play as target language phenomena and relevant for formulating a fluent translation.

Resolving reference, for instance, is especially relevant when translating into a language whose pronominal anaphors reflect gender from a language whose anaphors are gender neutral. For example, the Spanish translation of:

> The gardeners used dull chain saws to prune the trees and so several of them were damaged.

might be:

> Los jardineros utilizaron motosierras desafiladas para podar los arboles así que **varios de ellos** fueron dañad**os**.

or it might be:

> Los jardineros utilizaron motosierras desafiladas para podar los arboles así que **varias de ellas** fueron dañad**as**.

The choice depends on whether it was the saws or the trees that were damaged, that is to say, depending on the referent for 'several of them'. What is crucial however is that the decision (saws or trees) is based on one's knowledge of typical scenarios related to pruning trees. It is likely, for instance, that it was the trees rather than the saws that were damaged since dull chain saws tend to rip and burn limbs rather than cut them. But such reasoning can easily be overridden should later (or previous) context happen to shed additional light on the situation (see Farwell and Helmreich 2000: 1–11 for further discussion).

As for recovering ellipted information, it is often necessary when translating from a language that favors nominalization and/or noun compounding into a language that tends to favor oblique or sentential complements. For example, the translation into Spanish of:

… *the expansion of US small business investment* …

is either:

… *la expansión de las inversiones* ***de las pequeñas*** **empresas** en los EEUU …

or

… *la expansión de las inversiones* ***en las pequeñas*** **empresas** en los EEUU …

Here, the choice depends on whether *small business investment* is understood as 'investment on the part of small business' or 'investment in small business'. That interpretation, in turn, will depend on what the translator understands to be the state of the world at the time that the description was intended to apply.

Pragmatics is equally crucial for interpreting metonymy and metaphor. For instance, in order to translate *is rippling across* in the following text, it is prerequisite that the translator understand that the expression is being used metaphorically to express the notion that a particular practice (not a phenomenon that is literally given to wavelike behavior) is spreading from school to school just as small waves spread out from some initial epicenter. It would not be good practice to translate:

… *the story of how it [incorporating leadership skills into core curricula] all started and why* ***it is rippling across*** *the globe* ….

as:

… *la historia de cómo empezó todo y por qué* ***se está ondulando por*** *todo el mundo* ….

since the reader might well become confused about the meaning. Rather, the translator would better serve the audience of the translation by unpacking the metaphor along the lines of:

… *la historia de cómo empezó todo y por qué* ***se está extendiendo por*** *todo el mundo* ….

This is perhaps more prosaic but certainly communicates more clearly the original intent of the source text.

The need for pragmatics can similarly be motivated in situations involving the interpretation of metonymy or lexical, syntactic or semantic disambiguation. In fact, there is ample theoretical motivation for developing pragmatics-based approaches to MT.

Empirical motivation

Still, some would argue that the sorts of phenomena discussed here are not all that common in the mundane world of newspaper articles, parliamentary proceedings, business communication and so on and that the know-how and labor required for implementing such a system far exceeds the benefits that would be derived. Nonetheless, there is important empirical motivation for pragmatics-based approaches to MT as well.

Such motivation can be found in the comparative analysis of multiple translations of a given text by different translators. Placing such translations side by side, it is possible to identify

equivalent translation units and to observe the addition or omission of information and variations in the translators' perspectives on what the author of the original text intended to communicate and which aspects of that information are viewed as more or less important. It is possible to identify as well any outright errors on the part of the translators.

For a limited range of texts, such an analysis has been carried out (see Helmreich and Farwell 1998: 17–39, for a more detailed account). From a corpus of 125 Spanish language newswire articles and their translation into English by two independent professional translators, three sets of articles and translations were aligned on the basis of translation units and the translations were compared in terms of literal graphological equivalence.[1] If comparable translation units were graphologically the same, these units were set aside. If the units were graphologically different, the difference was attributed to one of three categories: error on the part of one of the translators, alternative paraphrastic choices, or variation in the underlying beliefs of the translators about the author and audience of the source language text, the audience of the translation, or about the world in general.

By way of example, the following is a headline from one of the articles in the data set:

Source language text:

Acumulación de víveres por anuncios sísmicos en Chile

Translation 1:

Hoarding Caused by Earthquake Predictions in Chile

Translation 2:

STOCKPILING OF PROVISIONS BECAUSE OF PREDICTED EARTHQUAKES IN CHILE

First the data is segmented into translation units and equivalents: (1) *Acumulación de víveres / Hoarding / STOCKPILING OF PROVISIONS*; (2) *por / Caused by /BECAUSE OF*; (3) *anuncios sísmicos / Earthquake Predictions / PREDICTED EARTHQUAKES*; (4) *en Chile / in Chile/ IN CHILE*. Next the translation equivalents are compared at the graphological level. Here, the first three pairs differ but the fourth pair, aside from capitalization, is the same. As a result the fourth pair is set aside.

In regard to each of the remaining pairs, the next step is to categorize them as either due to translator error (here there are no examples), or due to alternative choices of paraphrase (here, on the face of it, exemplified by the second pair) or due to differing beliefs about the source text author's intent or about the target audience's background knowledge (here exemplified by the first and third pairs).

Finally, differing chains of reasoning leading to the differing translations must be inferred. For instance, the choice of 'hoarding' indicates that the first translator believes the agents of the action are behaving selfishly and irrationally whereas as the choice of 'stockpiling' indicates that the second translator believes that the agents of the action are behaving calmly and rationally. The choice of 'earthquake predictions' indicates that the first translator believes that it is the predictions that are the reason for the action whereas the choice of 'predicted earthquakes' indicates that the second translator believes that it is possible earthquakes that are the cause of the action. Together, the differing translations set the reader up with rather different expectations of what the article is about.

The result of the analysis of a small data set consisting of roughly 352 phrasal translation units was that 142 units, or roughly 40 percent, differed with respect to at least one internal element. Altogether, there was a total of 184 differences of which 160 were lexical and 24 were syntactic. Of the 184 differences, the source of 27, or 15 percent, could be attributed to translator errors (9 to carelessness – roughly divided equally between the translators, 13 to source language interference – again roughly divided equally between the translators, and 5 to mistranslations). The source of 70 differences, or 38 percent, could be attributed to paraphrastic variation (including 60 lexical and 10 syntactic). Finally, the source of 75 differences, or 41 percent, could be attributed to differing assumptions on the part of the translators (67 being related to the interpretation of the source language text, and 8 being related to assumptions about the target audience's background knowledge).[2]

Thus, far from being uncommon, translation variations derived from differing beliefs on the parts of the translators account for 41 percent of all variations and 16 percent of all units translated. This is a significant portion of the translator's output and should be recognized as such.[3] The fact is that pragmatics is every bit as crucial in processing language and in translating from one language to another as lexis, morpho-syntax and semantics.

Theoretical framework

PBMT aims, at some level of abstraction, to model the human translation process. A human translator implicitly ascribes knowledge and beliefs to the author of the source language text or utterance, to the addressees of that text, and to the intended audience of the translation. The translator then uses these models as a basis for reasoning about the original intent of the source text and about the appropriateness of any translation. Therefore, a pragmatics-based approach needs to model the translator's ascription of relevant beliefs about the world (both shared and unshared) to (a) the author of the source text, including the author's beliefs about the beliefs of the addressees of that text, and (b) the audience of the translation.

Needless to say, this model is dynamic, changing with each new text segment processed, since the translator assumes that the author assumes that the addressees of the text have all updated their views of the author's view of the world (if not their own view of the world) in accordance with the author's intended interpretation of that text segment. In addition, the updated view will include any potential inferences that may be expected to follow from the locutionary content. The translator also assumes that the intended audience of the translation will have updated its view of the author's view of the world (if not their own view of the world) in accordance with the intended interpretation of the translation as well as any potential inferences that may be expected to follow from it.

Thus the goal is to model the translator's knowledge of the world (including the translator's knowledge of the two linguistic systems and associated sets of cultural conventions relevant to the translation at hand), the translator's view of the author's knowledge of the world and of the intended audience of the source text (including knowledge of the source language and associated cultural conventions) and the translator's view of the world knowledge of the target audience of the translation (including knowledge of the target language and associated cultural conventions).

These knowledge models are then used as background knowledge to support the two component tasks of the translation process, interpreting what the author intended to express by way of the source language text and then formulating an expression in the target language in such a way as to communicate that intended content to the degree possible to the audience of the translation in as a similar a manner as possible.

Discourse and utterance contexts

This background knowledge (or *discourse context*) is the source of all the beliefs that enter into play during interpretation and restatement. It may include beliefs about specific people, places and past events, about types of people, objects and events, about the source language and communicative strategies, about the social and cultural conventions of the setting of the communication and so on.

By way of concrete illustration, consider the following text and translations from the data set described previously that was used for empirically motivating PBMT. It comes from an article about the booming real estate market in Moscow in the early 1990s and here the author quotes a Moscow real estate agent who is talking to a prospective buyer. The agent says:

> *… los 300 metros cuadrados d**el tercer piso** estaban disponibles pero fueron aquilados …, sólo queda **el segundo piso** ….*

While one translator rendered this excerpt as:

> *… the 300 square meters of **the third floor** were available …, but they were rented …. All that is left is **the second floor** ….*

the second translated it as:

> *… the 300 square meters on **the fourth floor** were available, but they were rented …; only **the third floor** remains ….*

It is important to note that, despite appearances, both translations are appropriate and both are potentially accurate. They may even be expressing the same information. The reason for the apparent contradiction is that the two translators have differing beliefs about the story naming conventions that enter into play either during interpretation or during restatement or both. The specific conventions in question are:

> **Convention 1:** In Europe and elsewhere, people refer to the story that is at ground level as the ground floor, to the next level up as the first floor, to the level after that as the second floor and so on.

> **Convention 2:** In the United States and elsewhere, people refer to the story that is at ground level as the first floor, the next level up as the second floor, the level after that as the third floor and so on.

Very briefly, while the first translator (*tercer piso | third floor*, *segundo piso | second floor*) assumes that the floor naming convention is the same for both the source language participants and the audience of the translation, the second translator (*tercer piso | fourth floor*, *segundo piso | third floor*) assumes that Convention 1 applies during interpretation (possibly because the building is in Moscow) and that Convention 2 applies during restatement (possibly because the intended audience of the translation is American).

The knowledge that has been introduced during the actual discourse is the *utterance context*. It also is a source for the beliefs that enter into play during interpretation and restatement. It constitutes the foreground knowledge as well as serving as the interpretation of the discourse

thus far. It is the knowledge with which new information must be made consistent and coherent. It includes beliefs about the objects and events mentioned or implied during prior discourse and about the current state of the discourse, especially any unresolved issues (those objects, situations and events whose connection to the interpretation have yet to be established).

With respect to our example, the utterance context includes information that the author had mentioned in the article prior to the current text fragment such as:

- the commercial real estate market in Moscow is rapidly expanding
- there is a great demand for commercial properties
- properties are renting at $700 to $800 per square meter per year
- properties are renting at the highest prices in the world apart from Tokyo and Hong Kong
- the market is dominated by legal uncertainty and the usual result that the rich get richer.

While it is not entirely clear how this knowledge affects the interpretation or restatement of the text fragment in question, it may be that the second translator decided to apply Convention 1 during interpretation because he or she assumed that it is the convention that Muscovites follow.

Interpretation

The process of interpretation may be described in a semi-formal manner as follows.

To begin, a source language expression, E_i (e.g., *el tercer piso*), is provided with a semantic representation, p (i.e., third floor), which needs to be interpreted, i.e., added to the current source-language utterance context (call it $SLUC_h$) in a coherent fashion.[4]

Next, any beliefs that provide information relevant to the interpretation are inferred on the basis of beliefs drawn from the utterance and discourse contexts. For instance, it may be that b_4 (that the author is referring to the fourth story of the building) follows from p (third floor) and b_1 (Convention 1), resulting in one particular updated source language interpretation, $SLUC_{i1}$, or that b_5 (that the author is referring to the third story) follows from p (third floor) and b_2 (Convention 2), resulting in a second updated source language interpretation, $SLUC_{i2}$. Or it may be that b_6 (that the author is referring to some other story) follows from p (third floor) and b_3 (some other convention), resulting in yet a third updated source language interpretation, $SLUC_{i3}$. Finally, from these possible interpretations, that which is most compatible with the prior utterance context is selected as the intended updated interpretation ($SLUC_i$), e.g., $SLUC_{i1}$ by the second translator and either $SLUC_{i1}$ or $SLUC_{i2}$ (or neither) by the first translator.

Restatement

The process of restatement is more complex. First, the beliefs used to support the interpretation of the source text segment need to be identified. Next these beliefs must be integrated into the target language utterance context in order to produce the intended interpretation of the translation. Then, a translation is formulated which expresses that intended interpretation.

Somewhat more formally, the source language utterance context as it stood prior to interpreting E_i is first subtracted from the utterance context resulting from the interpretation of E_i, i.e., $SLUC_{i1}$ or $SLUC_{i2}$, thus isolating only those beliefs that were used to arrive at $SLUC_{i1}$, or $SLUC_{i2}$, (e.g., b_4 – that the author is referring to the fourth story – and b_1 – Convention 1, or b_5 – that the author is referring to the third story –and b_2 – Convention 2, respectively). That is to say, what are left are only those additional beliefs and inferences that were necessary in order to establish $SLUC_{i1}$ or $SLUC_{i2}$ given p.

Next, in order to formulate the translation, the intended interpretation of the yet to be formulated translation is partially created by updating the target language utterance context using the set of new beliefs (i.e., b_4 and b_1, or b_5 and b_2) so long as these are compatible with the target language discourse and utterance contexts. Note that, in the case of the second translation, b_1 may not be compatible with the discourse context of the audience of the translation.

At this point a target language expression, $E_i^{\star}$, having the semantic representation, $p^{\star}$, is formulated such that its interpretation in the context of the target language interaction is equivalent to that derived in the source language interaction, i.e., b_4 – that the author is referring to the fourth story, or, alternatively, b_5 – that the author is referring to the third story. This then results in the updated utterance context for the translation, $TLUC_i$, as well as the expression used for the translation, $E_i^{\star}$, along with its semantic interpretation, $p^{\star}$.

Equivalence

Within a pragmatics-based approach, there are two basic notions of equivalence. First, there is ***core equivalence***. This is measured in terms of the similarity of authors' intentions as represented by the source language interpretation that was used as a basis for the translation, e.g., $SLUC_i$, and the translator's intentions as represented by the preferred interpretation of the translation, e.g., $TLUC_i$ (for example, the levels above ground level and the beliefs and inferences required to identify it). The greater the overlap and consistency of the beliefs making up these two interpretations, the greater the core equivalence between the source language text and translation.

In addition to core equivalence, there is a broader notion of ***extended equivalence*** which is measured in terms of the pairwise similarity (and difference) between the other possible interpretations of the source language text and translation (e.g., $SLUC_{i1}$, $SLUC_{i2}$, … *vs* TLI_1, TLI_2, …). Here, equivalence would be based not simply on the primary interpretations of the source language text and translation but on all the other possible interpretations of the source language text and of their translations as well.

Computational platform

The computational platform needed for a PBMT system is necessarily complex. Since the novel part of the system involves inferencing to and from the intended perlocutionary effect using contextual information, the source text input will be represented as a set of logical propositions. This, in turn, suggests an interlingual approach to MT as a substrate for any pragmatic reasoning, since such approaches provide a source text with a meaning representation. Thus, the basic requirements include:

- a knowledge base
- a beliefs ascription component
- a default reasoning component
- an analysis component
- a generation component.

The knowledge base

There are a number of interlingual MT systems that could serve as a substrate, including Nyberg and Mitamura (1992: 1069–1073); Dorr *et al.* (1995: 221–250); Levin *et al.* (2000: 3–25); Mitamura and Nyberg (2000: 192–195), etc., which differ somewhat in the nature of

the interlingual representation, but more so in the semantic depth of that representation. The deeper the representation, the more useful it is for pragmatic reasoning. One of the deeper systems is that of Mahesh and Nirenburg (1995) and Nirenburg and Raskin (2004). It includes the following modules, which are assumed as an appropriate basis for a PBMT system:

- Ontology – a conceptual knowledge base of objects, properties, relations and activities
- Onomasticon – a database of proper names of people, places and other objects
- Fact Database – episodic knowledge of people, places and events.

Together, these modules determine the discourse context. Linguistic knowledge bases for each of the relevant languages used for providing meaning representations for expressions and generating expressions given meaning representations include the lexicons and lexical, syntactic and semantic rules.

Pragmatic reasoning components

Logical inferencing for speech act analysis

Given a semantic interpretation of an input (or set of such interpretations), the pragmatic reasoning components aim to identify the illocutionary and the perlocutionary intents, that is, what the author was *doing* in producing the input expression and how the author intended the audience to react.

To reach these interpretive levels requires a default reasoning mechanism. For example, the application of a story-naming convention in the example in the section on theoretical motivation must be the result of default reasoning, since there is neither concrete factual information about the particular convention used by the author or audience of the translation nor is there some absolute generalization that can allow certainty in this specific case. However, there are default generalizations ('Europeans generally use Convention 2') as well as other default inferences ('The article was written in Castilian, so the author is likely to be European') that would allow a default conclusion to be reached about what convention to adopt in interpreting the source text. Similar reasoning would exist for choosing the convention to apply for the target audience. One such default inferencing system is Att-Meta (Barnden *et al.* 1994a: 27–38, 1994b), which also deals with metaphorical input. Other possibilities might include Nottelmann and Fuhr (2006: 17–42), Motik (2006), Hustadt *et al.* (2007: 351–384), to name a few more recent systems.

A speech act identifier such as Levin *et al.* (2003) is used as a first step in identifying a plausible speech act. However, a PBMT system must be able to identify indirect speech acts, and be alert to unusual sequences of overt speech acts, such as, for instance, two successive questions by two different interlocutors. If a question such as 'Do you know where bin Laden is?' is answered with another question, it might be a request for clarification of the question (e.g., 'Do you mean Osama bin Laden or Mohammed bin Laden?'), or it might be a request for clarification of the questioner's perlocutionary intent ('Why do you want to know?' or 'What makes you think I know the answer?'), or it might even be a refusal to answer ('What do you take me for, a GPS system?').

In addition, it should be clear that the 'speech act' level of representation of a source language expression might be several degrees removed from that of its semantic content. For example, in interpreting a yard sign that says, *'*Only an ass would walk on the grass', not only must the speech act be recognized as a command rather than an informative statement, but the content of the command must be clear as well: *Please do not walk on the grass*.

Belief ascription for participant modeling

Finally, PBMT requires that the translator's view of the differing beliefs of the source language author and addressees as well as those of the target language audience guide translation choices.

Therefore, a simple, univocal reasoning system is insufficient. The system must have the capacity to attribute different beliefs to at least these three participants in the translation process: the source language author and addressees and the target language audience. In some translation scenarios, it may even be necessary to attribute differing beliefs to the people described in the source language text. Otherwise, if all the beliefs of all the participants are merged into a single model, it would undoubtedly be inconsistent, allowing anything to be deduced.

One possible belief ascription system that could be used for participant modeling is the ViewGen system (Ballim and Wilks 1991). This system creates complex embedded belief contexts, or viewpoints, that allow each participant to reason independently. Within the viewpoint of each participant, a different set of assumptions can be held about each topic of conversation, allowing for conflicting beliefs. In addition, participants may have their own views of other participants' viewpoints (which may or may not correspond to that participant's viewpoint). Such embedded viewpoints allow, for instance, for a translator to reason about a source text author's reference to a story in a building and then later reason about whether or not the audience of the translation could work out the same referent. Other approaches have been described by Maida (1991: 331–383), Alechina and Logan (2002: 881–888) and more recently Alechina *et al.* (2009: 1–15), Alechina and Logan (2010: 179–197), Wilks (2011: 337–344), and Wilks *et al.* (2011: 140–157).

Generator

Once the system has identified the meaning of an input text, the likely speech act and illocutionary intent, and the perlocutionary intent of the author, this information is then used to generate a translation.

Given the lack of information about how to incorporate pragmatic information in the production of an output sentence (Goodman and Nirenberg 1991; Hovy 1993: 341–386), the PBMT system falls back on a generate-and-test paradigm.

The system first attempts a translation of the semantic content of the utterance, using the IL substrate. The semantic representation of this target text is then placed in the target audience's belief space and the attempt is made to deduce an illocutionary and perlocutionary intent. If the results are the same as those of the original text, the translation may stand.

However, if unsuccessful, a translation is produced on the basis of the illocutionary level interpretation and the test is repeated. If the representation of the perlocutionary intent for the translation is equivalent to that of the source language document, it is assumed to be an acceptable translation.

If not, as a last resort, a translation based on the representation of the perlocutionary intent would be sent to the generator.

Benefits of pragmatics-based machine translation

Evidence provided thus far indicates that human translators translate on the basis of the communicative intent behind the source language text, as opposed to solely on the basis of its meaning. The previous section discussed how a computational platform could be constructed that would emulate this translation process. However, the question remains whether this

system could solve some of the difficult problems that arise from pragmatic phenomena such as reference resolution, the interpretation of metaphor and metonymy, the recovery of ellipted information, or morpho-syntactic and lexical disambiguation. For each of these problem areas a PBMT system should be expected to perform more adequately than other types of MT.

In addition, quantitative analysis of multiple translations of the same texts indicates that approximately 16 percent (1/6) of all translation units in a text differ because of differing beliefs. This holds true whether the study is of entire texts (Farwell and Helmreich 1997a: 125–131, 1997b) or of selected aspects of multiple texts (Helmreich and Farwell 2004: 86–93). In such situations, neither translation is wrong, but each variant is based on differing analyses of the intent of the author or of the common knowledge of the audience of the translation. The remainder of this section focuses on such situations, showing how a PBMT system would deal with them.

User-friendly translation

In some cases, the analysis of the source document may provide reasonable illocutionary and perlocutionary intents, but the reasoning that led to that analysis cannot be reconstructed in the target language due to the lack of crucial beliefs on the part of the audience of the translation. Compensating for that missing information in the formulation of the translation is what has been referred to as *user-friendly translation* (Farwell and Helmreich 1997a: 125–131).

An example of this is the translation of *On va pas jouer un scène de Feydeau*. As indicated in the introduction, reaching the appropriate level of interpretation requires inferencing from the playwright Feydeau to the kinds of plays he wrote, namely bedroom farces. Such information is likely to be lacking on the part of the English-speaking audience. Therefore, following the generation plan outlined in the previous section, the PBMT system would replace the text based on the locutionary content with a text based on the representation of the illocutionary intent, i.e., 'this is not a bedroom farce'.

A similar situation arises when the illocutionary and perlocutionary intents depend on an inference from the form of the utterance, rather than the content. Examples would include poetry (alliteration, rhyme, assonance) and, frequently, jokes such as puns.

In both these cases, the intent of the source language author is two-fold. In poetry, the author, on the one hand, intends that the content is understood and yet, on the other, the author intends to create an emotional (perhaps subliminal) effect through the poetic devices used. If the corresponding content of the target language text does not employ words with a similar effect, then that intended poetic effect fails to be captured. A PBMT system that could translate poetry would be able to identify such devices and select words and phrases in the target language that produce the same effect to the degree possible, perhaps even sacrificing aspects of the locutionary content.

In the case of puns and other jokes, the humor may depend on a reference to two different scenarios at the same time (Raskin 1984), often through the use of ambiguous words or phonetically similar patterns. For example, the joke:

Why do sharks swim in salt water?

Because pepper water makes them sneeze.

depends on the dual language-specific oppositions between salt water/fresh water and salt/pepper. In French, the expression for 'salt water' is *eau de mer* (sea water) and its opposite is *eau*

douce (sweet water), so a literal translation of 'pepper water' would fail to be humorous. A PBMT system should be capable of recognizing this dual intent, and providing an alternative translation of 'pepper water' as, perhaps, 'eau de cologne' while at the same time substituting *stink* for *sneeze* (see Farwell and Helmreich 2006 for further discussion).

Alternatively, inferencing in the target language belief space might not fail and yet produce a different communicative intent from that originally intended by the author of the original text. In such cases, a PBMT system needs to adjust the translation in such a way as to insure that the desired inferences are made, blocking any unwanted inference while accommodating an alternative but appropriate default belief.

An example of this phenomenon is the switching of story-naming conventions during target language generation as described earlier in the section on restatement. The differing default beliefs in the source and target cultures (not necessarily languages) result in differing illocutionary intents in regard to which story is actually being referred to, given the same semantic content. In such cases, the PBMT system would be capable of replacing the translation based on the locutionary content with one that would have the same illocutionary intent, given the different default conventions.

A generalization of this case involves any system of measurement or counting which differs across languages or cultures: metric versus English measures, various temperature scales (Fahrenheit, Centigrade, Kelvin), shoe sizes, dress sizes, numbering of Biblical Psalms and the Ten Commandments, musical keys, etc.

A measuring system includes a number of conceptual items: the content (what is to be measured, e.g., temperature), a starting point (e.g., the freezing point of water), a scale itself (for continuous measures, e.g., the temperature shift measured by one unit), a name for the system (e.g., Fahrenheit, Centigrade, etc.), and names for the scales themselves (e.g., for musical scales, the letters A, B, C, … G). Most frequently these conceptual items are not named explicitly in the text. Temperature system, for instance, is usually not mentioned so it must be inferred from the discourse and utterance contexts.

If a reference is made to a temperature of 32 degrees, the PBMT system could, for example, determine by default reasoning that the source language audience would interpret the temperature as degrees Fahrenheit, while the target language audience prefers to use Centigrade measurements. Then two translation options are available: (1) clarify the source language usage by inserting 'Fahrenheit' into the translation, or (2) translating the reference from '32 degrees' to '0 degrees' (Centigrade).

Translation based on high-level beliefs

The situations described in the previous section required adjustments of the translation in order to accommodate differing beliefs on the part of the audience of the translation and to support the inferencing needed for arriving at the desired interpretation. However, some of the more interesting variations between human translations may be due to the translators' attempt to maintain cohesion at the level of the text as a whole. Because of this, differences in the translators' assessments of what the overarching perlocutionary intent of the author of the source text is result in different patterns of lexical choice, each of which tends to cohere on the basis of connotation.

For instance, one article from the corpus described in the section on empirical motivation concerned the murder trial in Brazil of a person who was alleged to have killed a union leader (see Farwell and Helmreich 1997b for details). In this case, the two translations differed according to whether the translator believed the trial involved a straightforward murder or

involved a political assassination. Thus, the words derived from the Spanish *asesinar* (i.e., *asesinado*, *asesino*, etc.) were consistently translated by words derived from 'murder' or 'kill' by one translator and by words derived from 'assassinate' by the other, depending on their global viewpoint. Similarly, the victim was described as a 'union member' by the first translator or as a 'labor leader' by the second. The surrounding conflict was described as between 'landholders' and 'small farmers' by the first translator or as between 'landowners' and 'peasants' by the second. In each case the connotations of the words selected for the first translation imply a simple criminal trial whereas the connotations of the words selected for the second translation imply a trial due to and influenced by politics. Overall, these two patterns of lexical selection account for some 60 percent of the beliefs-based differences in the translations.

Another article from the same corpus described people buying up supplies in response to press stories about a possible future earthquake (see Farwell and Helmreich 1997a: 125–131 for details). In this case, the translators seemed to differ as to whether the article was about a government lapse in which an overly precise statement about an impending earthquake led to a rational response by people to stockpile supplies or about a reasonable government statement that was overblown by the local media, resulting in irrational hoarding. These two views are epitomized by the translations of the article's headline *Acumulación de víveres por anuncios sísmicos* as either: *Hoarding Due To Earthquake Prediction*, on the one hand, or as: *Stockpiling because of Predicted Earthquakes*, on the other. As mentioned before, the choice of 'hoarding' indicates that the first translator believes the agents of the action are behaving selfishly and irrationally while the choice of 'earthquake predictions' indicates it is a prediction that is the reason for their action. On the other hand, the choice of 'stockpiling' indicates that the second translator believes that the agents of the action are behaving calmly and rationally, while the choice of 'predicted earthquakes' implies that it is a possible earthquake that is the cause of the action. These general tendencies to select expressions that, on the one hand, exaggerate the panic of the people, to minimize the likelihood of an earthquake, and place the role of the press in a negative light by the first translator or that, on the other, convey the measured reaction of the people, maximize the likelihood of an earthquake, and place the role of government in a negative light by the second translator, are prevalent throughout the translations. In fact, the two differing assumptions about the source text author's general perlocutionary intent account for roughly half the beliefs-based differences in the translations.

In both these texts, the translators reached different conclusions about the over-arching perlocutionary intent of the author of the source text. These conclusions were arrived at through inferencing largely on the basis of implicit (not explicit) propositions or beliefs. In fact, there are several reasonable chains of reasoning that might link semantic content to illocutionary intent to perlocutionary intent. Because of this, PBMT systems must be capable of finding several differing inference chains since any one of them might potentially guide the generation of an alternative coherent translation.

Finally, the translator reaches conclusions not only about the beliefs of the source language author and addressees and target language audience, but also about the purpose of the translation itself. The translator will take into account whether the translation is for assimilation or dissemination, whether or not a particular detail is vital to an argument, whether or not the target audience will be scrutinizing the translation for each detail (as in the case of a piece of evidence or a legal document) or whether or not some details are irrelevant in the end. In this latter case, it is likely that some inferencing would be avoided altogether (such as, for instance, exactly which story of the Moscow building was being rented in the example in 'Discourse and utterance contexts').

Evaluation

In regard to the evaluation of MT quality, ideally it should be based on the notion of equivalence discussed earlier. As noted, this is defined by the similarity and difference between the interpretations of the source language text and the translation in terms of their component beliefs and inferences. But if the goal is to measure pragmatic equivalence, objective evaluation is complicated to say the least. Automatic evaluation at this point is completely implausible. Subjective evaluation might appear to be more plausible but, even so, it will be exceedingly problematical on the one hand, given the open-endedness of the inferencing process, and will clearly require a great deal of human effort on the other. Nevertheless, to broach the subject, it is first important to understand what constitutes pragmatic equivalence and then attempt to develop evaluation methodologies that are sensitive to that specific notion.

The actual comparison of interpretations, whether for core equivalence or extended equivalence, may be quantitative, based simply on the number of beliefs and inferences used to arrive at the interpretations of the source text or its translation. Or, the comparison may be qualitative, based on the propositional content of those beliefs and inferences used, their 'currency', the 'transparency' of their connection to or their consistency with the source language and target language utterance and discourse contexts.

In a perfect world, evaluation would consist of automatically or manually inspecting the number and similarity of content of the beliefs and the inferences used. To do this requires that those beliefs and inferences be explicitly represented. Yet, to date, no pragmatics-based MT system exists and, even if there were one, no reference translations exist whose underlying beliefs and inferences have been made explicit.

Currently, automated evaluation methodologies are entirely inappropriate for capturing any notion of equivalence based on the similarity of interpretations of the source language text and translation. These techniques are based on comparing MT output with human produced reference translations in terms of overlapping character sequences. This tells us nothing at all about beliefs and inferences that make up the interpretations compared. Still, were there a PBMT system that produced translations, no doubt its output could be compared alongside any other MT output. The problem would merely be that there would be no way of knowing whether a given translation having a relatively low rank was in fact inferior to a translation with a higher rank. For instance, clearly 'We are not acting out a scene from Feydeau' will not score as well against 'This is not a scene from a bedroom farce' as a reference translation, or vice versa, and yet neither is obviously an inferior translation.

As for manual evaluation, there is the very time consuming and human intensive sort of analysis that was carried out during the comparative analysis of multiple translations described in the section on empirical motivation. This process could in theory be made more objective by increasing the number of evaluators, although even this may not be entirely feasible – there are simply too many possible alternative interpretations and inferred connections to expect the analysts to always agree on a particular set. Still, traditional human evaluation methodologies such as those developed by such organizations as the American Translation Association (http://www.atanet.org/certification/aboutexams_overview.php), though human intensive, should at least be amenable to tolerating differing translations arising from differing interpretations on the part of the translators or translation systems. Perhaps not altogether practical and probably not so objective, they are nonetheless more reliable and flexible (for further discussion see Farwell and Helmreich 2003: 21–28).

Critical analysis

In addition to the problems concerning evaluation described in the previous section, pragmatics-based MT systems face a number of other serious challenges that must be met before they will have sufficient coverage to be deployed. These challenges are found both in the representation of linguistic phenomena and knowledge of the world as well as in the modeling of the translation process and the implementation of various crucial components.

To begin, the representation of many linguistic phenomena, particularly semantic and pragmatic, is often weakly motivated and incomplete. For example, one important area of overlap between semantics and pragmatics is reference resolution. In part the task of resolving reference requires deciding whether a given reference is to specific entity (real or imaginary) in the context of a description of a particular event as opposed to some generic entity in the context of a generalization. To wit, in:

The elephant (that we keep in our back yard) *is eating the grass*.

the elephant is being used to refer to some particular individual. By contrast, in:

Elephants eat grass.

elephants is being used to refer to some indefinitely large set of arbitrary individuals. Where the former sentence is being used to describe a particular event, the latter is being used to make a generalization, to describe a situation that, in the abstract, could happen.

The problem for representation systems is how to capture the semantics and pragmatics of reference. There are many possible taxonomies which might be offered but none that is generally accepted. One system, that was motivated by referring expressions in English (Farwell and Wilks 1991: 19–24), entails a fourfold classification that includes: (1) the pragmatic distinction between referring to a specific entity or entities, as in the first example above, or referring generically, as in the second example; (2) the semantic distinction between referring to a particular individual or individuals (*the elephant*, *the elephants*) or to an arbitrary individual (*an elephant*, *elephants*) and (3) the pragmatic distinction between whether a reference is to individuals that the speaker presumes the addressee can identify (*the elephant*, *that elephant*) or to individuals that the speaker presumes the addressee cannot identify (*an elephant*, *some elephants*). Finally, a fourth deictic distinction is made in the case of references to specific, particular, identifiable individuals which is whether that individual is nearby (*this elephant*), remote (*that elephant*) or neutral (*the elephant*).

Although this system is motivated by English referring expressions, the intent is to provide a general framework for classifying any reference made in any language and deciding whether or not corresponding references in a text and its translation are equivalent. Unfortunately, the extent to which the framework is successful is unclear since it can be difficult to identify where it is correct and where it fails, or whether its distinctions are too fine-grained or not fine-grained enough. Thus, while perhaps useful for comparing references in languages which generally mark types of reference explicitly, such as English or Spanish, it is difficult to apply and leads to a good deal of ambiguity in languages in which most types of reference are implicit, such as Chinese. It simply is not well enough defined or well enough evaluated for deciding such issues. Unfortunately, many other crucial descriptive taxonomies are only generally but vaguely understood, including those for time/tense, aspect, modality, voice, case, grammatical relations, functional relations, rhetorical relations, speech acts and so on.

In addition to being weakly motivated and incomplete, representations of linguistic phenomena and knowledge of the world are essentially arbitrary and often vague. As a result, they are not very amenable to annotation or evaluation. It can be rather difficult to confidently categorize concrete instances with the result that training annotators, preparing training materials, as well as the task of annotation itself are at best arts, not sciences. For instance, suppose you are attempting to categorize the meaning of *proposed* in:

> The Czech Minister of Transportation, Jan Strasky, proposed on Saturday that the State buy back the shares held by Air France.

given DECLARE_A_PLAN, PRESENT_FOR_CONSIDERATION, INTEND or ASK_TO_MARRY as possible choices. While ASK_TO_MARRY is relatively easy to rule out, a case might be made for any of the other three choices with the result that different annotators might well make different choices. This results in lower inter-annotator agreement, which, in turn, reduces our confidence in the quality of the annotated corpus. For an extensive discussion of text annotation for interlingual content and the evaluation of inter-annotator agreement, see Dorr *et al*. 2008: 197–243.

In addition to the challenges of adequate representation, PBMT faces a number of serious deficiencies in terms of the computational infrastructure[5] available today. Among the more important are the size and granularity of the Ontology, the ability to reason with uncertainty within an open-world context and the open-endedness of beliefs ascription. While progress has clearly been made in the last decade in all these areas, yet there remains much to be done.

Notes

1 The translations were done as part of the evaluation corpus for a US government run Machine Translation evaluation (White *et al*. 1994) and so were carefully supervised, done professionally, and the translators were given identical instructions about the translation – they were neither to add nor delete any information in the translation.

2 In addition there were 12 differences that were a consequence of translator choices about other elements within the same translation unit.

3 This is especially interesting since the translators were instructed to keep the translations as 'close' as possible, maintaining lexical and structural equivalence to the degree possible.

4 Actually it may have various semantic representations that will be filtered down to one, partly on the basis of establishing a coherent connection to prior text.

5 It should be pointed out that no PBMT systems exist today. A small prototype system, ULTRA (Farwell and Wilks 1991), was abandoned prior to implementation of the required context modeling and inferencing mechanisms. The Mikrokosmos KBMT system (Mahesh and Nirenburg 1995) was certainly an important step in the direction of PBMT but has never been fully developed nor thoroughly tested.

References

Alechina, Natasha and Brian Logan (2002) 'Ascribing Beliefs to Resource Bounded Agents', in *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '02)*, 15–19 July 2002, Bologna, Italy, 2: 881–888.

Alechina, Natasha and Brian Logan (2010) 'Belief Ascription under Bounded Resources', *Synthese* 173(2): 179–197.

Alechina, Natasha, Brian Logan, Hoang Nga Nguyen, and Abdur Rakib (2009) 'Reasoning about Other Agents' Beliefs under Bounded Resources', in John-Jules Ch. Meyer and Jan Broersen (eds) *Knowledge Representation for Agents and Multi-Agent Systems: First International Workshop, KRAMAS 2008, Sydney, Australia, September 17, 2008, Revised Selected Papers*, Berlin: Springer Verlag, 1–15.

Austin, J.L. (1975) *How to Do Things with Words*, Cambridge, MA: Harvard University Press.

Ballim, Afzal and Yorick Wilks (1991) *Artificial Believers: The Ascription of Belief*, Hillsdale, NJ: Lawrence Erlbaum.

Barnden, John, Stephen Helmreich, Eric Iverson, and Gees Stein (1994a) 'An Integrated Implementation of Simulative, Uncertain and Metaphorical Reasoning about Mental States', in Jon Doyle, Erik Sandewall, and Pietro Torasso (eds) *Principles of Knowledge Representation and Reasoning: Proceedings of the 4th International Conference*, 24–27 May 1994, San Mateo, CA: Morgan Kaufmann Publishers, 27–38.

Barnden, John A., Stephen Helmreich, Eric Iverson, and Gees C. Stein (1994b) 'Combining Simulative and Metaphor-based Reasoning about Beliefs', in *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum, 21–26.

Dorr, Bonnie J., Joseph Garman, and Amy Weinberg (1995) 'From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT', *Machine Translation* 9(3–4): 221–250.

Dorr, Bonnie J., Rebecca J. Passonneau, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith J. Miller, Teruko Mitamura, Owen Rambow, and Advaith Siddharthan (2008) 'Interlingual Annotation of Parallel Text Corpora: A New Framework for Annotation and Evaluation', *Natural Language Engineering* 16(3): 197–243.

Farwell, David and Yorick Wilks (1991) 'ULTRA: A Multilingual Machine Translator', in *Proceedings of the Machine Translation Summit III*, 1–4 July 1991, Seattle, WA, 19–24.

Farwell, David and Stephen Helmreich (1997a) 'User-friendly Machine Translation: Alternate Translations Based on Differing Beliefs', in Virginia Teller and Beth Sundeim (eds) *Proceedings of the MT Summit VI: Machine Translation: Past, Present, Future*, 29 October – 1 November 1997, San Diego, CA, 125–131.

Farwell, David and Stephen Helmreich (1997b) 'Assassins or Murderers: Translation of Politically Sensitive Material', in *Proceedings of the 26th Annual Meeting of Linguistics Association of the Southwest*, October 1997, University of California at Los Angeles, Los Angeles, CA.

Farwell, David and Stephen Helmreich (2000) 'An Interlingual-based Approach to Reference Resolution', in *Proceedings of ANLP/NAACL 2000 Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP*, 30 April 2000, Seattle, WA, 1–11.

Farwell, David and Stephen Helmreich (2003) 'Pragmatics-based Translation and MT Evaluation', in *Proceedings of Workshop on Machine Translation Evaluation: Towards Systematizing MT Evaluation* at the 9th Machine Translation Summit (MT Summit IX), 23–27 September 2003, New Orleans, LA, 21–28.

Farwell, David and Stephen Helmreich (2006) 'Pragmatics-based MT and the Translation of Puns', in Jan Tore Lønning and Stephen Oepen (eds) *Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT)*, 19–20 June 2006, Department of Computer Science, Oslo University, Oslo.

Goodman, Kenneth and Sergei Nirenberg (1991) *The KBMT Project: A Case Study in Knowledge-based Machine Translation*, San Mateo, CA: Morgan Kaufmann Publishers.

Helmreich, Stephen and David Farwell (1998) 'Translation Differences and Pragmatics-based MT', *Machine Translation* 13(1): 17–39.

Helmreich, Stephen and David Farwell (2004) 'Counting, Measuring, Ordering: Translation Problems and Solutions', in Robert E. Frederking and Kathryn B. Taylor (eds) *Machine Translation: From Real Users to Research: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004*, 28 September – 2 October 2004, Washington, DC/Berlin: Springer Verlag, 86–93.

Hovy, Eduard (1993) 'Automated Discourse Generation Using Discourse Structure Relations', *Artificial Intelligence: Special Issue on Natural Language Processing* 63(1–2): 341–386.

Hustadt, Ullrich, Boris Motik, and Ulrike Sattler (2007) 'Reasoning in Description Logics by a Reduction to Disjunctive Datalog', *Journal of Automated Reasoning* 39(3): 351–384.

http://www.atanet.org/certification/aboutexams_overview.php.

Jésus de Montréal (1989) Dir. Denys Arcand, Orion Classics, film.

Levin, Lori, Alon Lavie, Monika Woszczyna, Donna Gates, Marsal Galvadà, Detlef Koll, and Alex Waibel (2000) 'The Janus-III Translation System: Speech-to-speech Translation in Multiple Domains', *Machine Translation* 15(1–2): 3–25.

Levin, Lori, Chad Langley, Alon Lavie, Donna Gates, Dorcas Wallace, and Kay Peterson (2003) 'Domain Specific Speech Acts for Spoken Language Translation', in *Proceedings of the 4th SIGdial Workshop of Discourse and Dialogue*, 5–6 July 2003, Sapporo, Japan.

Mahesh, Kavi and Sergei Nirenburg (1995) 'A Situated Ontology for Practical NLP', in *Proceedings on the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence-95*, 19–21 August 1995, Montreal.

Maida, Anthony S. (1991) 'Maintaining Mental Models of Agents Who Have Existential Misconceptions', *Artificial Intelligence* 50: 331–383.

Mitamura, Teruko and Eric Nyberg (2000) 'The KANTOO Machine Translation Environment', in John S. White (ed.) *Envisioning Machine Translation in the Information Future: Proceedings of 4th Conference of the Association for Machine Translation in the Americas*, 10–14 October 2000, Cuernavaca, Mexico, 192–195.

Motik, Boris (2006) 'Reasoning in Description Logics Using Resolution and Deductive Databases', unpublished doctoral thesis, Universität Karlsruhe (TH), Karlsruhe, Germany.

Nirenberg, Sergei and Victor Raskin (2004) *Ontological Semantics*, Cambridge, MA: MIT Press.

Nottelmann, Henrik and Norbert Fuhr (2006) 'Adding Probabilities and Rules to Owl Lite Subsets Based on Probabilistic Datalog', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 14(1): 17–42.

Nyberg, Eric H. and Teruko Mitamura (1992) 'The Kant System: Fast, Accurate, High-quality Translation in Practical Domains', in *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, 23–28 August 1992, Nantes, France, 1069–1073.

Raskin, Victor (1984) *Semantic Mechanisms of Humor*, Dordrecht: D. Reidel.

Searle, John R. (1969) *Speech Acts: An Essay in the Philosophy of Language*, Cambridge: Cambridge University Press.

White, John S., Theresa A. O'Connell, and Francis E. O'Mara (1994) 'The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches', in *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas: Technology Partnerships for Crossing the Language Barrier* (*AMTA-1*), Columbia, MD, USA, 193–205.

Wilks, Yorick (2011) 'Protocols for Reference Sharing in a Belief Ascription Model of Communication', in *Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium (FS-11-01)*, 4–6 November 2011, Arlington, VA, 337–344.

Wilks, Yorick, Simon Worgan, Alexiei Dingli, Roberta Catizone, Roger Moore, Debora Field, and Weiwei Cheng (2011) 'A Prototype for a Conversational Companion for Reminiscing about Images', *Computer Speech and Language* 25(2): 140–157.

10
RULE-BASED MACHINE TRANSLATION

Yu Shiwen

PEKING UNIVERSITY, CHINA

Bai Xiaojing

TSINGHUA UNIVERSITY, CHINA

Introduction

In the field of computational linguistics, Rationalism and Empiricism prevailed alternately as the dominant method of research in the past few decades (Church 2011). The development of machine translation, accordingly, features the shift between these two methods.

Rule-based Machine Translation, abbreviated as RBMT and also known as Knowledge-based Machine Translation, relies on morphological, syntactic, semantic, and contextual knowledge about both the source and the target languages respectively and the connections between them to perform the translation task. The linguistic knowledge assists MT systems through computer-accessible dictionaries and grammar rules based on theoretical linguistic research. This rationalist approach contrasts with the empiricist approach, which views the translation process as a probabilistic event and therefore features statistical translation models derived from language corpora.

MT systems were generally rule-based before the late 1980s. Shortly after Warren Weaver issued a memorandum in 1949, which practically stimulated the MT research, Georgetown University and IBM collaborated to demonstrate a Russian–English machine translation system in 1954. The system was reported to work with six rules instructing operations such as selection, substitution, and movement, which applied to words and sentences in the sample. Despite being merely a showcase designed specifically for a small sample of sentences, this demonstration is still the first actual implementation of RBMT. In the years that followed, the rule-based approach dominated the field of MT research and was further explored, leading to a deeper division among the direct model, the transfer model, and the interlingua model. There was a change of direction from the direct model to the other two 'indirect' models as a result of the ALPAC report in 1966. Particularly in the latter half of the 1970s and the early 1980s, MT research revived from the aftermath of ALPAC and featured the predominance of the syntax-based transfer model (Hutchins 1994; 2004; 2010).

Whereas statistical analysis was found in the background, assisting rationalists when they took the center of the stage, corpus-based statistical methods emerged with considerable effectiveness at the end of the 1980s (Hutchins 1994). The rule-based approach, though

retreating to the background, developed its new framework using unification and constraint-based grammars (Hutchins 2010). The new millennium has seen more efforts on the development of hybrid MT, leveraging and combining the strengths of statistical models and linguistic rules. There are recent reflections on the rationalist positions being overshadowed in the past two decades and expectations for richer linguistic representations to be made better use of in machine translation (Church 2011).

The three models

The three basic models of RBMT start from different linguistic premises. All of them require source language (SL) analysis and target language (TL) synthesis, but with varying types, amounts, depths of analysis and accordingly different bases for synthesis. Figures 10.1, 10.2, and 10.3 depict the general design of the three models, in which dictionaries may include more than just lexical equivalents, and grammars refer broadly to all the necessary morphological, syntactic, semantic, and contextual rules.

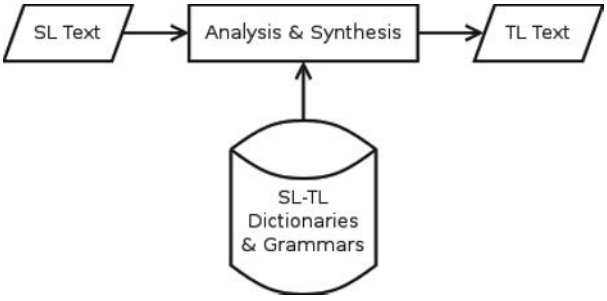


Figure 10.1 The direct model

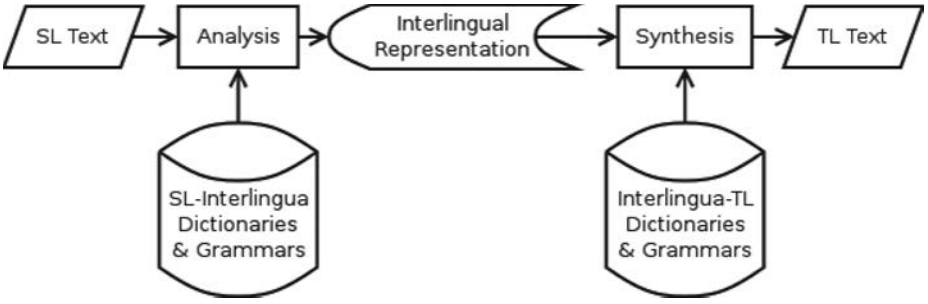


Figure 10.2 The interlingual model

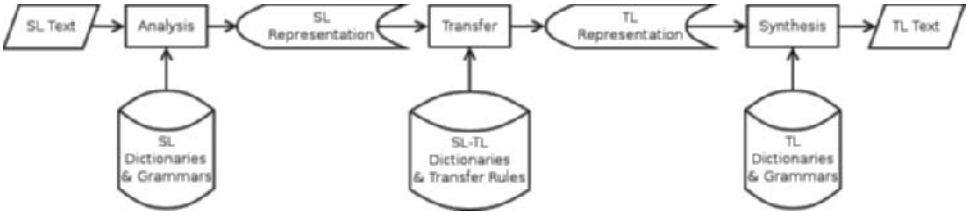


Figure 10.3 The transfer model

The direct model

The direct model is based on the assumption that translation tasks mainly require lexical transfer between the languages involved. It takes an SL sentence as a string of words, retrieves their lexical equivalents in TL from accessible bilingual dictionaries, and reorganizes these equivalents into the corresponding TL sentence. A minimal amount of morphological (and occasionally syntactic) analysis and word reordering are included, with relevant linguistic knowledge stored in the dictionaries or simply described by algorithms and then expressed by program codes. The direct translation model can be employed to handle a finite number of sentence patterns for language pairs with similar syntactic features.

Early designs of MT systems generally adopted this model, a typical example of which is the system developed at the University of Washington mainly during the 1950s. Sponsored by the U.S. Air Force, the Washington MT system produced word-for-word translation from Russian to English, with inadequate results. Bilingual dictionaries were so constructed that they assisted not only the selection of lexical equivalents but also the solution of other problems related to SL analysis and TL synthesis. For example, by searching the dictionary, the system was able to resolve the ambiguity arising from homographs. The Russian verb *dokhodyat*, which could correspond to *reach*, *ripen*, or *are done*, was only translated as *ripen* if followed by a noun labeled in the dictionary as fruit or vegetable. There were also entries in the dictionaries that gave rules for reordering the English output. (Hutchins 1986)

Nevertheless, the translation process was still simplified in the direct model. When the complexity of language became more recognized, and particularly after the ALPAC report was issued, more attention was diverted to the other two models that promised a closer look at the linguistic problems in translation.

The interlingual model

The interlingual model starts from the premise that semantico-syntactic intermediary representations can be found to link different languages, which take the form of interlingual symbols independent of both SL and TL. The translation process consists of two language-specific steps: SL analysis that leads to the conversion from SL texts to their interlingual representations and TL synthesis that produces TL texts based on the interlingual representations. The representations are expected to be unambiguous and express the full content of SL texts, which include their morphological, syntactic, semantic and even contextual information. Real projects and systems vary in their focus on the semantic or syntactic aspects of the texts.

A typical example of the interlingua-based MT is the Machine Translation System for Japan and its Neighboring Countries, the research and development of which started in 1987. Five languages were involved, including Chinese, Indonesian, Malaysian, Thai, and Japanese. The interlingua in this system was used to represent four kinds of information: events and facts, speaker's view, intension, and sentence structure (Tanaka *et al.* 1989). In most cases, interlinguas are designed for specific systems, but the DLT translation system developed in the Netherlands during the 1980s also adopted a modified form of Esperanto as the interlingua (Hutchins 1994).

The interlingual model offers an economical way to develop multilingual translation systems. It allows SL analysis and TL synthesis to work separately. Therefore, for a translation task among N languages, an interlingua-based MT system handles $2*N$ language pairs between the interlingua and the N languages, while the same translation task requires a direct translation system or a transfer-based system to handle $N*(N-1)$ language pairs. The advantage of the interlingua-based system increases when N is larger than 3. However, the difficulty in designing

an adequate interlingua is also evident, as the language-independent representations are supposed to cover various language-specific phenomena and categories.

The transfer model

The transfer model takes a sentence as a structure other than a linear string of words as it is taken in the direct model, and the syntactic view of the structure is commonly adopted. A sentence 他隨手寫了個字 (S1) in Chinese, for instance, may be treated as a combination of a noun phrase (NP) and a verb phrase (VP) as shown in Figure 10.4, where the noun phrase consists of a single possessive pronoun (PN), and the verb phrase consists of an adverb phrase (ADVP) formed by a single adverb (AD) and a verb phrase formed by another verb phrase and a noun phrase.

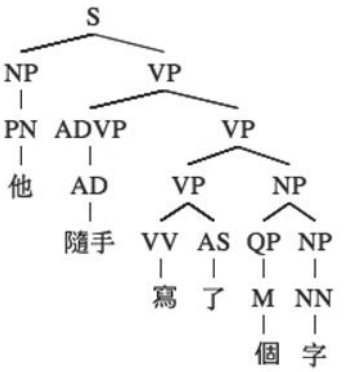


Figure 10.4 The syntactic tree for S1: 他隨手寫了個字

It is not possible to enumerate all the sentences in a natural language, which are actually infinite in number, but it is possible to find a finite number of structures that represent the vast majority of the sentences. Accordingly, it is not feasible to build an MT system that stores the TL translation for all possible SL sentences, but it is feasible to find for a finite number of structures in one language their equivalent structures in another language. Figure 10.5 presents the syntactic structure of the sentence *He roughly wrote a word* (S2) in English, which corresponds to the syntactic structure of S1. An important basis for the transfer model is the structural correspondence between the languages involved.

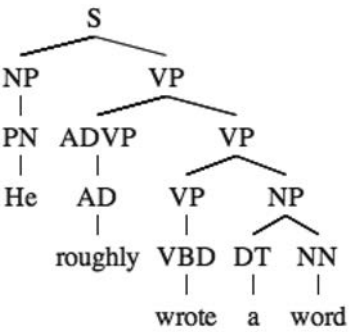


Figure 10.5 The syntactic tree for S2: *He roughly wrote a word*

It is not necessary, however, for the corresponding structures of two languages to be exactly the same. For example, the verb phrase 寫了 in S1 (see Figure 10.4) consists of a verb (VV) and an aspect marker (AS), while its corresponding English verb phrase in S2 (see Figure 10.5) is formed by a single verb in its past tense (VBD) *wrote*, which also illustrates the difference between Chinese and English in expressing aspect and tense. In S3, a more appropriate English equivalent of S1, the verb *scribbled* means "wrote (something) roughly", carrying the meaning of the adverb 隨手 as well, and the syntactic structure of S3 (see Figure 10.6), therefore, diverges further from that of S1.

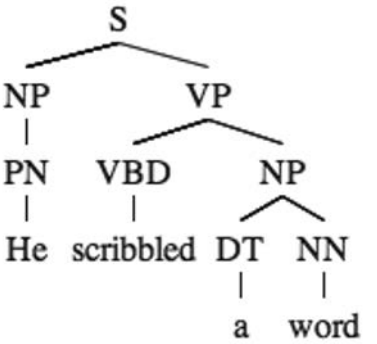


Figure 10.6 The syntactic tree for S3: *He scribbled a word*

The non-linear structural view leaves more room for the transfer model, which operates in three main steps: analysis, transfer, and synthesis. An SL sentence is first analyzed to yield an abstract representation of its internal structure, which is then converted to the equivalent TL representation through the transfer stage. The last step is to generate a surface structure for the TL representation, which is the TL translation for the SL input sentence. Representations in this model are specific to either SL or TL, and are thus different from those language-independent ones in the interlingual model. SL analysis carries much weight in the transfer model, which can be done on different levels – morphological, syntactic, semantic, and contextual.

Morphological analysis helps identify SL words, which is the first step toward any deeper analysis. In English and most European languages, words are generally separated by spaces. In English, tokenization is performed for contracted forms like *don't* and *we're*, for words followed by punctuation marks like commas and periods, for abbreviations like *U.S.A.*, etc. Lemmatization is performed to find the lemma for a given word in its inflected forms, which, for example, determines that *flies* is the third person singular form in the simple present tense of the lemma *fly*. Ambiguity arises, however, as *flies* is also the plural form of the noun lemma *fly*, which can either be resolved by the dictionary information on collocation or left for syntactic rules to handle. In Chinese, word boundaries have to be detected in the first place, as there are no spaces between words. Thus, S1 is segmented as 他/隨手/寫/了/個/字. Ambiguity arises again, for example, when the Chinese string 一個人 is analyzed, which can be either 一/個/人 or 一/個人, since 個人 is a justified word entry in the dictionary. To segment it properly, the occurrence of the numeral 一 before the measure word 個 has to be described either in a dictionary entry or by a rule. The grammatical categories of particular words are determined in syntactic analysis, but some morphological clues are also collected for that use. For example, the endings *-ing* and *-ed* in English generally signal verbs, and the endings *–ness* and *–ation* signal nouns.

Syntactic analysis helps identify the syntactic structure of an SL sentence: the grammatical categories of words, the grouping of them, and the relation between them. For S1, the MT

system consults the dictionary to determine the grammatical category of each word: 他 being a pronoun, 隨手 an adverb, 寫 a verb, 了 an aspect marker, 個 a measure word, and 字 a noun. Syntactic rules are then applied to build a syntactic tree for the sentence. Context-free grammar (CFG) is a formal grammar commonly used to describe the syntactic structure of sentences, which can be coded into computer program languages. The following are two sets of CFG rewrite rules describing the structures of S1 and S3 respectively.

Rule set 1:
S → NP+VP
NP→PN
VP→ADVP+VP
ADVP→AD
VP→VP+NP
VP→VV+AS
NP→QP+NP
QP→M
NP→NN

Rule set 2:
S → NP+VP
NP→PN
VP→VBD+NP
NP→DT+NN

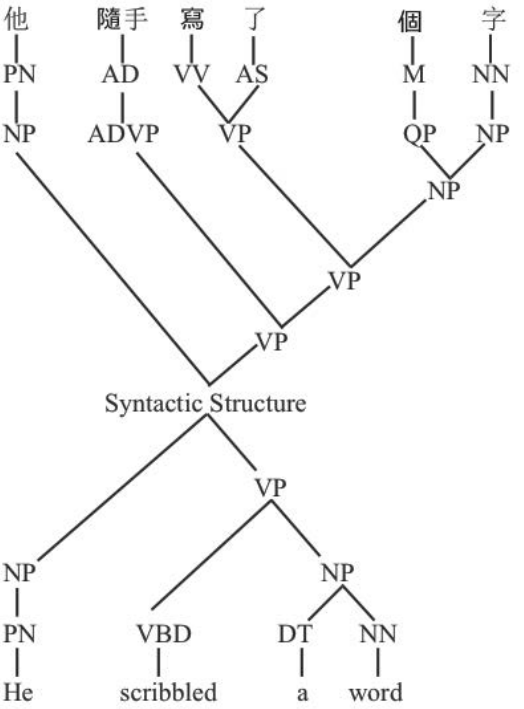


Figure 10.7 Syntactic transfer from S1 to S3

If SL analysis stops at the syntactic level, the SL syntactic tree will then be transferred to the corresponding TL syntactic tree. Figure 10.7 offers a better view of the structural correspondence between the two syntactic trees in Figures 10.4 and 10.6. Based on the TL syntactic structure, equivalent TL words are finally retrieved from the dictionary to form the TL translation. In most cases, appropriate morphological forms of TL words are to be derived.

There are, however, ambiguities left unresolved on the syntactic level, for which the semantic analysis of SL sentences is needed. S4 and S5, for example, have the same syntactic structure (see Figure 10.8). Ambiguity remains in the second noun phrase of each sentence, as the noun (NN) therein can be an argument of the verb (VV) but not always: in S4, 扒手 (pickpocket) is the patient of 破獲 (to capture), while in S5, 幻覺 (delusion) is not an argument specified by 迫害 (to persecute) at all. The two phrases are therefore rendered quite differently in English.

S4: 他像個被破獲的扒手。*(He looks like a captured pickpocket.)*

S5: 他有種被迫害的幻覺。*(He has the delusion of being persecuted.)*

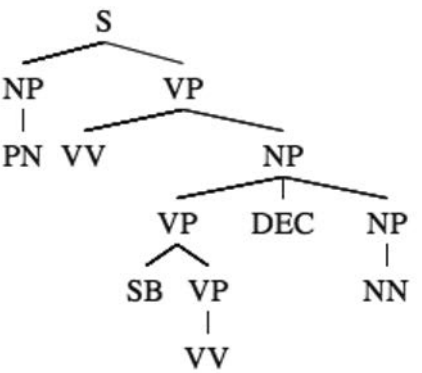


Figure 10.8 The syntactic tree for S4 and S5

To resolve such an ambiguity, dictionaries and rules are required in order to describe the semantic constraints of the verbs and the possible semantic connection between the verbs and the nouns in their context.

Difficulties and problems

Several decades of MT research and development have unarguably demonstrated the complexity of human natural languages and the complexity of the translation tasks between them. Difficulties and problems in RBMT also arise from these complexities. As the first substantial attempt at the non-numerical use of computers, MT received great attention and aroused high expectations. It is the ALPAC report that gave this attempt the first official evaluation and serious reflection, and as a consequence, disillusion with RBMT research appeared. Though the report was later viewed as biased (Hutchins 2010), it promoted more investigations on the structure and meaning of language and more research on the indirect rule-based models that involved closer syntactic and semantic considerations.

Frequently mentioned shortcomings of RBMT are mainly concerned with the sufficiency of rules and dictionaries, the method and cost of building them, the handling of ambiguities

and idiomatic expressions in language, the system adaptability in new domains, etc. These concerns have attracted a great deal of attention and deserve more if RBMT is to remain on the stage with greater effectiveness.

Formalized linguistic knowledge: great demand, high cost, and appropriate description

Linguistic knowledge is accessible to computers only when it is formalized. In RBMT, this is achieved through dictionaries and grammar rules.

With dictionaries, the initial concern was about the limited storage capacity and the slow accessing speed of computers, which turned out to be much easier to settle. The persisting concern, by contrast, is the huge and varying demand for linguistic knowledge to be formalized in dictionaries.

Take inflection as an example. There are different types of inflection for different grammatical categories, which complicates the task of morphological analysis. In Russian and German, nouns, pronouns, and numerals are inflected for gender, number, and case (six cases in Russian and four in German); adjectives, when modifying a noun, have to be inflected depending on the gender, number, and case of the noun, and additionally, inflected to indicate comparative and superlative meanings. English, although an inflected language, presents a lower degree of inflection on nouns: there are singular and plural forms for nouns; the case of nouns only has a remnant marker – the possessive indicator '*s*; few nouns are inflected for gender. The word *actor* usually refers to a man in plays or films whose job is acting, while the word *actress* refers to a woman for the same job. These, however, do not denote the grammatical gender. Pronouns are inflected for case (e.g. *I* as the nominative, *me* as the accusative, *my* as the possessive), number (e.g. *he/him* as the singular forms, *they/them* as the plural forms), and gender (e.g. *he/him* as the masculine forms, *she/her* as the feminine forms). The inflection of adjectives in English is not dependent on nouns. Instead, they have comparative and superlative forms of their own. The inflection of verbs is also complicated. When used as predicates, English verbs have finite forms that agree with the subjects in person and number. Inflection makes morphological analysis easier and more reliable, but it requires well-designed frameworks, development guidelines, knowledge representation schemes, and a huge amount of manual work to formalize all the above inflectional information and thereby to make use of them in morphological and syntactic analyses. For less inflected languages such as Chinese, other types of linguistic knowledge are formalized to assist morphological and syntactic analyses, with comparable demand and cost.

With grammar rules, there are further concerns. The balance between the number and the coverage of rules has to be considered in the first place. While a small number of rules usually fail to cover diversifying linguistic phenomena, a large number of them may give rise to conflict among themselves. In addition, the generalization of rules has to be appropriate in order to maximize their coverage of linguistic phenomena but minimize errors. The inappropriate description of linguistic knowledge in this sense will work negatively on the effectiveness of RBMT.

Semantic and contextual ambiguities: the high-hanging fruit

In regular communication, people rely on strings of words to exchange information. To translate the information from one language to another, an RBMT system has to recognize the underlying structure of the strings, theoretically through morphological, syntactic, semantic, and contextual analysis. While there is more physical evidence for morphological and syntactic

relations in some languages, semantic and contextual relations are harder to identify and define in all languages. Here are three more examples in Chinese.

S6: 猴子[Monkeys] 吃[eat] 香蕉[bananas]。Monkeys eat bananas.

S7: 學生[Students] 吃[eat] 食堂[dining hall]。Students have their meals in the dining hall.

S8: 老鄉[The folks] 吃[eat] 大碗[big bowls]。The folks eat with big bowls.

Syntactically, 吃 *chi [to eat]* is a verb, taking nouns like 香蕉, 食堂, and 大碗 to form predicator-object constructions. But to translate these sentences into English, semantic information is needed to specify that the verb 吃 indicates an action by the animal, and therefore it requires an agent and a patient in the sentence; and that the agent is usually an animal specified by a noun, and the patient a kind of food specified by another noun. Further, semantic markers are needed to distinguish 猴子 (animal), 學生 (animal, human), 老鄉 (animal, human) and 蘋果 (food), 食堂 (location), 大碗 (instrument) respectively. In the examples, therefore, the semantic roles of 食堂 and 大碗 as the objects of the verb 吃 are not patients, but the location and the instrument respectively.

Similarly, to understand a sentence correctly, the context beyond sentence boundaries is also essential. For instance, whether the sentence 小張打針去 is translated into *Xiao Zhang has gone to take an injection* or *Xiao Zhang has gone to give an injection* is decided by the contextual fact that Xiao Zhang is a patient or a nurse. Contextual information is more dynamic than semantic information, the formalization of which, accordingly, will be more complex in design and implementation.

Research on MT, particularly studies on RBMT, has been making greater use of the morphological and syntactic evidence. A moderate amount of semantic and contextual analysis has been explored, which helps to resolve some ambiguities left behind by morphological and syntactic analyses. An important reason behind these is the assumption that semantic and contextual information of language is in general less detectable and more difficult to formalize. However, after the low-hanging fruit has been picked up during the past decades of MT research (Church 2011), it will be strategically important and practically necessary to have more focused and collaborated efforts on semantic and contextual analysis.

Domain adaptability

One of the judgments on the 1954 Georgetown-IBM demonstration involves the domain restriction of the system, which was designed specifically to handle a particular sample of a small number of sentences mainly from organic chemistry based on a limited vocabulary of 250 words. The six rules worked well on the sentence patterns in the sample. (Hutchins 2010) It is possible for the system to work on sentences outside the sample, but it requires an expansion of the embedded dictionary in the first place. Further, either these new sentences have to conform to the patterns of those in the sample, or new rules have to be added to cover the new input sentences.

Although organic chemistry is a subfield with a considerably small number of lexical items and typical sentence patterns, the Georgetown-IBM demonstration did not cover all of them (Hutchins 2004). It is now well understood that MT systems are developed to meet widely differing translation needs. This is particularly true with RBMT systems, which rely heavily on dictionaries and grammar rules. The adaptation of RBMT systems, therefore, is more concerned

with the adaptation of the corresponding dictionaries and grammar rules covering the domain-specific morphological, syntactic, semantic, and contextual information.

Failure to adapt to new domains has been listed as one of the weaknesses of RBMT systems, but the other side of the coin is the acknowledgement that the performance of RBMT systems can be greatly improved through domain-specific adaptation – the adaptation of domain-specific dictionaries and grammar rules. An example of domain-specific RBMT can be found in the well-known Météo system, which was developed at Montreal to translate weather forecasts (Hutchins 2010).

Formalized description of linguistic knowledge

The significance of linguistic knowledge in MT has been repeatedly reflected upon, which leads to an ever-growing understanding of the role that linguistic knowledge plays in MT. In the case of RBMT particularly, difficulties and setbacks help to reveal, one after another, the necessity of morphological, syntactic, semantic, and contextual information. A more fundamental issue, however, is how to represent the linguistic knowledge so that it can be processed and utilized by MT systems. Basically, there are two types of formalized knowledge representations: dictionaries and grammar rules on the one hand, and corpora on the other. As explicit representations, the former adopt formal structures, such as relational databases and rewrite rules; as implicit representations, the latter use linear strings of words.

RBMT is in principle working with the explicit type. To handle an infinite number of sentences, RBMT systems rely on dictionaries that store the information about the finite number of words (in SL and TL respectively) and on grammar rules that describe the relationship between words.

Dictionaries

Being an indispensable component in almost all RBMT systems, dictionaries may store morphological, syntactic, semantic, and contextual information about the languages involved. There are several important considerations when a dictionary is designed and constructed for RBMT.

Purpose

To design a language knowledge base, it is necessary, above all, to decide whether it is to serve the special purpose of a specific system, as the adaptation of the knowledge base for new tasks requires significant investment of resources. A general-purpose dictionary is independent of any particular processing system and irrelevant even to any computational theory or algorithm. It is supposed to record the basic linguistic facts. Adaptation is needed if such a dictionary is to work in a RBMT system designed for another specific domain.

Structure

The structure of a dictionary determines its way of storing and managing the linguistic information. A suitable structure ensures the efficient use of a dictionary. The earliest dictionaries in RBMT systems were only consulted to find the TL equivalents of SL words, and the dictionary structure then was therefore quite simple. As understanding grows concerning the complex mechanism of natural languages and the complicated process of

translation, the structure of dictionaries evolves accordingly. Relational databases, for example, provide a means to manage large amounts of morphological, syntactic, semantic, and lexical equivalence information efficiently. A database can be a collection of tables designed for particular grammatical categories respectively. The attribute-value system makes it easy to describe a range of attributes for each linguistic entry. This can also be achieved by complex feature structures, but relational databases, by contrast, are more convenient for manual input and more efficient for computer access. In the sample table below, the classifiers that collocate with the nouns in the four entries are clearly specified.

Table 10.1 A sample table for nouns

Word	*POS*	*Individual classifier*	*Container classifier*	*Measure classifier*
人	n	個		
書	n	本，冊	箱	
鎖	n	把		
糖	n	塊	袋，罐	克，斤

A relational database can be converted to other forms of knowledge representation conveniently, to suit the specific purpose of application systems.

Word classification vs. feature description

To build a dictionary, it is necessary to integrate word classification with feature description. Theoretically, to describe the features of words is another way of classifying and distinguishing them. But due to the complexity of language, words of the same grammatical category, with many shared features, may still have their distinctive ones. Similarly, words from different grammatical categories may also have shared features. Such being the case, word classification and feature description complement each other to achieve a better coverage of the facts of real language use.

Relational databases provide an efficient solution to the combination of the two processes – defining the grammatical category of each word on the one hand and more importantly, adding elaborate description of various linguistic features for each word on the other.

Expert knowledge vs. computer-aided corpus study

The development of a dictionary requires enormous resources. Developed countries or regions enjoy the financial advantages, but find it unrealistic to engage high-level linguists in the tedious and tiresome work because of the high costs of manpower. Therefore, the development of dictionaries in those countries relies mainly on technology to automate the acquisition of knowledge. With the advance of computer science and the Internet, there are more machine-readable dictionaries and texts available, which benefit the development of new dictionaries. Obviously, this development also relies heavily on the engagement of linguists, the progress of theoretical linguistic research, and the collaboration between computer science and linguistics.

In addition to expert knowledge, computer-aided corpus study is also necessary. Different from the linguistic evidence based on linguists' reflections upon language use, corpus data come from communication in real contexts, which may add greatly to the existing understanding of language. Annotated corpora, with the imposed explicit linguistic

annotations, can also assist the learning of linguistic features. There are computer tools that help analyzing corpus data and thereby retrieving linguistic knowledge efficiently and accurately. Further, corpus data can also be used to verify the content of dictionaries, thus increasing their credibility.

Selection of entries and their attributes

The selection of entries and their attributes to be included in a dictionary is based on a clear understanding of the goal that the dictionary is to achieve. In RBMT, it is to assist SL analysis and TL synthesis, so syntactic and semantic considerations are usually valued. For example, a Chinese dictionary will have an entry for the verb 花 as in 花錢 *[to spend (money)]* and another entry for the noun 花 as in 鮮花 *[(fresh) flower]*. For the verb entry, the dictionary may describe its transitivity, the feature of its collocating nouns, adverbs, auxiliaries, etc.; and for the noun entry, the dictionary may describe its collocating classifiers, its function as subjects and objects, etc.

Beside the above-mentioned, there are other considerations when dictionaries are constructed for RBMT, which include, for example, the standardization of language, the change of language, word formation, etc.

Grammar rules

Another important component in RBMT systems is the set of grammar rules that account for the main procedures of analysis and synthesis. There are also transfer rules, particularly in the transfer model, which link the representations of two languages together.

Research on RBMT has greatly encouraged the development of formal linguistics. An important focus of theoretical linguistic research in RBMT is placed on formal grammars, examples of which are Phrase Structure Grammar (PSG), Generalized Phrase Structure Grammar (GPSG), Head-driven Phrase Structure Grammar (HPSG), Lexical Functional Grammar (LFG), etc. Rules of these formal grammars have been used to describe the structure of natural language sentences precisely.

In PSG, a classic formal grammar first proposed by Noam Chomsky in 1957, a grammar consists of: (i) a finite set of nonterminal symbols, none of which appear in sentences formed from the grammar; (ii) a finite set of terminal symbols, which appear in strings formed from the grammar; (iii) a start symbol, which is a distinguished nonterminal; and (iv) a finite set of rewrite rules (also called production rules), each in the form of $a \rightarrow b$ (a being a string of an arbitrary number of symbols with at least one nonterminal; and b being a string of an arbitrary number of symbols, including an empty string) (Chomsky 1957). Constraints on rewrite rules lead to different varieties of PSG. In the other grammars mentioned above, feature structures and unification are introduced, adding more precision to the procedures of analysis and synthesis. For example, rules in LFG present simultaneously two distinct but interrelated levels of structures: constituent structure and functional structure, the former being a conventional phrase structure tree and the latter involving syntactic functions or features like subject, object, complement, adjunct, etc. The following is an example of the functional structure for the sentence *A girl handed the baby a toy* (Kaplan and Bresnan 1982).

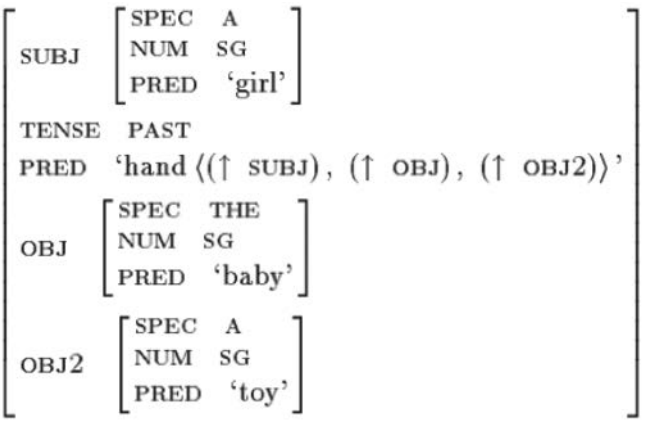


Figure 10.9 The functional structure in LFG

A pair of an attribute and its value may represent either a syntactic feature such as past tense (TENSE PAST) or a semantic feature such as the predicate-argument specification (PRED 'hand<(↑ SUBJ), (↑ OBJ), (↑ OBJ2)'), which defines the mapping between, for instance, the argument SUBJ and the function SUBJ in this functional structure.

The complexity of language in general makes it difficult for grammar rules to capture the nuances of genuine word usages. The new and promising corpus-based statistical approach emerged, making it possible to cover the detailed language use, and particularly the collocation between words. While bringing vigorous development to MT research, the new method requires the support of powerful computers and the time-consuming construction and processing of large-scale corpora. A balance can be achieved between formal grammars and corpora by adopting a new form for the traditional dictionaries – using relational databases to record more specific linguistic information.

Language knowledge base

The term *language knowledge base* offers a more inclusive and consistent way to refer to a machine-readable repository of linguistic knowledge collected, represented, organized, and thereafter utilized. It is usually more sophisticated, designed, developed, and integrated compared with the traditional dictionaries and rule sets in RBMT. The Comprehensive Language Knowledge Base (CLKB) developed at Peking University in China, for instance, is a collection of a grammatical knowledge base of word entries, a phrase structure knowledge base, an annotated monolingual corpus, a bilingual parallel corpus, a multilingual concept dictionary, and a term bank, which embodies the knowledge expansion from words to sentences and texts, from syntactical level to semantic level, from monolingual to multilingual, and from general domain to specific domain (Yu *et al.* 2011).

Rule-based automatic MT evaluation

The evaluation of MT output quality supports the advance of MT research, and the automation of this task ensures a higher level of efficiency, objectiveness, and consistency. There has been MT evaluation ever since the start of MT research, but for quite a long time, evaluations were

done manually, the most famous example of which delivered the ALPAC report. The report compared the translations of MT systems with the human translation to evaluate their intelligibility and fidelity.

Methods for the automatic evaluation of MT output quality also split between rule-based and statistical ones. MTE, the first automatic evaluation system for MT, was developed at Peking University, China during the 1980s and 1990s to evaluate the output quality of English–Chinese translation systems. Six classes of test points (Yu 1993) were defined: words, idioms, morphology, elementary grammar, moderate grammar, and advanced grammar. The following are some examples:

Spring is the first season in a year.
It is a spring bed.
Test point: word sense ambiguity

Are the students playing football?
Are the students playing football your classmates?
Test point: garden path sentence

A context-free formal language TDL was designed to describe the specific test pointes and their corresponding marking criterion. For example:

SL sentence: They got up at six this morning.

R→(492:1)★$A[的]$B$C★
R→(492:0)★
$A→早晨/上午
$B→六/6
$C→點鐘/點/時
##

where 492 is the code for the test point, 1 and 0 are the scores, and / separates alternatives. If an MT system produces 早晨六點 instead of 六點早晨 as a translation for the SL sentence, it scores 1; otherwise, it scores 0.

This rule-based method using test points can clearly locate the strengths and flaws of a system, but the definition of test points requires a huge amount of manual work, of which corpus-based automatic extraction has been implemented (Zhou 2008).

In contrast, the commonly used statistical method for MT evaluation is based on n-gram, with which all grams of a sentence are treated equally. The evaluation does not help distinguish the strengths and flaws of an MT system. In this respect, the rule-based approach can be more effective, focusing on particular test points – linguistic problems or difficulties – in SL analysis and TL synthesis.

Conclusion

Research on RBMT has been playing an important role in promoting the overall progress of MT. Despite the dominating influence of the statistical approach since the end of 1980s, the linguistic premises and assumptions of RBMT are still valued, and the insufficiency of theoretical linguistic research is being realized more than ever. As a result, more practical MT

systems or solutions are relying on both the rule-based and the statistical approaches to achieve a more satisfactory performance.

References

Chomsky, Noam (1957) *Syntactic Structure*, The Hague: Mouton.

Church, Kenneth (2011) 'A Pendulum Swung Too Far', *Linguistic Issues in Language Technology – LiLT* 6.

Dong, Zhendong 董振東, Dong Qiang 董強, and He Changling 郝長伶 (2011)〈下一站在哪裡?〉(Where Is the Next Stop?),《中文信息學報》(*Journal of Chinese Information Processing*) 25(6): 3−11.

Hutchins, W. John (2010) 'Machine Translation: A Concise History', in Chan Sin-wai (ed.) *Journal of Translation Studies: Special Issue on The Teaching of Computer-aided Translation*, 13(1−2): 29−70.

Hutchins, W. John (2004) 'The Georgetown-IBM Experiment Demonstrated in January 1954', in *Machine Translation: From Real Users to Research: The 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, 28 September – 2 October 2004, Washington, DC.

Hutchins, W. John (1994) 'Research Methods and System Designs in Machine Translation: A Ten-year Review, 1984-1994', in *Proceedings of the International Conference 'Machine Translation: Ten Years on'*, 12−14 November 1994, Cranfield University, UK.

Hutchins, W. John (1986) *Machine Translation: Past, Present, Future*, Chichester, UK: Ellis Horwood.

Kaplan, Ronald and Joan Bresnan (1982) 'Lexical-functional Grammar: A Formal System for Grammatical representation', in Joan Bresnan (ed.) *The Mental Representation of Grammatical Relations*, Cambridge, MA: MIT Press, 173−281.

Tanaka, Hozumi, Shun Ishizaki, Akira Uehara, and Hiroshi Uchid (1989) 'A Research and Development of Cooperation Project on a Machine Translation System for Japan and Its Neighboring Countries', in *Proceedings of the MT Summit II*, 16−18 August 1989, Munich, Germany, 146−151.

Yu, Shiwen 俞士汶, Sui Zhifang 穗志方, and Zhu Xuefeng 朱學鋒 (2011)〈綜合型語言知識庫及其前景〉(A Comprehensive Language Knowledge Base and Its Prospect)《中文信息學報》(*Journal of Chinese Information Processing*) 25(6): 12−20.

Yu, Shiwen 俞士汶 (ed.) (2003) 《計算語言學概論》(*Introduction to Computational Linguistics*). Beijing: The Commercial Press 商務印書館.

Yu, Shiwen (1993) 'Automatic Evaluation of Output Quality for Machine Translation Systems,' *Machine Translation* 8: 117−126.

Zhou, Ming, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao (2008) 'Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points', in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, August 2008, Manchester.

11
STATISTICAL MACHINE TRANSLATION

Liu Yang

TSINGHUA UNIVERSITY, CHINA

Zhang Min

SOOCHOW UNIVERSITY, CHINA

Overview

Statistical machine translation (SMT) is a machine translation paradigm that generates translations based on a probabilistic model of the translation process, the parameters of which are estimated from parallel text.

SMT was first introduced by Warren Weaver in 1949. He suggested that statistical techniques from Claude Shannon's information theory might make it possible to use computers to translate between natural languages automatically. However, this idea could hardly turn into reality at the time due to limited computer resources. Thanks to the improvement in computer power and the increasing availability of machine-readable text, a group of researchers at IBM TJ Watson Research Center launched the "Candide" project to re-introduce statistical techniques to machine translation in 1991. Since then, SMT has seen a resurgence in popularity and become one of the most widely studied machine translation methods.

The major difference between SMT and conventional rule-based MT lies in the acquisition of translation knowledge. Rule-based translation systems often require the manual development of linguistic rules, which can be costly, time-consuming, and hardly generalizable to other languages. Alternatively, SMT pursues a data-driven approach to acquiring translation knowledge. SMT systems are usually based on statistical models whose parameters, namely translation knowledge in SMT systems, can be derived from the analysis of machine-readable parallel text automatically. Therefore, SMT systems are language independent because they are not tailored to any specific pair of languages.

Generally, there are three fundamental problems in statistical machine translation:

1 **Modeling**. The heart of statistical machine translation is the probabilistic modeling of the translation process. Early statistical machine translation systems are based on *generative* translation models where a generative story is designed to describe how a computer translates natural languages step by step. Significant advances have been made by the introduction of *discriminative* models in 2002. As discriminative models are capable of incorporating a great deal of diverse and overlapping knowledge sources as features, they

have become mainstream in modern SMT systems. From the perspective of the basic translation unit, statistical machine translation has evolved from modeling flat structures (i.e., word, phrase) to hierarchical structures (i.e., syntactic tree) in the past two decades.

2 **Training**. As a data-driven approach, SMT estimates the parameters of translation models from parallel corpus automatically. This is called training or parameter estimation. The parameters of generative models are usually probability distributions on unobserved latent variables such as word-to-word translation sub-models, distortion models, etc. While the Expectation Maximization (EM) algorithm is widely used for word-based models, phrase-based and syntax-based models usually resort to simple and efficient heuristic methods for parameter estimation. To estimate the parameters of discriminative models, which are usually real-valued feature weights of log-linear models, the most widely used algorithm is minimum error rate training that can directly optimize feature weights with respect to the final evaluation metric.

3 **Decoding**. Given estimated translation model parameters and an unseen source language text, the goal of decoding is to find a target language text that maximizes translation probability. Due to the diversity of natural languages, the search space of SMT is often prohibitively large. Therefore, SMT systems have to use approximate search algorithms instead of exhaustive search in practice. The decoding algorithms in SMT can be roughly divided into two broad categories with respect to the order of generating target language words: left-to-right and bottom-up. Left-to-right decoding algorithms, which run in quadratic time, are mainly used in phrase-based systems where stacks are maintained to store promising partial translations. Bottom-up decoding algorithms (e.g., the CYK algorithm) are mainly used in syntax-based systems and generally run in cubic time.

Word-based SMT

The initial statistical machine translation system was based on word-based models, in which the basic unit of translation is the word (Brown *et al.* 1993).

The basic idea of word-based models is to design a generative story for the translation process: predicting the length of translation, deciding the permutation of words, and choosing appropriate words. Each decision is associated with a probability. The decision sequence with the highest overall probability is chosen to generate the optimal translation. Each type of decision corresponds to a sub-model in word-based models, the parameters of which are estimated from parallel corpus automatically.

Given a source language sentence $f_1^J = f_1 \cdots f_J$, how likely a target language sentence $\mathrm{e}_1^I = e_1 \cdots \mathrm{e}_I$ is a translation of the source language sentence can be denoted by a probability distribution $P(e_1^I | f_1^J)$. Therefore, the goal of statistical machine translation is to build a translation model, train the model parameters, and search for the optimal translation with highest translation probability.

Originated from Shannon's information theory, word-based translation models apply the Bayes theorem to make the search of optimal translations dependent on two models: an inverse translation model $P(f_1^J | e_1^I)$ that assigns a probability that the source language sentence f_1^J is a translation of the target language sentence e_1^I and a language model $P(e_1^I)$ that assigns a probability of the target sentence e_1^I:

$$P\left(e_1^I \mid f_1^J\right) = \frac{P\left(f_1^J \mid e_1^I\right) \times P\left(e_1^I\right)}{P\left(f_1^J\right)}$$

Intuitively, the inverse translation model $P(f_1^J \mid e_1^I)$ evaluates the fidelity of a translation while the language model $P(e_1^I)$ evaluates the fluency.

Brown *et al.* (1993) propose five translation models with increasing expressive power, namely IBM models 1–5. All IBM models are based on an important notion in statistical machine translation: word alignment. Word alignment indicates the correspondence between the words of source and target language sentences. It is introduced into translation models as a latent variable. Figure 11.1 shows a word alignment for a Chinese–English sentence pair. The dashed lines denote alignment links.

Figure 11.1 Word alignment

Therefore, the translation probability that a target language sentence is translated into a source language sentence is equal to the sum of alignment probabilities over all possible word alignments between the two sentences:

$$P\left(f_1^J \mid e_1^I\right) = \Sigma_{a_1^J} P\left(f_1^J, a_1^J \mid e_1^I\right)$$

As the IBM models are generative models, each of them is based on a generative story that describes how to transform a target language sentence to a source language sentence step by step. The generative story for IBM models 1 and 2 is as follows:

1 Given a target language sentence e_1^I, decide the length J of the corresponding source language sentence f_1^J.
2 For each source language position j ranging from 1 to J:
   (a) decide which target language word e_{a_j} is aligned to the current source language position j;
   (b) decide what the source language word f_j is given the aligned target language word e_{a_j}.

The above generative story can be exactly described by a probabilistic model in a mathematical way. The three types of decisions in the generative story correspond to three sub-models in the translation model: length sub-model, alignment sub-model, and translation sub-model. While IBM models 1 and 2 share the same length and translation sub-models, they differ in the choice of alignment sub-models. IBM model 1 assumes the alignment distribution is uniform. In contrast, IBM model 2 uses an alignment sub-model that depends on positions of words.

IBM models 3–5 are based on more sophisticated generative stories. They are different from simpler IBM models 1–2 because of the introduction of fertility, which explicitly describes the fact that a target language word can be aligned to multiple source language words. The generative story for IBM model 3 is as follows (Knight 1999):

1 For each target language word e_i, choose a fertility Φ_i, which is the number of source language words that will be generated from the target language word and depends only on the target word.
2 Generate source language words from the NULL target language word.
3 Generate source language words from the non-NULL target language words according the corresponding fertilities.
4 Move all the non-spurious words in the source language sentence.
5 Insert spurious words in the remaining open positions.

IBM models 3–5 are usually called fertility-based models. They have more parameters than IBM models 1–2. The most important parameters are the fertility, distortion, and translation sub-models.

The parameters of IBM models can be estimated from a given parallel training corpus consisting of a set of sentence pairs. Often, the unknown parameters are determined by maximizing the likelihood of the parallel training corpus using the expectation maximization (EM) algorithm. Note that the parameters of the statistical translation models are optimized using maximum likelihood estimation (MLE), which is not related to alignment and translation evaluation metrics used in practice.

Training IBM models often involves the computation of the alignment with highest probability, namely the Viterbi alignment. While there exist simple polynomial algorithms for IBM models 1 and 2, computing Viterbi alignments for the fertility-based models is non-trivial. As suggested by Brown *et al.* (1993), an efficient hill-climbing algorithm is widely used in finding Viterbi alignments for fertility-based models. The basic idea is to first compute the Viterbi alignment of a simple model. Then, this alignment is iteratively improved with respect to the alignment probability of fertility models by modifying the current alignment.

As the decoding problem for word-based models is NP-complete (Knight 1999), a sensible strategy is to examine a large subset of promising translations and choose just one from that. The stack-based decoding algorithm, which was first introduced in the domain of speech recognition, has been widely used in word-based SMT decoders. By building translations incrementally and storing partial translations in a stack (i.e., a priority queue), the decoder conducts an ordered best-first search in the search space. Other decoding algorithms include greedy and integer programming algorithms (Germann *et al.* 2001).

Manning and Schutze (1999) point out a number of drawbacks of word-based models:

1 No notion of phrases. The models relate only to individual words and do not model relationships between phrases.
2 Non-local dependencies. The models fail to capture non-local dependencies that are important in translation.
3 Morphology. Morphologically related words (e.g., *like, likes, liked*) are treated as separate symbols.
4 Sparse data problems. Estimates for rare words are unreliable.

Although word-based models are not widely used today, word alignments generated by word-based models still play an important role in training more advanced phrase-based and syntax-based translation models.

Phrase-based SMT

While word-based models only consider how each individual word is translated, phrase-based models are based on the intuition that a better way is to translate and move phrases as a unit in machine translation. A phrase in phrase-based models is usually a sequence of consecutive words. It is not necessarily a phrase in any syntactic theory. As phrases memorize local word selection and reordering, phrase-based models are capable of handling idiom translation, word insertion and deletion.

The generative story of phrase-based models is as follows:

1 Given a target language sentence, segment it into a sequence of phrases. Suppose that the number of source language phrases is identical to that of target language phrases.
2 Permutate the target language phrases.
3 Translate each target language phrase into a source language phrase one by one and form the source language sentence.

Figure 11.2 shows an example of phrase-based translation. Each block represents a phrase. The dashed lines denote the correspondence between Chinese and English phrases.

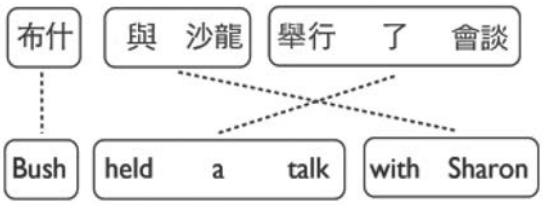


Figure 11.2 Phrase-based SMT

Therefore, there are three sub-models in phrase-based models: phrase segmentation, phrase reordering, and phrase translation. For simplicity, most phrase-based models typically assume a uniform distribution over segmentations.

Unlike word-based models that use an EM algorithm for parameter estimation, phrase-based models often resort to an efficient heuristic way to learn model parameters. Typically, training phrase-based models begins by training word-based models to generate word alignments for the parallel corpus. Then, the word-aligned parallel corpus is used to extract aligned phrase pairs. As IBM models assume that each source language word is aligned to at most one target language word, they cannot align a multiword phrase in the source language with a multiword in the target language. Therefore, a method called symmetrization is proposed to produce many-to-many word alignments (Och and Ney 2004). First, two separate word-based aligners are trained to produce a source-to-target alignment and a target-to-source alignment, respectively. Then, the two alignments are combined in a heuristic way to get an alignment that maps phrases to phrases. After getting symmetrized word alignments, all phrase pairs that are consistent with word alignments are extracted. A consistent phrase pair is one in which all words are aligned only to each other and not to any external words in the training instance. Once all the aligned phrase pairs are collected from the entire training corpus, the maximum likelihood estimates for the phrase translation probability of a particular pair can be computed as relative frequencies in two translation directions, respectively.

The distortion sub-model is an important component in phrase-based statistical machine translation. It models the distortion between source and target language phrases resulting from the divergence of word orders in natural languages. For example, while subject-verb-object (SVO) languages such as English often put the object after the verb (e.g., *I like you*), subject-object-verb (SOV) languages such as Japanese place object before the verb. A simple distortion sub-model widely used in phrase-based SMT is distance-based model (Koehn *et al.* 2003). It measures the distance between positions of a phrase in two languages. The distortion probability thus means the probability of two consecutive target language phrases being separated in the source language by a span of a particular length. Often, this simple distortion model penalizes large distortions by giving a lower probability. A problem with the distance-based model is that it is only conditioned on movement distance while some phrases are reordered more frequently than others. Therefore, lexicalized distortion models conditioned on actual phrases are proposed to alleviate the problem. Lexicalized distortion models consider three types of orientation of a phrase: monotone, swap, and discontinuous. The probability distribution can be estimated from the word-aligned parallel corpus. When extracting each phrase pair, the orientation type of a phrase pair is also extracted in that specific occurrence. A variation to the way phrase orientation statistics are collected is that phrase-based orientation models use phrases both at training and decoding time. A further improvement is a hierarchical orientation model that is able to detect swaps or monotone arrangements between very large blocks. Although lexicalized distortion models are more powerful than the distance-based model, they have the problem of sparse data, as a particular phrase pair may occur only a few times in the training data. Therefore, it is hard to obtain reliable estimates from training data.

Phrase-based SMT uses a stack decoding algorithm to search for optimal translations. The basic intuition is to maintain a sequence of priority queues with all partial translation hypotheses together with their scores. The decoding algorithm begins by searching for a phrase-translation table to collect possible translation options. Each of these translation options consists of a source language phrase, the target language phrase, and phrase translation probabilities. The decoder needs to search through combinations of these options to find the best translation. The target language sentence is generated from left to right in the form of partial translation hypothesis (hypothesis for short). Each hypothesis is associated with a cost for guiding the search. The cost combines the current cost of the phrase with an estimate of the future cost. The current cost is the product of translation, distortion, language model probabilities. The future cost is an estimate of the cost of translating the uncovered words in the source language sentence. As it is too expensive to estimate the future cost of distortion models, phrase-based SMT typically uses a Viterbi algorithm to compute the product of translation and language model probabilities. As the search space is exponential, most phrase-based decoders use pruning techniques to constrain the search space. For every stack, only the most promising hypotheses are kept and unlikely hypotheses are pruned. An important risk-free pruning technique is hypothesis recombination. It can safely discard degenerate hypotheses that cannot be part of the best translation.

While SMT originated from the noisy channel model, discriminative models have become the mainstream nowadays. In a discriminative model such as a log-linear model (Och and Ney 2004), the language model and the translation model can be treated as features:

$$P\left(e_1^I \mid f_1^J\right) = \frac{exp\left(\sum_{m=1}^{M} \lambda_m h_m\left(e_1^I, f_1^J\right)\right)}{Z}$$

where $h(\cdot)$ are feature functions, λ's are feature weights, and Z is a partition function.

In practice, sub-models of generative translation models are still the most important feature functions in the log-linear model. The flexible architecture of discriminative framework allows arbitrary overlapping features to be included in the translation process. In modern SMT systems, log-linear models are trained to directly optimize evaluation metrics using a method called minimum error rate training.

Quirk and Corston-Oliver (2006) summarize advantages and disadvantages of phrase-based SMT as follows:

1 Advantages
   - Non-compositionality. Phrases capture the translations of non-compositional phrases as a unit instead of reconstructing them word by word awkwardly.
   - Local reordering. Local reordering decisions are memorized in phrases.
   - Contextual information. Local context is incorporated in phrases.

2 Disadvantages
   - Exact substring match; no discontinuity. Discontinuous translation pairs are not allowed in most phrase-based SMT systems.
   - Global reordering. Phrases provide no effective global re-ordering strategy.
   - Probability estimation. Long phrases are most likely to contribute important translational and ordering information but are most subject to sparse data issues.
   - Partition limitation. Uniform distribution of phrase segmentation is problematic.

Despite these drawbacks, phrase-based models are still a simple and powerful mechanism for machine translation and have been widely used in commercial MT systems.

Syntax-based SMT

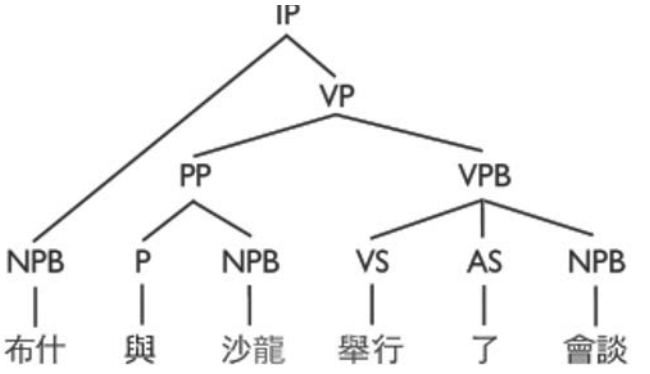


Figure 11.3 A phrase-structure parse tree

While both word-based and phrase-based models cast translation as a problem of permutating and concatenating flat structures, recent work in SMT has focused on modeling hierarchical syntactic structures. Motivated by the intuition that most natural languages are hierarchically structured, these syntax-based models attempt to assign a parallel syntactic tree structure to a pair of sentences in different languages, with the goal of translating the sentences by applying reordering operations on the trees (Figure 11.3). The mathematical model for these parallel structures is known as a **synchronous grammar** (also called a transduction grammar).

A synchronous grammar describes a structurally correlated pair of languages. From a generative perspective, a synchronous grammar is capable of generating pairs of related strings in two languages. A number of synchronous grammars and formalisms have been used since the late 1990s, most of which are generalizations of Context-Free Grammars (CFG) to the bilingual scenario:

- Synchronous Context-Free Grammar (**SCFG**)
- Inversion Transduction Grammar (**ITG**)
- Synchronous Tree Substitution Grammar (**STSG**)

A **Synchronous Context-Free Grammar (SCFG)** is like a CFG, but its productions have two related right-hand sides, namely the source language side and the target language side. The most well-known SCFG-based model is the hierarchical phrase-based translation model proposed by Chiang (2007). As a hierarchical phrase can contain other phrases, it is capable of capturing reordering of phrases. For example, the SCFG rule

$$X \rightarrow \left(\text{與}X_1\text{舉行了}X_2,\ \text{held a } X_2 \text{ with } X_1\right)$$

captures the different ordering of words in two languages, where X denotes a non-terminal and the subscripts denote the correspondence between non-terminals.

Formally, a hierarchical phrase pair corresponds to an SCFG production rule. Each right-hand side is a string of terminals and non-terminals. The model assumes that there is a one-to-one correspondence between non-terminal occurrences in the two right-hand sides. An SCFG derivation begins with a pair of linked start symbols. At each step, two linked non-terminals are rewritten by applying a production rule. Recursively applying SCFG rules generates a pair of sentences in two languages. As a data-driven approach, the bulk of SCFG consists of automatically extracted rules. The training corpus of the SCFG-based model is the same with phrase-based models: word-aligned parallel corpus. SCFG rules that are consistent with word alignment can be extracted from the training data in two steps. First, the extraction algorithm identifies initial phrase pairs in the same way as phrase-based SMT does. Second, the algorithm looks for phrases that contain other phrases and replaces sub-phrases with non-terminal symbols. As this scheme can generate a very large number of rules, a number of constraints are used to filter the grammar to achieve a reasonable grammar size. Besides SCFG rules learned from real-world data, Chiang (2007) also introduces glue rules to concatenate partial translations in a monotonic way. Unlike phrase-based decoders that use a stack algorithm, the decoder for SCFG is a CYK parser. Given a source language sentence, the decoder finds the yield on the target language side of the single best derivation that has the source yield of the input sentence. It organizes the hypotheses in a chart whose cells are sets of hypotheses. A problem faced by an SCFG-based decoder is that language model integration becomes more expensive because the decoder needs to maintain target language words at both ends of a partial translation, whereas a phrase-based decoder only needs to do this at one end because the translation is always growing from left to right. Therefore, the integration of the language model increases the decoding complexity of SCFG-based decoders. To alleviate this problem, Chiang (2007) proposes a method called *cube pruning* to discard most of the less promising hypotheses. Cube pruning has been widely used in modern phrase-based and syntax-based SMT systems. In summary, as a logical outgrowth of phrase-based model, the hierarchical phrase-based model is the first syntax-based model that empirically shows that moving from flat structures to hierarchical structures significantly improves translation quality.

Inversion Translation Grammar (ITG) (Wu 1997) is a synchronous grammar for synchronous parsing of source and target language sentences. It builds a synchronous parse tree to indicate the correspondence as well as permutation of blocks (i.e., consecutive word sequences). There are three types of production rules. A lexical rule $X \rightarrow f/e$ generates two words or phrases in two languages simultaneously. A non-terminal rule in square brackets $X \rightarrow [\mathrm{X}\ \mathrm{X}]$ generates two blocks in a monotone order. A non-terminal rule in angle brackets $X \rightarrow \langle \mathrm{X}\ \mathrm{X} \rangle$ generates two blocks in an inverted order. Generally, ITG can be seen as a special case of SCFG. Any ITG can be converted into an SCFG of rank two. Therefore, the decoder of an ITG-based SMT system is also a CYK parser. Xiong *et al.* (2006) introduce a maximum entropy-based reordering model for ITG. Instead of assigning a uniform distribution to non-terminal rules, they propose to make the decision on merging order dependent on the specific blocks.

Syntax-based models using SCFG and ITG only take the fundamental idea from syntax as they do not exploit any linguistically syntactic structures. By contrast, syntax-based models that use **Synchronous Tree Substitution Grammars (STSG)** usually leverage real linguistic parse trees. In an STSG, the productions are pairs of elementary trees, and the leaf non-terminals are linked just as in synchronous CFG. Depending on whether linguistic parse trees are used or not, syntax-based models can be roughly divided into four categories:

- String-to-String. No linguistic syntax is used. SCFG-based and ITG-based models are typically string-to-string.
- String-to-Tree. Linguistic syntax is used only on the target side.
- Tree-to-String. Linguistic syntax is used only on the source side.
- Tree-to-Tree. Linguistic syntax is used on both sides.

The string-to-tree models (Yamada and Knight 2001; Galley *et al.* 2004; Galley *et al.* 2006) exploit linguistic syntax only on the target side. Galley *et al.* (2004) propose an algorithm called GHKM to learn string-to-tree rules from word-aligned, target side parsed parallel corpus. Like phrases and hierarchical phrases, these syntactically motivated transformation rules must be consistent with word alignment. The GHKM algorithm distinguishes between two types of STSG rules: minimal and composed. While minimal rules are atomic and cannot be decomposed, composed rules can be formed out of smaller rules. String-to-tree models cast translation as a parsing problem: the decoder parses a source language sentence using the source projection of a synchronous grammar while building the target sub-translations in parallel. As string-to-tree rules usually have multiple non-terminals that make decoding complexity generally exponential, synchronous binarization (Zhang *et al.* 2006) is a key technique for applying the CYK algorithm to parsing with string-to-tree rules. This can be done by factoring each STSG rule into two SCFG rules. While phrase structure trees are successfully used in string-to-tree models, recent work on dependency trees further proves the benefit of exploiting linguistic syntax (Shen *et al.* 2010).

Tree-to-string models (Liu *et al.* 2006; Huang *et al.* 2006; Mi *et al.* 2008) explicitly use source parse trees and divide decoding into two separate steps: parsing and translation. A parser first parses a source language sentence into a parse tree, and then a decoder converts the tree to a translation on the target side. The decoding algorithm visits each node in the input source tree in a top-down order and tries to match each translation rule against the local sub-tree rooted at the node. Compared with the CKY algorithm used in string-to-string and string-to-tree decoders, tree-to-string decoding is much simpler and faster: there is no need for synchronous binarization and tree parsing generally runs in linear time. However, despite these

advantages, tree-to-string systems suffer from a major drawback: they only use 1-best parse trees to guide translation, which potentially introduces translation mistakes due to the propagation of parsing errors. This problem can be elegantly alleviated by using packed forests, which encode exponentially many parse trees in a polynomial space (Mi and Huang 2008). Taking a packed forest as input can be regarded as a compromise between taking a string and a single tree: decoding is still fast, yet does not commit to a single parse. In addition, packed forests can also be used for translation rule extraction, which helps alleviate the propagation of parsing errors into rule set.

Tree-to-tree models (Eisner 2003; Quirk *et al.* 2005; Zhang *et al.* 2008; Liu *et al.* 2009; Chiang 2010) explicitly use parse trees on both sides. The decoding algorithm for tree-to-tree translation can be either parsing or tree parsing. The tree parsing algorithm takes a source tree as input and produces a target tree (Eisner 2003; Quirk *et al.* 2005; Zhang *et al.* 2008; Liu *et al.* 2009). By contrast, the parsing algorithm takes a source sentence as input and generates source and target trees simultaneously (Chiang 2010).

The choice of syntax-based models depends on the availability of parsers. For example, string-to-tree models might be suitable for translating a resource-scarce language into a resource-rich language such as English. Similarly, tree-to-string models might work better for translating English into a resource-scarce language that has no high accuracy parsers.

The most frequently cited disadvantages of syntax-based SMT include

- Availability and accuracy of parsers. For most natural languages, there are no high-accuracy parsers. Even for resource-rich languages such as English, parsers usually only work well for limited domains. Therefore, the applicability of syntax-based SMT is severely limited.
- Huge grammar size. Syntax-based models usually learn a very large number of rules as compared with phrase-based models, which leads to high memory requirement.
- Decoding complexity. Syntax-based decoders are significantly slower than phrase-based decoders.

Despite these disadvantages, syntax-based models have undergone rapid development and have started to be used in commercial MT systems.

To conclude, the past two decades have witnessed the rapid development of statistical machine translation, moving from modeling flat structures (e.g., word, phrase) to hierarchical structures (e.g., tree). As the central goal of machine translation is to ensure the meaning equivalence between the input and output, semantics-based SMT is clearly an important future direction awaiting exploration. In addition, although SMT is claimed to be language independent, most systems are designed and tested for resource-rich languages such as English, Chinese, Arabic, and French. How to use SMT techniques to deal with other natural languages in the world, most of which are resource-scarce and significantly different from English, still remains a big challenge. More recently, the Do-it-yourself MT has emerged to give users a high degree of control over SMT systems. With the SMT systems accessible via the Web, the users are allowed to customize MT engines by uploading their own translation memories. Such cloud-based DIY SMT services are capable of providing high-quality, user-specific translations with much lower cost. We believe that the intersection between statistical methods and user engagement will be a promising direction for commercial SMT services.

References

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational Linguistics* 1(2): 233–312.

Chiang, David (2007) 'Hierarchical Phrase-based Translation', *Computational Linguistics* 33(2): 201–228.

Chiang, David (2010) 'Learning to Translate with Source and Target Syntax', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 11–16 July 2010, Uppsala, Sweden, 1443–1452.

Eisner, Jason (2003) 'Learning Non-isomorphic Tree Mappings for Machine Translation', in *ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 7–12 July 2003, Sapporo, Japan, 2: 205–208.

Galley, Michel, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer (2006) 'Scalable Inference and Training of Context-rich Syntactic Translation Models', in *Proceedings of ACL 2006*, 17–21 July 2006, Sydney, Australia, 961–968.

Galley, Michel, Mark Hopkins, Kevin Knight, and Daniel Marcu (2004) 'What's in a Translation Rule?' in *Proceedings of HLT-NAACL 2004*, 2–7 May 2004, Boston, MA, 273–280.

Germann, Ulrich, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada (2001) 'Fast Decoding and Optimal Decoding for Machine Translation', in *ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 9–11 July 2001, Toulouse, France, 228–235.

Knight, Kevin (1999) 'Decoding Complexity in Word-replacement Translation Models', *Computational Linguistics* 25(4): 607–615.

Koehn, Philipp, Franz J. Och, and Daniel Marcu (2003) 'Statistical Phrase-based Translation', in *NAACL '03 Proceedings of 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 27 May – 1 June 2003, Edmonton, Canada, 1: 49–54.

Liu, Yang, Qun Liu, and Shouxun Lin (2006) 'Tree-to-string Alignment Template for Statistical Machine Translation', in *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 20 July 2006, Sydney, Australia, 609–616.

Liu, Yang, Yajuan Lü, and Qun Liu (2009) 'Improving Tree-to-tree Translation with Packed Forests', in *ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2–7 August 2009, Singapore, 2: 558–566.

Manning, Chris and Hinrich Schutze (1999) *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Mi, Haitao and Liang Huang (2008) 'Forest-based Translation Rule Extraction', in *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 25–27 October 2008, Honolulu, HI, 206–214.

Mi, Haitao, Liang Huang, and Qun Liu (2008) 'Forest-based Translation', in *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, 15–20 June 2008, the Ohio State University, Columbus, OH, 192–199.

Och, Franz and Hermann Ney (2004) 'The Alignment Template Approach to Statistical Machine Translation', *Computational Linguistics* 30(4): 417–449.

Quirk, Chris and Simon Corston-Oliver (2006) 'The Impact of Parse Quality on Syntactically-informed Statistical Machine Translation', in *Proceedings of EMNLP 2006*, 22–23 July 2006, Sydney, Australia, 62–69.

Quirk, Chris, Arul Menezes, and Colin Cherry (2005) 'Dependency Treelet Translation: Syntactically Informed Phrasal SMT', in *ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 25–30 June 2005, University of Michigan, MI, 271–279.

Shen, Libin, Jinxi Xu, and Ralph Weischedel (2010) 'String-to-dependency Statistical Machine Translation', *Computational Linguistics* 36(4): 649–671.

Wu, Dekai (1997) 'Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora', *Computational Linguistics* 23(3): 377–403.

Xiong, Deyi, Qun Liu, and Shouxun Lin (2006) 'Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation', in *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 20 July 2006, Sydney, Australia, 521–528.

Yamada, Kenji and Kevin Knight (2001) 'A Syntax-based Statistical Translation Model', in *ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 9–11 July 2001, Toulouse, France, 523–530.

Zhang, Hao, Liang Huang, Daniel Gildea, and Kevin Knight (2006) 'Synchronous Binarization for Machine Translation', in HLT-NAACL '06 Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 4–9 June 2006, New York, 256–263.

Zhang, Min, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li (2008) 'A Tree Sequence Alignment-based Tree-to-tree Translation Model', in *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 15–20 June 2008, Columbus, OH, 559–567.

12

EVALUATION IN MACHINE TRANSLATION AND COMPUTER-AIDED TRANSLATION

Kit Chunyu

CITY UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Billy Wong Tak-ming

THE OPEN UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Introduction

Machine translation (MT) and *computer-aided translation* (CAT) both serve the same purpose of enhancing translation efficiency via utilization of the computer. MT refers specifically to the automation of translation by means of available computer technology. It keeps pursuing fully automatic high-quality translation (FAHQT) as its ultimate goal, which was criticized by some MT pioneers (e.g., Bar Hillel and Martin Kay) a few decades ago as an unrealistic objective to strive for but has been pursued since the very beginning of MT by so many other MTers through several generations of methodology and technology with notable but still limited successes. CAT is intended to provide suitable utilities with necessary language resources to assist human translation, aiming at maximizing the productivity of translation by means of combining the strengths of both sides.

Translation has become an industry that needs MT/CAT systems as 'machines' to facilitate translation production at various levels of automation. Typical utilities they provide to support human translation include monolingual/bilingual dictionaries and term banks (with terminology management tools), translation examples as in the form (and name) of translation memory, etc. Besides, it is even more fundamental that CAT incorporates MT as one of its facilities to provide an initial version of a translation of a certain quality, as high as possible, for human translators to post-edit into a final version up to their quality standards, unless it is less editable than its source. In principle, the higher the quality of MT output, the less human effort needed for post-editing and hence the higher productivity of translation.

The evaluation of MT/CAT deals with the issue of quantifying their effectiveness. In a broad sense, it is intended to systematically assess the quality, success and efficacy of any aspect of an MT/CAT system that gives rise to a concern about the degree of the system's usefulness. It has been developed into a unique discipline in the field despite the diversity of evaluation in terms of purposes and corresponding criteria used. Nevertheless, the quality of MT output is always at the core of evaluation. This explains why another term, *MT evaluation*, has been more

and more popular and become its default term. However, MT output is typically characterized by the lack of a widely recognized objective metric for quality quantification. Unlike a clear-cut correct output from a language processing system, e.g., word spelling from a spell checker, for an input, there is hardly an ideal or 'correct' one among so many possible translations for a source text. Besides, the productivity of translation with the aid of CAT facilities is also largely attributed to a user's proficiency in using them. In other words, the evaluation of MT/CAT inevitably involves the evaluation of its users. In response to the diversity and challenges of MT/CAT evaluation, different types of qualitative and quantitative measurement have been developed.

This chapter introduces the key issues and basic principles of MT/CAT evaluation, concerning MT systems with or without human intervention to finalize translation output. It begins with a brief review of the history of MT/CAT evaluation to outline the evolution of evaluation methodology and technology, along with the development of MT/CAT over the past several decades. The highly context-dependent multi-dimensional nature of the evaluation will then be described, including various applications of system output and different evaluation purposes. Then the existing evaluation methodologies will be presented and illustrated. On the one hand, an MT/CAT system is evaluated as a piece of software in terms of general parameters such as speed and number of supported file formats, subject to existing standards and criteria. On the other hand, its evaluation becomes a matter of text quality assessment because MT outputs are essentially in the form of text. The major approaches of MT evaluation, including both manual and automatic, will be presented with a discussion of their strengths and weaknesses.

A brief history

The evolution of MT evaluation is inseparable from the development of MT technology. Historically, the first MT demonstration, held in 1954 by a joint effort of Georgetown University and IBM, not only attempted the first application of non-numerical programming ever run on a computer but also, as a matter of fact, conducted the very first MT evaluation. It had a tremendous impact immediately, raising the awareness of MT greatly, attracting a substantial amount of investment of money and research effort into this new-born field in the subsequent years, and consequently starting a mushrooming period of MT for about a decade. It was later considered highly controversial, misleading and even deceptive, due to its simplicity, in that it used a set of 60 prepared or selected sentences, 250 lexical items merely covering these sentences, and 6 operational rules. However, we have to admit its success in achieving its preset goal to 'test' the feasibility of MT, instead of its robustness and output quality, and in arriving at the affirmative and even 'convincing' conclusion—'Yes, we can do it!'— despite some unrooted exaggeration and unrealistic over-expectation that followed it. From the current point of view, one can hardly find any significant methodology and technology of MT and MT evaluation in this piece of initial work.

One may consider that the earliest evaluation of MT begins from the criticism of the first generation MT 'technologies' and the MT research in the 1950s on the wrong track towards the unrealistic goal of FAHQT. The first formal, in a sense, and influential evaluation was conducted by ALPAC[1] in the mid-1960s, to examine the effectiveness of the funding to support MT research at that time, covering a variety of aspects such as the translation market in US, the speed and cost of producing MT outputs, and their quality. The ALPAC report (1966) also quotes some statistics from a study by Orr and Small (1967: 1–10) that compares MT outputs and human translations in terms of their comprehensibility, to illustrate that MT

outputs are not as accurate, readable and therefore not as useful as human translations.[2] It further provides a few samples of MT output containing 'unnatural constructions and unnatural word order', to support the claim that the goal of FAHQT was not achievable. A controversy that arises from this conclusion is that such an evaluation focused too much on regarding MT only as a production tool to meet users' translation needs in the US, and did not recognize the other potentials of MT and the expanding global translation market. Nevertheless, the impact of the ALPAC report to MT/CAT evaluation is long-lasting. Three of its final recommendations are made constructively to support 'practical methods for evaluation of translations', 'evaluation of quality and cost of various sources of translations', and 'evaluation of the relative speed and cost of various sorts of machine-aided translation' (1966: 34). In addition, its evaluation methodology, as described in detail in Carroll (1966: 55–66), greatly influenced many evaluation practices in subsequent years.

Despite all that, the ALPAC report did not bring about any immediate revolutionary change to MT evaluation. According to Hutchins and Somers (1992), most MT evaluations at that time were still carried out by nonprofessionals with very little or even no expertise in MT techniques. They were unable to judge what could be possible or unrealistic for MT or to provide any useful comments on system performance or constructive recommendations for the target audiences of evaluations. On the other hand, evaluations by system developers were often performed at a minimal scale and prone to misleading results, mostly due to carefully selected evaluation data for 'demonstration' of system performance in a positive way. This kind of evaluations not only failed to adequately reveal the performance of an MT system but also hindered the advancement of the whole field by hiding the real weaknesses and potentials of the technology in use.

MT evaluation started to develop into its own discipline when a good number of research systems were developed and more and more commercial systems entered into the market to compete against one another, demanding a fair and objective assessment of their performance and usability. In the late 1970s, Systran, one of the oldest commercial MT systems, was assessed for the European Community (EC), precursor of the EU. The assessment results were compiled in a report (Slype 1979), presenting the first comprehensive study on MT evaluation. This report covers all existing proposals and practices of MT evaluation at that time, presenting a critical assessment for each of them. Furthermore, it provides a holistic view of MT evaluation as a multi-faceted activity, comprising a range of dimensions including purpose, text typology, effectiveness and efficiency, micro and macro criteria, and methods. Accordingly, MT evaluation extends its boundary to encompass many more interrelated parameters of these kinds than before, and proper settings of such parameters require thorough considerations in different evaluation scenarios.

The DARPA[3] MT initiative (White and O'Connell 1994: 134–140; White *et al.* 1993: 206–210, 1994: 193–205) was a representative attempt at comparative MT evaluation in the 1990s. It was intended to assess the progress of sponsored MT research, involving a heterogeneity of language pairs, computational approaches and potential end-uses. A suite of evaluation methodologies was accordingly formulated with ambitious goals: to be applicable to contextual diversity, economical to administer and portable to other evaluations, with subjectivity minimized. The evaluation covered both research (on statistical, interlingual, and human-assisted MT) and commercial systems, assessing the translation quality and usability of system outputs. Furthermore, evaluation methods were studied and compared for their sensitivity of measurement, efficiency, and the expenditure of human time and effort demanded. Such dual foci on assessing both systems and evaluation methods, i.e., evaluation and meta-evaluation, have been followed by many subsequent practices.

In contrast to evaluations by human judges, a paradigmatic change in MT evaluation in the past decade is the prevalence of automatic evaluation metrics. With the aid of these metrics large-scale evaluations can be conducted on a large number of systems and language pair combinations within a reasonable time and cost. Examples include the IWSLT[4] series (Akiba *et al.* 2004: 1–12; Eck and Hori 2005: 11–32; Fordyce 2007; Paul 2006: 1–15, 2008: 1–17, 2009: 1–18; Paul *et al.* 2010: 3–27), HTRDP (Liu *et al.* 2005: 18–22), TC-STAR (Choukri *et al.* 2007), the CESTA,[5] NIST open MT evaluation,[6] and SMT workshop.[7] Through these evaluations, not only the performance of the state-of-the-art MT approaches but the validity and effectiveness of different evaluation methods, be they manual or automatic, are also examined.

Table 12.1 Types of translation purpose and their requirements of translation quality

Purpose of translation	*Required quality of translation*
Dissemination	Publishable quality
Assimilation	At a lower level of quality
Interchange	Translation between participants in one-to-one communication or of an unscripted presentation
Information access	Translation within multilingual systems of information retrieval, information extraction, database access, etc.

Source: Hutchins (2003)

Applications, purposes and criteria

In essence, evaluating an MT/CAT system is to assess how well it serves what it is aimed to serve. Towards this goal one has to answer at least three interrelated *what*-questions, i.e., the intended applications of the system in question, the purposes of evaluation and then the appropriate criteria to use, before moving ahead to deal with the matter of *how*, i.e., the methodology of evaluation. Answers to these questions determine the design of an evaluation. A clearly defined application entails a specific evaluation purpose, such that a system is assessed only for what it is designed to serve. This purpose guides the selection and definition of appropriate criteria, according to which suitable evaluation methodology can then be formulated.

An MT/CAT system is usually developed for some specific applications. While CAT tools are for restricted uses, e.g., translation memory primarily for supporting translators in the reuse of previous translations, MT systems have a wide range of potential applications, e.g., for use as a CAT facility to provide an initial translation for further post-editing, a utility for gisting/browsing foreign texts, a means for information dissemination, etc. Different purposes require different levels of system performance in terms of translation quality, as in Table 12.1 generalized by Hutchins (2003: 5–26). No MT system has been able to translate any kind of text in any subject at a publishable-quality level. The performance of an MT system is highly dependent on the subject domain(s) to which it is optimized and on the knowledge with which it is equipped. METEO, a specialized MT system, was used to translate weather forecasts between English and French successfully for two decades (1981–2001), using a sublanguage with a restricted lexicon and grammar. A general MT system is only capable of delivering translations in gistable quality in most cases, but can be very useful when translation quality is not the first priority (e.g., to get the rough idea or the subject of a text, so as to locate information or decide whether professional human translation is needed).

Therefore, we have to bear in mind that the usability judgment of MT varies according to the intended use of system output, besides translation quality. In a survey exploring the usefulness of MT from the users' perspective, Morland (2002) notes that 'those who feel comfortable with English do not want pure MT translations [from English to their native language], but those who are not as strong in English find it useful'. A fair comment from this survey is that 'pure MT is rough—often obscure, frequently humorous—but it can be useful'. MT can be selected as a good enough solution for a particular task despite its translation quality. As discussed in Church and Hovy (1993: 239–258), a well-chosen application helps determine how to evaluate a system and make it look good. In contrast, an inappropriate intended task makes it difficult to find a suitable evaluation paradigm, and may lead to bias in interpreting evaluation results.

Apart from a right application, the design of evaluation is also dependent on the purposes of interested parties to conduct the evaluation. Different parties involved in different stages of an MT/CAT system, from research and development to procurement, installation and operation, are interested in different aspects of a given system. Hutchins and Somers (1992) and White (2003: 211–244) discuss the special interests of typical parties in MT. For example, *researchers* would like to know whether, and to what extent, a particular method works for a hypothesis, or is extendable to a new domain. *Developers* have to identify errors that can be corrected within the capacity of a system, find out the limitations of the system such as its coverage of text types and subject domains, and decide what facilities should be provided to intended users. *Lay users* are only able to access a system's output and perform a 'black box' evaluation[8] to examine its capabilities, acceptability and cost-effectiveness in their own working environments. *Translators* are mainly concerned with the gain in productivity from using a system to help with translating, in particular, by way of revising MT output up to an acceptable quality standard.

Accordingly, a particular type of evaluation is needed for finding out the right kind of information to serve a specific purpose of each party. White (2003: 211–244) presents a categorization of MT evaluation including the following types. *Feasibility test* examines the possibility that a theory or method can be accomplished, and its potentiality for success after further research and implementation. *Internal evaluation* focuses on whether some components of an experimental, prototype, or pre-release system can work as planned, and if not, what causes and solutions there are to problems. *Usability evaluation* tests the usefulness of a system for end users, involving its utility and users' satisfaction with it, in terms of the extent to which it enables users to achieve their specific goals. *Operation evaluation* explores the cost-benefits of a system in a particular operational environment. *Declarative evaluation* assesses the ability of a system to translate texts for actual end use. *Comparison evaluation* investigates particular attributes (e.g., translation quality) shared by different systems, in order to find out the best one, the best implementation, the best theoretical approach, etc.

Once the purpose of evaluation is clear and the type of evaluation needed is determined, corresponding criteria can be defined and methodology implemented. For instance, criteria for evaluating an MT system to serve the needs of a translator may first include the quality of system output, because a system of poor output (i.e., below the required level for an intended application) is unlikely to be useful no matter how good it is in other aspects. Other criteria may include the amount and ease of work on pre- and post-editing system input and output, facilities for text editing, consistency of terminology, number of language pairs supported, cost of dictionary maintenance, etc. In general, an MT/CAT system is evaluated on the one hand in terms of the quality of its output, and on the other hand as a piece of software, using criteria such as usability, operability, speed, etc.

Software evaluation

As an MT/CAT system is by nature a computer program, its evaluation can be considered a stage of software engineering, to which existing standards of software evaluation are applicable. Jointly developed by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), the standard ISO/IEC 9126 (1991; 2001) provides a quality model and identifies several types of metric for software evaluation. It defines software quality as 'the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs'. Six generic characteristics are identified, namely, functionality, reliability, usability, efficiency, maintainability and portability, each of which is further broken down into a number of sub-characteristics.[9] A sub-characteristic is composed of a set of attributes, representing some verifiable or measurable features of a software product. This standard was followed by the ISLE[10] project to develop a criterion taxonomy for MT evaluation (ISLE 2001), resulting in FEMTI.[11] FEMTI gathers together and systematizes all possible contexts of evaluation once suggested, in line with the corresponding ISO/IEC characteristics/sub-characteristics and attributes, and collects all MT evaluation methods and metrics proposed so far, categorized under respective attributes. It helps practitioners of MT evaluation identify suitable evaluation methods for their own needs.

Central to the rationale of FEMTI is that MT evaluation is 'only a special, although rather complex, case of software evaluation in general' (Hovy *et al.* 2002: 43–75). The quality of a piece of software can be characterized and measured by (1) first defining the context of evaluation and corresponding characteristics, (2) then identifying attributes related to each characteristic, and (3) finally selecting appropriate methods to assess each attribute. In other words, the quality is compositional in nature. The meaning of some characteristics, like functionality, is operational and has to be specified by choices of attribute and evaluation method. Therefore, the notion of quality or any of its characteristics is ambiguous in isolation. Its semantics becomes clear only when a user has chosen what and how to evaluate, which is determined by factors such as purpose, task and text genre, or, in a more general term, by the context of evaluation.

Hence, there needs to be a specific set of evaluation criteria and corresponding methodology for each scenario of evaluation. For example, Kit and Wong (2008: 299–321) presents a comparative evaluation of online MT systems with a focus on finding out the best, out of the ones available, for lay users in legal translation. Recognizing the inability of MT to deliver high quality translation of legal texts without human post-editing, the purpose of this evaluation is to provide interested parties with substantial evidence for proper selection of online MT systems for different language pairs, assuming a need for practical use of MT for legal translation. Evaluation criteria then include coverage of supported language pairs (regarding how often users have to switch to another MT system), translation quality for a language pair on average (representing how well translation between languages can be done by various systems in general and how confident users can be with it) and translation quality for a language pair in particular (serving as an indicator for selecting the best MT system for a specific language pair). The evaluation of output quality is based on quantitative metrics using a large corpus of legal texts, in order to properly and concretely reveal the performance of online MT systems.

Note that MT/CAT evaluation is not limited to assessing a system as a whole only. It is also applied to a system component, especially in the area of research and development. For instance, an example-based MT system typically consists of three major components respectively responsible for matching, alignment and recombination of translation examples. Each of them

can be viewed as an independent component with its own functions, data input and output, and therefore has to be evaluated using different methods. For matching and alignment, measures such as precision and recall are commonly used to quantify system performance of retrieving relevant translation examples from the database and mapping source-target example fragments respectively, through comparison with a pre-defined set of gold-standard data. Evaluation of recombination is more complicated, because the recombined example fragments are final MT outputs whose quality is highly dependent on the performance of the previous stages, i.e., matching and alignment.

The evaluation of CAT systems is also challenging. There are a wide diversity of system types, such as optical character recognition (OCR), project management, termbank, translation memory, interactive MT, etc., with different designs and functionalities for different applications. Moreover, two systems of the same type may involve different kinds of features, so it may not be easily attainable to directly compare them in a feature-by-feature manner. A translation memory system may support some unique file types while another may have some proprietary technologies to match translation records. On the other hand, it is claimed that many CAT systems enhance translators' productivity, in comparison with not using them. However, such productivity gains, if any, also depend on users' proficiency in system operation. Without sufficient training, an inexperienced user may not gain any significant benefit from using them, or it may even inhibit his/her productivity. This kind of human factor has to be minimized or properly controlled in CAT evaluation.

Approaches to CAT evaluation can be categorized into two types, namely, automatic vs. manual. Automatic evaluation uses objective metrics to quantify the measurable aspects of a system such as speed of execution and usability of system output. A widely used metric to evaluate translation memory and interactive MT systems is keystroke ratio, the ratio between (1) the number of keystrokes required to modify a given system output into a reference translation and (2) the number of characters in the reference translation. A reference translation is a human translation of the same source text in use as a 'model answer'[12] for comparison with system output. The keystroke ratio helps to estimate the amount of human effort required to produce the final translation from the output of TM and/or MT. A ratio larger than one means that revising a system output takes more effort in number of keystrokes than typing the whole reference translation from scratch—literally speaking, the system output in question brings no productivity gain.

Manual evaluation relies on users' subjective judgments and experiences in assessing an MT/CAT system. Focuses may be put on system features, on their quality (e.g., how well they are designed and implemented) and suitability (e.g., to what extent they suit a user's particular needs). Typical evaluation methods of this kind include average user rating on a scale, e.g., a 5- or 7-point scale, and a user trial involving a group of users to test a system under controlled conditions. For example, to investigate the productivity gain by using an MT/CAT system, a user trial can be carried out to measure and compare the difference of time between using and not using the system in question to translate a test set.

Quality of MT output

As the main function of an MT system is to provide a translation service, quality of MT output is of particular interest to all parties, and usually regarded as a primary criterion of MT evaluation. Nevertheless, the notion of 'translation quality' was considered indefinable by some scholars (Slocum 1985: 1–17), accounting for the long-time absence of a universally accepted standard and method for its quantification. Accordingly, multiple quality criteria have

been proposed, demonstrating various interpretations of the notion with an emphasis on different evaluation aspects.

Central to the notion is the question of how text quality is characterized. Even though MT output is still far from reaching the quality of human written text, it is grasped by readers in a similar manner as texts written by humans. Moreover, one of the main purposes of MT evaluation is to assess 'to what extent the makers of a system have succeeded in mimicking the human translator' (Krauwer 1993: 59–66). In other words, a piece of MT output has to show its quality in terms of its success in approximating human translation, and therefore shares many features with the latter, although the technical details of their assessment may be different.

Text quality can be characterized from both monolingual and bilingual perspectives. The quality of a monolingual text is multifaceted, as shown in different evaluations with their own sets of criteria. For example, for language learners' writing assessment a holistic grading can be based on the general impression of their effort with regard to writing, or on the overall success of their written communication, without attending to any particular individual element involved. Alternatively a detailed assessment may use an analytical scale with multiple parameters for scoring, such as those in Diederich (1974), including ideas, organization, wording and phrasing, style, grammar and sentence structure, punctuation, spelling, and legibility of handwriting, whose weights in the final grade can be adjusted to fit different situations.

Quality of bilingual text has much in common with that of a monolingual one. Both text types have to adhere to some general quality criteria such as grammaticality, readability, coherence, etc. What distinguishes a translation from a monolingual text is its correspondence with a source text in another language, demanding an equivalence relation in terms of meaning, in particular. This is a unique feature not required by other text types, and it is central to many translation theories characterizing the notion of translation quality.

Equivalence is also a controversial property of a translation, however, which has been defined and interpreted in different ways over the years. Some notable definitions include what Nida (1964) calls *formal* and *dynamic* equivalence, Catford's (1965) *textual* equivalence, Newmark's (1982, 1988) *semantic* and *communicative* equivalence, among many others. From the perspective of evaluation, House (2009: 222–224) makes the criticism that many treatments of translation equivalence, such as 'faithfulness to the original' or 'the natural flow of the translated text', are 'atheoretical in nature', offering poor operationality, and solely dependent on the knowledge, intuition and competence of a translator. It is difficult to establish general principles and develop, accordingly, a systematic procedure to assess the features needed to characterize translation relationship.

In practice, translation assessment commonly relies on an error-based grading. An example of this is the ATA[13] certification examination. Errors in a translation are identified and rated in terms of their consequence to the meaning, understanding, usefulness, and/or content of a translation. Its overall quality is graded in four dimensions, namely, usefulness/transfer, terminology/style, idiomatic writing, and target mechanics. Each of them is further divided into four ranks for detailed characterization of performance variation.

Mostly evaluation of translation quality requires comprehension as a prerequisite, for determining various kinds of equivalence relation and/or identifying errors in a translation. It is put forth in Hayes *et al.* (1987: 176–240) that 'reading to comprehend' is the basis of 'reading to evaluate'. These two cognitive processes differ in their purposes, in that the former attempts to construct an integrated representation of a text to understand how the ideas in the text work as a whole, while the latter aims at identifying problems in the text and sometimes also at

finding solutions. It is worth noting that readers may also detect problems, whilst endeavoring to understand a text, but they usually do not devote much thinking or conscious attention to them unless the problems are bad enough to hinder their reading.

Evaluation of MT output falls somewhere between reading to comprehend and reading to evaluate. Depending on the purpose of evaluation, evaluators may want to know how comprehensible an MT output is, without any need to diagnose its problems, or to perform a detailed error analysis to examine a system's strengths and weaknesses. However, what is complicated here is that both comprehension and evaluation of an MT output demand a judgment of its correspondence with the source text. Both of them reflect the intelligibility and fidelity of the output, two common criteria for assessing the quality of MT output, referring, respectively, to the extent to which an output can be understood and is accurate in meaning. The quality of a translation may be uneven in these two aspects, because a translation may be strong in following the rules and conventions of the target language but weak in preserving the meaning of its source. However, the reverse is hardly conceivable. Although one may artificially list examples that are 'perhaps optimally faithful, but far less intelligible than a translation' (White 2001: 35–37), such cases rarely occur in reality. It is reasonable to consider that evaluation of fidelity subsumes that of intelligibility, and the former cannot be isolated from the latter. In other words, determination of translation equivalence requires understanding of a text.

A more general challenge for MT evaluation lies in the idiosyncratic difference between MT and human translation. The quality of human translation is in general expected to be publishable, but the best quality of MT in the general domain is for 'gisting' only, except that of the outputs from tailor-made systems for specialized domains. It is not unusual to find an MT output of extremely poor quality with unusual word choices, garbled characters, or unreadable word order, not to mention an appropriate judgment of its quality, such as the following outputs from three MT systems[14] for the same source text.

MT1: o???? face deliver information the hope resumes talk
MT2: Han to will restore the discussion towards the transmission hope the information
MT3: the rok will provide the dprk transfer hopes to resume talks and Information

It is difficult to apply to them any higher-order criteria of evaluation such as functional appropriateness or stylistic elegancy of a text. Thus even though the ultimate goal of MT is to attain human translation quality, a different profile of evaluation criteria and methods needs to be used in order to cope with these kinds of text characteristics of MT output.[15]

Manual evaluation

MT evaluation has relied on human judges since its inception, and will inevitably continue to do so in the future. It is the end users of both MT systems and languages who determine the usefulness of an MT system and judge the quality of its output. Although human judgments of text quality are usually perceived and described as subjective and inconsistent, they are nevertheless the ultimate 'gold standard' that cannot be overridden by any automatic measure.

Manual evaluation of MT output entails two aspects: intrinsic and extrinsic. The former focuses on judgment of language quality, while the latter aims to test the usability of MT output in a specific task that MT is expected to facilitate.

Intrinsic

Quality assessment

In quality assessment evaluators are asked to rate, in terms of their intuitive judgment, the 'goodness' of a translation, which is normally presented sentence by sentence or as a sequence of even smaller syntactic constituents. Two most commonly used criteria in the assessment are *fidelity* and *intelligibility*. Fidelity is about whether the transfer of meaning from a source to a target text is accurate and adequate without loss, addition or distortion. Evaluators have to be bilingual if they need to work on both source texts and MT outputs for comparison, but they can be monolingual if human translation is available as a reference. Intelligibility, on the other hand, is a monolingual attribute of target text, referring to 'the ease with which a translation can be understood' (Slype 1979), regardless of whether the content of a source is accurately translated. It is a key indicator for how easily a reader can grasp key message from a translation.

Both fidelity and intelligibility are rated with a scale. The 5-point scale in Table 12.2 has been widely used in many open MT evaluations in the past decade. For example, in the NIST open MT evaluations, evaluators were instructed to spend no more than 30 seconds on average on assessing both the fidelity and intelligibility of a segment of MT output (LDC 2002). Their qualitative judgments need to be based on an instant intuition, rather than a thorough understanding of translation candidates.

In principle, fidelity and intelligibility are independent of each other. However, there are existing findings to show that they are in fact highly correlated (White 2001: 35–37). It is thus possible to devise an evaluation method by measuring just one of them, e.g., fidelity, and then inferring the other.

Table 12.2 The 5-point fidelity/intelligibility scale

Fidelity	*Intelligibility*
5 All	Flawless
4 Most	Good
3 Much	Non-native
2 Little	Disfluent
1 None	Incomprehensible

Source: LDC (2002)

Translation ranking

Translation ranking resorts to human preference, by ranking a number of translation candidates instead of rating with respect to any of their quality attributes. Evaluators are instructed to 'rank translations from Best to Worst relative to the other choices (ties are allowed)' (Callison-Burch *et al.* 2009: 1–28), given a list of several system outputs each time. A variant of translation ranking is to pick, after a pairwise comparison, a preferred version, or none if the quality of two outputs is indistinguishable. The overall performance of an MT system is then reflected in the average number of times its outputs are ranked higher than the others. Translation ranking has been the official human evaluation method in the statistical MT workshops since 2008, replacing the conventional fidelity/intelligibility judgment (Callison-Burch *et al.* 2008: 70–106).

The formulation of this method is driven by the poor inter-annotator agreement on traditional quality rating. Note that in many cases of MT evaluation what system developers

need most is a system ranking, not the details of a system's quality in terms of particular features such as fidelity and/or intelligibility.

Error analysis

While quality assessment appraises the 'goodness' of a translation, error analysis judges a translation from the opposite perspective, i.e. measuring its 'badness'. It starts with identifying translation errors, and ends with an estimation of 'the amount of work required to correct [a] "raw" MT output to a standard considered acceptable as a translation' (Hutchins and Somers 1992: 164). It seeks to pinpoint the feasibility, potentials and limitations of a system, by examining the contexts in which it is most likely to be effective or prone to failure (Lehrberger and Bourbeau 1988). It is considered a more reliable method than quality assessment, because identifying errors is in general more objective and consistent among evaluators than rating goodness of translation (Schwarzl 2001). Furthermore, its results are usually more meaningful and interpretable to interested parties like system developers and users.

Errors can be classified in different ways according to the variety and complexity of MT systems, grammatical features of texts in various domains or languages, and user demands (Lehrberger and Bourbeau 1988). Table 12.3 illustrates an excerpt of error classification from Vilar *et al.* (2006: 697–702), which can be used to count the numbers of error types in MT output, so as to obtain an overall distribution of error frequencies.

The difficulties of error analysis lie in the identification and classification of errors. First, apart from obvious mistakes in syntax and lexical choices, what constitutes an error is a subjective matter, involving human factors like evaluators' tolerance of imperfections in a sentence and their preference of expression. Second, classifying errors into pre-defined categories is often problematic, because of the unclear boundaries of error types that are closely interlinked in nature (Arnold *et al.* 1994; Flanagan 1994: 65–72). For example, a missing auxiliary verb could be classified as a missing word or an incorrect form of main verb (Trujillo 1999).

Table 12.3 Excerpt of error classification

Word order	*Incorrect word*
Word level	Sense
– Local range	– Wrong lexical choice
– Long range	– Incorrect disambiguation
Phrase level	Incorrect form
– Local range	– Extra word
– Long range	– Style
	– Idiom

Source: Vilar *et al.* (2006: 699)

Extrinsic

Information extraction

A typical use of MT is to assist in extracting key information from foreign texts. According to a user study by Taylor and White (1998: 364–373), the kinds of information of interest to users range from name entities to relationships between participants in events presented in a text. The extent to which users can completely and correctly identify such key information in an MT output is thus a direct indicator of the usability of the output.

This idea is formulated into an evaluation metric in Lo and Wu (2011: 220–229). They defined the usability of MT as the extent to which it can help human readers successfully grasp essential event information: *who did what* to *whom*, *when*, *where*, *why* and *how*. Such event information is closely associated with semantic roles (see Table 12.4) in sentences of MT output, reference translation and source text. Human judges can compare the content of semantic roles on both sides and determine whether those ones in the MT output are correct, partially correct, or incorrect.

Table 12.4 Correspondence of semantic roles and event information

Semantic role	*Event*	*Semantic role*	*Event*
Agent	who	Location	where
Action	did	Purpose	why
Experiencer	what	Manner	how
Patient	whom	Degree or extent	how
Temporal	when	Other adverbial arguments	how

Source: Lo and Wu (2011: 225)

Comprehension test

The ultimate goal of MT is to enable users to comprehend foreign texts. Their understanding of MT output can be examined by reading comprehension tests. Evaluators are given passages of MT output and human translation for the same source text to read, and then a set of questions about the passages to answer. Their performance in the comprehension tests using the two types of passage reflects the degree to which MT can accurately and comprehensibly translate the source texts.

The MT evaluation reported in Tomita (1992) and Tomita *et al.* (1993: 252–265) used TOEFL[16] test materials to see whether the materials translated by humans and by different MT systems result in different degrees of understanding by examinees, in terms of the TOEFL scoring of their answers to questions. In this evaluation, TOEFL passages were rendered into Japanese by MT and human translators, and TOEFL questions translated manually for a group of Japanese students to answer.

Reeder (2001a: 67–71, 2001b) presented another evaluation method based on language learner tests. Following her observation that evaluators were able to distinguish between texts written by native speakers and language learners in fewer than 100 words, Reeder studied their performance on differentiating 'authors' of translations, i.e., human or MT. Native speaker subjects were given a short passage to read and then identified whether it was a human or machine translation. How well the subjects recognized MT output was measured in terms of the accuracy of identification and the number of words required for a judgment.

Another evaluation method is based on the 'cloze procedure' (Somers and Wild 2000). For a passage of MT output, some words (e.g., every word in ten) are masked and subjects are asked to guess the missing words. The underlying idea of this method comes from an observation in Gestalt psychology, i.e., 'the human tendency to complete a familiar but not-quite-finished pattern … by mentally closing up the gaps' (Taylor 1953: 415). The quality of MT output, in terms of readability or intelligibility, is indicated by how much it helps the subjects to guess the missing words correctly.

Post-editing

Post-editing involves a human editor revising an MT output up to an acceptable level of quality. Quality of MT output is assumed to have an inverse correlation with the amount of effort needed for the revision. In this way MT is assessed as a means of raising translators' productivity in terms of the cost-effectiveness of post-editing its output as a usable initial draft of translation. In the worst case, this draft may take a translator even longer to post-edit than to translate its source text from scratch.

A direct measurement of post-editing effort is the amount of time required to revise an MT output. This provides a clear indication of the usability of MT output, for the cost of post-editing is directly reflected in the amount of revision time. Its drawback is also obvious, however. Post-editing time depends on many factors, especially such external ones as post-editors' concentration and working environment. It is unlikely that a post-editor can maintain the same concentration upon the revision of every sentence, nor that two post-editors spend the same time on the same MT output. Therefore it is rather questionable whether different post-editors' time on post-editing can be comparable.

A more objective measure is the *edit-distance*, which counts the number of changes required to revise an MT output, including addition, deletion, substitution and transposition of words. To allow a fair comparison, the number of changes is then normalized by the number of words in the revised translation. Provided with a source text or human translation for reference, post-editors are asked to carry out a minimal number of edits to make an MT output understandable and accurate in meaning.

Automatic evaluation

Automatic evaluation of MT outputs involves the use of quantitative metrics without human intervention in runtime. It is intended to meet the demand from the MT community to overcome the shortcomings of manual evaluation, which is inevitably prone to personal biases of human judges and to inconsistency of subjective judgments. There are different views of what should be considered errors in translation and different levels of acceptability that revision has to achieve. Furthermore, manual evaluation is usually too costly in terms of time and monetary cost, and can hardly cope with the enormous scale of MT evaluation, as evidenced by the common practice of the past decade in the field.[17] Automatic metrics thus serve as a desirable solution providing a quick and cost-effective means for trustable estimation of the quality of MT output.

Text similarity metrics

Most automatic measures for MT evaluation rely on available human translations as reference for comparison with MT outputs. In this way MT evaluation is turned into a problem of computing monolingual text similarity. Rudimentary ideas of this kind can be dated back to the 1950s. Miller and Beebe-Center (1956: 73–80) suggest that 'the fact that a grader can recognize errors [in a student's translation] at all implies that he must have some personal standard against which he compares the student's work … this might consist of his own written translation; more often it is probably a rather vague set of translations that would be about equally acceptable'. A primitive evaluation method for assessing 'the relation between the test translation and the criteria' is thus 'to ask if they use the same words' and 'to compare the order of the words which were common to the test and the criterion translations'. Although there is

rarely only one correct translation for a source text, different versions of translation may share certain common words or phrases. This provides a basis for statistical comparison of an MT output with a reference in terms of common textual features.

BLEU and NIST

BLEU[18] (Papineni *et al.* 2001) is widely recognized as one of the first and the most influential metrics for automatic MT evaluation, well known for its rationale that 'the closer a machine translation is to a professional human translation, the better it is'. It is based on counting the number of n-grams, namely sequences of consecutive word(s) of varying length, co-occurring in an MT output and in one or more versions of corresponding reference, usually each in the form of a sentence. This idea is illustrated in Papineni *et al.* (2001) with the following exemplary translation candidates (*C1–2*) and references (*R1–3*) for the same source text.

> *C1*: It is a guide to action which ensures that the military always obeys the commands of the party.
> *C2*: It is to insure the troops forever hearing the activity guidebook that party direct.
> *R1*: It is a guide to action that ensures that the military will forever heed Party commands.
> *R2*: It is the guiding principle which guarantees the military forces always being under the command of the Party.
> *R3*: It is the practical guide for the army always to heed the directions of the party.

Human readers can easily identify *C1* as a better translation than *C2*. The results of n-gram matching,[19] as presented in Table 12.5, confirm that *C1* shares more n-grams than *C2* with all versions of reference, giving an evaluation result in agreement with human judgment.

Table 12.5 Number of common n-grams in translation candidates (*C1–2*) and references (*R1–3*)

	C1			*C2*		
n-gram length	1	2	3	1	2	3
R1	14	8	6	8	1	0
R2	11	4	1	5	1	0
R3	9	3	1	6	1	0

Upon counting the number of common n-grams, precision,[20] namely the proportion of matched n-grams in a candidate, can then be calculated, by dividing this number by the total number of n-grams in the candidate. For an entire test set, the n-gram precision p_n is defined as

$$p_n = \frac{\sum_{c \in C} \sum_{w^n \in c} \text{match}\left(w^n\right)}{\sum_{c \in C} \sum_{w^n \in c} w^n}$$

where c refers to each sentence in the candidate set C, and w^n an n-gram of the length n. In Papineni *et al.*'s proposal, p_n is computed for n-grams up to length 4 and then averaged to the geometric mean P_{avg} as

$$P_{avg} = \sum_{n=1}^{N} \alpha \log p_n$$

where the max n-gram length $N = 4$ and the weight $\alpha = 1/N$ for each p_n. As length discrepancy between candidate and reference is concerned, the n-gram precision can penalize a candidate for being longer than its reference. However, it cannot properly deal with a too short candidate. For this, a *brevity penalty* factor BP_{BLEU} is introduced into BLEU. It is a decaying exponential with respect to candidate length L_c and reference length L_r, defined as

$$BP_{BLEU} = \begin{cases} 1, & \text{if } L_c > L_r; \\ e^{\left(1-\frac{L_r}{L_c}\right)}, & \text{if } L_c \le L_r. \end{cases}$$

Then BLEU score is calculated as

$$BLEU = BP_{BLEU} \cdot \exp\left(P_{avg}\right)$$

resulting in a number between 0 and 1, with a larger one to indicate higher similarity of candidates to respective references.

NIST[21] is another metric revised from BLEU with a number of modifications (Doddington 2002: 138–145). While BLEU weights each n-gram equally, NIST gives more weight to n-grams that are more informative. The fewer occurrences of an n-gram in reference translation, the more informative it is considered to be. For an n-gram w^n of length n, its information weight Info(w^n) is computed following the equation

$$\text{Info}\left(w^n\right) = \log_2 \frac{\sum_{r\in R}\sum_{w^{n-1}\in r} w^{n-1}}{\sum_{r\in R}\sum_{w^n\in r} w^n}$$

where r refers to a sentence in the reference set R and w^{n-1} an n-gram of length n-1. In other words, it estimates the information of w^n given the first w^{n-1}. Other modifications in NIST include a revised version of brevity penalty, to minimize the penalty on small variations in translation length. Furthermore, the geometric average of n-gram precision in BLEU scoring is changed into an arithmetic average, to avoid possible counterproductive variance due to low occurrence frequency of long n-gram. Accordingly, the NIST metric is formulated as

$$NIST = P_{avg} \cdot \exp\left(BP_{NIST}\right)$$

where

$$P_{avg} = \sum_{n=1}^{N} \frac{\sum_{c\in C}\sum_{w^n\in c}\text{match}\left(w^n\right)\text{Info}\left(w^n\right)}{\sum_{c\in C}\sum_{w^n\in c} w^n}$$

and

$$BP_{NIST} = \beta \log_2 \min(L_c/L_r, 1)$$

In Doddington's proposal, $\beta = -\log_2 2/\log_2 3$ and $N = 5$.

BLEU and NIST have been widely used as *de facto* standard measures to monitor system performance in research and development, especially for the comparison of systems, and adopted as official measures in many open MT evaluations, such as the NIST open MT evaluation, IWSLT, and the statistical MT workshop.

METEOR

METEOR[22] is proposed in Banerjee and Lavie (2005: 65–72) as a recall-oriented evaluation metric, standing out from such precision-oriented ones as BLEU and NIST.[23] It begins with an explicit word-to-word alignment to match every word (i.e., a unigram) in a candidate with a corresponding one, if any, in a reference. To maximize the possibility of matching, it uses three word-mapping criteria: (1) exact character sequences, (2) identical stem forms of word, and (3) synonyms. After the word alignment is created, the unigram precision P and unigram recall R are calculated as

$$P = \frac{m}{L_c} \text{ and } R = \frac{m}{L_r}$$

where m is the number of matched unigrams, and L_c and L_r are the lengths of candidate and reference, respectively. A harmonic mean of P and R is then computed, with a parameter α to control their weights, as

$$F_{mean} = \frac{PR}{\alpha P + (1-\alpha) R}$$

There is a fragmentation penalty for each candidate sentence in a problematic word order. It is applied with respect to the number of 'chunks' in a candidate, formed by grouping together matched unigrams in adjacent positions. In this way, the closer the word order of matched unigrams between candidate and reference, the longer and fewer chunks are formed, and the lower the penalty. The fragmentation penalty is calculated as

$$Penalty = \gamma \left(\frac{ch}{m} \right)^{\beta}$$

where ch is the number of chunks, γ a parameter determining the maximum penalty ($0 \leq \gamma \leq 1$), and β another parameter determining the functional relation between fragmentation and the penalty.

The METEOR score for a candidate is computed as its word matching score after deduction of fragmentation penalty, as

$$METEOR = F_{mean} \left(1 - Penalty\right)$$

METEOR is also characterized by its high flexibility in parameter weighting, which can be straightforwardly optimized to new training data in a different context of evaluation. In Lavie and Agarwal (2007: 228–231), the parameters of METEOR are optimized to human judgments of translation adequacy and fluency, resulting in the following setting (adequacy | fluency): α = (0.82 | 0.78), β = (1.0 | 0.75) and γ = (0.21 | 0.38).

TER

TER[24] (Snover *et al.* 2006: 223–231) is an evaluation metric based on the quantification of edit-distance between two strings of words, i.e., the minimal number of operations required to transform one string into another. It can be used to measure the needed post-editing effort to revise a candidate into a reference.

TER is formulated as the minimal number of insertions *INT*, deletions *DEL*, substitutions *SUB*, and shifts *SHIFT* (reordering) of words that are required for each word in a candidate, i.e.,

$$\text{TER} = \frac{INT + DEL + SUB + SHIFT}{N}$$

where N is the average number of words in the reference(s) in use. It returns a score between 0 and 1, quantifying the difference of two sentences in the range of no difference (no edit is needed) to entirely different (every word has to be changed). A later version of TER, namely TER-Plus (TERP) (Snover *et al.* 2009: 259–268), further extends the flexibility of edit beyond the surface text level to allow edit operations on word stems, synonyms and paraphrases.

ATEC

ATEC[25] (Wong and Kit 2008, 2010: 141–151, 2012: 1060–1068) is a lexical-informativeness based evaluation metric formulated to quantify the quality of MT output in terms of word choice and position, two fundamental aspects of text similarity. It goes beyond word matching to highlight the fact that each word carries a different amount of information contributing the meaning of a sentence, and provides a nice coverage of MT evaluation at the word, sentence and document level.

The assessment of word choice first maximizes the number of matched words in a candidate in terms of word form and/or sense, and then quantifies their significance in terms of informativeness. Words in a candidate and a reference are matched with the aid of various language techniques and resources, including stemming and phonetic algorithms for identifying word stems and homophones, and thesauri and various semantic similarity measures for identifying synonyms and near-synonyms, respectively. Informativeness of each matched and unmatched word is calculated using a term information measure to estimate the significance of each bit of information in a reference that is preserved or missed in a candidate, so that a higher weight is assigned to a more informative word.

Following the observation that position similarity of matched words also critically determines the quality of a candidate, two distance measures are formulated to quantify the divergence of word positioning and ordering between a candidate and a reference. They are used to adjust, by way of penalty, the significance of the information load of matched words.

The basic formulation of ATEC is based on the precision P and recall R of the adjusted matched information $m(c,r)$ in a candidate, which are defined as follows:

$$P = m(c,r) / i(c), \quad R = m(c,r) / i(r), \text{ and}$$

$$m(c,r) = i(c,r)\,Penalty(c,r)$$

where $i(c,r)$, $i(c)$ and $i(r)$ are the information load of matched words, candidate and reference, respectively, and *Penalty* (c,r) is defined in terms of the positioning and ordering distance, their

weights and respective penalty limits. The final ATEC score is computed as the parameterized harmonic mean of P and R with a parameter to adjust their relative weights. It differs from other F_{mean}-based metrics (e.g., METEOR) in a number of ways, including the information-based precision and recall, the penalty by word positioning and ordering distance, and other technical details of parameterization.

The ATEC formulation has undergone several versions, testing the effectiveness of different features and information measures. The version in Wong and Kit (2010), despite its *ad hoc* fashion of formulation and parameter setting, illustrates an impressive performance comparable to other state-of-the-art evaluation metrics.

For evaluation at the document level, connectivity between sentences in a document is further recruited as a new feature, approximated by a typical type of cohesion, namely lexical cohesion. It is a factor to account for the critical difference between human translation and MT output, in that the former uses more cohesive devices than the latter to tie sentences together to form a highly structured text. The lexical cohesion measure *LC* is defined in Wong and Kit (2012) as the ratio of lexical cohesion devices to content words in a candidate. It is integrated into ATEC as

$$ATEC_{doc} = \alpha \cdot LC + (1 - \alpha) \cdot ATEC$$

where α is a weight to balance *LC* and *ATEC*. In this way ATEC extends its granularity of evaluation from the sentence to the document level for a holistic account of translation quality.

Quality estimation

In contrast with the similarity-based MT evaluation, quality estimation (QE) is intended to 'predict' quality of MT output without referencing to any human translation. Potential use of QE includes providing feedback to MT developers for system tuning, selecting the best MT output from multiple systems, and filtering out poor MT outputs that can hardly be comprehended or that require a considerable amount of effort for post-editing.

The rationale, or assumption, of QE is that quality of MT output is, to a certain extent, determined by a number of features of the source text and source/target language. For instance, the length and the structural complexity of a source sentence usually have an inverse correlation with the quality of its output. Also, the lengths of source and target sentence are normally close to a particular ratio, and a significant deviation from this ratio may signal a problematic output. It is thus possible to train a QE predictor with certain features, using available machine learning techniques to capture the relationship between these features and quality ratings of MT output in training data. The number of features may range from dozens to several hundred, as illustrated in the previous works by Blatz *et al.* (2003), Rojas and Aikawa (2006: 2534–2537) and Specia *et al.* (2010: 39–50), among many others.

QE can be categorized into *strong* and *weak* forms according to precision of estimation. The former gives a numerical estimate of correctness, as a probability or a score in a given range, whereas the latter only a binary classification of correctness (e.g., 'good' vs. 'bad' translation). Recent progress in QE has demonstrated a comparable performance to that of commonly used automatic evaluation metrics in the field, in terms of correlation with human judgments (Specia *et al.* 2010: 39–50).

Meta-evaluation

Automatic evaluation metrics and QE measures both need to be evaluated by what is called *meta-evaluation*, in order to validate their reliability and identify their strengths and weaknesses. It has become one of the main themes in various open MT evaluations. The reliability of an evaluation metric depends on its consistency with human judgments, i.e., the correlation of its evaluation results with manual assessment, to be measured by correlation coefficients. Commonly used correlation coefficients for this purpose include Pearson's r (Pearson 1900: 157–175), Spearman's ρ (Spearman 1904: 72–101) and Kendall's τ (Kendall 1938: 81–93). The magnitude of correlation between evaluation scores and human judgments serves as the most important indicator of the performance of a metric.

Using the correlation rate with human judgment as objective function, parameters of an evaluation metric can be optimized. Two parameters that have been extensively studied are the amount of test data and the number of reference versions needed to rank MT systems reliably. Different experiments (Coughlin 2003: 63–70; Estrella *et al.* 2007: 167–174; Zhang and Vogel 2004: 85–94) give results to support the idea that a minimum of 250 sentences are required for texts of the same domain, and 500 for different domains. As there are various ways to translate a sentence, relying on only one version of reference may miss many other possible translations. Hence multiple references are recommended for a necessary coverage of translation options. Finch *et al.* (2004: 2019–2022) find that the correlation rate of a metric usually rises along with the number of references in use and becomes steady at four. No significant gain is then further obtained from more references. Furthermore, Coughlin (2003: 63–70) shows that even a single reference can yield a reliable evaluation result if the size of test data is large enough, i.e., 500 sentences or above, or the text domain is highly technical, e.g., computer.

Nevertheless, the reliability of evaluation metrics remains a highly disputed issue. Although the evaluation results of automatic metrics do correlate well with human judgments in most cases, there are still discordant ones. For instance, Culy and Riehemann (2003: 1–8) show that BLEU performs poorly on ranking MT output and human translation for literary texts, and some MT outputs even erroneously outscore professional human translations. Callison-Burch *et al.* (2006: 249–256) also give an example in a 2005 NIST MT evaluation exercise in which a system ranked at the top in human evaluation is ranked only sixth by BLEU scoring. Thurmair (2005) attributes the unreliable performance of evaluation metrics, especially BLEU, to the way they score translation quality. Since most evaluation metrics rely heavily on word matching against reference translation, a direct word-to-word translation is likely to yield a high evaluation score, but a free translation would then be a disaster. Babych *et al.* (2005: 412–418) state that the evaluation metrics currently in use in the field cannot give a 'universal' prediction of human perception of translation quality, and their predictive power is 'local' to a particular language or text type. The Metrics for Machine Translation Challenge[26] which aims at formally evaluating existing automatic MT evaluation technology results in the following views on the shortcomings of current metrics (NIST 2010):

- They have not yet been proved able to consistently predict the usefulness, adequacy, and reliability of MT technologies.
- They have not demonstrated that they are as meaningful in target languages other than English.
- They need more insights into what properties of a translation should be evaluated and into how to evaluate those properties.

Currently, MT evaluation results based on automatic metrics are mainly used for ranking systems. They provide no other useful information about the quality of a particular piece of translation. Human evaluation is still indispensable whenever an in-depth and informative analysis is needed.

Summary

MT/CAT evaluation is characterized by its multi-dimensional nature. Focusing on addressing the issue of interpretation of translation 'quality', different modes of evaluation have been developed, with different criteria for different purposes, applications and users. Methodologically they belong to the categories of software evaluation, using existing standards and criteria, and text quality assessment, using a diversity of manual and automatic methods.

A genuine challenge of evaluating MT output lies in the critical difference between MT and human translation. While the quality of human translation is expected to be publishable in general, the best quality of MT in the general domain is at most suitable for 'gisting' only, except those systems specifically tailor-made for specialized domains. Evaluation methods for human translation are not directly applicable to MT without necessary modification. Both the standard of acceptable translation quality and evaluation criteria have to be adjusted accordingly, in order to avoid over-expectation from MT and to cope with the context of the practical use of MT.

The current trend of MT evaluation is shifting from human assessment towards automatic evaluation using automatic metrics. Despite the overall strong correlation between automatic and manual evaluation results as evidenced in previous works, automatic metrics are not good enough to resolve all doubts on their validity and reliability. In practice, they do not directly assess the quality of MT output. Rather, they measure how similar a piece of MT output is to a human translation reference. Theoretical support is yet to be provided for basing the evaluation on the relation between text similarity and translation quality, although there has been experimental evidence to support the correlation of these two variables. It has been on our agenda to further explore whether such a correlation would remain strong, constant and even valid across evaluation contexts involving different language pairs, text genera and systems.

Notes

1 ALPAC: The Automatic Language Processing Advisory Committee.
2 However, the original interpretation of the data in Orr and Small (1967) is generally positive, concluding that MT outputs 'were surprisingly good and well worth further consideration', while noting the poorer quality of MT outputs in comparison to human translations. Comparing this with the contrasting reading in the ALPAC report, we can see that the same evaluation result can be inconsistently interpreted with regard to different expectations, perspectives and purposes.
3 DARPA: the US government Defense Advanced Research Projects Agency.
4 IWSLT: the International Workshop on Spoken Language Translation.
5 Campagne d'Evaluation de Systèmes de Traduction Automatique (Machine Translation Systems Evaluation Campaign), at http://www.technolangue.net/article.php3?id_article=199.
6 http://nist.gov/itl/iad/mig/openmt.cfm.
7 Workshop on Statistical Machine Translation, at http://www.statmt.org.
8 This term is coined to refer to the kind of system testing that focuses only on the output without any examination of the internal operations producing such an output.
9 An extended standard, ISO/IEC 25010, was released in 2011 as a replacement of ISO/IEC 9126, defining 8 quality characteristics and 31 sub-characteristics.
10 ISLE: International Standards for Language Engineering.

11 FEMTI: Framework for the Evaluation of Machine Translation in ISLE, at http://www.isi.edu/natural-language/mteval/.
12 In essence, a source text may have multiple versions of acceptable translation, and any of them can hardly be regarded as the sole 'model answer'. In practice, however, given a test set of sufficient size, using single translation as reference usually yields reliable evaluation result.
13 American Translators Association, at http://www.atanet.org.
14 Quoted from the Multiple-Translation Chinese (MTC) dataset part-2 (Huang *et al.* 2003).
15 This is in contrast with the early thought that MT outputs should share the same quality scale as human translations. For instance, it was once conceived in Miller and Beebe-Center (1956) that 'a scale of the quality of translations should be … applicable to any translation, whether produced by a machine or by a human translator'.
16 TOEFL: Test of English as a Foreign Language.
17 For example, the NIST Open MT Evaluation 2009 (http://www.itl.nist.gov/iad/mig/tests/mt/2009/) includes four sections, each of 31,000−45,000 words of evaluation data and 10−23 participating systems, i.e. a total of 2.4 million words of MT output to evaluate, indicating the impracticality of resorting to any manual evaluation approach at a reasonable cost and in a reasonable time.
18 BLEU: BiLingual Evaluation Understudy.
19 For instance, the co-occurring n-grams in *C1* and *R1* are as follows.
1-gram: It, is, a, guide, to, action, ensures, that, the, military, the, commands, the, party;
2-gram: It is, is a, a guide, guide to, to action, ensures that, that the, the military;
3-gram: It is a, is a guide, a guide to, guide to action, ensures that the, that the military.
20 Also see the chapter of Information Retrieval and Text Mining.
21 NIST: the National Institute of Standards and Technology.
22 METEOR: Metric for Evaluation of Translation with Explicit Ordering.
23 In general, a recall-oriented metric measures the proportion of reference content preserved in a translation candidate, whereas a precision-oriented one measures the proportion of candidate content matched with a reference.
24 TER: Translation Edit Rate.
25 ATEC: Assessment of Text Essential Characteristics.
26 http://www.nist.gov/itl/iad/mig/metricsmatr.cfm.

References

Akiba, Yasuhiro, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii (2004) 'Overview of the IWSLT04 Evaluation Campaign', in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-04)*, Kyoto, Japan, 1−12.

ALPAC (1966) *Languages and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC: National Academy of Sciences, National Research Council.

Arnold, Doug, Lorna Balkan, Siety Meijer, R. Lee Humphreys, and Louisa Sadler (1994) *Machine Translation: An Introductory Guide*, London: Blackwells-NCC.

Babych, Bogdan, Anthony Hartley, and Debbie Elliott (2005) 'Estimating the Predictive Power of N-gram MT Evaluation Metrics across Language and Text Types', in *Proceedings of Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12−16 September 2005, Phuket, Thailand, 412−418.

Banerjee, Satanjeev and Alon Lavie (2005) 'METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments', in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, 65−72.

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing (2003) *Final Report of Johns Hopkins 2003 Summer Workshop on Confidence Estimation for Machine Translation*, Baltimore, MD: Johns Hopkins University.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (2006) 'Re-evaluating the Role of BLEU in Machine Translation Research', in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Trento, Italy, 249−256.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder (2008) 'Further Meta-evaluation of Machine Translation', in *Proceedings of the ACL Workshop on Statistical Machine Translation (WMT-08)*, Columbus, OH, 70–106.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder (2009) 'Findings of the 2009 Workshop on Statistical Machine Translation', in *Proceedings of the EACL Workshop on Statistical Machine Translation (WMT-09)*, Athens, Greece, 1–28.

Carroll, John B. (1966) 'An Experiment in Evaluating the Quality of Translations', *Mechanical Translation and Computational Linguistics* 9(3–4): 55–66.

Catford, John Cunnison (1965) *A Linguistic Theory of Translation: An Essay in Applied Linguistics*, London: Oxford University Press.

Choukri, Khalid, Olivier Hamon, and Djamel Mostefa (2007) 'MT Evaluation and TC-STAR', in *Proceedings of MT Summit XI: Workshop on Automatic Procedures in MT Evaluation*, Copenhagen, Denmark.

Church, Kenneth W. and Eduard H. Hovy (1993) 'Good Applications for Crummy Machine Translation', *Machine Translation* 8(4): 239–258.

Coughlin, Deborah (2003) 'Correlating Automated and Human Assessments of Machine Translation Quality', in *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches*, 23–27 September 2003, New Orleans, LA, 63–70.

Culy, Christopher and Susanne Z. Riehemann (2003) 'The Limits of N-gram Translation Evaluation Metrics', in *Proceedings of Machine Translation Summit IX: Machine Translation for Semitic Languages: Issues and Approaches*, 23-27 September 2003, New Orleans, LA, 1–8.

Diederich, Paul Bernard (1974) *Measuring Growth in English*, Urbana, IL: National Council of Teachers of English.

Doddington, George (2002) 'Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics', in *Proceedings of Human Language Technology Conference (HLT-02)*, San Diego, CA, 138–145.

Eck, Matthias and Chiori Hori (2005) 'Overview of the IWSLT 2005 Evaluation Campaign', in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-05)*, Pittsburgh, PA, 11–32.

Estrella, Paula, Olivier Hamon, and Andrei Popescu-Belis (2007) 'How Much Data Is Needed for Reliable MT Evaluation? Using Bootstrapping to Study Human and Automatic Metrics', in *Proceedings of Machine Translation Summit XI*, 10–14 September 2007, Copenhagen Business School, Copenhagen, Denmark, 167–174.

Finch, Andrew, Yasuhiro Akiba, and Eiichiro Sumita (2004) 'How Does Automatic Machine Translation Evaluation Correlate with Human Scoring as the Number of Reference Translations Increases?' in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 26–28 May 2004, Lisbon, Portugal, 2019–2022.

Flanagan, Mary (1994) 'Error Classification for MT Evaluation', in *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas: Technology Partnerships for Crossing the Language Barrier (AMTA-94)*, Columbia, MD, 65–72.

Fordyce, Cameron S. (2007) 'Overview of the IWSLT 2007 Evaluation Campaign', in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-07)*, Trento, Italy.

Hayes, John R., Linda Flower, Karen A. Schriver, James F. Stratman, and Linda Carey (1987) 'Cognitive Processes in Revision', in Sheldon Rosenberg (ed.) *Advances in Applied Psycholinguistics, Volume II: Reading, Writing, and Language Processing*, Cambridge: Cambridge University Press, 176–240.

House, Juliane (2009) 'Quality', in Mona Baker and Gabriela Saldanha (eds) *Routledge Encyclopedia of Translation Studies*, 2nd edition, London and New York: Routledge, 222–224.

Hovy, Eduard, Margaret King, and Andrei Popescu-Belis (2002) 'Principles of Context-based Machine Translation Evaluation', *Machine Translation* 17(1): 43–75.

Huang, Shudong, David Graff, Kevin Walker, David Miller, Xiaoyi Ma, Chris Cieri, and George Doddington (2003) *Multiple-Translation Chinese (MTC) Part 2*, Linguistic Data Consortium. Available at https://catalog.ldc.upenn.edu/LDC2003T17

Hutchins, W. John (2003) 'The Development and Use of Machine Translation Systems and Computer-based Translation Tools', *International Journal of Translation* 15(1): 5–26.

Hutchins, W. John and Harold L. Somers (1992) *An Introduction to Machine Translation*, London: Academic Press.

ISLE (2001) *Evaluation of Machine Translation*. Available at: http://www.isi.edu/natural- language/mteval.

ISO/IEC 9126 (1991). *Information Technology – Software Product Evaluation, Quality Characteristics and Guidelines for Their Use*. International Organization for Standardization and International Electrotechnical Commission, Geneva, Switzerland.

ISO/IEC 9126 (2001). *Software Engineering – Product quality, Part 1: Quality Model*. International Organization for Standardization and International Electrotechnical Commission, Geneva, Switzerland.

Kendall, Maurice George (1938) 'A New Measure of Rank Correlation', *Biometrika* 30(1–2): 81–93.

Kit, Chunyu and Tak-ming Wong (2008) 'Comparative Evaluation of Online Machine Translation Systems with Legal Texts', *Law Library Journal* 100(2): 299–321.

Krauwer, Steven (1993) 'Evaluation of MT Systems: A Programmatic View', *Machine Translation* 8(1–2): 59–66.

Lavie, Alon and Abhaya Agarwal (2007) 'METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments', in *Proceedings of the Second ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic, 228–231.

LDC (2002) 'Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations'. Available at: http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf.

Lehrberger, John and Laurent Bourbeau (1988) *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*, Amsterdam and Philadelphia: John Benjamins.

Liu, Qun, Hongxu Hou, Shouxun Lin, Yueliang Qian, Yujie Zhang, and Isahara Hitoshi (2005) 'Introduction to China's HTRDP Machine Translation Evaluation', in *Proceedings of Machine Translation Summit X: 2nd Workshop on Example-based Machine Translation*, 12–16 September 2005, Phuket, Thailand, 18–22.

Lo, Chi-kiu and Dekai Wu (2011) 'MEANT: An Inexpensive, High-accuracy, Semi-automatic Metric for Evaluating Translation Utility via Semantic Frames', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-11)*, Portland, OR, 220–229.

Miller, George A. and J.G. Beebe-Center (1956) 'Some Psychological Methods for Evaluating the Quality of Translations', *Mechanical Translation* 3(3): 73–80.

Morland, D. Verne (2002) 'Nutzlos, bien pratique, or muy util? Business Users Speak out on the Value of Pure Machine Translation', in *Translating and the Computer 24*, London: ASLIB.

Newmark, Peter (1982) *Approaches to Translation*, Oxford: Pergamon Press.

Newmark, Peter (1988) *A Textbook of Translation*, New York and London: Prentice Hall.

Nida, Eugene A. (1964) *Towards a Science of Translating*, Leiden: E. J. Brill.

NIST (2010) 'The NIST Metrics for MAchine TRanslation 2010 Challenge (MetricsMATR10): Evaluation Plan'. Available at: http://www.nist.gov/itl/iad/mig/metricsmatr10.cfm.

Orr, David B. and Victor H. Small (1967) 'Comprehensibility of Machine-aided Translations of Russian Scientific Documents', *Mechanical Translation and Computational Linguistics* 10(1–2): 1–10.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2001) *BLEU: A Method for Automatic Evaluation of Machine Translation*, IBM Research Report RC22176 (W0109–022).

Paul, Michael (2006) 'Overview of the IWSLT06 Evaluation Campaign', in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-06)*, Kyoto, Japan, 1–15.

Paul, Michael (2008) 'Overview of the IWSLT 2008 Evaluation Campaign', in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-08)*, Hawaii, 1–17.

Paul, Michael (2009) 'Overview of the IWSLT 2009 Evaluation Campaign', in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-09)*, Tokyo, Japan, 1–18.

Paul, Michael, Marcello Federico, and Sebastian Stüker (2010) 'Overview of the IWSLT 2010 Evaluation Campaign', in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-10)*, Paris, France, 3–27.

Pearson, Karl (1900) 'On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling', *Philosophical Magazine* 50(5): 157–175.

Reeder, Florence (2001a) 'In One Hundred Words or Less', in *Proceedings of the MT Summit Workshop on MT Evaluation: Who did What to Whom?* Santiago de Compostela, Spain, 67–71.

Reeder, Florence (2001b) 'Is That Your Final Answer?' in *Proceedings of the 1st International Conference on Human Language Technology Research (HLT-01)*, San Diego, CA.

Rojas, David M. and Takako Aikawa (2006) 'Predicting MT Quality as a Function of the Source Language', in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, 2534–2537.

Schwarzl, Anja (2001) *The (Im)Possibilities of Machine Translation*, Frankfurt: Peter Lang Publishing.

Slocum, Jonathan (1985) 'A Survey of Machine Translation: Its History, Current Status, and Future Prospects', *Computational Linguistics* 11(1): 1–17.

Slype, Georges van (1979) *Critical Study of Methods for Evaluating the Quality of Machine Translation*, Tech. rep. Bureau Marcel van Dijk / European Commission, Brussels.

Snover, Matthew, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006) 'A Study of Translation Edit Rate with Targeted Human Annotation', in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Visions for the Future of Machine Translation (AMTA-06)*, Cambridge, MA, 223–231.

Snover, Matthew, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz (2009) 'Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric', in *Proceedings of the 4th Workshop on Statistical Machine Translation (WMT-09)*, 30–31 March 2009, Athens, Greece, 259–268.

Somers, Harold L. and Elizabeth Wild (2000) 'Evaluating Machine Translation: The Cloze Procedure Revisited', in *Proceedings of the 22nd International Conference on Translating and the Computer*, London, UK.

Spearman, Charles Edward (1904) 'The Proof and Measurement of Association between Two Things', *The American Journal of Psychology* 15: 72–101.

Specia, Lucia, Dhwaj Raj, and Marco Turchi (2010) 'Machine Translation Evaluation versus Quality Estimation', *Machine Translation* 24(1): 39–50.

Taylor, Kathryn and John White (1998) 'Predicting What MT Is Good for: User Judgements and Task Performance', in *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas: Machine Translation and the Information Soup (AMTA-98)*, 28–31 October 1998, Langhorne, PA, 364–373.

Taylor, Wilson L. (1953) 'Cloze Procedure: A New Tool for Measuring Readability', *Journalism Quarterly* 30: 415–433.

Thurmair, Gregor (2005) 'Automatic Means of MT Evaluation', in *Proceedings of the ELRA-HLT Evaluation Workshop*, Malta.

Tomita, Masaru (1992) 'Application of the TOEFL Test to the Evaluation of Japanese-English MT', in *Proceedings of the AMTA Workshop on MT Evaluation*, San Diego, CA.

Tomita, Masaru, Masako Shirai, Junya Tsutsumi, Miki Matsumura, and Yuki Yoshikawa (1993) 'Evaluation of MT Systems by TOEFL,' in *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation: MT in the Next Generation (TMI-93)*, 14–16 July 1993, Kyoto, Japan, 252–265.

Trujillo, Arturo (1999) *Translation Engines: Techniques for Machine Translation*, Springer-Verlag.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney (2006) 'Error Analysis of Statistical Machine Translation Output', in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, Genoa, Italy, 697–702.

White, John S., Theresa A. O'Connell, and Lynn M. Carlson (1993) 'Evaluation of Machine Translation', in *Proceedings of the Workshop on Human Language Technology (HLT-93)*, Plainsboro, NJ, 206–210.

White, John S. and Theresa A. O'Connell (1994) 'Evaluation in the ARPA Machine Translation Program: 1993 Methodology', in *Proceedings of the Workshop on Human Language Technology (HLT-94)*, Plainsboro, NJ, 134–140.

White, John S., Theresa A. O'Connell, and Francis E. O'Mara (1994) 'The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches', in *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-94)*, Columbia, MD, 193–205.

White, John S. (2001) 'Predicting Intelligibility from Fidelity in MT Evaluation', in *Proceedings of the MT Summit Workshop on MT Evaluation*, Santiago de Compostela, Spain, 35–37.

White, John S. (2003) 'How to Evaluate Machine Translation', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadephia: John Benjamins Publishing Company, 211–244.

Wong, Tak-ming and Chunyu Kit (2010) 'ATEC: Automatic Evaluation of Machine Translation via Word Choice and Word Order', *Machine Translation* 23(2–3): 141–151.

Wong, Tak-ming and Chunyu Kit (2008) 'Word Choice and Word Position for Automatic MT Evaluation', in *Proceedings of the AMTA 2008 Workshop – NIST MetricsMATR 08*, Waikiki, HI.

Wong, Tak-ming and Chunyu Kit (2012) 'Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, Jeju, Korea, 1060−1068.

Zhang, Ying and Stephan Vogel (2004) 'Measuring Confidence Intervals for the Machine Translation Evaluation Metrics', in *Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, 4–6 October 2004, Baltimore, MD, 85−94.

13

THE TEACHING OF MACHINE TRANSLATION

The Chinese University of Hong Kong as a case study

Cecilia Wong Shuk Man

INDEPENDENT SCHOLAR

Introduction

The boost in technological development has led to a need to include the teaching of technology in traditional disciplines technically, practically and theoretically. Translation training is no exception and occupational needs must be met to cope with the technological changes. As Ignacio (2010) states, 'Translation education needs to give graduates not only the ability to use the technology, but also the frame through which to understand such change.'

According to Hutchins (1986), the idea of using mechanical devices to overcome language barriers was first suggested in the seventeenth century. However, all proposals required human translators to use the tools and involved no construction of machines. After the invention of mechanical calculators in the nineteenth and twentieth centuries, pioneering activities were initiated by Charles Babbage. The first proposal for 'Translating Machines' appeared in 1933. Two patents for mechanical dictionaries have been issued: French: Georges Artsrouni (July 1993) and Russian: Petr Petrovich Smirnov-Troyanskii (September 1993). Georges Artsrouni's idea, 'Mechanical Brain', was a device, worked by electric motor, for recording and retrieving information on a broad paper band, which could store several thousand characters. Each line on the tape contained the entry word (SL word) and equivalents in other languages (TL equivalents). Perforations were coded on a second paper or metal band as a selector of correspondences. Petr Petrovich Smirnov-Troyanskii created a machine for the selection and printing of words while translating from one language into another or into several others simultaneously. He envisaged three stages in the translation process, and the machine was involved in the second stage as an 'automated dictionary'. It included both bilingual and multilingual translation and became the basic framework for subsequent machine translation systems.

In 1946–1949, an electronic digital computer was created following World War II. The start of development of the MT system was initiated by conversations and correspondence between Andrew D. Booth (a British crystallographer) and Warren Weaver of the Rockefeller Foundation in 1947. Andrew D. Booth and Richard H. Richens collaborated in developing a

strict word-by-word dictionary translation by using punched card machinery on a wide variety of languages in 1948. In their approach, they segmented words into stems and endings in order to reduce the size of the dictionaries and introduce grammatical information into the system. It became the fundamental approach applied in machine translation systems in early development – the 'Rule-based approach'.

In 1949, Warren Weaver suggested that 'some reasonable way could be found of using the micro context to settle the difficult cases of ambiguity'. He believed that the translation problem could largely be solved by 'statistical semantic studies' as logical structures of language are with probabilistic uniformities. He thought the investigation of language invariants or universals was the most promising approach – thus the basic idea of the statistical approach applied in machine translation systems used nowadays.

In 1950, Erwin Reifler, head of Department of Far Eastern and Slavic Languages and Literature at the University of Washington in Seattle, introduced the concepts of 'pre-editor' to prepare the text for input into the computer (to indicate the grammatical category of each word in SL) and 'post-editor' to resolve problems and tidy up the style of the translation (to select the correct translation from the possibilities found and to rearrange the word order of the TL). Pre-editing and post-editing processes are still necessary at present so as to make the translation results generated by machine translation systems usable.

In 1951, the first full-time researcher on Machine Translation (MT), Yehoshua Bar-Hillel, produced a survey of the (con)current position of MT at the end of 1951. He suggested the use of machines in different processes: (1) analyzing each word into stem and grammatical categories, (2) identifying small syntactical units, (3) transforming a sentence into another that is the logical equivalent to it. He suggested a second stage – building an explicit, programmable method for syntactic analysis. He also considered the possibilities of constructing a universal grammar or 'transfer grammars … in which the grammar of one language is stated in categories appropriate to some other language'. In 1952, the Rockefeller Foundation sponsored the first conference on MT, held at the Massachusetts Institute of Technology (MIT) and organized by Bar-Hillel. After the 1952 conference, an MT research team was established by Leon Dostert at Georgetown University. In collaboration with IBM, by the end of 1953 the team had developed the first MT program with Russian–English translation. Its public demonstration in January 1954 was the first real demonstration of MT on a computer. It was the first implementation of translation beyond word-for-word translation. There was no pre-editing required. In the MT system was a vocabulary of just 250 Russian words, a mere six rules of grammar and a carefully selected sample of Russian sentences. It might be regarded as the first practical machine translation system using a rule-based approach. Although there were limitations, it showed that MT was open to further development.

Machine Translation has been developing for over 60 years. Since the start of the government-motivated and military-supported Russian–English translations in the US in the 1950s and the 1960s, there has been intensive research activity. After the ALPAC (Automatic Language Processing Advisory Committee) report, Machine Translation was considered a 'failure' and no longer worthy of serious scientific consideration. From the mid-1960s, Machine Translation research was ignored. However, in the 1970's multilingual problems prevailed in the European Community. A change in attitudes toward Machine Translation arose in Europe. The Commission of the European Community (CEC) purchased the English–French version of the Systran system in 1976 (a greatly improved product from the earliest systems developments at Georgetown University in Washington, DC). In the years that followed, different systems developed in the United States and in European countries for

translating texts in various areas including the military, scientific, industrial and medical areas as well as in weather forecasting. Then in the 1980s, it spread among the Asian countries. With Japanese being an isolated language with no similarities to any other language, the need for good translation services was crucial to Japan's commercial and economic growth. Therefore, a great demand for translation to and from English and other languages, led to rapid Machine Translation development in Japan. In the 1980s, the Japanese 'fifth generation' project had established a major position in the future world economy. In recent years, different machine translation systems have produced readable translations.

As Hutchins stated, 'Machine Translation is the application of computers to the translation of texts from one natural language into another' (1986). Arnold *et al.* put it as 'the attempt to automate all, or part of, the process of translating from one human language to another' (1994). The Machine Translation System tends to automate all of the translation process whereas the Computer-aided Translation System is a partly automatic machine translation system equipped with computational storage tools and matching algorithms for past translation reuse. The translation process requires human intervention and the translation decisions are mainly made by humans.

Throughout its development, machine translation systems have been evolving from the first generation direct approach – doing word-to-word translation with the computer as electronic dictionary lookup – to the practical transfer approach, putting emphasis on the differences between language pairs in translating natural languages. This evolution was primarily in the systems using linguistic rules and analyses as the translation framework. On the other hand, some systems employ a corpus-based approach in order to fully utilize the information resources available through access to the World Wide Web. With the abundant text data available from the internet, through the use of statistical calculation, machine translation systems can generate usable translation results from the calculated probabilities. Alternatively, systems can also directly make use of aligned translated texts as built-in translation memory for generating new translations. With the example-based machine translation approach, machine translation systems work like a computer-aided translation system. However, the translation memory is built in by developers instead of accumulatively created by users. Various approaches were invented as relatively new techniques for machine translation. Pattern-based as well as knowledge-based approaches are examples of such techniques and they can be considered as extensions of the basic linguistic transfer rule-based approach and/or corpus-based approach. Nowadays, most machine translation systems use a hybrid approach so that they can take advantage of the different methods of translation.

The Teaching of Machine Translation: The Chinese University as a case study

The curriculum design of the Master of Arts in Computer-Aided Translation (MACAT) Programme in The Chinese University of Hong Kong includes both theoretical and practical courses. Emphasis is being placed on both the machine translation and computer-aided translation areas. After several decades' development of machine translation, we can introduce the students to the current scenario in the field, teaching them what machine translation is nowadays and equipping them with hands-on experience of the systems. In this chapter, an account of the experience of teaching the machine translation related courses *Editing Skills for Computer Translation* and *Computer Translation* is given.

Different course structures based on different aims and objectives of the computer translation courses (frameworks of the courses)

As introduced by Chan Sin-wai (2010: 86), the Master of Arts in Computer-Aided Translation of The Chinese University of Hong Kong 'is a graduate program that places equal emphasis on computer-aided translation and machine translation'. In contrast to the required course on Computer-Aided Translation (CAT), *Introduction to Computer-Aided Translation*, we have the elective course, *Computer Translation* for introducing basic concepts and knowledge in computer translation. *Editing Skills for Computer Translation* is a required course for introducing techniques in making the best use of the results generated by machine translation systems.

Computer Translation – an elective course

Computer Translation is set as an introductory course for teaching basic concepts and theory concerning computer translation (with another complementary course, *Introduction to Computer-Aided Translation* introducing concepts concerning CAT). It therefore focuses more on knowledge transmission. According to Somers (2003), there are different perspectives in using machine translation (what we refer to as 'computer translation' here) in the classroom depending on the types of 'student'. The perspectives include (1) teaching about computers and translation, (2) teaching of the software to trainee translators, (3) teaching languages, and (4) educating end users to use machine translation software. The computer translation course belongs to the first perspective, teaching about computer and translation and the fourth perspective, educating students on the use of machine translation systems.

Curriculum design

In curriculum design, the *Computer Translation* course involves both an introduction to theoretical concepts and acquisition of practical skills. Machine translation is different from computer-aided translation. The whole translation process is automatically done by the computer. The quality of the translation is highly dependent on the design and implementation of the machine translation systems. Basic concepts on how computers translate then have to be introduced in *Computer Translation*. Students then have an idea of the steps (including natural language processing steps) undergone within the computer when making the translations. How the source language is analyzed and how the target translation output is generated is explained. Machine translation systems generate translation through various approaches using different computational algorithms. How the systems make their translations through these approaches is also introduced in the course. Different approaches employed in the translation systems play a significant role in making accurate translation output which also influences the quality of the output. Through understanding the approaches, students get to know the strengths and weaknesses of the translation systems, and so use suitable software for translating different specific genres. Through taking the course, students should be able to evaluate different translation systems by their performance in various aspects, including accuracy, speed, and algorithm employed. Moreover, according to their different developmental strategies and target customers, various translation systems' strengths may lie with particular types of texts. In their group presentation done at the end of the course, each group of students will have to evaluate the performance of different machine translation tools on specific genres (i.e. different genres for different groups) so that they can explore various performances of the software on different genres or text types in a collaborative evaluation. Through such an evaluation,

students can make an informed decision when selecting certain systems for their own use. In preparing the presentation, they would have a lot of hands-on experience on different machine translation systems provided by the programme. The topics covered in *Computer Translation* include the following:

> Introduction to Computer Translation, which includes the basic concepts of computer translation and the differences between computer translation and computer-aided translation.
>
> Different approaches to Computer Translation, which include
> - Corpus-based
> - Example-based
> - Rule-based
> - Knowledge-based
> - Memory-based
> - Pattern-based and
> - Statistical approaches
> - Natural language text processing in computer translation, namely, word segmentation, part-of-speech tagging and parsing
> - Hands-on experience in translation systems in the lab sessions, as well as
> - Evaluation of translation systems as the group presentation project.

Practical hands-on experience

Hands-on experience of the machine translation software is also one of the core aspects of the course, which can help to deepen relevant knowledge acquired by the students. We provide lab sessions in class for students to work on some of the state-of-the-art machine translation systems. Coincidently, Somers also stated that hands-on experience is essential in teaching machine translation to trainee translators. Although not all of the students in our course are trainee translators, it is still of value for some of the students to try the software and have real experience in using it. They may as a result locate a suitable tool for their own use and buy a copy of it. This involves the purchase and selection of suitable licenses of translation systems. In addition, there are a lot of online versions of various machine translation systems for evaluation use. However, the quality generated varies. Although it is costly to keep updating the licenses of the translation systems, it is worth doing so in order to equip students with the necessary techniques in using the machine translation systems and to meet practical needs in real life. It also encourages respect for property rights and innovations, which can also boost the exchange of information. Besides, our department also provides resources for remote accessing of some of the translation systems, so that students can conveniently make use of the software to do testing and prepare their presentation at a remote location. Students can then benefit by having more hands-on experience with the systems.

Class interactive participation

Class exercise and discussion is another strategy used to encourage active learning through exchanges among students. By the end of the course, students have to evaluate different software systems and understand their weaknesses. (This is also one of Somers' suggestions on teaching trainee translators about machine translation.) The students are encouraged to think

of ways to improve the performance of the systems so that creativity cultivation and problem-solving skills can be fostered. Furthermore, it is hoped that interest in further research in the area can be sown.

Means of learning activities and assessments

A total of two and a quarter hours of combined lectures and tutorials is provided weekly. Students are given small-scale class exercises for group discussion at the end of each lesson.

Assessments are based on students' performance in class exercises, two written assignments, which ask the students to show their understanding of how a computer translates and to compare the different approaches employed in machine translation systems, as well as, through group presentation, evaluating the performance of different machine translation systems.

Research by Bisun D. and Huy P. P. (2006) found that South Pacific tertiary students have two main orientations when approaching study. They are Meaning and Reproducing, as Richardson stated in 1994.[1] '[S]tudents are directing their effort to understanding the materials studied, and on the other hand it is about reproducing materials for academic assessment purposes' (p.16). The result also correlated to data found in Hong Kong.[2,3] Since self-motivated further study is one of the goals in tertiary education, memorization should not be advocated. Assessment by examination is avoided in both courses in order to promote active learning instead of passive memorizing.

As a whole, the course provides students with adequate knowledge in machine translation through understanding it, experiencing it and aiming at improving it in further study.

Editing Skills for Computer Translation – a required course

Editing Skills for Computer Translation, on the other hand, encourages practical implementation of editing skills on the computer translation outputs. Compared with computer translation, it is set with different content focus and a different nature. *Computer Translation* equips the students with the techniques needed for using different computer translation systems and understanding the rationale behind them. They are able to have reasonable expectation of the pattern that particular systems generate in their translations and how the systems perform. They may even be able to anticipate certain errors in translations generated by specific computer translation systems. In *Editing Skills for Computer Translation,* students then learn the techniques involved in correcting those errors in an effective and efficient way, which helps them to make best use of the outputs generated by the computer translation systems.

Curriculum design

For *Editing Skills for Computer Translation*, the acquisition of more practical skills is emphasized. Purposes and strategies in editing the translations generated by computer translation systems are discussed. As there is still no fully automatic high-quality computer translation output, editing is still inevitable when using computer translation applications. The course introduces the concepts and skills essential to the editing of the source and target texts before, during and after computer translation, so as to optimize efficiency and translation quality. The three main types of editing processes: pre-editing, interactive editing and post-editing are introduced. Editing skills on different aspects including various linguistic levels are described in the course. Real examples are also employed for illustration. By the end of the course, students are required to try formulating some practical editing guidelines on a specific type of text generated by any

one specific computer translation system. Students will have to be able to practically apply the skills of editing the translation output generated by the computer translation software.

The topics covered in *Editing Skills for Computer Translation* include the following:

- Computer translation editing: purposes and strategies
- Editing skills: methods of translation
- Pre-editing: methods
- Pre-editing: data customization
- Interactive editing
- Post-editing: lexical aspects
- Post-editing: grammatical aspects
- Post-editing: semantic aspects
- Post-editing: pragmatic aspects
- Post-editing: cultural aspects
- Computer Translation Editing and Computer-Aided Translation: an integrated system

Practical hands-on experience

Hands-on experience of the machine translation systems is also encouraged in this course. Through remote accessing of the machine translation systems, students may use any of the systems provided by the department for preparation of their assignments and presentations. They may select any specific software for their translation works.

Class interactive participation

Group presentation is a means of encouraging practical application of editing skills on output generated by different software. At the same time, peer discussion is encouraged. Through the preparation of the presentation, students can familiarize themselves with at least one of the translation software systems among those provided for their use and practically try to implement different editing skills to the output generated by the software on different genre types of text. Each group of students is responsible for a different type of text to enable comparison of the performance output of the systems on different types of texts. They can share their implementation results with the other classmates, so that the learning process of every student can be enriched by the various ways of applying the skills and the specific methods used in handling a particular type of text. Experience in verifying editing guidelines is useful and impressive to them. Specifically in this course, doing post-editing for the first time can be a tedious task. It can sometimes make editors feel frustrated. However, after the students have tried it, they can, on the one hand, gain experience in it and use it as a stepping stone to getting the ball rolling. On the other hand, through peer support, they can overcome obstacles through discussion and become more interested in it. To give an experience from my class as an example: a group of students showed their experience in post-editing the texts in the presentation. They were showing signs of desperation at the very beginning of the preparation process. They were, however, enthusiastic by the end of it. They even quoted Dale Carnegie, when sharing their feelings with their classmates: 'If you act enthusiastic, you'll be enthusiastic'.

With the *Computer Translation* setting taken in the first term and *Editing Skills for Computer Translation* in the second, the former prepares students for the use of different software and familiarization with their operations and weaknesses, while the latter helps them to overcome the weaknesses through editing skills. They are therefore supplementary to each other.

Means of learning activities and assessments

As with *Computer Translation*, two and a quarter hours of combined lectures and tutorials are conducted each week.

Assessments are based on a written essay, class participation and group presentations showing how they apply the editing skills to the computer-generated translations of different genres.

Resources and technical support for the courses

The MACAT Programme provides a variety of electronic resources to its students in order to facilitate their learning of the courses offered by the Programme. For the computer translation related course, the provision of computer translation systems access through both in-class lab sessions and remote accessing are important in helping students familiarize with the use of the software. The Computer Terminal Room of the department also serves as a venue where students can have hands-on experience of different computer translation systems, installed on the computer platform of the machines in the laboratory. The Translation Software Library provides the user manuals and documentations of the computer translation systems for students' reference. Up-to-date licensing of the computer translation systems is also one of the essential components in facilitating effective teaching and learning of the courses.

In deciding which software is suitable for use in the course, we have considered different factors as follows:

1 The popularity of the software used in corporations and organizations can be one of the factors affecting the choice,so as to help the students meet their occupational needs.
2 Translation quality can be another consideration. It is difficult to estimate, however, as the accuracy of translation results varies from different genres. Judgement on the quality of translation is highly dependent on the needs of the users.
3 Language pairs supported by the systems and the functionality of the systems are among the concerns when selecting the software for rule-based systems.
4 Cost and maintenance of the machine translation systems including administrative costs are concerns in the running of the courses too.

Any special offer for educational and/or remote access licenses can be another factor affecting the decision.

In our courses, we have employed software from different vendors in order to provide more variety of choice and a wider picture of the field. As a result, the following software systems are employed in our teaching,

1 Systran, which provides a wide range of language pairs for translation, including most European and Asian languages, and is the one with the longest developmental history in the field. It also provides an online version of the tool at http://www.systransoft.com.
2 Dr. Eye, originating in Taiwan. This is a comprehensive language learning software with translation capability, which is particularly good with literary texts.
3 Transwhiz, originating as an English/Chinese translation tool. A special feature on parsing the tree structure of the sentences can be shown in the process of translation. Its online version is provided at http://www.mytrans.com.tw/tchmytrans/Default.aspx.
4 Kingsoft Translation Express, a dictionary-based software for English/Chinese translation, reasonably priced and with some free download versions, such as, http://ky.iciba.com.

5 LogoMedia Translate, specifically designed for European languages. Promt is one of the software in their family, which is mainly designed for rapid translation for idea gisting.
6 Yaxin, a computer-aided translation system, however, is equipped with a comprehensive list of dictionaries for dictionary lookup and string matching even without a translation memory imported. It can therefore also be classified as a dictionary-based machine translation tool.

There are abundant resources providing online translations. However, some are backboned by several identical software engines. In our courses, we include systems like, Google translate: http://translate.google.com/, SDL powered FreeTranslation.com: www.freetranslation.com/, Babelfish: http://www.babelfish.com/, WordLingo: http://www.worldlingo.com/en/products/text_translator.html and Microsoft Translator: translation2.paralink.com/ OR free-translator.imtranslator.net/ as some of the testing software for the students to choose from. We provide flexibility to the students to try any online translation software as testing tools. Online systems have the advantage of having updated information included in the translation outputs. For example, terms newly created or arisen like 'blue tooth', 'unfriend' and 'sudoku' can be correctly rendered by online translation tools. Commercial product versions of the machine translation system sometimes may not be able to update their word list or glossary within the system at the same pace. On the other hand, online systems, for the same reason, may fail to translate accurately because of too frequent updates of information from the World Wide Web fed into the systems, particularly in the case of systems supported by search engines. One example is the translation of proper names that may be changing all the time depending on the frequency of the appearance of the names found in any format of news feed. Students can benefit from assessing the performance of such different types of machine translation systems in various aspects.

Different teaching strategies according to different learning processes

A common difficulty when setting the 'target' for the master degree programme of Computer-Aided Translation in the Chinese University of Hong Kong is the diverse background of the students. This might be applicable to most master degree programmes of any international institutes. The diversity is not only with regard to intellectual disciplines, but also regional cultures. This section discusses how we try to make use of various teaching strategies aiming at striking the balance among the interests of different types of students, and how we optimize the learning processes of the students.

Bloom's Taxonomy of Learning Domains (Clark 1999)

According to Bloom's Taxonomy of Learning Domains, there are three types of learning: (1) cognitive: mental skills (knowledge); (2) affective: growth in feelings or emotional areas (attitude); and (3) psychomotor: manual or physical skills (skills).

Both *Computer Translation* and *Editing Skills for Computer Translation* involve the learning processes in cognitive and psychomotor domains. As a postgraduate course, it inevitably involves a development of intellectuality. Knowledge transmission is essential in both courses; therefore, cognitive learning is evoked. For the psychomotor counterpart, the skills applied in using the computer systems and editing the results are a form of skills transmission.

Cognitive domain

With regard to cognitive learning, there are six levels: from concrete to abstract, from basic to advanced. They are (i) knowledge, (ii) comprehension, (iii) application, (iv) analysis, (v) synthesis, and (vi) evaluation. In the course, *Computer Translation,* concepts and theories in computer translation are introduced whereas in *Editing Skills for Computer Translation,* previous research on how editing is done on computer translation is introduced. Such knowledge transition belongs to the first level, ((i) knowledge: recall data). In the written assignment, students have to show their understanding of the theories in *Computer Translation*. Examples of editing guidelines are discussed in *Editing Skills for Computer Translation*. Level (ii) comprehension involves understanding and interpretation of theories. Students have to be able to evaluate the performance of the computer translation systems in their group work so that they can show their application and analyzing technique in the *Computer Translation* course. Students in the *Editing Skills for Computer Translation* course have to apply the relevant editing guidelines to certain texts taken from computer translation in their group works. They show their application learning process through applying what they have learned in a real situation. During the group project, they also have to analyze and relate what relevant guidelines are to be applied so that they can perform the analysis stage of the cognitive learning process. The individual written assignment requires students to set up rules for the computer translation systems so as to generate better translation results, to analyze an edited text, and to create a set of editing guidelines based on the raw output of computer translation and the post-edited version of the text respectively in both *Computer Translation* course and *Editing Skills for Computer Translation* course. The synthesis learning process takes place. In both courses, students have to evaluate computer translation systems and the editing guidelines and skills to computer translation. The evaluation learning process completes the cognitive learning phenomena of the two courses. In general, the courses cover every learning behaviour in the cognitive domain.

Psychomotor domain

The *Computer Translation* course provides opportunities for students to have extensive hands-on experience with computer translation Systems in a classroom setting which trains them with the specific skills for operating the systems. With regard to skills development or training in *Editing Skills for Computer Translation*, practical implementation is the most crucial element. Therefore, real application of editing skills on texts done through group presentation also provides students with essential experience of learning in a psychomotor process. The experience can on the one hand deepen the knowledge they have acquired, and on the other hand, enhance their enjoyment of editing. They can also show their team spirit and share their results with one another.

Outcome-based teaching and learning

What is outcome-based education? Outcome-based teaching and learning has been widely adopted in various countries such as Singapore, the United Kingdom, the United States and Australia. According to Spady (1994: 12), outcome-based education is 'clearly focusing and organizing everything in an educational system around what is essential for the students to be able to do successfully at the end of their learning experiences'.

We have to consider what abilities are important for students to have and to organize the curriculum, instruction and assessment in order to make sure the learning ultimately happens.

As shown in the last section, consideration of the organization of the courses and assessment is based on Bloom's taxonomy and so covers every learning behaviour in the cognitive domain and the psychomotor domain. In outcome-based teaching and learning, 'the outcomes are actions and performances that embody and reflect learner competence in using content, information, ideas, and tools successfully' (Spady 1994: 13).

Learning outcomes of computer translation

In the *Computer Translation* course, the learning outcomes are as follows:

1 Students can understand basic concepts and reasons for computer translation.
2 Students have explored different grammar frameworks employed in computer translation.
3 Students can learn what the basic text processing steps in computer translation are, specifically, sentence identification, word segmentation, POS tagging and parsing.
4 Students can understand different approaches employed in translation systems, including rule-based, knowledge-based, example-based, memory-based, pattern-based and statistical approaches.
5 Students can understand the typical ambiguities generated in computer translation.
6 Students can learn how to evaluate and analyze the approaches applied in different computer translation software.
7 Students have hands-on experience in using different computer translation tools and become familiar with them.
8 Students are able to present on the analysis of the performance of and the approach applied in different computer translation systems.

After taking the course, they should be able to identify the weaknesses and strengths of different computer translation approaches that are applied in translation systems and evaluating as well as analyzing them.

Learning outcomes of editing skills for computer translation

In general, students on the course can learn how to make the best use of the output generated by computer translation software through editing. They are required to practice the editing skills on different genres translated by available translation software and present in groups. At the end of the course, they have to develop some editing guidelines on a specific text translated by a computer translation system. The learning outcomes are as follows.

1 Students can understand the basic concepts in computer translation editing.
2 Students can understand the purposes and basic classifications of computer translation editing.
3 Students can understand different editing strategies for different computer translation approaches.
4 Students can use the general rules for pre-editing on examples given to them as reference for practical uses in editing.
5 Students can understand the concepts of pre-editing, and practical ways in doing data customization in pre-editing.

6 Students can learn how to do interactive editing.
7 Students can understand how to do post-editing with the focus on various aspects, including lexical, grammatical, semantic, pragmatic and cultural aspects.
8 Students can learn the differences between computer translation editing and computer-aided translation. They can explore the editing capabilities in different translation systems.

In the *Editing Skills for Computer Translation* course, students have to apply the relevant editing guidelines that they have learned in working on their group project. They are expected to be able to practically implement the skills they have learned.

In order to obtain the optimal goal of generating a learning environment, setting appropriate learning outcomes can help to increase students' learning and ultimate performance abilities. One of the main purposes of outcome-based education, as stated, is '[e]nsuring that all students are equipped with the knowledge, competence, and qualities needed to be successful after they exit the educational system' (Spady 1994: 20). In the courses, although knowledge is delivered through traditional lecturing, assessment and projects are ways of ensuring that the students are able to accomplish the learning outcomes, such as applying what they learn in real situations of computer translation systems usage, and of editing computer translated texts.

As outcome-based teaching and learning is the trend in tertiary education in Hong Kong, various universities are also beginning to adopt such a framework in their courses. Improvement in organization of the courses, curriculum definition, instruction and assessment, based on outcome-based teaching, are foreseeable.

Overcoming difficulties generated by student diversity

Accommodation of students' characteristics is one of the key components in the teaching–learning process. Students in the courses are from diverse backgrounds with various disciplinary differences, including different subjects in arts, sciences and engineering, such as computer study, translation, communication, chemistry, language, and education, etc. Disciplinary differences mean students have different expectations of the course and have a different pace in study. Therefore, it is inevitable that a middle line be taken in setting the coverage and comprehensiveness of the course content, in order to facilitate different students' interests. In addition, relatively more optional reading is provided as references if the students want to pursue more reading. Extra research case studies are also suggested for students who have an interest in pursuing research studies.

On the other hand, different types of targeted careers (e.g. executives, teachers, officials, students, engineers, translators, speech theraptists, news reporters, etc.) also influence the students' interest in different topics. However, working hands-on with the software is one of the few common interests, as students can easily note the positive effect of gaining experience in using the tools. Even students from an Arts-subject background, usually reluctant to get involved in technical matters, are very pleased to learn how to use new software and to have more hands-on work.

Regionally the students are mainly coming from Hong Kong, Taiwan, Macao, Singapore and mainland China. In recent years, some have come from European countries. Though most of these regions use the same language, Chinese, both writing systems and speaking systems can be different. In terms of language medium, we follow the university's regulation of using English (an international language) for ease of communication. As our subject involves translation, the regional differences can sometimes be obstacles. In contrast, we can make beneficial use of the differences among the students. For example, we can benefit from the

regional variations among the students to enrich our discussion in the classroom through encouraging them to suggest regional different translations to cultural specific terminologies.

Reminders on the resources provision issue

It is worth noting that updating adequate licenses of the computer translation systems is important in guaranteeing the carrying out of hands-on experience by the students. Choosing software is among the issues to be reviewed and reconsidered periodically so as to keep pace with the rapid development of translation technology. Adequate technical support for both teachers and students is obligatory so as to facilitate reliable provision of the technological resources. In our case, a designated technician has responsibility for this.

Future trend of computer translation and its teaching

With decades of development on computer translation, it has now reached a bottleneck in further improving the accuracy and quality in the translation outputs. Recent developments tend to combine the advantages of both computer translation and computer-aided translation. One idea is taking advantage of computer translation's automation through providing computer translation results from public vendors. This tailor-made, high-quality data can suggest more choices of translation to those doing Computer-Aided Translation. Another option is to use the translation memory in computer-aided translation products for reviewing results generated by computer translation systems. It is hoped that in the coming era, almost fully automatic high-quality translation can be achieved by combining both computer translation and computer-aided translation technologies. Alternatively, in my opinion, by combining the strengths of different translation approaches incorporated in computer translation systems, specific newly developed hybrid approach systems, particularly designed for handling a particular genre, can improve practically the effectiveness and efficiency in the future performance of computer translation systems development. However, specific types of the systems may be restricted to specific usages.

The unwelcome old idea of using feedback comments to improve the design of computer translation systems is nowadays employed extensively in different online systems, such as Google Translate. The change of attitude is driven by the improved performance of the computer translation results. The extensive use of mobile devices also serves to make more convenient the updating and uploading of feedback for the systems which gives users a greater incentive to make a response to the system. A successful example is Google which provides a rating option and 'revise translation/suggest translation' options making it possible for users to contribute to the improvement of the translation generated by the system. However, the control of quality remains an unsolved problem. Whether the reviewer is qualified or authorized to make the judgements may affect the quality of translation outputs.

The rapid spread of the use of convenient mobile devices brings some controversy in web-based, cloud-based and even open source translation technology too. Security – or we should more precisely call it intellectual property rights – still remains one of the unsettled concerns. Should something uploaded, whether it be web-based, cloud-based or even open source, be adhered to, or protected by, the law regulating intellectual property rights? Does a translation carry property rights? 'A good translator is a lazy translator', as we have to maintain consistency in the use of language in our translation, specifically in professional translation. How and where should we locate the entity of translation copyright? Should we share without any requirement or regulation? Or how can we protect our own rights? How to make it so that everything can

be governed regularly? It still needs much discussion. If a translation is treated as an 'art' or 'cultural product' like the creation of works of art, such as poem creation, we should indeed show great respect towards property rights.

Conclusion

One of the main goals of tertiary education is to stimulate students' interest in self-learning or further education. Research shows that different attitudes can engender different effects/results in the outcome. Maintaining good relationships with students can help students to have a better attitude towards their learning. They are more willing to ask questions when they have this. A majority of experienced teachers would agree that sometimes it can be difficult to predict what topics are of interest to all the students and what areas/aspects would be 'too' difficult for the students. Although evaluation questionnaires can help to allay some of the doubts, it would be more effective if the students were to voice out immediately any difficulties they may have in understanding and/or to ask any questions. Communication is crucial in providing a good environment for the teaching and learning process.

The various teaching approaches employed in the courses described in this chapter can generate different effects in different ways. Through lecturing, knowledge is introduced, students' interest is stimulated and attention to/concern with specific aspects is aroused. Class exercise and discussion can help students to brainstorm ideas and exchange information. Group presentation can address real practice and application of the knowledge acquired. A question-and-answer session follows each presentation, serving as a platform for students to pose any questions to the presenters. It can help students to think about any aspects they may have neglected. 'When there is competition, there is improvement.' It also acts as a reminder to students of those areas to which they have not paid attention but to which others have. At the same time, it can be a good opportunity for teachers to remind them of those areas that they might have overlooked. They can also learn to rethink in detail and in depth, in order to clarify their ideas. It would be an opportunity for them to learn about getting to the point and putting correct emphasis on the respective area next time. An assignment can be an assessment at the same time, a real attempt to compare and evaluate the state-of-the-art translation software, and also to develop guidelines for practical editing (even for future reference). In general, different teaching strategies can be employed for different courses with different aims and intended learning outcomes. Communication, as a result, is the key for setting appropriate intended learning outcomes for our students.

Notes

1 As quoted in Bisun Deo and Huy P. Phan (2006: 16–17), Richardson, J. T. E. (1994), 'Cultural specificity of approaches to studying in higher education', *Higher Education*, 27, 449–468.

2 As quoted in Bisun Deo and Huy P. Phan (2006: 16–17), D. Kember and D. Y. P. Leung (1998) 'The Dimensionality of Approaches to Learning: An Investigation with Confirmatory Factor Analysis on the Structure of the SPQ and LPQ', *British Journal of Educational Psychology* 68: 395–407.

3 As quoted in Bisun Deo and Huy P. Phan (2006: 16–17), N. Y. Wong, W.Y. Lin, and D. Watkins (1996) 'Cross-cultural Validation of Models of Approaches to Learning: An Application of Confirmatory Factor Analysis', *Educational Psychology* 16: 317–327.

References

Arnold, Doug J., Lorna Balkan, Siety Meijer, R. Lee Humphreys, and Louisa Sadler (1994) *Machine Translation: An Introductory Guide*, London: Blackwells-NCC.

Babelfish (28 March 2014). Babelfish. Available at: http://www.babelfish.com.
Deo, Bisun and Huy P. Phan (2006) 'Approaches to Learning in the South Pacific Region: A Confirmatory Factor Analysis Study', in *Proceedings of the AARE 2006 International Education Research Conference*: *Adelaide Papers Collection*, 26–30 November 2006, Adelaide, Australia.
Chan Sin-wai (ed.) (2010) Journal of Translation Studies: Special Issue on the Teaching of Computer-aided Translation 13(1–2): 275–282.
Clark, Donald R. (1999) 'Bloom's Taxonomy of Learning Domains: The Three Types of Learning'. Available at: http://www.nwlink.com/~donclark/hrd/bloom.html.
Google (28 March 2014) Google Translate. Available at: http://translate.google.com.
Hutchins, W. John (1986) *Machine Translation: Past, Present, Future*, Chichester: E. Horwood and New York: Halsted Press.
Ignacio, Garcia (2010) 'Translation Training 2010: Forward Thinking, Work Ready', in Chan Sin-wai (ed.) *Journal of Translation Studies: Special Issue on the Teaching of Computer-aided Translation* 13(1–2): 275–282.
Kember, D. and D. Y. P. Leung (1998) 'The Dimensionality of Approaches to Learning: An Investigation with Confirmatory Factor Analysis on the Structure of the SPQ and LPQ', *British Journal of Educational Psychology* 68: 395–407.
Kingsoft (28 March 2014) Translation Express. Available at http://ky.iciba.com.
Microsoft (28 March 2014) Microsoft Translator. Available at: http://translation2.paralink.com/ OR free-translator.imtranslator.net.
Richardson, J. T. E. (1994), 'Cultural specificity of approaches to studying in higher education', *Higher Education*, 27, 449–468.
SDL (28 March 2014) FreeTranslation.com. Available at: http://www.freetranslation.com.
Somers, Harold L. (2003) 'Machine Translation in the Classroom', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 319–340.
Spady, William G. (1994) *Outcome-based Education: Critical Issues and Answers*, American Association of School Administrators.
Systran (28 March 2014). Available at: http://www.systransoft.com.
Transwhiz (28 March 2014). Available at: http://www.mytrans.com.tw/tchmytrans/Default.aspx.
Wong, N. Y., W.Y. Lin, and D. Watkins (1996) 'Cross-cultural Validation of Models of Approaches to Learning: An Application of Confirmatory Factor Analysis', *Educational Psychology* 16: 317–327.
WorldLingo (28 March 2014). Available at: http://www.worldlingo.com/en/products/text_translator.html.

This page intentionally left blank

PART II

National/regional developments of translation technology

This page intentionally left blank

14
TRANSLATION TECHNOLOGY IN CHINA

Qian Duoxiu

BEIHANG UNIVERSITY, CHINA

A historical sketch

In 1946, the world's first practical fully electronic computer ENIAC machine came into use in the University of Pennsylvania (Hutchins 1986: 23–24). China was then on the brink of a full-scale civil war after its anti-Japanese war ended in the previous year. In 1947, when Warren Weaver (1894–1978) came up with the idea of using computers in translating, China was divided by a civil war between forces led by the Kuomintang and forces led by the Communist Party. In 1949, the war was won by the latter and the People's Republic of China was founded. Then efforts began to be made in an organized way in almost everything.

Following the United States and the former Soviet Union, scholars in China started research in machine translation in 1956. China demonstrated its achievements in machine translation for the first time in 1959 and thus joined the exclusive club in this field (Dong 1995: 85–91; Fu 1999). Even though Chinese characters output device was not available then, this first system was for the automatic translation of twenty sentences of different syntactic structures between Russian and Chinese, with the output of Chinese characters in coded form (Chan 2004: 295).

Later, on the basis of enlarged corpora and store of structures, research for English–Chinese machine translation started. Forerunners include the Institute of Scientific and Technical Information of China (ISTIC), Harbin Institute of Technology (HIT), Beijing Foreign Studies University (BFSU), South Chinese University of Technology, and other institutions. Due to the devastating consequences of the Great Cultural Revolution (1966–1976), research and development in this field entered into a 10-year-long stagnation. It was resumed in 1978 when the opening-up and reform policy was implemented.

Communication in this field began to increase. In 1980, the First Seminar on Machine Translation was held in Beijing. In 1982, the Second Seminar was held in Shanghai. There were only a few dozen participants at that time. However, since 2000, there have been more than 100 people for each China workshop on Machine Translation (CWMT). In 2011, the 7th Workshop was held in Xiamen, Fujian province. The theme was to evaluate different systems dealing with different language pairs in different fields. Multilingualism and wide coverage of domains are now the trend (http://www.cas.cn/xw/yxdt/201110/t20111008_3359421.shtml). The 8th CWMT was held in Xi'an, Shaanxi province in September 2012. Its themes included MT models, techniques and systems, multi-lingual MT system evaluation, and other topics. The 9th CWMT was held in Kunming, Yunnan province in October 2013. Its theme was to test MT systems developed for different domains based on a unified standard.

Communication with the outside world is also flourishing. Before its dissolution in 2011, LISA (Localization Industry Standards Association) had held its annual China Focus Forum several times. Now GALA (http://www.gala-global.org) is involved in various activities related to localization and globalization in the Chinese context.

Besides this, academia, industry, and professional organizations have all come to realize the importance of translation technology in the process of globalization and localization. In 2009, the Localization Service Committee of the Translators Association of China (TAC) was set up. Its members are all leading translation technology and language/localization service providers (LSP, http://www.taclsc.org/index.asp).

At present, with cloud computing technology, many more parties are involved and small-scale conferences are numerous. In the meantime, more languages, more domains, and more approaches related to machine translation have been under research.

Though approaches in the development of translation technology in China have not been very different from those adopted in other countries, there are some distinctive features.

First, research and development in China has been chiefly sponsored by the government since the very beginning. Early in 1956, 'Machine Translation/Mathematical Theories for Natural Language' was already an item listed in the Government's Guidelines for Scientific Development. Later it was among the major national scientific and technical projects, such as 'The Sixth Five-Year Plan', 'The Seventh Five-Year Plan' and '863 Plan'. Though there was a 10-year stagnation in translation research in China from 1966 to 1976, it was not because of the shortage of funding, but because of political and social upheavals during the Cultural Revolution.

Second, scholars from various fields and institutions have been involved in the development of translation technology in China ever since its start. The collaboration, which is common in this field, among people from computer sciences, mathematics, and linguistics has spurred on the development of translation technology greatly in China. It was also in this early period (1956–1976), not necessarily through the impact of the 1966 ALPAC Report, that people in China realized that machine-aided translation is more feasible, at least in the foreseeable future.

In the mid-1970s, translation technology research regained its original momentum and resumed its rapid growth on the basis of collective efforts of many ministries and institutions, with the Institute of Linguistics of the Chinese Academy of Social Sciences acting as the spearhead. A 5-year-long collaboration yielded some rudimentary systems and helped to train many researchers, who would continue their work in places all over China. In the meantime, researchers were sent abroad or recruited to do postgraduate study in this area. National conferences or seminars on translation technology were held regularly and related journals were published (e.g., *Journal of Chinese Information Processing*).

The 1980s and 1990s witnessed the second important phase in the development of translation technology in China. During this period, two milestone practical systems came into being. One is the KY-1 English–Chinese Machine Translation System developed by the Academy of Military Sciences in 1987, which won the second prize of the National Scientific and Technical Progress Award and was later further refined into TranStar, the first commercialized machine translation system in China. The other is the 'IMT/EC-863' English–Chinese Machine Translation System developed by the Institute of Computing Technology of Chinese Academy of Sciences. This system won the first prize of the National Scientific and Technical Progress Award in 1995 and has brought about tremendous profits. These two systems are the children of collaborative efforts of various institutions and people. Another system worth mentioning is the 'MT-IR-EC' developed by the Academy of Posts and Telecommunications, which is very practical in translating INSPEC (Information Service in Physics, Electro-Technology,

Computer and Control) titles from English into Chinese. Not mentioned here are many other efforts made in this period, including the joint programme between China and Japan, which introduced translation technology in the Chinese context to the outside world. This helped to train talents and promoted the transmission of technology and the accumulation of resources. Consequently, some Japanese–Chinese machine translation systems came into being, such as those developed by Tsinghua University, Nanking University, and the China University of Science and Technology. In the middle 1990s, for the first time the world over, a research group led by Yu (1993: 117–126) at the Institute of Computational Linguistics of Peking University, constructed a quite reliable evaluation system of translation technology.

From the 1990s onwards, translation technology in China has undergone a rapid growth. Many commercial systems are available on the market. All these systems share some common features. For example, most of them are equipped with very big multi-disciplinary and domain specific dictionaries, operational through the network and user-friendly. New technologies, such as human–machine interface, began to be developed. So in a sense, translation technology in China is not far behind in its PC (personal computer) products and network system development. The dominant technology strategy and guideline of translation technology in the Chinese context then were not very different from those adopted in other parts of the world. They are mainly transformation-based, rule-based and very practical (such as ECMT-78 developed by Liu Zhuo in 1978), many of which are still in use today (for more information, read Chan 2004: 66; Feng 2007; Fu 1999).

In recent years, substantial efforts have been made in developing translation technology in China. In 1999, Yaxin CAT 1.0 was publicized. It is China's first all-in-one computer-aided translation (CAT) system which combines translation memory, human–machine interaction and analysis. Now Yaxin CAT 4.0 is commercially available and has been very popular among Chinese CAT users.

There are several academic organizations active in translation technology on the Chinese Mainland. For example, the Chinese Information Processing Society (CIPSC) has been organizing several international and national conferences since it was founded in 1981 and is an active participant of international exchanges. However, much is still to be done because the exchanges are mainly done in Chinese, while little effort has been made to have the conversation conducted in English in order to be recognized by a larger audience beyond the Chinese context.

Up to now, more than 50 years has passed since its incidence and translation technology has witnessed great progress in the Chinese context (for information about Taiwan, Hong Kong, and Macau, read related sections).

Translation technology: principles, strategies, and methodology

Basic and dominating methods in machine translation research and application in China are, not surprisingly, consistent with the approaches adopted in other countries. For example, Chinese researchers tried the transfer approach, where conversion was through a transfer stage from abstract (i.e. disambiguated) representations of SL texts to equivalent TL representations. Three stages are involved: analysis, transfer, and generation (or synthesis), such as KY-1 system mentioned in the previous section (for more information, read Hutchins, http://www.nlp.org.cn/docs/20030724/resource/Machine%20Translation%20over%20fifty%20years.htm).

They also tried the rule-based approach, which was most obvious in the then dominant transfer systems. It was also at the basis of the various interlingua systems, both those which were essentially linguistics-oriented, and those which were knowledge-based, such as

HansVision developed by Beijing Creative Next Technology Ltd. (Chan 2004: 94) (for more information, read Hutchins, op. cit.).

Example-based and corpus-based approaches were later adopted as a more feasible way to tackle the problems encountered by previous approaches. It is now widely acknowledged and used as the best method in this field in China. For example, at Beijing Foreign Studies University (BFSU), a research project on the design and construction of a bilingual parallel corpus has been going on for several years and one of its goals is to shed some light on the bi-directional translation between Chinese and English (Wang 2004: 73–75). A lot more work has been done for this purpose and is briefly mentioned in the next section.

Major participants and achievements

There are many active participants in the research and development of MA and CAT. One leading organization is the Chinese Information Processing Society of China (CIPSC, http://www.cipsc.org.cn/index.php). It was established in June 1981, its mission being to develop methods for processing Chinese with the aid of computer technology, including automatic input, output, recognition, transfer, compression, storage, concordance, analysis, comprehension, and generation. This is to be done at different linguistic levels (character, lexical, phrasal, sentential, and textual). The field has developed into an interdisciplinary subject area in a very robust way with collaborative work by scholars from fields like philology, computer sciences, artificial intelligence, cognitive psychology, and mathematics. This organization has been in close contact with the outside world, playing a very active role in the world MT-Summits.

Chinese Linguistic Data Consortium (CLDC, http://www.chineseldc.org/cldcTest.html) is an organization affiliated to CIPSC. Its mission is to build up databases of Chinese at different linguistic levels. The databases can be used in the fundamental research and development in Chinese information processing. The Consortium has been supported financially by several national-level 863 research projects such as General Technical Research and Fundamental Databases for a Chinese Platform (2001AA11401), Evaluation Technology Research and Fundamental Databases for a Chinese Platform (2004AA114010), a national key 973 fundamental research project called Comprehension and Knowledge Data-mining of Image, Phonemes and Natural Language (G19980305), and many other similar endeavours.

Another purpose of this Consortium is to provide standards and regulations for Chinese language information processing to be used by different institutions both at home and abroad so that they can communicate with each other with the same criteria. For example, CLDC-2009-004 is a very large bilingual (English–Chinese) parallel corpus covering a variety of fields and text types. It contains 20,000,000 sentence pairs (http://www.chineseldc.org/index.html).

CLDC-LAC-2003-003 is an annotated and segmented (lexical level) Chinese corpus. There are 500 million Chinese characters in this balanced corpus and the data are all POS tagged and segmented with human validation.

CLDC-2010-006 is also known as CIPS-SIGHAN CLP 2010, which is a corpus for evaluating lexical segmentation of simplified Chinese. The emphasis of this evaluation system is to see how well algorithms can do segmentation of Chinese words and expressions across different fields and text types. There are four sub-corpora, namely, literature, computer sciences, medicine, and business, each with 50,000 Chinese characters. This database also contains reference corpus, untagged training corpora (literature and computer sciences, each with 100,000 Chinese characters), evaluation guidelines, and an overall evaluation report.

CLDC-2010-005 is a bilingual Chinese–Mongolian parallel corpus of 60,000 sentences pairs. The texts belong to several types, including political, legal, daily usage, literature, and other types.

CLDC-2010-001 is an ICT web-based Chinese−English parallel corpus. The data was collected in 2009 from the web. It uses XML language to mark up the information and the encoding is UTF-8. There are 1,075,162 sentence pairs in this corpus.

Not many journals related to translation technology in China are available. *Journal of Chinese Information Processing* (CN11-2325, ISSN1003-0077) was created in 1986. It is the official publication of CIPSC, co-sponsored by China Association for Science and Technology and Institute of Software, Chinese Academy of Sciences. Papers published in this journal are all related to Chinese information processing and machine translation between Chinese and other languages. Other journals in the broad field of translation and linguistics may accept papers on this topic occasionally.

It is also worthy of note that, in this region, much attention has been paid to the teaching and study of translation technology over the past years. The world's first Master of Arts programme in Computer-aided Translation was offered by the Department of Translation, The Chinese University of Hong Kong in 2002. The enthusiasm demonstrated by students admitted into this programme in the past ten years is symptomatic of the growing demand of the society at large. Later, more and more universities on the Chinese Mainland began to offer programmes related to this. In 2006, the Ministry of Education decided that translation and interpretation should be set up as an independent degree programme. In the curriculums for both undergraduates and postgraduates, it is stipulated that computer-aided translation should be either required or elective. It is believed that such programmes will attract more and more talents to join their hands in this field.

Application and mainstream tools

Early attempts to develop practical tools were many, sponsored by either the Government or private organizations. In Table 14.1 is an incomplete list of systems and their developers.

Some of these tools have been further developed and been successful commercially into the present day. Huajian Group (http://www.hjtek.com/en/index.html) is an example here. This high-tech enterprise is affiliated to the Chinese Academy of Sciences and is mainly engaged in technological research, product development, application integration, and technical services in the field of computer and language information processing. It has provided the government, businesses, and individuals with solutions to computer information system applications such as computer information processing, systems integration, and information services. It has developed around 60 translation tools. Solutions like a multi-lingual application service have been adopted by many domestic organizations.

The core technologies of the Huajian series include a solution to the problem of translation quality and knowledge acquisition. They combine the advantages of rule-based and corpus-based methods; multi-level attributive character system and integrated SC (semantic and case) syntax system; pre-analysis and feedback in rule-based contextual testing; integrated analysis of grammar, semantics and general knowledge in multi-route dynamic selection; a solution to polysemy using special rules; real mode expression based on multi-level abstract characteristics; semantic similarity calculation based on compatibility of multi-level characteristics, and intelligent machine translation technology (http://www.hjtek.com/en/index.html).

Table 14.1 Early attempts at CAT research and development

System	*Developer (surname, given name)*
Transtar English-Chinese MT system	Dong Zhendong
JFY (Gaoli) English-Chinese MT system	Liu Zhuo
IMT/EC English-Chinese MT system	Chen Zhaoxiong
TYECT English-Chinese MT system	Wang Guangyi
TECM English-Chinese MT system	Liu Xiaoshu
TH(Tsinghua) English-Chinese MT system	Chen Shengxin
NetTrans English-Chinese MT system	Wang Huilin
SuperTran English-Chinese MT system	Shi Xiaodong
HansBridge English-Chinese MT system	Creative Next Technology Ltd.
Ji Shi Tong English-Chinese MT system	Moon Computer Company
TongYi English-Chinese MT system	Tongyi Institute of MT Software
East Express English-Chinese MT systems	Shida-Mingtai Computer Ltd.
Yaxin English-Chinese MT system	Yaxincheng Computer Software Ltd.
FCAT system	(Feng, Zhiwei)
KEYI-1 English-Chinese system	Mars Institute
Kingsoft Powerword	Kingsoft
Oriental Express	SJTU Sunway Software Co Ltd.

The original interactive hybrid strategies machine translation method is adopted in most of its systems and system implementation algorithms are independent of specific language and open development platforms with multi-user consistency protection mechanisms. There are nine translation systems dealing with seven languages (Chinese, English, Japanese, Russian, German, French and Spanish). In this way, massive multilingual language information and corpus has been accumulated (http://www.hjtek.com/en/index.html).

Since the early 1990s, tools with different brand names have become commercially available. More recently, there are Youdao, Lingoes, Iciba (PowerWord), among others, which are leading online and offline tools for automatic translation based on corpus.

Computer-aided translation technology has been developed mainly by several leading companies. Two mainstream tools are introduced here.

One is Yaxin CAT series, developed by Beijing Orient Yaxin Software Technology Co., Ltd (http://www.yxcat.com/Html/index.asp). It has been regarded as one of the best professional translation tools produced by a domestic developer. Its products include Single-user Version (English−Chinese Version and Multilingual Version), office-aided Translation Teaching System (Multilingual Two-way), and Computer-aided Translation Teaching System (Multilingual Two-way). The products have the following major advantages:

- combination of machine translation with computer-aided translation;
- improving translation speed and quality by pre-translation, in-translation and post-translation processing;
- built-in term banks based on more than 80 domain specific dictionaries;
- working from bilingual corpora to multilingual translation suggestions, integration of embedded, external, and stand-alone translation systems, and operating from single-user and network-based processing to cloud service.

The other one is TRANSN (http://www.transn.com). Its cloud translation technology combines cloud computing with traditional translation technology and is an internationally advanced fourth generation language processing technology. The core technology includes the following components:

- fragmentation cloud translation technology which is based on cloud computing and enables large-scale high-speed parallel processing of translation tasks;
- workflow technology capable of infinitely flexible translation process;
- configuration and carrying out different forms of automatic translation workflow control;
- fuzzy TM engine technology to improve precision in fuzzy TM matching and raising efficiency more than threefold;
- synchronized translation technology;
- real-time synchronization technology which provides technical support for internet-based collaborative translation;
- corpus processing technology to help realize large-scale and low-cost corpus
- processing;
- data-mining technology based on dash board model and high-speed data engines to produce translation data reports that meet requirements at different levels;
- search engine technology that offers translators maximum support in obtaining authentic interpretations for difficult terms and expressions, and machine translation technology.

Discussions about translation technology

There are many discussions about translation technology in the Chinese context. With his numerous publications, Chan Sin-wai (2002, 2004, 2008, 2009) from the Chinese University of Hong Kong has been regarded as one of the doyens. The works authored or edited by him provide a panoramic picture, as well as helpful resources, for anyone interested in this topic. Publications by Feng (1999, 2004, 2007) and Qian (2005, 2008a, 2008b, 2009, 2011a, 2011b) can also be taken as useful references to the research, application and teaching in this field.

More specifically, according to Wen and Ren (2011: 58–62), there are altogether 126 articles on CAT collected by China National Knowledge Infrastructure (CNKI) from 1979 to 2010. They can be divided into four major categories—theory, teaching, technology and tool, and industry.

When theory is concerned, three aspects are the focus of attention, namely, explanation and differentiation of terms (Zhang 2003: 56–57), comparison of MT and CAT (Zhang and Yu 2002: 54–58), and attempts at using a multi-modal approach to CAT (Su and Ding 2009: 84–89).

Articles on CAT teaching are many, noteworthy among which are the pioneering ideas of setting up CAT courses at the tertiary level (Ke and Bao 2002: 61–68), pedagogical reflections on curriculum design of CAT as a course (Qian 2009: 49–53, 2010: 13–26) and a master's programme in engineering with CAT as its orientation (Yu and Wang 2010: 38–42).

Technology and tools are what CAT is both theoretically and practically about. People from both fields have contributed to this topic. MT, together with TM (translation memory), its history, development, application, limitations, and prospects are discussed in many papers (e.g., Qian 2005; Su 2007). The corpus-based approach is recognized as the most promising method in MT and CAT (Liu *et al.* 1997: 61–66; Liu 2006: 84–85). As for tools, there is an array of papers, mostly on a single tool or on CAT tools on general.

Industry has always played the key role in the research and development of MT and CAT. Though publications in this area are mainly about conferences and news reports, they provide important information on the latest activities in both research and development of translation technology.

Prospects

Rapid growth and remarkable achievements, however, don't mean that the technologies involved are quite mature. The history of MT research and development indicates that MT and CAT require the collective efforts of people from various fields. In the past, induction was done manually and was time-consuming and very costly. It is also problematic because consistency is very difficult to arrive at. Once some new rules are added to improve the translation of certain sentences, it would be very difficult to handle other sentences which didn't present any problems before the addition. New errors would appear when new formations are made, which has led to the growing complexity of the system and the growing difficulty in maintenance. This has been a universal bottleneck for MT system in the past several decades.

For Chinese, another problem is word segmentation, which is the first, yet a key step in Chinese information processing. So far, there has not been a perfect solution though many advances have been made (Sun 2001; Yu 2002). On the one hand, research conducted at Peking University demonstrates that there is no need for an absolute definition of word boundary for all segmenters, and that different results of segmentation shall be acceptable if they can help to reach a correct syntactic analysis in the end (Duan *et al.* 2003). On the other hand, the testable online tool it has developed cannot yet segment words with ambiguous meanings in most cases (http://www.icl.pku.edu.cn/icl%5Fres/segtag98).

Translation technology is a fast developing area. With mobility and innovation becoming the keywords of this era, it is no wonder that new tools emerge every now and then and activities are multi-faceted and based in different places. Take Dr Eye suite tools (http://www.dreye.com/en) as an example. It is originally from Taiwan, but now has its headquarters based in Shanghai. Like other mainstream tools of the world, it includes instant dictionary, translation engine, multi-lingual voices, multi-lingual dictionary provided by Oxford University Press, and many other user-friendly functions. Similar tools are many, such as Lingoes (http://www.lingoes.cn), Youdao (http://www.youdao.com), PowerWord (http://www.iciba.com), to name only a few.

With the rapid growth of Internet Technology, the future of MT and CAT research and development is quite promising and more advances are to be expected. But as was pointed out in the previous sections, the quality of MT translations has not been substantially improved. One thing that is clear is that MT is not only a problem of language processing, but also one of knowledge processing. Without the accumulation of knowledge and experience over the years, it is hardly possible to develop an MT system which is practical. The short cycle of development at present is the result of many years' hard work and the accessibility of shared resources.

Looking forward, it is apparent that there is still a long way to go before MT can truly meet the demands of the users. Generally speaking, things to be done for both MT and CAT research and development between Chinese and other languages should include the following:

1 Though the notion of a 'text' has been lost because the translation tools now available operate primarily at sentential level (Bowker 2002: 127), the analysis of the source language

(Chinese in most cases) should be done in the context beyond the present sentential level which is isolated and based on the comprehension of the original. Future analysis should take the sentence cluster or even the entire text into consideration. While analysis today seeks to find out the syntactic relationship tree, semantic relationship of the concepts involved at most, future analysis should be on the textual meaning instead. Once this is arrived at, meaning transfer could be done more accurately than what is done by the present systems (Dong 2000).

2 Basic research needs to be deepened and strengthened, especially the construction of common-sense databases. Scholars have even suggested that a knowledge dictionary should be built up to facilitate comprehension-based analysis, such as the one developed by Dong Zhendong, a leading Chinese scholar in MT, and his colleagues (Dong 1999; http://www.how-net.com), which has shed some light on the comprehension-based analysis and explorations of disambiguation.

3 The stress of research and development should be more and more on the parameterized model and a corpus-based, statistically oriented and knowledge-based linguistic approach. Accumulation of bilingual and multi-lingual language data/corpora will make it more feasible to develop more fully automated domain-specific machine translation systems. Efforts should be made to develop a method for semantic disambiguation and an objective evaluation of it. Automatic learning (acquisition, training) strategies of the computer and a bi-directional system design should be strengthened. A more user-friendly feedback control function should be developed so that the user can adjust the behavior of the system.

4 As is pointed out by Hutchins (1999) and applicable to MT and CAT in the Chinese context, translation software now available is still expensive. How to develop an efficient system that is of low cost, high reliability and required less work on constructing the translation memory (TM) for individual translators is another emerging problem. Besides, translation systems into minor languages and spoken language should also be further explored.

5 It is necessary for scholars in the Chinese context to learn from and exchange with others and to have closer contact with the industry. The collaboration, led by Yu Shiwen from the Institute of Computational Linguistics, Peking University, between Peking University and Fujitsu has been fruitful. They have managed, to a great extent, to produce a tagged corpus of 13,000,000 Chinese characters in order to find out some statistical rules and parameters for processing this language (http://www.icl.pku.edu.cn/WebData_http-dir-listable/ICLseminars/2002spring). Organizations in China have made efforts to have their voices heard by joining the international community. For example, EC Innovations (http://www.ecinnovations.com) is now a member of TAUS (Translation Automaton User Society, http://www.translationautomation.com), which held its 2012 Asia Translation Summit in Beijing.

6 Attention should be paid to 'spoken language translation', which still eludes us and could be a very ambitious project (Somers 2003: 7).

7 Attention should also be paid to network teamwork, from stand-alone systems, so that multiple users can share the same resources.

Translation technology is now the trend in every aspect of the industry. One manifestation of this is that training programmes of varying durations have been offered, while more universities on the Chinese Mainland are starting to have courses on translation technology. Topics for the programmes cover approaches to CAT and MT, localization, tools, translation project

management, and so on (Qian 2009). The total number of trainees enrolled is on the rise. There are strong reasons to believe that translation technology will have a promising future.

References

Bowker, Lynne (2002) *Computer-aided Translation Technology: A Practical Introduction*, Ottawa: University of Ottawa Press.

Chan, Sin-wai (ed.) (2002) *Translation and Information Technology*, Hong Kong: The Chinese University Press.

Chan, Sin-wai (2004) *A Dictionary of Translation Technology*, Hong Kong: The Chinese University Press.

Chan, Sin-wai (2008) *A Topical Bibliography of Computer(-aided) Translation*, Hong Kong: The Chinese University Press.

Chan, Sin-wai (2009) *A Chronology of Translation in China and the West*, Hong Kong: The Chinese University Press.

Dong, Zhendong (1995) 'MT Research in China', in Dan Maxwell, Klaus Schubert, and Toon Witkam (eds) *New Directions in Machine Translation*, Dordrecht-Holland: Foris Publications, 85–91.

Dong, Zhendong (1999) 'Review of MT in China in the 20th Century'. Available at: http://tech.sina.com.cn.

Dong, Zhendong 董振東 (2000)〈中國機器翻譯的世紀回顧〉(Review of Machine Translation in the 20th Century). 《中國計算機世界》(*China Computer World*), Issue 1.

Duan, Huiming, Bai Xiaojing, Chang Baobao, and Yu Shiwen (2003) 'Chinese Word Segmentation at Peking University'. Available at: http://acl.upenn.edu/w/w03/w03-1722.pdf.

Feng, Zhiwei 馮志偉 (1999) 〈中國的翻譯技術:過去，現在和將來〉(Translation Technology in China: Past, Present and Future), in Huang Changning 黃昌寧 and Dong Zhendong 董振東 (eds) 《計算機語言學文集》(*Essays on Computational Linguistics*), Beijing: Tsinghua University Press 清華大學出版社, 335-340.

Feng, Zhiwei (2004). *Studies on Machine Translation*. Beijing: China Translation and Publishing Corporation.

Feng, Zhiwei 馮志偉 (2007) 《機器翻譯今昔談》(*Machine Translation: Past and Present*), Beijing: Language and Culture Press 語文出版社.

Fu, Aiping (1999) 'The Research and Development of Machine Translation in China', in *MT Summit VII: MT in the Great Translation Era: Proceedings of the Machine Translation Summit VII*, 13–17 September 1999, Kent Ridge Digital Labs, Singapore, 86–91.

http://www.cas.cn/xw/yxdt/201110/t20111008_3359421.shtml.

http://www.chineseldc.org/cldcTest.html.

http://www.chineseldc.org/index.html.

http://www.cipsc.org.cn/index.php.

http://www.dreye.com/en.

http://www.ecinnovations.com.

http://www.gala-global.org.

http://www.hjtek.com/en/index.html.

http://www.how-net.com.

http://www.iciba.com.

http://www.icl.pku.edu.cn/icl%5Fres/segtag98.

http://www.icl.pku.edu.cn/WebData_http-dir-listable/ICLseminars/2002spring.

http://www.lingoes.cn.

http://www.nlp.org.cn/docs/20030724/resource/Machine%20Translation%20over%20fifty%20years.htm.

http://www.taclsc.org/index.asp.

http://www.translationautomation.com.

http://www.transn.com.

http://www.youdao.com.

http://www.yxcat.com/Html/index.asp.

Hutchins, W. John (1986) *Machine Translation: The Past, Present and Future*, West Sussex, England: Ellis Horwood Limited.

Hutchins, W. John (1999) 'The Development and Use of Machine Translation Systems and Computer-aided Translation Tools', in *Proceedings of the International Symposium on Machine Translation and Computer Language Information Processing*, 26–28 June 1999, Beijing, China.

Ke, Ping 柯平 and Bao Chuanyun 鮑川運 (2002) 〈世界各地高校的口筆譯專業與翻譯研究機構〉(Interpreting and Translation Programmes of Tertiary Institutions and Translation Research in the World), 《中國翻譯》*Chinese Translators Journal* 4: 61–68.

Liu, R. (2006) 'Establishing Large-scale Chinese and English Bilingual Parallel Corpus', 《太原理工大學學報》(社會科學版) (*Taiyuan Science and Technology*) (Social Sciences Edition) 10: 84–85.

Liu, Xiaohu 劉小虎, Li Sheng 李生, and Wu Wei 吳葳 (1997) 〈機器輔助翻譯中模糊查詞典和快速錄入單詞〉(Fuzzy Dictionary Look-up and Fast Input Word in Machine Aided Translation), 《中文信息學報》(*Journal of Chinese Information Processing*) 4: 61–66.

Qian, Duoxiu (2005) 'Prospects of Machine Translation in the Chinese Context', *Meta* 50(4).

Qian, Duoxiu and Teng Xiong (2008a) 'Localization and Translation Technology in the Chinese Context', in *Proceedings of the 18th World Congress of the International Federation of Translators: Translation and Cultural Diversity,* 4–7 August 2008, Shanghai, China.

Qian, Duoxiu (2008b) 《科技翻譯質量評估—計算機輔助的〈中華人民共和國藥典〉英譯個案研究》(Computer-aided Quality Assessment in Scientific and Technical Translation – The Pharmacopoeia of the People's Republic of China as a Case Study), Changchun: Jilin University Press 吉林大學出版社.

Qian, Duoxiu (2009) 〈計算機輔助翻譯課程教學思考〉(Pedagogical Reflection on the Design of a Course in Computer-aided Translation), 《中國翻譯》(*Chinese Translators Journal*) 4: 49–53.

Qian, Duoxiu (2010) 'Pedagogical Reflections on Computer-aided Translation as a Course', in Chan Sin-wai (ed.) *Journal of Translation Studies: Special Issue on the Teaching of Translation Technology* 13(1–2): 13–26.

Qian, Duoxiu 錢多秀 (2011a) 《計算機輔助翻譯》(*Computer-aided Translation: A Coursebook*), Beijing: Foreign Languages Teaching and Research Press.

Qian, Duoxiu (2011b) 'Applications of Translation Technology in Interpreting', *Minority Translators Journal* 4: 76–80.

Somers, Harold L. (2003) 'Introduction', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 1–12.

Su, Keh-Yih (1997) 'Development and Research of MT in Taiwan'. Available at: http://www.bdc.com.tw/doc/twmtdvp.gb.

Su, Mingyang (2007) 'Translation Memory: State of the Art and Its Implications',
Foreign Languages Research, 5.

Su, Mingyang 蘇明陽 and Ding Shan 丁山 (2009) 〈翻譯單位研究對計算機輔助翻譯的啟示〉(Research on Translation Unit and Its Implications on Computer-aided Translation), 《外語研究》(*Foreign Languages Research*) 6: 84–89.

Sun, Maosong 孫茂松 (2001) 〈漢語自動分詞研究的若干最新進展〉(New Advances in the Study of Automatic Segmentation of Chinese Language), in 《輝煌二十年—中國中文信息學會二十周年學術會議論文集》(*Proceedings of the Conference of the 20th Anniversary of the Chinese Information Processing Society*), Beijing: Tsinghua University Press 清華大學出版社, 20–40.

Wang, Kefei 王克非 (2004) 〈雙語對應語料庫：研製與應用〉(The Design and Construction of Bilingual Parallel Corpus), 《中國翻譯》(*Chinese Translators' Journal*) 6: 73–75.

Wen, Jun 文軍 and Ren Yan 任艷 (2011) 〈國內計算機輔助翻譯研究述評〉(Review of Computer-aided Translation (1979-2010) in China), 《外語電化教學》 (*Computer-assisted Foreign Language Education*) 3: 58–62.

Yu, Jingsong 俞敬松 and Wang Huashu 王華樹 (2010)〈電腦輔助翻譯碩士專業教學探討〉(A Master Programme in Computer-aided Translation), 《中國翻譯》(*Chinese Translators Journal*) 3: 38–42.

Yu, Shiwen (1993) 'Automatic Evaluation of Output Quality for Machine Translation Systems', *Machine Translation* 8: 117–126.

Yu, Shiwen 俞士汶 (ed.) (2002) 《第二屆中日自然語言處理專家研討會論文集》(*Proceedings of CJNLP 2002, the 2nd China-Japan Natural Language Processing Joint Research Promotion Conference*), 30 October – 2 November 2002, Beijing: Institute of Computational Linguistics, Peking University.

Zhang, Zheng 張政 (2003) 〈「機器翻譯」、「計算機翻譯」還是「電子翻譯」？〉(Machine Translation, Machine Translation or Electronic Translation?), 《中國科技翻譯》(*Chinese Science and Technology Translators Journal*) 2: 56–57.

Zhang, Zhizhong 張治中 and Yu Kehuai 俞可懷 (2002) 〈「機器翻譯」還是「機器輔助翻譯」—對「機器翻譯」之管見〉(Machine Translation or Machine-aided Translation? —Our View on "'Machine Translation"'《大連理工大學學報》(社會科學版), *Journal of Dalian University of Technology* (Social Sciences)) 3: 54–58.

Zhuang, Xiaoping (2007) 'The Integration of Machine Translation and Human Translation', *Journal of Yibin University*, 8.

15

TRANSLATION TECHNOLOGY IN CANADA

Elliott Macklovitch

FORMER PRESIDENT OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS

Introduction

Technology – understood here as machinery and equipment developed from the application of scientific knowledge for the solution of practical problems[1] – is clearly an evolving and historically conditioned notion. As our scientific knowledge advances, new technologies are routinely developed that render previous technology obsolete. The devices that were considered 'high-tech' for one generation may often wind up as quaint antique store objects in the succeeding generation. Allow me to illustrate with a little bit of personal history.

Although I had previously trained and worked as a linguist, I accepted a job as a translator with the Canadian Translation Bureau[2] in 1981. The Bureau is the largest employer of language professionals in the country and the government's principal centre of expertise on all matters involving translation and linguistic services. Upon my arrival at the Bureau, the equipment I was given to support me in my work was an electric typewriter. Compared to the manual machines that had long been in use, this was considered to be significant technological advance. Powered by a humming electric motor, it produced cleaner, more even copy, and demanded less manual force on the part of the typist. Moreover, my machine had a self-correcting key which allowed me to backspace and white over incorrect characters. And of these, there were a great many in my first texts, because I had never learned how to type. Seeing that my salary depended on the number of words I produced each day, it wasn't long before I purchased a teach-yourself-to-type manual and eventually became a semi-proficient typist.

Not all my colleagues at the Translation Bureau in those years used a typewriter to draft their target texts. A fair number preferred to dictate their translations, using a hand-held recording device commonly known as a dictaphone.[3] Once the translator had finished dictating her target text, she handed over the resulting cassette to a typist for transcription. The women in the typing pool had another set of impressive skills: not only were they speed typists, but they also needed to have a strong mastery of the spelling and grammar of the target language in order to transform the spoken translation on the recording into a correctly transcribed written version.[4] And they too worked with specialized equipment: headphones and a tape player they could control with a foot pedal. The dictating translators were among the most productive at the Bureau, even when the time and cost of transcription were taken into account.

Aside from these two basic pieces of equipment, the dictaphone and the electric typewriter, Canadian translators, as far as I could tell, had access to very little else in the way of technology

at the time. The section of the Translation Bureau I worked in was ringed with rows of filing cabinets in which all our past translations were stored and in which we routinely rummaged in an effort to locate texts similar to the new ones we were assigned; this, with varying degrees of success. And in our section library, there was a large card index containing drawers full of file cards on which we were urged to record the results of our terminological research. There too, searching for a term equivalent was frequently a hit-or-miss affair, since translators did not always have the time between assignments to register new terms. And even when they did, each concept rarely received more than one record, making it nearly impossible to locate the appropriate equivalent via a synonymous, alternate or abbreviated term. In short, the practice of translation in Canada some 30 years ago benefited very little from what we would consider technology today.

The advent of the first computerized tools

Things began to change just a few years later, with the arrival at the Translation Bureau of the first dedicated word processing machines. These bulky monsters, which were actually invented by a Canadian,[5] featured a keyboard and a video screen, and were dedicated in the sense that word processing was all they were designed to do, unlike the general-purpose programmable microcomputers that later supplanted them. Nevertheless, this early word processing technology proved to be invaluable to the employees in the Bureau's typing pools, greatly simplifying the job of introducing corrections into the final version of the text to be delivered. But perhaps these machines' most significant technical innovation was the removable magnetic disk on which texts were stored. Not only did these 8-inch disks greatly reduce the space required to store the Bureau's enormous production volume, they also made it much easier to locate and retrieve previous texts. If one had the right disk in hand, the operator simply had to type in the file name on the keyboard – far more efficient than rummaging through countless paper files.

The other technological innovation that had a significant impact on Canadian translators in the mid 1980s was facilitated access to *Termium*, the government's large computerized terminology bank. Originally developed at the University of Montreal in the early 1970s, *Termium* was acquired by the government of Canada in 1975, in an effort to help standardize the technical terminology and official appellations in the officially bilingual federal public service. Shortly after its acquisition, the government began a major scale-up of the database, as well as a fundamental revamping of the underlying software. By the time *Termium III* was released in 1985, the bank contained over a million records and its network of users numbered about 2500 people, the great majority of whom were government translators and other civil servants.[6] In the translation section I worked in, the arrival of a single dedicated terminal which allowed us direct access to the term bank, without having to address our requests to an overtaxed team of terminologists, was a major event; although by today's standards, interaction with the bank was anything but convivial.

Mention was made in the previous paragraph of the fact that the public service in Canada is officially bilingual. Before pursuing our examination of translation technology, it may be worthwhile to clarify just what this means, since official bilingualism has had extremely important consequences for the translation industry in this country. In 1969, the Canadian Parliament passed the Official Languages Act, which consecrated English and French as the country's two official languages, both having equal status. As a result of this Act, all Canadians have the right to services from the federal government in the official language of their choice, and all federal public servants have the right to work in one or other official language.[7] Furthermore, all the laws, regulations and official documents of the federal government must

be published simultaneously in both official languages, and both versions of these documents have equal weight before the law. Since it was first passed in 1969, the Official Languages Act has fueled much heated debate, and it underwent significant amendments in 1988. But one indisputable consequence of the legislation was to vastly increase the demand for English–French translation in Canada. Indeed, the federal government became, and today still remains, the largest source and client of translation in the country; and for a long time, the Translation Bureau, the agency responsible for translation, interpretation and official terminology in the federal public service, was one of the largest translation services in the world.[8] Yet even when its workforce surpassed a thousand full-time employees, the Bureau still had difficulty in meeting the continually rising demand for translation, while restraining its operating costs. In an effort to find solutions to both aspects of this problem, the Translation Bureau was impelled to search for innovative ways of streamlining the translation process, and it soon became an active partner in the development and evaluation of translation technology.

Machine translation

Before joining the Translation Bureau in 1981, I worked for four years at the TAUM group, TAUM being an acronym for 'Traduction Automatique de l'Université de Montréal', or machine translation at the University of Montreal. At the time, TAUM was one of the foremost MT research groups in the world. A year before I arrived at the university, TAUM had delivered to the federal government a first operational version of an MT system specifically designed for the translation of weather bulletins. Known as *TAUM-Météo*, this system was long considered to be one of the great success stories in machine translation; and to this day, the successor of *Météo* continues to translate the weather bulletins published by Environment Canada at a rate of more than 5 million words a year.[9] What exactly were the factors responsible for this unprecedented success?

To begin with, the weather bulletins published by the government's meteorological service constitute a highly restricted sublanguage which employs a small number of short sentence patterns and a limited vocabulary of a few thousand words (including place names). In itself, this serves to eliminate many of the ambiguities that are so pervasive in ordinary texts and which make machine translation such a difficult task. What's more, *TAUM-Météo* was designed to handle only the telegraphic portion of the bulletins describing weather conditions in specific localities of the country. It wasn't meant to translate the synopses that introduce these short bulletins, which describe major meteorological developments in larger regions, using a language that is far more free-ranging in form. These synopses were left to the Bureau's translators, who were also asked to revise the *Météo* system's output. The translators were more than willing to do this, for two reasons: first, the quality of the machine translations was generally quite good, with less than 5 per cent of the system's output requiring modification; and second, the translators were actually grateful to be relieved of such a boring and repetitive translation task. As for the Bureau, the introduction of the *Météo* system meant that it was able to meet the client department's requirements for rapid turnaround time of a large volume of text without having to incur the cost of hiring a large number of translators.[10]

Shortly after *TAUM-Météo* was delivered, the Translation Bureau was advised of another enormous translation task that it would be receiving. The government was about to purchase a new coastal patrol aircraft and, in accordance with the Official Languages Act, it would be required to translate into French the training and maintenance manuals, not just for the airplane but for all the sophisticated tracking equipment it carried as well. Flush from the success of the *Météo* project, the Bureau turned to the TAUM group and asked it to develop a new machine

translation system that would help it handle this daunting workload. TAUM agreed to take on this challenge, although retrospectively some group members now view that decision as foolhardy, or at least somewhat naïve. This was the birth of the *TAUM-Aviation* project, on which I came to work as an English-language linguist in 1977.

What was it that made *TAUM-Aviation* such an ambitious project – perhaps even an overly ambitious one? For one thing, the aviation manuals that we were undertaking to translate by machine bore absolutely no resemblance to the simple syntax and limited vocabulary of weather bulletins. These manuals may have belonged to a well-defined sublanguage; i.e. they did exhibit certain recurrent characteristics that distinguished them from ordinary, everyday English.[11] That said, this sublanguage was an exceedingly complex one, the description of which required the creation of very large dictionaries and a full computational grammar of English for the analysis of the texts to be translated. In no way could the linguists and lexicographers working on the project rely on the sublanguage to simplify their task, as their colleagues had been able to do on the *Météo* project.

Grammars and dictionaries had to be developed for *TAUM-Aviation* because we were of course working in the rule-based paradigm of machine translation; no other paradigm was available at the time, except perhaps for simplified word-for-word translation, which was clearly not up to the task. More precisely, *TAUM-Aviation* could be characterized as a second generation, rule-based system. Unlike earlier MT systems, those of the second generation proposed higher-level formal languages designed specifically for linguistic descriptions; and these descriptions of linguistic knowledge were clearly distinguished from the programming languages used to actually implement the system. Furthermore, second generation systems broke down the translation operation into three distinct linguistic phases: source-language analysis, which generated a syntactico-semantic representation of the text to be translated; bilingual transfer, which mapped that representation into its target-language equivalent; and a monolingual generation phase, which transformed the target-language tree structure into a correctly ordered and inflected target sentence. Details aside, what is important to realize is that all this linguistic description had to be undertaken by human specialists. In order to have the system map a source-language sentence into its target-language equivalent, they needed to hand-craft hundreds, if not thousands of linguistic rules. Like other types of expert systems (as they were called at the time), we were endeavoring to formalize and implement what we understood to be the mental operations of a qualified human translator. The problem, however, is that a qualified human translator, in grappling with the pervasive ambiguities inherent in natural language, routinely draws on vast amounts of linguistic and extra-linguistic knowledge – far more that we could ever hope to code into an MT system.[12] We slowly came to realize this on the *TAUM-Aviation* project. We rationalized it by telling ourselves that while our system was not meant to replace human translators, it might nevertheless render them more productive by providing them with a first draft of reasonable quality, which they could post-edit cost-effectively.

Such at least was our hope. Over the four years of the *Aviation* project, I believe it is fair to say that we succeeded in developing one of the most sophisticated MT systems in existence at the time.[13] When the system finally came to be evaluated, however, it was found to fall well short of its ambitious objectives. The translations produced by *TAUM-Aviation* were generally judged to be of very good quality, based as they were on a deep linguistic analysis. The problem, unfortunately, was that for too many sentences, the system produced no output at all, usually because the input didn't conform in some way to *Aviation*'s analysis grammar or to the lexical information contained in its dictionaries. These had been developed through the painstaking study of what was deemed to be a representative corpus: a 70,000-word hydraulics

manual. But when tested on material from outside the hydraulics domain, the system simply didn't generalize gracefully; which is another way of saying that it wasn't robust enough. Moreover, extending *Aviation*'s deep linguistic analyses to new domains would require a significant investment of time and effort; i.e., a full-fledged system based on this approach would turn out to be exceedingly costly. In 1981, the *TAUM-Aviation* project was concluded and, unable to find other sources of funding, the TAUM group was forced to disband.

As it happens, the weaknesses of the system developed on the *TAUM-Aviation* project were shared by most, if not all MT systems in the late 1970s and 1980s, including the major commercial systems that were trying to break into the translation market in the USA. The best of these systems were far too expensive for individual translators and could only be afforded by large corporations or translation services. Moreover, the quality of their output was not consistently good enough to allow for the cost-effective production of translations destined for publication or broad dissemination.[14] The federal Translation Bureau conducted trials of several such systems in the 1980s, but none was able to satisfy its requirements. The effort to create wide-ranging, general-purpose MT systems through the rule-based approach was simply too difficult a challenge for computational linguistics at the time. A radically different approach to the problem was required, and it finally emerged in the early 1990s with the advent of Statistical Machine Translation (SMT), as we will see below.

Machine-aided human translation

Yehoshua Bar-Hillel, who was the first full-time MT researcher in history, was also the first to demonstrate (Bar-Hillel 1960: 45–76) that fully automatic, high quality machine translation of unrestricted texts – sometimes abbreviated as FAHQTUT – is in fact impossible. The ingenious thought experiment by which he arrived at this conclusion need not concern us here; however, we can invoke the three parameters of his famous acronym to help characterize the state of the art in MT in the late 1980s and early 1990s. For MT's ultimate objective encompasses just these parameters: full automation, high quality, and general applicability. In the period under consideration, it was often said that, while the ultimate goal remained unattainable, it was still possible to develop systems which achieved two of these three desiderata. Fully automatic MT systems could produce high quality, but only in restricted domains, as demonstrated by the *Météo* system. Otherwise, when fully automatic systems were applied to unrestricted texts, it was high quality that would have to be forfeited. On the other hand, if high quality was a *sine qua non*, particularly for the translation of texts in wide-ranging domains, then a compromise would have to be made on full automation. To achieve this last sub-set of the desiderata, the only reasonable approach was to develop sub-optimal systems designed to assist – and not replace – the human translator.

The demand for high-quality translation was growing dramatically during this period, which was a time of expanding globalization, to the point that many professional translators were having increasing difficulty in coping with larger workloads and shorter deadlines. Not surprisingly, many of their large-scale corporate clients began to look to machine translation, hoping to find in that technology a solution to their pressing practical problems. It was in this context that Martin Kay (1980) published his seminal paper, 'The Proper Place of Men and Machines in Language Translation', in which he reiterated Bar-Hillel's arguments on the unfeasibility of FAHQTUT, and advanced instead a more modest program of machine-aided human translation. While this may not have been the message that many of the large clients of translation wanted to hear, it was also the approach adopted by the research group directed by Pierre Isabelle at the Canadian Workplace Automation Research Centre in Montreal,[15] where

I went to work in 1984. Like Martin Kay, Pierre contended that research on machine translation was fully justified and indeed necessary in advancing our understanding of natural language processing. But this remained a *research* goal and, as such, it was unlikely to provide practical solutions for working translators in the short term. For this problem, Pierre's group took a radically different, extremely original approach, setting as its goal the development of a whole new generation of computer-assisted translator tools.

Before turning to the CITI's program in machine-aided translation, allow me a short digression on statistical machine translation. As we mentioned at the end of the preceding section, the advent of SMT represented a revolutionary paradigm shift. It was first proposed by a team of researchers at IBM (Brown *et al.* 1990: 79–85) who were intent on applying to translation the same statistical techniques which had proven so successful in automatic speech recognition. A key feature of this new 'empirical' approach was its reliance on large amounts of previously translated text. This was the data from which their machine learning algorithms automatically acquired its translation knowledge, as opposed to the traditional, rationalist approach, in which linguists and lexicographers relied on their intuitions to craft declarative rules. It is interesting to note that the translated corpus that proved critical to the IBM group – both because of its size and its quality – came from the Canadian House of Commons, where all debates were required by law to be translated into the other official language. Electronic versions of those debates had in fact existed for some time. What allowed the IBM group to actually exploit that data was the development of automatic alignment programs, which calculate formal links between corresponding sentences in two files, one containing the source text and the other, its translation. Texts that are linked in this way were first called *bitexts* by Brian Harris, a professor of translation at the University of Ottawa, who was also among the first to appreciate their potential usefulness for human translators. (See Harris 1988: 8–11.) For the machine learning algorithms used in SMT to work effectively, the bitexts to which they are applied must be extremely large – in the millions of words – far more than anyone could ever align by hand.

In terms of the quality of the translations they produced, the early SMT systems did not really achieve a great leap forward; it wasn't until several years later that they finally overtook the traditional rule-based systems in the public competitions organized by US government. What they did do, however, was radically reduce the time and effort required to develop a new MT system; for their automated learning algorithms could be applied to any language pair for which sufficient translation data was available. But even then, it seemed clear that SMT was not yet the panacea that struggling translators and their overwhelmed clients were hoping for. This is why Pierre Isabelle's team at the CITI, while pursuing its own research into statistical MT, also undertook two major projects in machine-*aided* translation, designed to provide shorter-term solutions to the hard-pressed corps of professional translators, both in Canada and elsewhere.

The first of these was called Translator's Workstation project, and it was developed with the support of the federal Translation Bureau. Its goal was to integrate, within a user-friendly interface, various off-the-shelf programs, some of which were not even intended for translators. This may sound simple enough today, but it must be remembered that at the time most translators had only recently migrated to personal computers, whose hardware and operating system imposed serious limitations on the sharing and transfer of data between different applications. Following the suggestion made by Martin Kay in his 1980 paper, the central component of the successive workstations that were developed at the CITI remained a word processing program, to which a number of ancillary applications were added, including programs for glossary management, grammar and spell-checking, French verb conjugation, file

comparison, etc.[16] Attempts were also made to provide translators with a full-page monitor and to link their workstations together in a local area network. For further details on the Workstation project, the historically curious reader is referred to Macklovitch (1991).

The CITI's other major project in machine-aided translation was more original and certainly more ambitious, in that it set out to develop a whole new set of translator support tools. The project's starting point, or credo, was famously formulated by Pierre Isabelle as follows: 'existing translations contain more solutions to more translation problems than any other available resource' (Isabelle *et al.* 1993: 205). The challenge, of course, is how to make all that knowledge readily available to working translators; and the answer, it turned out, lay in the recently discovered concept of bitextuality. In 1993, William Gale and Kenneth Church (1993: 75–102), two brilliant researchers at AT&T Bell laboratories, published a paper containing an algorithm for automatically aligning sentences in large, parallel corpora, i.e. for creating arbitrarily large bitexts. At the CITI, Pierre Isabelle, George Foster and Michel Simard improved on the Gale-Church algorithm by exploiting the presence of cognates in the set of parallel texts (Simard *et al.* 1993: 1071–1082). And the CITI researchers went one big step further: they developed an interface and a database structure that allowed users to query the resulting bitext – the queries, in the case of translators, normally corresponding to a translation problem. *TransSearch*, as the resulting system was called, would retrieve all sentences containing the submitted query; and because this was a *bitextual* database, along with each retrieved sentence came the corresponding sentence in the other language, where the translator could often find the solution to her problem.[17] In 1996, the CITI made a version of *TransSearch* freely available on the Internet. It included a parallel corpus composed of tens of millions of words of Canadian Parliamentary debates; once again, that same data on which SMT had been spawned. The system proved so popular with translators that it was soon transferred to a private sector partner, who now manages subscriptions that are sold at a very reasonable price and ensures that the databases are regularly updated. Other bilingual concordancers have since become available – imitation is often said to be the ultimate compliment – but *TransSearch* was the first such tool that allowed translators to take advantage of all the richness that up to then had lain dormant in past translations.

The CITI's two other projects in translator support tools did not meet with the same commercial success as *TransSearch*, although they were probably even more innovative. The *TransCheck* project, as its name suggests, set out to develop a translation checker: similar in conception to a spelling or grammar checker, with the important difference that *TransCheck* focussed on *bilingual* errors, i.e. errors of correspondence between two texts that are mutual translations.[18] The system began by automatically aligning the sentences in the source and target files; and then it verified each aligned sentence pair to ensure that it respected certain obligatory translation correspondences, while containing no prohibited correspondences. An example of a prohibited correspondence would be an instance of source language interference, such as a deceptive cognate. (*TransCheck* incorporated an open list containing many common translation interdictions between English and French.) An example of a compulsory correspondence would the correct transcription of numerical expressions (including dates, monetary expressions, measurements, etc.), or certain terminological equivalences which had to be respected. Terminological consistency, however, turned out to be much more difficult to enforce than we had anticipated, owing to pronominalization, natural omissions and other forms of translator licence. This was probably one of the principal reasons why *TransCheck* never achieved widespread adoption.[19]

The *TransType* project proposed a radically new type of editing environment for the translator, designed to reduce the time and effort required to key in a target text by exploiting

the proposals made by an embedded SMT system. The way the system operated was basically as follows. For each source sentence, the SMT system, operating in the background, would generate a host of potential translations. When the user began to type her translation of a given sentence, the system would select, among its automatically generated candidates, the most likely of those that were compatible with the prefix the user had keyed in and propose an extension to the user's draft. The user could either accept that proposal, ignore it by continuing to type, or modify it in some way. With each new keystroke the user entered, however, the system would immediately revise its predictions and propose a new compatible extension. These interactions between the user and the system would continue until a satisfactory target equivalent for the sentence was generated. Shortly after the CITI's MT group moved to Université de Montréal (where it became the RALI Laboratory), the *TransType* project was awarded a European Commission research grant, allowing several prominent research groups to join the TransType2 research consortium. Two translation companies, one in Canada and one in Spain, also participated in the project, providing invaluable end-user feedback to the developers. The TT2 project was pursued until 2005, and gave rise to a series of sophisticated prototypes for a number of language pairs and exploring intriguing research questions – e.g. how can an SMT engine learn from the user's interactions in real time – for which practical solutions are only now beginning to emerge.[20]

The current situation

At this point, I want to shift the perspective somewhat, moving away from a historical account in order to focus on the current situation of translation technology in Canada. But first, a few words on the translation market in this country and the place of Canadian translators in it.

For a country with a relatively small proportion of the world population – about 0.5 per cent – Canada accounts for a surprisingly large proportion of the world's translation production: approximately 10 per cent, according to a recent study by PricewaterhouseCoopers (2012). It is difficult to obtain recent, accurate figures on the number of Canadian translators, but one federal government website mentions an average 10,250 persons (including terminologists and interpreters) between the years 2008 and 2010.[21] Another government study puts the number of firms working in this sector at about 800, with most of these employing five or fewer people. The government's own Translation Bureau is by far the largest service in the country, with over 1,200 full-time employees. And the great bulk of translation that is done in Canada is still between the two official languages, English and French. Finally, translators are relatively well-paid in Canada compared to their colleagues in other countries; the Service Canada website puts their average annual income at $50,000. So in principle, Canadian translators can afford to invest modestly in technology. But do they?

Once again, there is not a great deal of recent and reliable data available on this question, but we can begin by looking to our neighbours to the south. In 2008, the American Translators Association published a study which showed that the three most commonly used technology tools among its members were word processing applications (98 per cent usage), translation memory applications (47 per cent usage), and terminology management software (27 per cent usage).[22] I strongly suspect the situation is quite similar among Canadian translators, but we can sharpen the picture somewhat by examining the results of a smaller-scale survey conducted by AnneMarie Taravella in 2011 for the Language Technologies Research Centre (LTRC) in Gatineau, Quebec. Almost all of the 380 respondents to this survey were translators, terminologists or students enrolled in a translation program, and there was some disagreement among them as to whether word processing belonged to the

category of language technologies, or whether the latter should be restricted to technologies that are used only by language professionals. On the other hand, virtually everyone surveyed used a word processor. Moreover, 97 per cent of the respondents said they regularly consulted what Mme Taravella called a 'passive' language technology, i.e. terminology banks like *Termium* and *Le grand dictionnaire terminologique*,[23] correctors like *Antidote*, or bilingual concordancers and online dictionaries like *Linguee*. And much like their American counterparts, 54 per cent of the Canadian respondents claimed to use at least one 'active' language technology, these being essentially various types of translation memory. In short, the picture that emerges from this survey is that, as a group, Canadian translators are certainly computer literate today. To this, I would add my own personal observation that those who work in larger translation services or companies are even more likely than their freelance colleagues to work with a translation memory tool. Mme Taravella also asked her respondents to assess their use of language technologies:

> 97% of the respondents who indicated that they used language technologies claimed that they helped save time, 90% claimed that they improved the quality of their work and 90% claimed that they increased the uniformity of their work. 44% of the respondents … claimed that it was a requirement of their employer or their clients.
>
> (Taravella 2011: 10)

Looking briefly at the supply side of the equation, the Canadian language industry boasts a number of small but innovative companies which develop various types of translation technology. Perhaps the best known of these sell translation memory systems; they include MultiCorpora (and its *MultiTrans* product), Terminotix (and its highly-regarded *LogiTerm* system), and JiveFusion. Several Canadian companies have developed sophisticated systems for managing translation workflow, including MultiCorpora's *Prism Flow* and Logosoft's *TransFlow*. The latter company also offers a bilingual concordancer called *Tradooit*, while Terminotix has developed *SynchroTerm*, a bilingual term extraction tool, as well as *AlignFactory*, an automatic alignment program that facilitates the creation of bitexts.

Returning to Mme Taravella's survey, it is interesting to note that none of her respondents mentioned machine translation, which is now used by millions of persons every day on public websites like Google Translate and Microsoft Translator. But what of MT for the production of publication-quality translation? The PricewaterhouseCoopers report (2012) has the following to say on the question:

> Machine translation, although a key productivity enhancing tool, is generally not considered to produce a level of quality sufficient to correctly convey a full message in another language, and its output must be reviewed by a qualified translator. As a result of the significant post process editing, machine translation is not widely adopted. It is generally used for large volume translations with an accuracy rate of 75% to 85%.
>
> (2012: 19)

The Canadian Translators, Terminologists and Interpreters Council (CTTIC) is cited as the source of the opinion in the first sentence; no source is cited for the startling across-the-board figures given in the final sentence. If it were indeed true that MT systems were capable of achieving 85 per cent accuracy on arbitrary texts,[24] I'm quite sure that there would be tremendous interest in the technology, given the ever-increasing demand for translation worldwide and the strong market pressure to lower costs and shorten turnaround

times. As for the need to have the MT output reviewed by a qualified human translator, this in itself is not sufficient reason to discard the technology. Rather, the real question today is the following: given the impressive improvements in statistical MT in recent years, and the possibility of training such systems for well-defined domains where large volumes of past translation are now available, has the performance of such specialized engines reached a level where their output can be cost-effectively post-edited? We are hearing more and more evidence from various quarters that the answer to this question may well be yes. In Canada, the largest private-sector translation provider has been using *Portage*, the NRC's highly regarded SMT system, for over two years to help it produce some of its technical translations, and other major translation firms are actively exploring the possibility of integrating machine translation into their operations.

Another promising technology which hasn't yet been mentioned is automatic speech recognition (ASR). As we stated in the Introduction, dictation used to be the preferred mode of text entry for many translators in Canada; preferred, not only because it is fast – everyone speaks faster than they can type – but also because it allows the translator to focus on her specialization, relegating such mundane matters as layout and format to a typist. Two factors combined to change this situation in the mid 1980s: the advent of the personal computer, equipped with sophisticated word processing programs; and the increasing difficulty of finding competent typists who knew their grammar and spelling well enough to produce an error-free text from a recording. In many services, translators were instructed to turn in their dictaphones and were told that henceforth they would have to type their own target texts on a PC. Moreover, this was often presented as the inevitable march of progress, although for many translators – particularly those who were not proficient typists – the concrete benefits were not immediately obvious.

In May 2011, I conducted a series of consultation sessions with the employees of the federal Translation Bureau which focussed on the technologies they were currently using and those they would ideally like to have. In the course of those sessions, I was surprised to discover that a fair number of translators continue to dictate their texts. These included older employees who had never given up their dictaphone, as well as younger translators who were using commercial ASR systems (almost always Dragon *NaturallySpeaking*), some in response to health problems. In principle, this technology has the potential to resolve the difficulties alluded to in the previous paragraph. For the translator, it allows a return to the more comfortable mode of dictation; only now, instead of having to wait for the typist to complete the transcription, the target text magically appears on the computer screen almost as fast as she can speak it. And for the translation manager, the elimination of the typist should help lower operating costs.

Except that we're not quite there yet … Between this idealized scenario and the real-world performance of the best of today's ASR systems, there remains a gap that is populated by speech recognition errors which the translator is obliged to correct, thereby reducing her productivity. These systems have made remarkable progress in recent years and that gap is certainly closing; but for many translators, particularly those who work in languages other than English, the word error rates remain too high to allow automatic dictation to be cost-effective. This situation is likely to improve in coming years, as will the other major problem with ASR: the fact that the technology is not yet satisfactorily integrated with the other support tools that translators have come to rely on, particularly translation memory systems.

Notes

1 Definition drawn from the online version of the Oxford English Dictionary.
2 http://www.btb.gc.ca/btb.php?lang=eng.
3 Dictaphone was actually the registered trademark of an American company, but I am employing the term informally here to refer to any tape recorder used in translation.
4 In fact, transcription often involved several iterations of correction between typist and translator.
5 Stephen Dorsey, the founder of AES and later Micom Data Systems.
6 For more on *Termium* at the time, see Landry (1987).
7 The application of the latter clause is subject to certain geographical restrictions, i.e. to areas where the minority language has a certain minimum density.
8 Until it was overtaken by the European Commission's translation service.
9 The volume used to be higher. It has declined somewhat in recent years, since Environment Canada is now generating in parallel certain bulletins that used to be drafted in one language and then translated.
10 For a short article on the development of the *Météo* system, see Kittredge (2012: 19–20).
11 See Lehrberger (1982: 81–106) for a detailed discussion of this sublanguage question.
12 It wasn't just the number of rules that was problematic; the rules often conflicted with one another in unpredictable ways.
13 For a detailed description and assessment of *TAUM-Aviation*, see Isabelle and Bourbeau (1985: 18–27).
14 On the other hand, they could and were used for other purposes, notably for information gathering by military and intelligence services.
15 The centre later changed its name to the CITI. Several members of the machine-aided translation team there, including Pierre Isabelle, had previously worked at the TAUM group at Université de Montréal.
16 At the time, many of these components were not yet included within the word-processing program.
17 For a detailed description of *TransSearch*, see Macklovitch *et al.* (2000: 1201–1208).
18 For more on *TransCheck*, see Macklovitch (1995).
19 Although some commercial products now exist which do offer a similar type of bitextual quality assurance, e.g. *ErrorSpy* by D.O.G. GmbH.
20 For more on the TT2 project, see Casacuberta *et al.* (2009: 135–138).
21 http://www.servicecanada.gc.ca/eng/qc/job_futures/statistics/5125.shtml. Of this number, about 74 per cent are said to work full-time.
22 These figures are cited in PricewaterhouseCoopers (2012).
23 Like so many other linguistic resources, these two large-scale term banks are now accessible over the Internet. According to the Internet World Stats site, Internet penetration in Canada was about 83 per cent in 2012.
24 Although much depends on the linguistic units these accuracy figures are applied to. If 85 per cent of the *sentences* in a machine translation are accurate, then this would undoubtedly enhance production; if it's 85 per cent of the *words*, then the impact on productivity is far less certain. (My thanks to Pierre Isabelle for pointing this out to me.)

References

Bar-Hillel, Yehoshua (1960) 'The Present Status of Automatic Translation of Languages', in *Advances in Computers*, 1, New York: Academic Press. (Reprinted in Sergei Nirenburg, Harold L. Somers, and Yorick Wilks (eds) *Readings in Machine Translation*, Cambridge, MA: MIT Press, 45–76.)

Brown, Peter, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John Lafferty, Robert L. Mercer, and Paul S. Rossin (1990) 'A Statistical Approach to Machine Translation', *Computational Linguistics* 16(2): 79–85.

Casacuberta, Francisco, Jorge Civera, Elsa Cubel, Antonio L. Lagarda, Guy Lapalme, Elliott Macklovitch, and Enrique Vidal (2009) 'Human Interaction for High-quality Machine Translation', *Communications of the ACM – A View of Parallel Computing* 52(10): 135–138.

Gale, William A. and Kenneth W. Church (1993) 'A Program for Aligning Sentences in Bilingual Corpora', *Computational Linguistics* 19(1): 75–102.

Harris, Brian (1988) 'Bi-text, a New Concept in Translation Theory', *Language Monthly* 54: 8–11.

Isabelle, Pierre and Laurent Bourbeau (1985) 'TAUM-AVIATION: Its Technical Features and Some Experimental Results', *Computational Linguistics* 11(1): 18–27.

Isabelle, Pierre, Marc Dymetman, George Foster, Jean-Marc Jutras, Elliot Macklovitch, Francois Perrault, Xiaobo Ren, and Michel Simard (1993) 'Translation Analysis and Translation Automation', in *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japan, 201–217.

Kay, Martin (1980) 'The Proper Place of Men and Machines in Language Translation', Xerox Corporation. (Reprinted in Sergei Nirenburg, Harold L. Somers, and Yorick Wilks (eds) *Readings in Machine Translation*, Cambridge, MA: MIT Press, 221–232.)

Kittredge, Richard (2012) 'Reflections on TAUM-MÉTÉO', *Circuit* 117: 19–20.

Landry, Alain (1987) 'The Termium Termbank: Today and Tomorrow', in Catriona Picken (ed.) *Proceedings of the 9th Translating and the Computer conference*, London: ASLIB, 130–144.

Lehrberger, John (1982) 'Automatic Translation and the Concept of Sublanguage', in Richard Kittredge and John Lehrberger (eds) *Sublanguage: Studies of Language in Restricted Semantic Domains*, Berlin: de Gruyter, 81–106.

Macklovitch, Elliot (1991) 'The Translator's Workstation … in Plain Prose', in *Proceedings of the 32nd Annual Conference of the American Translators Association*, 16–19 October 1991, Salt Lake City, UT.

Macklovitch, Elliot (1995) 'TransCheck – Or the Automatic Validation of Human Translations', in *Proceedings of MT Summit V*, 10-12 July 1995, Luxembourg, Geneva: EAMT European Association for Machine Translation.

Macklovitch, Elliot, Michel Simard, and Philippe Langlais (2000) 'TransSearch: A Free Translation Memory on the World Wide Web', in *Proceedings of LREC 2000*, 31 May – 2 June 2000, Athens, Greece, 1201–1208.

Nirenburg, Sergei, Harold L. Somers, and Yorick Wilks (2003) *Readings in Machine Translation*, Cambridge, MA: MIT Press.

PricewaterhouseCoopers LLP (2012) 'Translation Bureau Benchmarking and Comparative Analysis: Final Report'. Available at: http://www.btb.gc.ca/publications/documents/rapport-report-benchmarking-eng.pdf.

Simard, Michel, George Foster, and Pierre Isabelle (1993) 'Using Cognates to Align Sentences in Bilingual Corpora', in *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research: Distributed Computing*, 2: 1071–1082.

Taravella, AnneMarie (2011) 'Preliminary Summary Report on the Results of the Survey Conducted among Users of Language Technologies in April-May 2011'. Available at: http://www.crtl.ca/publications_LTRC.

16

TRANSLATION TECHNOLOGY IN FRANCE

Sylviane Cardey

UNIVERSITY OF FRANCHE-COMTÉ, FRANCE

Introduction

In respect of natural language processing, machine translation is without doubt the application which requires the most in terms of knowledge at all the linguistic levels (lexico-morpho-syntactic), and this without talking of oral machine translation which is even more difficult and which we mention very succinctly here.

Languages when confronted with machines have already been the object of much research and criticism in respect of their analysis. Languages are codes which transmit information by

1 'words' present in dictionaries (words are here between quotation marks as we do not have a precise definition of this concept even if it is regularly used to describe all sorts of phenomena) and which are the conventional way of representing things and ideas;
2 inflexions which add information to the message constituted by words; and
3 the rules of syntax which add precision in their turn in respect of the individual meaning of words and their role in relation to other words.

Problems concerning variously homophony, homography, polysemy, and all the ambiguities that are possible at the levels of lexis, syntax and, above all, composition from the least to the most frozen are still far from being solved for processing by machine.

> One can say that translation is only possible by means of an analysis of all the linguistic elements used to represent meaning – semantic values, inflexions and grammatical values, syntactic values – which are entangled in the words and the relations between them. This analysis is followed by the synthesis of the linguistic elements of language B, or output language, chosen because they enable expressing approximately the same content and are combined according to language B's own laws.
>
> (translated from Delavenay (1959))

Brief history

The first translation machines appeared in the 1930s notably with the work of the Russian scientist Troyanskji. The first experiments in machine translation involving computers date from 1948 in Britain and the USA; the USSR started in 1954 and Italy followed in 1959 (Léon 1998: 55–86).

The beginnings of machine translation in France

France started MT research in 1959–60 with the creation of the ATALA (l'Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée) in 1959, and the CETA (Centre d'Études pour la Traduction Automatique) in December 1959 within the Institut Blaise Pascal (IBP) with two sections, one in Paris directed by Aimé Sestier (CETAP), and the other in Grenoble directed by Bernard Vauquois (CETAG).

One wonders why France started so late, that is thirteen years after the first MT demonstration on a computer in New York in January 1954 at the instigation of the Georgetown University team directed by Léon Dostert (who was French), and above all after the Bar-Hillel report (1959–1960) and subsequently the ALPAC report in 1966. In fact, it was in 1967 that the CETA organized its second conference on Natural Language Processing where the first effective demonstration of French–Russian translation by computer was presented.

The end of the 1950s

In 1954, there were no computers in France whilst there were several in Britain. 'Informatique' (Computer Science) was unknown at the epoch (the term only appearing in 1962, coined by Ph. Dreyfus), and one talked simply of experiments in the USA. In this context, Sestier, who became director of the CETAP (Centre école des troupes aéroportées) was one of the rare persons in the defence sector who was interested in computing.

The various companies working with the French Ministry of Defence on electronic and high precision mechanical problems had all refused to take the technological risk of starting to build a French computer. The only company which had agreed to undertake technology studies was IBM. After two unsuccessful attempts to construct a French computer, France purchased a British machine in 1955, an Elliot 402, for the IBP. Also, despite the presence of A.D. Booth, one of the British pioneers of machine translation, this latter subject was not discussed at a conference organized by the IBP in January 1951 entitled 'Les machines à calculer et la pensée humaine' (Computing machines and human thought); interest in machine translation did not appear to be echoed in France. It has to be noted that French linguists manifested no specific interest in formal languages.

However, Emile Delavenay (founder of the ATALA), because of his responsibility for translation and editing services at the United Nations, New York up to 1950, took a close interest in the problems of translation at the international level. Thus it is not surprising that he was the instigator of MT in France; in his memoirs (1992) Delavenay evokes the lack of receptivity by linguists and academics in general concerning the idea of creating MT systems in France. However, a MT work group was constituted around Delavenay which kept abreast of progress in the work of the Americans, the British and the Russians. This group took the name 'groupe international d'études sur la traduction automatique' and met regularly at UNESCO; the group was at the origin of the ATALA. In 1953, UNESCO took stock of the growing global need for scientific and technical translations, reporting the lack of training of translators and the excessive costs of translation. Finally, numerous MT papers were presented at the first IFIP (International Federation for Information Processing) congress organized by UNESCO in Paris in June 1959. The creation of CETA at the IBP resulted in associating MT closely with the development of numerical methods, computers and automated documentation. An important role was given to the interaction between applied mathematics, formal languages and linguistics. As in many countries, the defence

sector was the stakeholder in the development of this discipline, where mathematicians, engineers and linguists worked together in CETA from 1961 onwards. The language in which the work was done was Russian.

The Sestier Report (1959)

However, for Sestier, mass production of translations was the priority and CETA ought to offer certain services: rough translations and studies on indexing and automatic extraction. Sestier also proposed the name *Centre d'études et d'exécution de traductions automatiques* which underlined the centre's vocational response to social needs. The method recommended by the Sestier report is especially centred on analysis of the source language, this being Russian as for most of the US research, and the task to be achieved was the translation of scientific and technical articles. The objectives were in fact linked in part to defence and to counter espionage.

The report pointed out the lack of personnel provided with a 'fundamental' linguistic training. Grenoble was made responsible for morphology and Paris with syntax. The Grenoble group decided to take on as well lexical polysemy problems. This decision was declared as a temporary and unstable step at the first Scientific Council meeting held 20 February 1960. Martinet and Benveniste, members of the Council, were sharply critical of this division between morphology and syntax. They said that it was not pertinent when the objective was to compare two structures; it would be more interesting to start from a solution which was less graphical and more linguistic. As well as this, this division very rapidly became irksome. The pretexts concerning it were due as much to differences in computers as differences in methods. According to Sestier (July 1960) the Grenoble group developed a morphological system uniquely for a binary machine, and which was thus strictly unusable by CETAP which was using a decimal machine with a small memory. Furthermore, for Sestier, the CETAG system appeared unnecessarily complicated. The Parisians thus decided to take on the morphological analysis. The members of CETAG showed in their project report their intention also to do research in syntax concerning the translations Russian–French, Japanese–French (Makato Nagao was invited) and German–French, adopting the model Sydney Lamb had developed at the University of California, Berkeley.

In 1963, Bernard Vauquois, very interested in formal languages and with a Russian group of which Igor Mel'cuk was a member, was working on an intermediary language which he called a pivot language (Vauquois 1975).

Bar-Hillel's report became known in 1962. CETAP was dissolved and Maurice Gross and Yves Gentilhomme went back to the Laboratoire de calcul numérique of the Institut Blaise Pascal. Following this crisis, the name of ATALA was changed to *Association pour le Traitement Automatique des Langues*, and its review *La Traduction Automatique* to *TA Informations, Revue internationale des applications de l'automatique au langage*.

The applied mathematics section of the IBP encouraged linguists and logicians to collaborate in carrying out a detailed and accurate study of natural languages. In other words, problems in MT are due variously to linguistics, to logic, to electronics and to programming. As for research, it was the development from 1963 onwards of a syntactic rules language; then from 1965 of a pivot language which constituted the most original research by CETA and thus opted for an MT method using an intermediate language. The pivot language being a syntactico-semantic model ensured independence of the translation process's analysis and synthesis phases. By 1970, as for most MT endeavours, CETA considered MT as the transfer of the meaning of a text written in a source language to a target language. At this epoch pivot languages were an attempt to formalize this intermediate level that was called the 'semantic level'.

Bernard Vauquois, Maurice Gross and Yves Gentilhomme have all, in one way or another, contributed to current MT research at respectively Grenoble, Paris and Besançon, all of which we address in the third part of the chapter.

Machine translation in France at the present time

Methodologically one can say that linguists and computer scientists share the MT scene. Sometimes they work together, but even so, as elsewhere in the world, they have great difficulty in listening to each other. This is to be compared with the outset of MT when mathematicians and linguists succeeded perhaps better in collaborating.

There is a great need for translation; in addition, the point is to know how the different types of translation are divided up. Three types can be distinguished:

1 Rapid and crude translation with a view to knowing very approximately the content of some document. This type concerns scientific and industrial organizations where researchers and engineers, not being able to read texts in the original language, need to inform themselves of research or other work conducted elsewhere. The users, who are familiar with their proper scientific domains, do not require perfect quality. This type of translation is also relevant for multilingual organizations when working documents of a temporary nature are involved.
2 Translations of texts of a general or specialized scope which have to be of good quality.
3 Accurate translations; this concerns for example standards, prescriptive texts of multilingual organizations, or in safety and security critical domains as we will see with Centre Tesnière's work at Besançon.

There exists also a range of tools such as translator aids and dictionaries.

We present here six research centres and companies which are currently active in France in MT each with their methodology and their products. We start with those organizations whose work is principally based on rule-based methodologies and finish with those which use rather statistically-based methodologies.

Centre Tesnière

We start with the Centre de recherche en linguistique et traitement automatique des langues, Lucien Tesnière (in brief Centre Tesnière), which is a research laboratory in the Université de Franche-Comté, Besançon. Since its foundation in 1980, research has been and continues to be done by linguists, mathematicians and computer scientists working together. The Centre was created by Professor Yves Gentilhomme and has been directed since 1994 by Professor Sylviane Cardey.

Centre Tesnière has many MT systems involving particular methodologies. Two of them, Korean–French and Chinese–French (Cardey *et al.* 2003: 22–44), use the transfer method with very fine grained linguistic analyses. Another, French–Arabic (Cardey *et al.* 2004: 37–47) uses a double pivot and a gradual generation involving both languages at the same time. A third methodology, pivot + transfer (Cardey *et al.* 2008: 322–329), has been developed which involves controlled French to variously controlled Arabic, Chinese, English and Thai, and controlled French too (identity translation).

All these machine translation systems are based on Centre Tesnière's constructive micro-systemic linguistic analysis approach in which traceability is inherent (systemic quality

model) (Cardey 2013). The systemic approach that has been mentioned here is based on logic, set theory, partitions and relations, and also the theory of algorithms. A theoretical approach which is mathematically based, whatever it is, ought to be able to accept linguistic formalisms. For this reason such an approach has to be sufficiently flexible so as to enable the construction of models themselves founded on model theory using a constructive logic approach. These models must adapt themselves as the analysis proceeds and when new problems are uncovered. The linguistic approach involves the delimitation of sets by choosing only and uniquely those elements which serve to solve the problem, the sets so involved being able to function together in terms of relations.

As Centre Tesnière works with safety critical domains which cannot admit any error, in their pivot + transfer methodology (Cardey *et al.* 2008: 322–329), the source controlled language must not only conform to normal controlled language constraints, but it must also be able to be machine translated, without manual pre- or post-edition (no time available during emergencies) to target languages which are themselves controlled. The methodology is based on linguistic norms and a supporting mathematical model for the construction of a single source controlled language to be machine translated to specific target controlled languages.

1 The source and target languages are controlled as controlled languages *per se*, that is for human use, the traditional raison d'être for controlled languages.
2 The source and target languages are controlled in a mutual manner so as to ensure reliable machine translation. The authors of the messages only know the source language.

The first step thus consists in detecting what is common and what is divergent in the languages concerned. Equivalence tables are established in micro-system form in order to solve divergence problems and for finding the 'equivalent form' for each concept in the other languages. This presents a real challenge as texts have to be machine translated without error into several target languages and without post-editing.

As well as for end-user applications, such computational processing can be useful, for instance, for the mechanical verification of linguistic data representations, for grammatical concordances and traceability and also for automated case-based benchmark construction, etc.

Given a defined domain and a specific need to be processed, the equivalences and divergences between the languages concerned can be represented in the following way. With three (or more) languages, whatever they are, the systems that are common to the three languages in question are constructed, to these are added the systems that are common to all the pairs of languages, and finally are added the systems specific to each language (see Figure 16.1).

Certain of the systems will be common with inflexional languages, others with agglutinative languages and still others with isolating languages.

As said at the outset, Centre Tesnière's model can be applied to all languages. Figure 16.2 illustrates the potential for extraction from **ข้อต่อรอง** (Thai). This is the sort of problem that can be solved using their methodology.

Centre Tesnière's methodology working in intension allows the detection, tagging and disambiguation of neologisms, and also automatic acronym detection.

Centre Tesnière has coordinated amongst others the French (ANR) project LiSe (Linguistique et Sécurité) and the European MESSAGE project which concern security in general, and in particular where communication involving humans ought to be rapid and correct. Generation of information without ambiguity, rapidly and in several languages being the need in the case of emergencies and crises, using micro-systemic linguistic analysis Centre Tesnière has classified and organized the language equivalences and divergences in the form of

a compositional micro-system structure expressed in a declarative manner by means of typed container data structures together with their contents so as to be incorporated in the machine translation process (Cardey *et al.* 2008: 322–329). This has resulted in a model based on language norms and divergences with inherent tracing.

The controlled languages mirror each other. The architecture of the machine translation system is thus based on the variants being divergences between the controlled target languages and the canonical controlled French source language, these divergences being organized in such a manner as to affect the translations during the translation process.

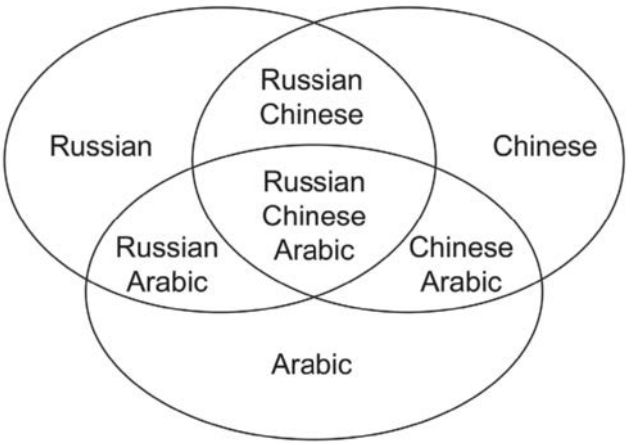


Figure 16.1 Common and specific systems between languages

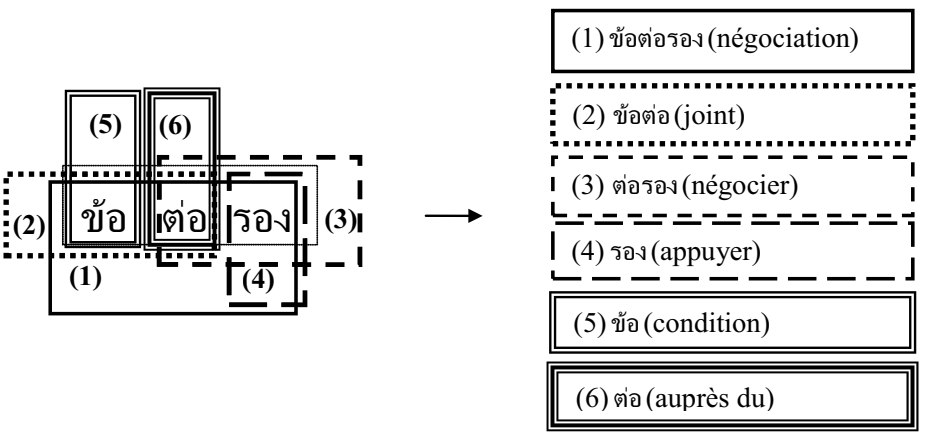


Figure 16.2 Illustration of the potential for extraction from **ข้อต่อรอง** (Thai)

GETALP

We now present GETALP (Groupe d'Étude pour la Traduction/le Traitement Automatique des Langues et de la Parole) with as principal actors Professor Bernard Boitet and Professor Hervé Blanchon, successors to Bernard Vauquois.

In MT, GETALP distinguishes three types of architecture:

1 the linguistic architecture described in the famous Vauquois triangle which comprises different phases or steps;
2 the computational architecture of these different steps, whether they be produced by:
   - expert approaches (using experts in linguistics for writing programs, rules and dictionary entries),
   - empirical approaches (performing automatic learning on more or less annotated data) or
   - hybrid approaches (combining both expert and empirical approaches);
3 the operational architecture (determining in particular where and when a human may intervene, whether control by back-translation is enabled, or whether, in oral situations, the translation is 'consecutive' or 'quasi-simultaneous').

Both the first and second types are independent, whilst the third often impacts on the choices made for the first two.

In an international context which, as regards research, now in the main promotes purely empirical methods in MT, GETALP has chosen not to set aside the expert approach (the one historically used in Grenoble under Bernard Vauquois), in particular the semantico-pragmatic pivot approach, whilst also progressing towards the construction of a Lingware Development Environment (LDE), without ignoring the empirical approaches for which original approaches are proposed. Whilst participating in the emergence of the hybrid approaches we note that an LDE is a Computer-Aided Software Engineering tool for Natural Language Processing. This provides for instance specialized languages for linguistic programming (SLLP) which enable making specific operations on manipulated objects.

As to international research now in the main promoting purely empirical methods in MT, from the point of view of GETALP, two phenomena explain this tendency: (1) the sponsors of MT projects want increasingly frequent evaluations and these evaluations, founded upon the availability of parallel corpora, do encourage empirical approaches; (2) the availability of free software toolboxes, exploiting aligned corpora, has promoted their development.

In terms of linguistic architecture, since late 2000, GETALP mainly focuses on pivot-based approaches (using the expert approach for both analysis and generation), and the direct approach (using the empirical approach). As for pivot-based MT the main contributions of the GETALP team are speech-to-speech translation (partner of the C-STAR II consortium (Blanchon and Boitet 2000: 412–417) and the NESPOLE! Project, using an ontological interlingua) and text translation (using the UNL semantico-linguistic interlingua). As for the direct approach the main contributions of the GETALP team are two-fold: news and speech within the WMT and IWSLT competitive evaluation campaigns (Potet *et al.* 2012).

In terms of operational architecture, GETALP focuses mainly on Dialogue-Based MT (DBMT), where the targeted user is a monolingual author, or a writer who is not fluent in the target language, and Interactive Multilingual Access Gateways.

As for DBMT, using a multilevel transfer linguistic approach, GETALP has proposed and evaluated a technique to produce interactive disambiguation questions. The components involved in the translation process cooperate within a distributed architecture by means of a

'light' document processing environment. GETALP has also proposed the idea of Self-Explaining Documents (SED) (Choumane *et al.* 2005: 165–167). An SED is a document enriched with the answers provided by the author during interactive disambiguation. It gives readers, on demand, explanations about its intended meaning, in order to avoid any misunderstanding due to ambiguities.

The concept of iMAG allows external multilingualization of websites (Boitet *et al.* 2010: 2–12). Unlike present translation gateways like the one provided by Google, this consists in associating with an elected website a dedicated iMAG gateway containing a translation memory and a specialized lexical database, and enables Internet users to improve translations, thus the translation memory, through on-line post-edition, and to enrich the lexical database, wherein these data may then be used for constructing MT systems specialized in the sub-language of the selected website.

Systran

Let us turn now to an industrial machine translation system, the oldest, with as its principal scientific actor Jean Senellart; he was a pupil of Maurice Gross of whom we have spoken in the first part of the chapter.

Systran is the supplier covering the largest range of machine translation (MT) methodologies in France.

The company was founded in 1968 in La Jolla, California by Peter Toma. It was acquired by Gachot S.A., a French company, in 1986. Systran's headquarters is located in Paris. It has a subsidiary in San Diego, California. Systran has numerous customers all over the world.

Today Systran exploits several methods ranging from rule-based MT, hybrid MT to purely statistical MT for new language pairs.

The Systran system is traditionally classified as a rule-based system. However, over the decades, its development has always been driven by pragmatic considerations, progressively integrating many of the most efficient MT approaches and techniques. Nowadays, the baseline engine can be considered as a linguistic-oriented system making use of dependency analysis and decision trees, general transfer rules as well as of large manually encoded dictionaries (100,000–800,000 entries per language pair). In recent years Systran has developed a hybrid approach which consists of the combination of rule-based translation and statistical post-edition (Dugast *et al.* 2007). Based on translation memories or other parallel corpora the system learns statistical models to correct the MT output and renders the final translation more fluent. In contrast to statistical MT (SMT) systems the hybrid approach sharply reduces the amount of data required to train the software. It also reduces the size of the statistical models whilst maintaining a high performance.

Systran's most recent products allow users to customize fully their translation tasks (including terminology creation and management, domain adaption of the analysis, and building dedicated translation models and monolingual domain language models). The major products Systran offers today are the Systran Training Server and SystranLinks.

Systran Training Server

Systran Training Server with its two major components, the Corpus Manager, used to upload translation memories or other kinds of textual resources, and the Training Manager that makes use of this data, for:

- bilingual terminology extraction
- creation of hybrid MT systems based on statistical post-edition
- creation of statistical MT systems.

Systran Translation Server performs translations and comes with a set of plug-ins for various applications, administration and user tools, such as the Dictionary Manager, Translation Project Manager, and Document Aligner.

SystranLinks

SystranLinks, a website-translation solution that offer an innovative online CMS platform to launch and manage localization projects. It reproduces, translates and hosts the new sites. Users can leverage their translation memories and create their own translation resources.

Systran is used by its customers mostly in the domains of multilingual communication, defence and security, and technical documentation and localization.

What is remarkable is the terminology control with 20 specialized domains spread over four dictionaries:

- Economics/Business, Legal, Political Sciences, Colloquial, Automotive, Aviation/Space, Military Science
- Naval/Maritime, Metallurgy, Life Sciences, Earth Sciences, Medicine, Food Science
- Computers/Data Processing, Electronics, Mathematics, Engineering, Optics
- Nuclear, Chemical

An important point is that their system and platform have been built and designed to support additional language pairs. There are currently 90 working language pairs.

LIMSI

We will now see different methodologies using mostly statistics with the LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur) with as principal actor Professor François Yvon.

LIMSI's main research interests focus on Statistical Machine Translation (SMT) systems, which have, in the past few years, witnessed rapid progress, owing for instance to the success of the Google and Microsoft online translation engines.

Statistical Machine Translation systems rely on the statistical analysis of large bilingual corpora to train stochastic models aiming at approximating the mapping between a source and a target language. In their simplest form, these models express probabilistic relationships between source and target word strings, as initially formulated in the famous IBM models in the early 1990s. More recently, these models have been extended to capture more complex representations (e.g. chunks, trees, or dependency structures) and probabilistic mappings between these representations. Such models are typically trained from parallel corpora containing examples of source texts aligned with their translation(s), where the alignment is typically defined at a sub-sentential level; this computationally intensive training step delivers statistical models encoding a partial knowledge about translational equivalences, taking the form of huge tables containing up to hundreds of millions of numerical parameters.

In this context, LIMSI is developing its research activities in several directions, with the overall goals (1) to ground the development of such systems (from alignment to training to tuning and

inference) on robust and well-understood statistical techniques, rather than on a set of heuristic procedures and *ad hoc* recipes; and (2) to understand better these very complex mechanisms and develop useful evaluation and diagnostic tools. This, LIMSI believes, requires study in order to complete the SMT development process, from alignment to training and tuning, and to develop and maintain their own MT platform, which is available as open source software.

Regarding alignment models, most recent work deals with the design and training of discriminative alignment techniques (Tomeh *et al.* 2011: 305–312) in order to improve both word alignment and phrase extraction.

LIMSI's decoder, N-code, belongs to the class of n-gram based systems. In a nutshell, these systems define the translation as a two step process, in which an input source sentence is first non-deterministically reordered yielding an input word lattice containing several possible reorderings. This lattice is then translated monotonically using a bilingual n-gram model; as in the more standard approach, hypotheses are scored using several probabilistic models, the weights of which are tuned with minimum error weight training. This approach has been extended in many ways (with gappy units or with factored translation models) and the resulting system is now released as open source software (Crego *et al.* 2011: 49–58). An important line of research is the integration of approaches based on *discriminative training techniques* as a replacement for the standard training procedure for N-code. A first accomplishment in this line of research is a SMT system where Conditional Random Fields (CRF) are used as translation models (Lavergne *et al.* 2011: 542–553), an engineering tour de force requiring training CFRs with very large output label sets and billions of descriptors. Another successful development along the same lines has been the use of Neural Network Translation Models, which have proven effective in compensating for the lack of sufficient parallel data (Le 2012: 39–48). This work extends that carried out on large-scale NN statistical language models in speech recognition.

LIMSI's activities are not restricted to these core modules of SMT systems; many other aspects are also investigated such as tuning, multi-source machine translation, extraction of parallel sentences from comparable corpora, etc.

All these innovations need to be evaluated and diagnosed, and significant efforts are devoted to the vexing issue of quality measurements of MT output. LIMSI's SMT systems have thus taken part in several international MT evaluation campaigns. This includes a yearly participation in the WMT evaluation series (2006–2011), where LIMSI has consistently been ranked amongst the top systems, especially when translating into French. LIMSI has also partaken in the 2009 NIST MT evaluation for the Arabic–English task, as well as in the 2010 and 2011 IWSLT evaluations. As an alternative to standard evaluation protocols relying on very crude automatic comparison between a human and an automatic translation, LIMSI is also developing evaluation measures focusing on restricted facets of translation quality, as well as (self) diagnosis tools based on the computation of *oracle* scores (Sokolov *et al.* 2012: 120–129).

Finally, LIMSI is involved in a number of national and international projects, with both academic and industrial partners.

LIUM

Another research laboratory using a statistics based methodology is the LIUM (Laboratoire d'informatique de l'Université du Maine).

The LIUM, which has been carrying out research in MT since 2007, is directed by Professor Holger Schwenk who since 2004 was amongst the first researchers in France to support the statistical approach. Statistical methods called phrase-based and hierarchical are used to translate

English, German, Arabic and Chinese. One of the strong points is the use of innovative machine-learning techniques, notably continuous space modelling (Schwenk 2010: 137–146, 2012: 1071–1080) and non-supervised machine learning (Lambert *et al.* 2011: 284–293). The most recent research (Schwenk 2012) is aimed at improving the generalization of systems which currently fail due to morphological variants of the same 'word'. The LIUM also carries much research concerning variously adapting and specializing models to particular domains, using comparable corpora (Abdul-Rauf and Schwenk 2011: 341–375) and corrective learning. As well as a purely statistical approach, several ideas are being explored in order to include linguistic knowledge.

This work has allowed the LIUM to develop impressive MT systems which are systematically classed amongst the best in numerous international evaluations, notably OpenMT in 2008, 2009 and 2012. The LIUM has also developed a speech translation system which was ranked first in the IWSLT international evaluation in 2011.

The LIUM participates in developing translation systems in the framework of the DARPA GALE and BOLT programmes. The goal of these programmes is to obtain major advances in the machine translation of Arabic and Chinese with emphasis on processing informal language and dialects. The LIUM participates in the EuromatrixPlus project and the ANR Instar project. The LIUM is coordinator of the ANR Cosmat project which aims at putting into place a collaborative translation service for scientific documents deposited on the French HAL multi-disciplinary open archive. This involves developing technologies adapted to scientific documents. The user will have the possibility of correcting the translations done by machine and these former will be used for improving the translation system. This idea is taken further in the European MateCat project where translation systems are being developed for aiding human translators in their daily work with a CAT translation tool capable of including all the user's corrections in real time. The LIUM is in charge of developing techniques concerning adapting to the document domain and even to the translator's style.

The LIUM collaborates with enterprises and with public organizations.

Xerox

Xerox Research Centre Europe's activities in machine translation are mainly concentrated around the following aspects.

Phrase-based Statistical Machine Translation, where Xerox has developed its own translation environment TRANSLAB; this environment is mainly used for applications internal to Xerox, in particular:

- Production of automatic translations in technical domains (printers, automotive industry, etc.) for consumption by the CDLS (Content Development and Language Services) branch of Xerox, which produces high-quality human translations for internal and external clients. The automatic translations that are produced are post-edited by human translators, to complement the use of Translation Memories, in order to improve the efficiency of the whole translation process. In the majority of cases, this type of work is from English to various European languages.
- Translations of documents which are relevant to customer-relationship management. Xerox Services, today a major branch of Xerox, amongst other activities handles the outsourcing of services for many other companies, in such diverse domains as human resources, financial accounting and health. Machine Translation techniques allow English-speaking (say) Xerox agents to understand and to exploit certain of the documents

sent by end-customers who speak other languages and permit decoupling domain expertise from linguistic competencies.

Also in the context of customer relationships, it is sometimes necessary not only to understand foreign-language documents but also to respond to the end-customers in their own language. For this purpose, Machine Translation is typically not able to offer a sufficient guaranteed quality, and XRCE has thus developed a different approach, Multilingual Document Authoring, where an English-speaking agent (say) is guided into the interactive generation of both an English and also a foreign document. Whilst this approach does guarantee high semantic and syntactic quality of the produced documents, it requires careful design and development of the underlying representations which is cost-effective only when the discourse domains of the documents can be sufficiently circumscribed in advance.

XRCE has produced a number of publications around translation technologies, several of which have been influential, in particular in such domains as:

- Novel phrase-based SMT models (Simard *et al.* 2005: 755–762, Zaslavskiy *et al.* 2009: 333–341)
- Confidence estimation of translations (Specia *et al.* 2009: 28–35)
- Learning to predict the quality of a translation model (Kolachina *et al.* 2012: 22–30)
- Preserving the privacy of translation resources (Cancedda 2012: 23–27)
- Multilingual Document Authoring (Brun *et al.* 2000: 24–31)

Conclusion

In reality, what is interesting in France is that different technologies as well as quite different hybrid technologies are used giving prominence either to linguistics and mathematics or to computing. We see too the will in progressing to obtain better results together with the curiosity of researchers looking at other methodologies according to the domain and public addressed by their systems.

References

Abdul-Rauf Sadaf and Holger Schwenk (2011) 'Parallel Sentence Generation from Comparable Corpora for Improved SMT', *Machine Translation* 25(4): 341–375.

Blanchon, Hervé and Christian Boitet (2000) 'Speech Translation for French within the C-STAR II Consortium and Future Perspectives', in *Proceedings of ICSLP*, 16–20 October 2000, Beijing, China, 412–417.

Boitet, Christian, Hervé Blanchon, Mark Seligman, and Valérie Bellynck (2010) 'MT on and for the Web', in *Proceedings of IEEE NLP-KE '10*, 21–23 August 2010, Beijing, China, 2–12.

Brun, Caroline, Marc Dymetman, and Veronika Lux (2000) 'Document Structure and Multilingual Authoring', in *Proceedings of the 1st International Conference on Natural Language Generation, Association of Computational Linguistics (INLG '00)*, 12–16 June 2000, Mitzpe Ramon, Israel, 24–31.

Cancedda, Nicola (2012) 'Private Access to Phrase Tables for Statistical Machine Translation', in *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics*, 8-–4 July 2012, Jeju, Korea, 23–27.

Cardey, Sylviane (2013) *Modelling Language*, Amsterdam and Philadelphia: John Benjamins.

Cardey, Sylviane, Peter Greenfield, and Mi-Seon Hong (2003) 'The TACT Machine Translation System: Problems and Solutions for the Pair Korean – French', *Translation Quarterly* 27: 22–44.

Cardey, Sylviane, Peter Greenfield, and Wu Xiaohong (2004) 'Designing a Controlled Language for the Machine Translation of Medical Protocols: The Case of English to Chinese', in *Proceedings of AMTA-2004: The 6th Conference of the Association for Machine Translation in the Americas*, 28 September – 2 October 2004, Georgetown University, Washington DC, USA, in Robert E. Frederking and Kathryn

B. Taylor (eds) *Machine Translation: From Real Users to Research*, Berlin, Heidelberg: Springer-Verlag, 37–47.

Cardey, Sylviane, Peter Greenfield, Raksi Anantalapochai, Mohand Beddar, Dilber Devitre, and Gan Jin (2008) 'Modelling of Multiple Target Machine Translation of Controlled Languages Based on Language Norms and Divergences', in B. Werner (ed.) *Proceedings of ISUC2008*, 15–16 December 2008, Osaka, Japan, 2008, IEEE Computer Society, 322–329.

Choumane, Ali, Hervé Blanchon, and Cécile Roisin (2005) 'Integrating Translation Services within a structured Editor', in *Proceedings of DocEng 2005 (ACM Symposium on Document Engineering)*, 2–4 November 2005, Bristol, UK, 165–167.

Crego, Josep M., José M. Mariño, and François Yvon (2011) 'N-code: An Open-source Bilingual N-gram SMT Toolkit', *Prague Bulletin of Mathematical Linguistics* 96: 49–58.

Delavenay, Emile (1959) *La machine à traduire*, Paris: Presses Universitaires de France.

Dugast, Loïc, Jean Senellart, and Philipp Koehn (2007) 'Statistical Post-editing on Systran's Rule-based Translation System', in *Proceedings of the 2nd Workshop on Statistical Machine Translation, ACL Workshop on Statistical Machine Translation 2007*, 23 June 2007, Prague, Czech Republic, 220–223.

Kolachina, Prasanth, Nicola Cancedda, Marc Dymetman, and Sriram Venkatapathy (2012) 'Prediction of Learning Curves in Machine Translation', in *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics*, 8–14 July 2012, Jeju, Korea, 22–30.

Lambert, Patrik, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf (2011) 'Investigations on Translation Model Adaptation Using Monolingual Data', in *Proceedings of the 6th Workshop on Statistical Machine Translation*, 30–31 July 2011, Edinburgh, Scotland, UK, 284–293.

Lavergne, Thomas, Alexandre Allauzen, Josep Maria Crego, and François Yvon (2011) 'From N-gram-based to CRF-based Translation Models', in *Proceedings of the 6th Workshop on Statistical Machine Translation*, 30–31 July 2011, Edinburgh, Scotland, UK, 542–553.

Le, Hai Son, Alexandre Allauzen, and François Yvon (2012) 'Continuous Space Translation Models with Neural Networks', in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3–8 June 2012, Montréal, Canada, 39–48.

Léon, Jacqueline (1998) 'Les débuts de la traduction automatique en France (1959–1968): à contretemps ?' *Modèles linguistiques tome XIX*, fascicule 2, 55–86.

Potet, Marion, Laurent Besacier, Hervé Blanchon, and Marwen Azouzi (2012) 'Towards a Better Understanding of Statistical Post-editing Usefulness,' in *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, 6–7 December 2012, Hong Kong, China.

Schwenk, Holger (2010) 'Continuous Space Language Models For Statistical Machine Translation', *The Prague Bulletin of Mathematical Linguistics* 93: 137–146.

Schwenk, Holger (2012) 'Continuous Space Translation Models for Phrase-based Statistical Machine Translation', in *Proceedings of COLING 2012: Posters*, December 2012, Mumbai, India, 1071–1080.

Simard, Michel, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser (2005) 'Translating with Non-contiguous Phrases', in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics (HLT '05)*, October 2005, Vancouver, Canada, 755–762.

Sokolov, Artem, Guillaume Wisniewski, and François Yvon (2012) 'Computing Lattice BLEU Oracle Scores for Machine Translation', in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 23–27 April 2012, Avignon, France, 120–129.

Specia, Lucia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini (2009) 'Estimating the Sentence-level Quality of Machine Translation Systems', in *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, May 2009, Barcelona, Spain, 28–35.

Tomeh, Nadi, Alexandre Allauzen, and François Yvon (2011) 'Discriminative Weighted Alignment Matrices for Statistical Machine Translation', in Mikel Forcada and Heidi Depraetere (eds) *Proceedings of the European Conference on Machine Translation*, Leuven, Belgium, 305–312.

Vauquois, Bernard (1975) *La traduction automatique à Grenoble*, Paris: Dunod.

Zaslavskiy, Mikhail, Marc Dymetman, and Nicola Cancedda (2009) 'Phrase-based Statistical Machine Translation as a Traveling Salesman Problem', in *Proceedings of ACL/IJCMLP 2009: Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP: Proceedings of the Conference*, 2-7 August 2009, Suntec, Singapore, 333–341.

17

TRANSLATION TECHNOLOGY IN HONG KONG

Chan Sin-wai and Ian Castor Chow

THE CHINESE UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Billy Wong Tak-ming

THE OPEN UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Introduction

Translation technology includes typically machine translation (MT), computer-aided translation (CAT), and a wide array of supporting tools and computational techniques to facilitate translation. It has been deemed to be the solution to the ever-increasing amount of information waiting to be translated between languages, especially in recent decades as advances of information technologies have significantly reduced the cost of information dissemination and promoted international communication.

Hong Kong is a multicultural region in the world where inter-language interaction is prevalent, leading to a great demand for translation and a huge volume of available texts in different language versions. As a multilingual city, Chinese and English are the official languages in Hong Kong, and a number of Asian languages are spoken by various communities.[1] Furthermore, the two character sets of written Chinese in use, i.e., traditional and simplified, are different not only in graphemes of characters but also preferred vocabularies and syntactic structures due to the regional and cultural differences between Hong Kong and mainland China. Nowadays the provision of governmental and commercial documents online in traditional/simplified Chinese and English bilingual versions is essential, such as the Bilingual Laws Information System on the website of the Department of Justice[2] which contains a comprehensive collection of the laws of Hong Kong. In some cases multilingual versions covering a number of languages are also available.[3] They provide a huge number of multilingual texts serving as a valuable resource of translation technology, for instance, for developers of MT to perform system training, or users of CAT to build up translation memory.

A typical genre of translation in Hong Kong is practical writing, covering various fields such as governmental, legal and business. It is characterized by the use of domain-specific terminology, and highly repetitive wording and sentence patterns resulting from the requirement of standardized written style. Texts having these features are suitable to be processed by CAT tools like termbanks and translation memory systems to store standard and repeated entries for future reuse for the sake of consistency and efficiency.

The translation market in Hong Kong poses other challenges for translators, which prepares an appropriate environment for the growth of translation technology. As pointed out in Au

(2001: 185–192), for instance, financial documents are commonly either prepared in a hurry or revised frequently with stringent deadlines, leaving translators very limited time to work on and deal with the different versions while maintaining the translation quality, not to mention the linguistic challenges of specific terminology and syntactic structures. This is where technologies can be utilized to take over the routine and mechanical tasks for which they are designed.

This chapter describes the development of translation technology in Hong Kong. We use the term 'translation technology' to refer to all kinds of language technology which aid translation directly and indirectly, including, in addition to MT and CAT, those for Chinese language processing such as traditional-simplified Chinese translation and word segmentation, and those for corpus construction, etc. So far their development mainly focuses on research and teaching in tertiary institutes.[4] The former covers a wide range of topics and is conducted by scholars from three main disciplines including translation, computational linguistics and computer science and engineering. The latter includes programmes and courses offered in various universities. In addition, there are a few, though limited, number of applications worth mentioning.

Research and academic activities

The research and academic activities related to translation technology have been actively carried out in Hong Kong. Most of the universities have different kinds of ongoing or completed works in this area, including the establishment of research centres, the organization of international conferences, publication of academic journals, encyclopedias, and a large number of research outputs.

The Chinese University of Hong Kong (CUHK)

The Hung On-To Research Laboratory for Machine Translation (1960s–1970s), Machine Translation Laboratory (1999) and Centre for Translation Technology (2006), Department of Translation

The Chinese University of Hong Kong is the first tertiary institution to conduct research into translation technology. As early as in 1969, an MT system, *The Chinese University Language Translator (CULT)* (Loh 1972) was developed by the Hung On-To Research Laboratory for Machine Translation. According to Chan (2001: 205–218), translation output of CULT was found to be satisfactory for Chinese-to-English translation of scientific writings given some means of pre-editing techniques. CULT was later redesigned as an interactive online MT system with the construction of a new Chinese keyboard.

The Machine Translation Laboratory (MTL) was set up by the Department of Translation in 1999. It has five goals to achieve:

1 to serve as a centre for the collection of computer-related materials;
2 to serve as a centre for the study of the application and analysis of the existing software available in the market;
3 to build up a communication network of MT centres throughout the world and of active researchers in the field;
4 to propose some interdepartmental, interfaculty, or even intercollegiate projects that will contribute significantly to scholarship in the field or meet the needs of the local community; and
5 to build up terminological databases for various subjects or professions that will help to achieve standardization in the translation of specialized vocabularies (Chan 2001: 205–218).

An MT system, *TransRecipe* (Chan 2002a: 3–22) was developed by the MTL for translating Chinese cookbooks into English. It combines into its system design the corpus-based, example-based, pattern-based, and rule-based approaches. More importantly, it adopts a 'translational approach' to have translation methods coded into the system, in addition to linguistic and computational concepts, such that human translators can contribute their expertise in the process of MT development.

The Centre for Translation Technology (CTT) was established by the same department in 2006, for carrying out research into translation technology, making practical translation tools to serve the industry through collaboration with translation companies and sister institutions, and promoting the use of translation technology in society. In particular, the goals of CTT include:

1 the building of domain-specific translation corpora;
2 the construction of a database of works on computer-aided translation;
3 the creation of a software library; and
4 the organization of seminars on translation technology.

(Chan 2006: 12)

Besides, CTT also supports the teaching of CAT Programme offered by the Department, in terms of creating supporting resources to teaching, which include:

1 CAT literature archive – including monographs, anthologies, conference proceedings, academic journals, and electronic magazines, published since 1984, which are classified into categories and reposited in easy-to-read electronic format;
2 CAT system user manual archive – including manuals of more than 200 MT and CAT systems, which can be accessed and searched easily;
3 CAT system operation video archive – including a series of videos to demonstrate the operation of the various systems; and
4 CAT project archive – including the CAT projects conducted by the students of the MACAT programme.

(Chan 2008b: 2)

The Department of Translation also publishes different kinds of scholarly works related to translation technology. The *CAT Bulletin* is published regularly to facilitate dissemination of information about the MACAT programme to targeted readers. It includes information on the most updated programme structure, course contents, staff profiles, academic activities, public seminars, staff publications, and research findings. It also provides information on the new advances in CAT delivered through conference proceedings, seminar speeches, and students' translation projects.

The *Journal of Translation Studies* is a peer-reviewed international journal dedicated to the publication of research papers in all areas of translation. A special issue (Volume 13, Numbers 1 and 2, 2010) on the teaching of CAT is published, covering the practical experience, systems and facilities, curriculum and course design, and the future of CAT teaching.

A forthcoming *Journal of Translation Technology* will be the first international journal of this kind in Hong Kong. It will serve to promote the scholarly study of translation technology, publish academic articles on the history, theory, and practice of the discipline, and review articles of books on the field.

Other publications include a book *Translation and Information Technology* (Chan 2002b) which brings together experts from different disciplines to discuss how new technologies work on translator education and translation practice, as well as the conceptual gaps raised by the interface of human and machine. *A Dictionary of Translation Technology* (Chan 2004) is the first

dictionary in the field covering in total 1,375 entries which serves as a comprehensive reference for general readers as well as specialists in translation technology. *A Topical Bibliography of Computer(-aided) Translation* (Chan 2008a) provides a wide variety of information on the literature in the field, i.e., 8,363 entries of works written either in English or Chinese by 5,404 authors between 1948 and 2006, in the form of journal articles, conference papers, book chapters, project reports, software reviews, and newsletter features on or about documentary and speech MT/CAT. There are also conference presentations such as Chow (2012) which discusses how Web 2.0 and hybridity of MT-CAT change the design and use of translation tools, and Siu (2012) which illustrates an automatic pre- and post-editing approach to MT.

In the past decade a series of conferences oriented to translation technology were hosted by the Department, including:

- 2012 – The 10th Anniversary Conference of the MA in Computer-aided Translation Programme: New Trends in Translation Technology;
- 2009 – International Conference: The Teaching of Computer-aided Translation;
- 2006 – International Conference: Computer-aided Translation: Theory and Practice;
- 2004 – International conference: Translation Software – The State of the Art; and
- 2000 – International Conference: Translation and Information Technology.

These conferences drew together scholars, translation practitioners and software developers from different countries to exchange their knowledge, experiments and visions of various themes of translation technology.

The Human-Computer Communications Laboratory (1999) and the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies (2005), Department of Systems Engineering and Engineering Management

The Human-Computer Communications Laboratory (HCCL)[5] was established in 1999 with a mission to 'foster interdisciplinary research and education in human-centric information systems'. It supports research areas including but not limited to: speech recognition, spoken language understanding, speech generation, conversational systems development, audio information processing, multimodal and multimedia interface development, intelligent agents, mobile computing, and computer-supported collaborative work. Some representative research works related to translation technology conducted by HCCL include Chinese–English MT based on semi-automatic induced grammars (Siu and Meng 2001: 2749–2752; Siu *et al.* 2003: 2801–2804), and translingual speech retrieval (Lo *et al.* 2001: 1303–1306; Meng *et al.* 2004: 163–179).

The Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies[6] was established in 2005. This laboratory was recognized as a Ministry of Education of China (MoE) Key Laboratory in 2008. It conducts basis research and technology development in five strategic areas: (1) computer vision, (2) computer graphics, (3) speech processing and multimodal human-computer interaction, (4) multimedia processing and retrieval, and (5) wireless communications and networking. A piece of research on named entity translation matching and learning was conducted (Lam *et al.* 2007: 2).

The Hong Kong University of Science and Technology (HKUST)

Human Language Technology Center

The Hong Kong University of Science and Technology is the only university in Hong Kong that does not offer a translation programme. However considerable research on language and

speech technology has been conducted in the Department of Computer Science and Engineering and Department of Electronic and Computer Engineering. Led by the faculty members of both departments, the Human Language Technology Center (HLTC)[7] was founded in the 1990s, specializing in research on speech and signal processing, statistical and corpus-based natural language processing, machine translation, text mining, information extraction, Chinese language processing, knowledge management, and related fields. A number of systems have been built at HLTC, including automated language translation for the Internet, speech-based web browsing, and speech recognition for the telephone.

Two of the HLTC members, Professors Dekai Wu and Pascale Fung, have been extensively involved in the research into translation technology. Professor Wu is renowned for his significant contributions to MT, especially the development of inversion transduction grammar (Wu 1995: 69–82, 1996: 152–158, 1997: 377–404; Wu and Wong 1998: 1408–1414), a syntactically motivated algorithm for producing word-level alignments of parallel sentences, which pioneered the integration of syntactic and semantic models into statistical MT. Some of his recent representative works include semantic-based statistical MT (Wu and Fung 2009: 13–16; Wu *et al.* 2010: 236–252), and an automatic MT evaluation metric based on semantic frames named MEANT (Lo and Wu 2011: 220–229; Lo *et al.* 2012: 243–252).

Professor Fung is the founding Director of InterACT[8] at HKUST, a joint research and education centre between the computer science departments of eight leading universities in this field worldwide. One of their projects is EU-BRIDGE,[9] aiming at 'developing automatic transcription and translation technology that will permit the development of innovative multimedia captioning and translation services of audiovisual documents between European and non-European languages'. Fung also conducted researches in Chinese MT (Fung 2010: 425–454), translation disambiguation (Cheung and Fung 2005: 251–273), speech MT (Fung *et al.* 2004), term translation (Fung and McKeown 1996: 53–87, 1997: 192–202), etc.

City University of Hong Kong (CityU)

Department of Chinese, Translation and Linguistics

The Department of Chinese, Translation and Linguistics at The City University of Hong Kong is one of the university departments in Hong Kong most actively involved in the research and teaching of translation technology. A number of funded projects have been conducted, covering various aspects of translation technology, such as the following ones:

EXAMPLE-BASED MACHINE TRANSLATION (EBMT) FOR LEGAL TEXTS
(PI: JONATHAN J. WEBSTER)

> 'This project applies the "example-based" approach to the translation of the specialized language of legislation and legal documents … Research into the application of the example-based approach will be based on an aligned parallel corpus representing the work of top professionals in legal translation …'

A PILOT STUDY OF LEARNING FROM EXAMPLES TO TRANSLATE BETTER (PI: CHUNYU KIT)

> 'This project will explore advanced technologies and practical methodology to implement an online machine translation (MT) system that can learn to translate better. By providing an online translation service with a bilingual editor for manual

> post-editing, the system acquires translation knowledge from translators to enrich its example base and language models … A unique feature of this system is that it adapts its translation towards translators' expertise via learning …'

CONSTRUCTION OF AN ON-LINE PLATFORM FOR COMPUTER-AIDED TEACHING AND LEARNING OF BILINGUAL WRITING AND TRANSLATING IN/BETWEEN ENGLISH AND CHINESE (PI: CHUNSHEN ZHU)

> '[T]his project proposes to build an electronic platform for on-line teaching/(self-) learning of bilingual writing and translation in/between English and Chinese (traditional and simplified) … The products … help alleviate the pressure on the teaching of labor-intensive courses of translation and (bilingual) writing/editing …'

The research of CityU in translation technology is also fruitful. For example, Kit *et al.* (2002: 57–78) critically review the major stages of example-based MT and present a lexical-based text alignment approach for example acquisition. Kit *et al.* (2005: 71–78) illustrate how English–Chinese parallel texts of the laws of Hong Kong can be harvested on the Web and aligned at the subparagraph level. Song *et al.* (2009: 57–60, 2010: 62–65) propose a new method for transliteration of name entities based on statistical MT technology. Kit and Wong (2008: 299–321), Wong and Kit (2008, 2009a: 337–344, 2009b: 141–155, 2010: 360–364, 2011: 537–544, 2012: 1060–1068) and Wong *et al.* (2011: 238–242) conduct a series of works on developing an automatic metric, namely ATEC, for MT evaluation, and how automatic evaluation metrics can be used to aid MT users to opt for a proper MT system. Seneff *et al.* (2006: 213–222) present techniques to combine an interlingua translation framework with statistical MT. Zhu (2005, 2007a, 2007b) creates a computer platform, ClinkNotes, for assisting translation teaching.

The Hong Kong Institute of Education (HKIEd)

Research Centre on Linguistics and Language Information Sciences

The Research Centre on Linguistics and Language Information Sciences (RCLIS),[10] which was founded in 2010, aims to 'foster interdisciplinary research in the diverse areas of linguistics, natural language processing and information science'. It also provides a forum for scholars from the same or different institutes to work on problems of language and information technology in Chinese speech communities. The research of RCLIS is focused on '(1) the structures, as well as encoding and decoding, of information in the context of natural language, through which human beings acquire, and (2) mak[ing] use of knowledge, and… computational techniques to study and simulate the processes involved'.

During the years, RCLIS has conducted a number of projects funded by different agencies, such as the Research Grant Council of HKSAR, the Commerce and Economic Development Bureau's Innovation Technology Fund, the Judiciary, as well as private funding sources. Those projects related to translation technology include:

- *Parallel Classical-Colloquial Chinese Alignment Processing and Retrieval Platform*
- *A Computational Lexicon Based on Intrinsic Nature of Word Senses: A Bilingual Approach*
- *BASIS (project on Chinese Lexical Data Mining) Part of Speech POS Tagging to Simplified and Traditional Chinese Texts*

- *Bilingual Reference System*
- *Chinese Semantic Networks and Construction of a (Pan) Chinese Thesaurus*

The research outputs of RCLIS range from computer systems and language resource to publications. For instance, an online platform ACKER[11] was developed for Chinese language teaching and learning. It aligns classical Chinese texts with their modern Chinese counterparts, and provides a search engine that enables bi-directional retrieval of the processed texts. Besides, a Chinese–English parallel corpus of patent documents was constructed and used as a benchmark for the international competition on Chinese–English MT jointly organized by RCLIS and the National Institute of Information and Communication Technology (NICT) in Tokyo in 2010. Other relevant researches include anaphora resolution for MT (Chan and Tsou 1999: 163–190), and MT between Chinese and other languages (Lu *et al.* 2011: 472–479; Tsou 2007a, 2007b; Tsou and Lu 2011).

The Hong Kong Polytechnic University (PolyU)

Research into translation technology is conducted in various departments at The Hong Kong Polytechnic University, including the Department of Chinese and Bilingual Studies, Department of Computing, and Department of Electronic and Information Engineering. The related projects include:

- Cantonese-Putonghua Inter-dialect Machine Translation and its Integration with the World Wide Web;
- Evolving Artificial Neural Networks to Measure Chinese Sentence Similarity for Example-Based Chinese-to-English Machine Translation; and
- Building up a Computerized Mechanism for General Translation Business.

The relevant research outputs cover a number of topics. Lau and Zhang (1997: 379–382), Zhang (1998: 499–506, 1999: 40–47), Zhang and Lau (1996: 419–429) and Zhang and Shi (1997: 225–231) conduct a series of works to explore the design of an inter-dialect MT system particularly for the Cantonese–Mandarin dialect pair, which involves two dialects widely used in Chinese speech communities but with considerable differences in pronunciation, vocabulary and syntactic rules. Liu and Zhou (1998: 1201–1206), Wu and Liu (1999: 481–486) and Zhou and Liu (1997a: 65–72, 1997b: 520–525) present a hybrid MT model integrating rules and automatically acquired translation examples, and a resulting system prototype PolyU-MT-99 designed for Cantonese–English translation. Wang *et al.* (2004: 97–102) proposes a rule-based method for English-to-Chinese MT. Zhang (2005: 241–245) illustrates the design of what he calls ABCD concordancer and discuss its application to translation.

The University of Hong Kong (HKU)

The Pattern Recognition and Computer Vision Laboratory, Department of Computer Science

The Pattern Recognition and Computer Vision Laboratory[12] at the Department of Computer Science has several focused areas including Chinese computing and MT. It has an ongoing project on MT (i.e., *Marker Identification in Example-Based Machine Translation*) and is developing an English–Chinese MT system using a combination of knowledge- and example-based approaches.

Teaching

The teaching of translation technologies in Hong Kong is delivered within the disciplines of computer science and engineering, as well as arts and social sciences. The former focuses on algorithmic, programming, modeling and software engineering aspects, while the latter more on translation practices utilizing MT/CAT systems, and the general understanding of computational processing of languages.

Courses in translation technologies are usually elective ones, particularly in the curriculum at the undergraduate level. Surveying the translation technology courses offered at the higher institutions in Hong Kong, the teaching of this subject can be categorized into two modes – *specific* and *overview*. The *specific* mode takes translation technology as the core subject matter. It covers more in-depth issues concerning computer and translation and usually provides tutorials or workshops of translation tools application. Hands-on experience of translation tools and a deeper academic exploration of the subject can be anticipated. The *overview* mode takes a general approach introducing different advanced strategies in the use of computer technology to tackle linguistic issues, in which translation technology is only one of the selected topics of the entire course. This type of general language technology course usually introduces the basic concepts and background of translation technology but actual practice of translation tools may not be included.

The following outlines the overall situation of the translation programmes and translation technology courses in the universities and higher institutions in Hong Kong, which is also summarized in Table 17.1.

At present there are 18 degree-awarding universities and higher education institutions in Hong Kong.[13] Eight of them are government-funded, with six universities – CityU, CUHK, HKBU, HKU, LU and PolyU – offering bachelor degree programmes which major in translation, and four – CityU, CUHK, HKBU and PolyU – offering taught master degree programmes in translation. The other ten universities and institutions are either self-financing or publicly funded. For degrees majoring in translation, HSMC and OUHK offer a bachelor degree while OUHK also offers a taught master degree. At HKUST, language or translation technology related courses are offered from the computer science perspective by the Department of Computer Science and Engineering.

At undergraduate level, CityU, CUHK and HSMC include translation technology in their translation degree programmes. HKU and SYU introduce the subject through translation technology overview or language technology courses, which are respectively offered to students majoring in human language technology and English. HKUST teaches the subject from the computer programming perspective for students major in computer science.

At postgraduate level, four of the five taught master degree translation programmes include computer technology related courses in their curriculum. CityU, CUHK and PolyU offer translation technology specific courses and HKBU has a technology-related translation course. OUHK does not offer any course in this subject area at either undergraduate or postgraduate level. HKIEd offers a language technology course which overviews translation technology in a taught master degree programme. The postgraduate translation technology courses at HKUST are offered to research degree students at the Department of Computer Science and Engineering.

Besides the 18 universities and institutions, the post-secondary institute CUTW (see p. 306) offers a translation technology specific course in their associate degree programme in translation. Table 17.1 outlines the translation technology related courses at different universities and institutions in Hong Kong.

Table 17.1 Translation technology related courses at universities and higher institutions in Hong Kong

	Translation technology specific course		*Language technology or translation technology overview course*	
	Undergraduate	*Postgraduate*	*Undergraduate*	*Postgraduate*
CityU ⋆ ^	✓	✓	✓	✓
CUHK ⋆ ^	✓	✓	✓	✓
HKBU ⋆ ^				✓
HKIEd				✓
HKU ⋆			✓	✓
HKUST		✓	✓	
LU ⋆				
PolyU ⋆ ^		✓		
HSMC ⋆	✓			
OUHK ⋆ ^				
SYU			✓	
CUTW #	✓			

Notes:
 # Offering associate degree programme in translation
⋆ Offering bachelor degree programme in translation
^ Offering taught master degree programme in translation

There are currently two Master of Arts programmes majoring in translation technology. The Department of Translation at CUHK offers the Master of Arts in Computer-aided Translation, which was the first master's degree programme of this scope in the world. The Department of Chinese, Translation and Linguistics at CityU offers the Master of Arts in Language Studies, which provides four optional specializations including Translation with Language Information Technology. These two programmes specialize in close attention to subject classification concerning different aspects of translation technologies, with the main focus of each individual course ranging from theoretical to practical issues.

The following sections enumerate the courses of translation technologies at different universities and higher institutions in Hong Kong.

Department of Translation, The Chinese University of Hong Kong

The Department of Translation (TRAN)[14] at The Chinese University of Hong Kong (CUHK) is the pioneer of translation technology. It offered the world's first MA programme in Computer-aided Translation (MACAT)[15] in 2002. The programme aims to 'deepen students' understanding of the workings of language as an essential tool of communication and equip them with the knowledge of translation technology'. A number of translation software systems are available at the department computer lab and student accessible remote server for the teaching and learning of the practical use of computer(-aided) translation software, including:

SDL Trados Studio	SDL Trados MultiTerm	SDL Passolo
Déjà Vu	WordFast Pro	memoQ
Systran	Otek Transwhiz	Dr Eye
Logo Media Translate	YaXin CAT	Xueren CAT
XTM Cloud	PROMT	Kingsoft Fast AIT

At undergraduate level, the department offers two elective courses for BA students major or minor in translation.

TRAN2610 Introduction to Computer-aided Translation
TRAN3620 Machine Translation

In addition to essential topics in CAT and MT, TRAN2610 offers practical training in translation software tools and TRAN3620 includes approaches and evaluation MT systems.

At postgraduate level, there are six courses related to translation technology. Two of them are required courses for MACAT students. Students also have to choose at least one from the three machine translation related elective courses. The courses are also open to enrollment, as an elective course, for students from the Master of Arts in Translation, another postgraduate programme offered by the same department.

Postgraduate translation technology courses at the Department of TRAN, CUHK:

REQUIRED COURSES

TRAN6601 Introduction to Computer-aided Translation
TRAN6602 Editing Skills for Computer Translation

ELECTIVE COURSES (COMPUTER TRANSLATION)

TRAN6821 Computer Translation
TRAN6822 Natural Language Processing
TRAN6823 Terminology Management

ELECTIVE COURSES (TRANSLATION PRACTICE)

TRAN6812 Practical Translation Workshop

The two required courses cover the underlying rationale, basic principles and other essential concepts in translation technologies. Practical training in human–technology interactions are also highlighted in the two courses. TRAN6601 focuses on the use of CAT software tools, TRAN6602 on specialized editing skills for coping with different MT systems. The elective courses focus on different areas of translation technology. In TRAN6812, for example, students have to complete different types of translation projects assigned by course lecturer. The translation project on software localization and content localization was first introduced to the course as a partial fulfillment requirement in the spring term in 2013, being the first practical software localization course in Hong Kong.

The Department of Chinese, Translation and Linguistics, City University of Hong Kong

The Department of Chinese, Translation and Linguistics (CTL)[16] at City University of Hong Kong (CityU) offers BA and MA programmes focused on language technology. The BALLT[17] programme (Bachelor of Arts in Linguistics and Language Technology), though not centrally focused on translation technology, aims to produce language professionals who are familiar with language-related application of computers. A number of language technology and computational linguistics courses are offered for students major or minor to this profession.

There are three core language technology related courses in the BALLT curriculum. Students majoring this programme have to complete all three courses while students taking a minor in language technology are required to complete the first two courses.

Undergraduate language technology courses at the Department of CTL, CityU:

REQUIRED COURSES

CTL2231 Introduction to Language Technology
CTL3210 Electronic Publishing
CTL3233 Computational Linguistics

The department offers a wide range of elective courses with orientations towards language technology, applied linguistics or language studies. Elective courses specialized in language technology and computational linguistics are listed below:

ELECTIVE COURSES

CTL2206 Fundamentals of Mathematics and Statistics for Language Studies
CTL3219 Document Processing and Publishing
CTL3220 Corpus Linguistics
CTL3222 Machine Translation
CTL3224 Computational Lexicography
CTL3228 Chinese Computing
CTL3232 Computer Programming for Language Studies
CTL4218 Advanced Topics in Computational Linguistics
CTL4221 Natural Language Parsing
CTL4225 Computer Assisted Language Learning
CTL4234 Linguistic Computing
CTL4237 Introduction to Web-oriented Content Management

These courses involve the use of computer technology in different language issues. Courses such as Corpus Linguistics, Computational Lexicography, Chinese Computing and Natural Language Parsing provide fundamental training and background concepts in translation technology. The translation technology specialized course *CTL3222* Machine Translation is not offered to translation students but restricted to students pursuing linguistics and language technology. The same is the case for the Department of Translation at CUHK, and the Department of CTL at CityU offers a CAT elective course, *CTL3354* Introduction to Computer-Aided Translation for students major or minor in translation.

At postgraduate level, the Department of CTL offers a Master of Arts in Language Studies (MALS)[18] programme which provides four optional specializations for their students: Language and Law, Linguistics, Translation and Interpretation, and Translation with Language Information Technology. The MALS with specialization Translation with Language Information Technology (TLIT) was started in 2010. Together with the MACAT programme at CUHK, they are the only two MA programmes in Hong Kong which focus on translation technologies. The department has several computer(-aided) translation systems for the teaching and learning of translation technology, including

Déjà Vu	Systran
Otek Transwhiz	Dr Eye

Postgraduate translation technology courses at the Department of CTL, CityU:

REQUIRED COURSES (FOR THE TLIT SPECIALIZATION)

CTL5411 Computational Linguistics
CTL5620 Translation Technology
CTL5628 Human–Machine Interactive Translation

ELECTIVE COURSES

CTL5629 Translation Tools Development
CTL5631 Corpora and Translation
CTL5632 History of Machine Translation

The required courses are compulsory to students who take the specialization in TLIT while the elective courses are open to students of all specializations of the MALS programme.

School of Translation, Hang Seng Management College

The Bachelor of Translation with Business (BTB)[19] offered by the School of Translation at Hang Seng Management College (HSMC) is another undergraduate programme in Hong Kong which provides the teaching and learning of Computer-aided Translation. Different from the Department of Translation at CUHK and the Department of CTL at CityU, which offer one CAT and one MT course, the School of Translation offers two CAT courses. The two courses[20, 21] are designed as a series such that a deeper level of exploration and analysis are demanded in the second course.

TRA3105 Computer and Business Translation 1
TRA4104 Computer and Business Translation 2

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

The Department of Chinese and Bilingual Studies (CBS)[22] at The Hong Kong Polytechnic University (PolyU) offers translation programmes at both undergraduate and postgraduate levels. In contrast to the BA translation programmes at CUHK and CityU, there is no

computer(-aided) translation course offered to undergraduate students at PolyU. At postgraduate level, there are two translation technology specialized courses and a corpus linguistics course offered as electives to students of the Master of Arts in Translation and Interpretation (MATI):[23]

CBS517 Computer Tools for the Language Professionals
CBS560 Computer Assisted Translation
CBS580 Applied Corpus Linguistics

The course CBS517, as an elective course, is also offered to students of other Master of Arts programmes run by the department while CBS560 is restricted to MATI students only.

Department of Computer Science and Engineering, The Hong Kong University of Science and Technology

The teaching and learning of translation technology at the Hong Kong University of Science and Technology (HKUST) is offered by the Department of Computer Science and Engineering (CSE).[24] Their courses cover the topics of human language technology and machine translation, and are designed from the perspectives of programming and engineering. The teaching focus is hence quite different from the above-mentioned cases at other universities, which are designed for students without a computer science background and emphasize the strategic utilization of translation systems in translation practices. There are one undergraduate course and four postgraduate courses in the scope of computational linguistics at the Department of CSE. Courses identified from their department website[25, 26, 27, 28] are listed below:

UNDERGRADUATE COURSE

COMP4221 Introduction to Natural Language Processing

POSTGRADUATE COURSES (RESEARCH DEGREE)

COMP5221 Natural Language Processing
COMP621M Advanced Topics in AI: Structural Statistical Machine Translation
COMP621J Advanced Topics in AI: Statistical Machine Translation
COMP621F Advanced Topics I AI: Speech Recognition: Theory and Applications

Department of Linguistics, The University of Hong Kong

At the University of Hong Kong (HKU), no translation technology course is included in the curriculum of the translation programme[29] offered by the School of Chinese. The education of the interdisciplinary subject Language and Technology is found at the Department of Linguistics, which offers a major in Human Language Technology (HLT)[30] at undergraduate level. The Bachelor of Arts in Human Language Technology

> explores the theoretical and practical issues surrounding the ability to get technology, especially modern information communications technology (ICT), to interact with humans using natural language capabilities … [and] investigates how technologies, especially ICTs, can serve as useful adjuncts to humans in language understanding, including analysis, processing, storage and retrieval.

Students majoring in HLT may opt for taking designated courses from the Computer Science stream. The following are language technology related courses offered by the Department of Linguistics.

UNDERGRADUATE COURSES

LING1002 Lanugage.com
LING3101 Computational linguistics
LING3111 Language and literacy in the information age
LING3141 Language and information technology
LING3125 Corpus linguistics

POSTGRADUATE COURSE

LING6024 Computer-Assisted Language Learning (CALL)

Although the discipline Human Language Technology does not specialize in translation technology, two translation technology overview courses[31, 32] are identified. From the course description of *LING1002* Language.com, 'Some of the questions to address in this course include the following: Can computers and the internet do translations automatically and accurately? …' In *LING6024* Computer-Assisted Language Learning (CALL), the course covers topics including 'the use of E-dictionaries and thesauruses; and the use of corpus and concordancing program … Other related topics such as machine translation … will also be briefly introduced.'

Centre for Translation, Hong Kong Baptist University

The BA and MA Translation programmes offered by the Centre for Translation at the Hong Kong Baptist University[33] do not offer any translation technology course. At postgraduate level, there is an elective course focused on corpus and translation titled *Corpus-based Approach to Translation*.[34] According to the course description, this course is 'designed to introduce students to the application of corpora to the practice of and research on translation. It helps students to design, conduct research and report research findings using the corpora approach.'

Department of Linguistics and Modern Language Studies, The Hong Kong Institute of Education

The Hong Kong Institute of Education (HKIEd) does not offer any translation programmes, but the MA in Educational Linguistics and Communication Sciences[35] offers a course, *LIN6008* Computer Technology for Language Studies, which focuses on the use of the computer in language processing. According to the course information,[36] it covers a number of language technology topics including 'natural language processing applications such as machine translation … and evaluation of relevant software. They will also learn how to cultivate language corpus for linguistic analysis.'

The Department of English Language and Literature, Hong Kong Shue Yan University

The Hong Kong Shue Yan University (SYU) does not offer any translation degree programme, but the teaching and learning of translation is provided to BA students majoring English. The curriculum of BA in English,[37] offered by the Department of English Language and Literature, includes a number of translation courses. The role of technology in the translation industry, the concept of computer-aided translation and corpus-based translation are included as selected topics in two courses:

ENG440 Translation and Globalization
ENG487 Contemporary Translation Theory and its Applications

In *ENG440* Translation and Globalization, 'The role of modern technology and its influence on the translation industry will also be introduced'. In *ENG487* Contemporary Translation Theory and its Applications, concepts including Computer-aided translation and Corpus-based approaches to translation are covered in its course outline under the section 'Recent developments in translation studies'.

Others

According to the information from the course lists and descriptions of the academic programmes, there is no translation technology course offered by the Translation Department of Lingnan University (LU) or the School of Arts and Social Sciences of Open University of Hong Kong (OUHK).

In the scope of Associate Degree and Higher Diploma programmes offered by major institutions in Hong Kong, only one case of translation technology education is identified. It is the Associate of Arts Programme in Translation[38] offered by the School of Humanities and Social Sciences at The Chinese University of Hong Kong – Tung Wah Group of Hospitals Community College (CUTW). The curriculum includes a translation technology specific course,[39] *TR57003* Computers and Translation, offered as an elective course. The course covers important concepts in CAT, with emphasis on Chinese–English and English–Chinese translation. 'Topics to be discussed include language engineering, terminology management, translation memory systems, computerized term banks and translation software'.

Applications of translation technology

This section reviews the applications of translation technologies in Hong Kong in three areas: translation service providers, translation software companies, and translation technology practices.

Translation service providers

It is very difficult to survey the use of translation tools within the translation industry in Hong Kong. Translation companies may not explicitly state their use of translation tools in their company description or official website. Some of them, however, provide services such as software and website localizations or terminology extraction and standardization, implying that translation tools are utilized. Besides, as the option for translation tools may be determined by

translators, translation companies claiming the use of translation technology may not state explicitly which software is used. For the international language service providers, it is also arguable whether they should be categorized as part of the Hong Kong translation industry even if Hong Kong is one of their serving markets. The following companies are the few, if not all, translation service providers which are identified with explicit statement of the use of translation tools at their official website, and have a serving office or have registered in Hong Kong.

Chris Translation Service Company Limited[40] uses Trados as their CAT tool. According to their website, 'with 110 licenses for using Trados, a translation support tool, we respond adequately to translation and localization needs in the information and telecommunications technology industry that require translation memory control and a large volume of translation'. The company also uses Trados SDK to develop their own tools 'to check translations for breaches of style rules and a tag search tool'.[41]

KERN Global Language Services[42] uses a number of translation software systems and software localization tools including Across, Trados, Transit, DejaVu, APIS-FMEA, Visual Localize, Passolo and Catalyst. Besides translation and localization, their language services also include terminology database service with software-supported terminology extraction.

INTLINGO Global Language Solutions[43] uses SDL Trados, Passolo and Wordfast. Besides multimedia production and desktop publishing, their services also include software and website localizations.

CTC Translation and Localization Solutions Limited[44] has an extensive list of software tools in terminology management, CAT, QA and localization at their technology webpages. The software tools include SDL Trados, SDL Multiterm, IBM Translation Manager, Termstar, Wordfast, Language Weaver PET, Microsoft Localization Studio, Passolo, SDLX, Lingobit Localizer, etc.

BEAUHORSE Professional Translation Ltd does not state explicitly which CAT tool is used. It is noted on their website[45] that '[their] translation system is refined and backed up by a strong database and Translation Memory System built up over years and computer-aided translation software to ensure quality and a prompt turnaround time'.

TranslateMedia offers the STREAM[46] service which integrates translation memory and glossary management. The software tool used by the company is not made explicit.

Devon Financial Translation Services Limited[47] uses 'a combination of the latest translation technology, a unique project management system, and a multi-stage quality assurance scheme'. Their staff are trained in 'leveraging the latest in translation and localization technology'. The software tool used by the company is not made explicit.

Translation software companies

Two software developers in Hong Kong are identified for providing translation software.

The KanHan Technologies Limited[48] 'is an information technology solution provider targeting Hong Kong and China market'. They have developed the HanWeb software, for webpage translation between traditional and simplified Chinese.

Heartsome Technologies Ltd.[49] 'specialized in language translation technologies'. It is a company registered in Hong Kong with branches in South Korea, Singapore and China. It is also the only CAT developer with its corporate headquarters situated in Hong Kong. Their CAT products include Heartsome Translation Studio, Heartsome TMX Editor and Heartsome Dictionary Editor.

Translation technology practices

There are not many salient examples of translation projects in Hong Kong emphasizing the use of translation tools. The utilization and acceptability of translation tools in large-scale translation projects seems yet to attain an acknowledgeable status. The following are two examples identified where translation software systems were used significantly.

Traditional and Simplified Chinese Website Translation – HanWeb

With increasing collaboration between Hong Kong and mainland China, there is a growing demand for websites from Hong Kong providing both Traditional and Simplified Chinese versions. The demand is especially significant in the official information disseminated by the Government and other public sectors. The above-mentioned HanWeb Server[50] is a software application for webpage translation among Traditional Chinese, Simplified Chinese, Unicode and Cantonese Braille. HanWeb operates as a real time translation server which performs machine translation as well as the generation of resulting webpages. This software system is used by the HKSAR government and many NGOs, public utilities, institutions and companies.

Chinese Bible translation – The CSB translation project

The Chinese Standard Bible[51] (CSB) from the Asia Bible Society and Holman Bible Publisher is a recent translation project which highlights the application of computer technology in aiding the translation process. As stated in their website, 'a customized set of software tools was developed to create, revise and manage the translation at each stage. The revisions were aligned to the original language to facilitate cross-checking and consistency during the translation process – something never before done with a Bible translation.' Wu and Tan (2009) list a number of tree-based techniques supporting the CSB translation project:

- Tree alignment
- Tree-based translation memory
- Tree-based concordance
- Tree-based interlinear view
- Probabilistic Hebrew Synonym Finder
- Probabilistic Similar Verse Finder

Summary

Several decades have elapsed since the first MT system was invented in Hong Kong at The Chinese University of Hong Kong in 1969. What has been achieved so far in translation technology can be deemed substantial, at least in terms of research and teaching, notably starting from the late 1990s. A number of research centres were set up, a series of international conferences were organized, and plenty of research projects were funded and conducted. They have brought forth a wide variety of research outputs covering various aspects of translation technology, particularly in MT, including system development, approaches, evaluation, etc., and other specialized related areas such as CAT, terminology, lexicon and semantic network, parallel text retrieval and alignment.

Education in translation technology is well developed at both postgraduate and undergraduate level. The launch of the world's first master's degree programme majoring in CAT in 2002

significantly highlights this increasing academic pursuit and the demand for knowledge and practical skills in this profession. Translation technology has now become a typical elective course in the curriculum of translation programme in different tertiary institutes. Different related courses in computational linguistics, language technology and computer sciences and engineering have also prepared graduates with various specializations to support the research and development, and the use of translation technology.

Although the use of translation technology is not highly prevalent among translators according to available information, there are a number of translation companies employing MT and CAT in their production, and several software developers producing CAT tools adopted by commercial and governmental sectors. While more and more institute graduates have received professional training in translation technology, it is to be expected that what they have learnt will somehow be put into practice.

Translation technology is a multidisciplinary area involving translation, linguistics, computer science, information engineering, and human technology. Both research and teaching in this subject area may be expensive in terms of computer resources and academic staffing with multidisciplinary backgrounds. Interdisciplinary and intercollegiate collaborations should be encouraged, as we envisage, to foster knowledge exchange rather than isolated efforts, such that the duplication of expenses can be avoided and resources can be spent on the more important and significant issues in the realm of translation technology.

Notes

1 According to the Hong Kong 2011 Population Census Thematic Report: Ethnic Minorities (http://www.censtatd.gov.hk/hkstat/sub/sp170.jsp?productCode=B1120062), the population in Hong Kong other than Chinese constitutes 6.4% of the whole; within those 44.2% speak English at home, followed by Filipino (3.7%), Indonesian (3.6%), and Japanese (2.2%).

2 http://www.legislation.gov.hk.

3 For instance, other than traditional/simplified Chinese and English, the website of The Hong Kong Trade Development Council (http://www.hktdc.com) provides 11 language versions, including German, Spanish, French, Italian, Portuguese, Russian, Czech, Polish, Arabic, Japanese, and Korean; and the website of The Hong Kong Tourism Board (http://www.discoverhongkong.com) provides language versions of Dutch, French, German, Spanish, Russian, Arabic, Indonesian, Japanese, Korean, Thai, and Vietnamese. They represent the origins of major trade partners and visitors to Hong Kong respectively.

4 There are currently nine universities in Hong Kong, including, in alphabetical order, City University of Hong Kong (CityU), Hong Kong Baptist University (HKBU), Hong Kong She Yan University (HKSYU), Lingnan University (LU), Open University of Hong Kong (OUHK), The Chinese University of Hong Kong (CUHK), The Hong Kong Polytechnic University (PolyU), The Hong Kong University of Science and Technology (HKUST), and The University of Hong Kong (HKU). All of these are government-funded except HKSYU and OUHK which are self-financed. Besides, there are a number of institutes and colleges for those involved in research and teaching of translation technology include Hang Seng Management College (HSMC), The Hong Kong Institute of Education (HKIEd), and The Chinese University of Hong Kong – Tung Wah Group of Hospitals Community College (CUTW).

5 http://www.se.cuhk.edu.hk/facilities/lab_hcc.html.

6 http://sepc57.se.cuhk.edu.hk.

7 http://www.cs.ust.hk/~hltc.

8 http://interact.ira.uka.de.

9 http://www.eu-bridge.eu.

10 http://www.ied.edu.hk/rclis.

11 Aligned Chinese Knowledge Exchange Repository, at https://acker.chilin.hk/.

12 http://www.cs.hku.hk/research/pr.jsp.

13 Refer to footnote 4 for the names and abbreviations of the universities and institutions.

14 http://traserver.tra.cuhk.edu.hk/.
15 http://traserver.tra.cuhk.edu.hk/eng_programmes_macat.html.
16 http://ctl.cityu.edu.hk.
17 http://ctl.cityu.edu.hk/Programmes/334/Progs_Deg_BALLT.asp.
18 http://ctl.cityu.edu.hk/Programmes/Progs_ProgStruct_MALS.asp.
19 http://www.hsmc.edu.hk/en/academic_twb_pi.php.
20 http://www.hsmc.edu.hk/en/module_info_BTB.php?shortname=TRA3105.
21 http://www.hsmc.edu.hk/en/module_info_BTB.php?shortname=TRA4104.
22 http://www.cbs.polyu.edu.hk.
23 http://www.cbs.polyu.edu.hk/programmes/postgraduate-MATI.php.
24 http://www.cse.ust.hk.
25 http://www.cse.ust.hk/~dekai/4221.
26 http://www.cse.ust.hk/pg/courses/fall2005.html#621m.
27 http://www.cse.ust.hk/pg/courses/spring2004.html#621j.
28 http://www.cse.ust.hk/pg/courses/spring2002.html#621f.
29 http://web.chinese.hku.hk/handbook/2013_2014/handbook2013_2004.pdf.
30 http://www.linguistics.hku.hk/pro/major_hlt_2012_4yr.html.
31 http://www.linguistics.hku.hk/cou/fir/ling1002.html.
32 http://www.linguistics.hku.hk/cou/adv/ling6024.html.
33 http://www.tran.hkbu.edu.hk/.
34 http://www.tran.hkbu.edu.hk/EN/Taught_Postgraduate/Course_Description.asp.
35 http://www.ied.edu.hk/maelacs/.
36 http://www.ied.edu.hk/maelacs/view.php?secid=3140.
37 http://www.hksyu.edu/english/BA_in_English.html.
38 http://www.cutw.edu.hk/en/programme/ad/courses/hge/translation/info.
39 https://sss.cutw.edu.hk/cutw/students/download/ad_handbook_2012.pdf.
40 http://www.chris-translate.com/english/technology.
41 http://www.chris-translate.com/english/technology/support.html.
42 http://www.e-kern.com/en/translations/software-formats.html.
43 http://intlingo.com/technologies.
44 http://www.ctc-china.com/index.asp.
45 http://www.beauhorse.com/en/strengths.html.
46 http://www.translatemedia.com/stream-translation-workflow-technology.html.
47 http://www.devonhk.com/en/index.php.
48 http://www.kanhan.com/en/about-us.html.
49 http://www.heartsome.net/EN/home.html.
50 http://www.kanhan.com/en/products-services/hanweb-server.html.
51 http://www.chinesestandardbible.com/translation.html.

References

Au, Kim-lung Kenneth (2001) 'Translating for the Financial Market in Hong Kong', in Chan Sin-wai (ed.) *Translation in Hong Kong: Past, Present and Future*, Hong Kong: The Chinese University Press, 185–192.

Chan, Sin-wai (2001) 'Machine Translation in Hong Kong', in Chan Sin-wai (ed.) *Translation in Hong Kong: Past, Present and Future*, Hong Kong: The Chinese University Press, 205–218.

Chan, Sin-wai (2002a) 'The Making of TransRecipe: A Translational Approach to the Machine Translation of Chinese Cookbooks', in Chan Sin-wai (ed.) *Translation and Information Technology*, Hong Kong: The Chinese University Press, 3–22.

Chan, Sin-wai (ed.) (2002b) *Translation and Information Technology*, Hong Kong: The Chinese University Press.

Chan, Sin-wai (2004) *A Dictionary of Translation Technology*, Hong Kong: The Chinese University Press.

Chan, Sin-wai (2006) 'Centre for Translation Technology', *CAT Bulletin* 5: 12.

Chan, Sin-wai (2008a) *A Topical Bibliography of Computer(-aided) Translation*, Hong Kong: The Chinese University Press.

Chan, Sin-wai (2008b) 'CAT Teaching Resources at the Centre for Translation Technology', *CAT Bulletin* 9: 2.

Chan, Samuel and Benjamin Tsou (1999) 'Semantic Inference for Anaphora Resolution: Toward a Framework in Machine Translation', *Machine Translation* 14(3–4): 163–190.
Cheung, Percy and Pascale Fung (2005) 'Translation Disambiguation in Mixed Language Queries', *Machine Translation* 18(4): 251–273.
Chow, Ian Castor (2012) 'Technology Mashup for Translation – MT, CAT and Web 2.0: New Trends in Translation Tool and Website Localization', in *New Trends in Translation Technology: The 10th Anniversary Conference of the Master of Arts in Computer-aided Translation Programme*, Hong Kong: The Chinese University of Hong Kong.
Fung, Pascale (2010) 'Chinese Machine Translation', in Nitin Indurkhya and Fred J. Damerau (eds) *The Handbook of Natural Language Processing*, 2nd edition, Boca Raton, FL: Chapman and Hall/CRC Press, 425–454.
Fung, Pascale, Yi Liu, Yongsheng Yang, Yihai Shen, and Dekai Wu (2004) 'A Grammar-based Chinese to English Speech Translation System for Portable Devices', in *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju Island, Korea.
Fung, Pascale and Kathleen McKeown (1996) 'A Technical Word and Term Translation Aid Using Noisy Parallel Corpora across Language Groups', *Machine Translation*: *Special Issue on New Tools for Human Translators* 12(1–2): 53–87.
Fung, Pascale and Kathleen McKeown (1997) 'Finding Terminology Translations from Non-parallel Corpora', in *Proceedings of the 5th Annual Workshop on Very Large Corpora*, Hong Kong, 192–202.
http://ctl.cityu.edu.hk.
http://ctl.cityu.edu.hk/Programmes/334/Progs_Deg_BALLT.asp.
http://ctl.cityu.edu.hk/Programmes/Progs_ProgStruct_MALS.asp.
http://intlingo.com/technologies.
http://interact.ira.uka.de.
http://sepc57.se.cuhk.edu.hk.
http://traserver.tra.cuhk.edu.hk.
http://traserver.tra.cuhk.edu.hk/eng_programmes_macat.html.
http://web.chinese.hku.hk/handbook/2013_2014/handbook2013_2004.pdf.
http://www.beauhorse.com/en/strengths.html.
http://www.cbs.polyu.edu.hk.
http://www.cbs.polyu.edu.hk/programmes/postgraduate-MATI.php.
http://www.censtatd.gov.hk/hkstat/sub/sp170.jsp?productCode=B1120062.
http://www.chinesestandardbible.com/translation.html.
http://www.chris-translate.com/english/technology.
http://www.chris-translate.com/english/technology/support.html.
http://www.cs.ust.hk/~hltc.
http://www.cs.hku.hk/research/pr.jsp.
http://www.cse.ust.hk.
http://www.cse.ust.hk/~dekai/4221.
http://www.cse.ust.hk/pg/courses/spring2002.html#621f.
http://www.cse.ust.hk/pg/courses/spring2004.html#621j.
http://www.cse.ust.hk/pg/courses/fall2005.html#621m.
http://www.ctc-china.com/index.asp.
http://www.cutw.edu.hk/en/programme/ad/courses/hge/translation/info.
http://www.devonhk.com/en/index.php.
http://www.discoverhongkong.com.
http://www.e-kern.com/en/translations/software-formats.html.
http://www.eu-bridge.eu.
http://www.heartsome.net/EN/home.html.
http://www.hksyu.edu/english/BA_in_English.html.
http://www.hktdc.com.
http://www.hsmc.edu.hk/en/academic_twb_pi.php.
http://www.hsmc.edu.hk/en/module_info_BTB.php?shortname=TRA3105.
http://www.hsmc.edu.hk/en/module_info_BTB.php?shortname=TRA4104.
http://www.ied.edu.hk/rclis.
http://www.ied.edu.hk/maelacs.
http://www.ied.edu.hk/maelacs/view.php?secid=3140.
http://www.kanhan.com/en/about-us.html.

http://www.kanhan.com/en/products-services/hanweb-server.html.
http://www.legislation.gov.hk.
http://www.linguistics.hku.hk/cou/adv/ling6024.html.
http://www.linguistics.hku.hk/cou/fir/ling1002.html.
http://www.linguistics.hku.hk/pro/major_hlt_2012_4yr.html.
http://www.se.cuhk.edu.hk/facilities/lab_hcc.html.
http://www.tran.hkbu.edu.hk.
http://www.tran.hkbu.edu.hk/EN/Taught_Postgraduate/Course_Description.asp.
http://www.translatemedia.com/stream-translation-workflow-technology.html.
https://acker.chilin.hk.
https://sss.cutw.edu.hk/cutw/students/download/ad_handbook_2012.pdf.
Kit, Chunyu and Tak-Ming Wong (2008) 'Comparative Evaluation of Online Machine Translation Systems with Legal Texts', *Law Library Journal* 100(2): 299–321.
Kit, Chunyu, Haihua Pan, and Jonathan J. Webster (2002) 'Example-based Machine Translation: A New Paradigm', in Chan Sin-wai (ed.) *Translation and Information Technology*, Hong Kong: The Chinese University Press, 57–78.
Kit, Chunyu, Xiaoyue Liu, King Kui Sin, and Jonathan J. Webster (2005) 'Harvesting the Bitexts of the Laws of Hong Kong from the Web', in *Proceedings of the 5th Workshop on Asian Language Resources (ALR-05)*, 14 October 2005, Jeju Island, Korea, 71–78.
Lam, Wai, Shing-Kit Chan, and Ruizhang Huang (2007) 'Named Entity Translation Matching and Learning: With Application for Mining Unseen Translations', *ACM Transactions on Information Systems* 25(1): 2.
Lau, Chun-fat and Xiaoheng Zhang (1997) 'Grammatical Differences and Speech Machine Translation between Chinese Dialects', in *Proceedings of the 17th International Conference on Computer Processing of Oriental Languages (ICCPOL '97)*, 2–4 April 1997, Hong Kong Baptist University, Hong Kong, 379–382.
Liu, James N.K. and Lina Zhou (1998) 'A Hybrid Model for Chinese-English Machine Translation', in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC'98)*, 11–14 October 1998, San Diego, CA, 1201–1206.
Lo, Chi-kiu and Dekai Wu (2011) 'MEANT: An Inexpensive, High-accuracy, Semi-automatic Metric for Evaluating Translation Utility via Semantic Frames', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, 19–24 June 2011, Portland, OR, 220–229.
Lo, Chi-kiu, Anand Karthik Tumuluru, and Dekai Wu (2012) 'Fully Automatic Semantic MT Evaluation', in *Proceedings of the 7th Workshop on Statistical Machine Translation*, Montreal, Canada, 243–252.
Lo, Wai-Kit, Patrick Schone, and Helen Meng (2001) 'Multi-scale Retrieval in MEI: An English-Chinese Translingual Speech Retrieval System', in *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2: 1303–1306.
Loh, Shiu-chang (1972) 'Machine Translation at the Chinese University of Hong Kong', in *Proceedings of the CETA Workshop on Chinese Language and Chinese Research Materials*, Washington, DC.
Lu, Bin, Ka Po Chow, and Benjamin K. Tsou (2011) 'The Cultivation of a Trilingual Chinese-English-Japanese Parallel Corpus from Comparable Patents', in *Proceedings of Machine Translation Summit XIII*, Xiamen, China, 472–479.
Meng, Helen, Berlin Chen, Sanjeev Khudanpur, Gina-Anne Levow, Wai-Kit Lo, Douglas Oard, Patrick Schone, Karen Tang, Hsin-Min Wang, and Jianqiang Wang (2004) 'Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval', *Computer Speech and Language* 18(2): 163–179.
Seneff, Stephanie, Chao Wang, and John Lee (2006) 'Combining Linguistic and Statistical Methods for Bi-directional English-Chinese Translation in the Flight Domain', in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06)*, Cambridge, MA, 213–222.
Siu, Kai-Chung and Helen Meng (2001) 'Semi-automatic Grammar Induction for Bi-directional English-Chinese Machine Translation', in *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, 2749–2752.
Siu, Kai-Chung, Helen Meng, and Chin-Chung Wong (2003) 'Example-based Bi-directional Chinese-English Machine Translation with Semi-automatically Induced Grammars', in *Proceedings of the 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2801–2804.
Siu, Sai Cheong (2012) 'Automated Pre-editing and Post-editing: A Hybrid Approach to Computerized Translation of Initial Public Offering (IPO) Prospectuses', in *New Trends in Translation Technology: The*

10th Anniversary Conference of the Master of Arts in Computer-aided Translation Programme, Hong Kong: The Chinese University of Hong Kong.

Song, Yan, Chunyu Kit, and Xiao Chen (2009) 'Transliteration of Name Entity via Improved Statistical Translation on Character Sequences', in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, Suntec, Singapore, 57–60.

Song, Yan, Chunyu Kit, and Hai Zhao (2010) 'Reranking with Multiple Features for Better Transliteration', in *Proceedings of the 2010 Named Entities Workshop*, 16 July 2010, Uppsala University, Uppsala, Sweden, 62–65.

Tsou, Benjamin K. (2007a) 'Salient Linguistic and Technical Gaps between Chinese and other Asian Languages Relevant to MT', in *Translation Automation User Society (TAUS) Exec Forum*, Beijing, China.

Tsou, Benjamin K. (2007b) 'Language Technology Infrastructure for MT', in *Translation Automation User Society (TAUS) Exec Forum*, Beijing, China.

Tsou, Benjamin K. and Bin Lu (2011) 'Machine Translation between Uncommon Language Pairs via a Third Common Language: The Case of Patents', in *Proceedings of Translating and the Computer Conference 2011 (ASLIB-2011)*, London.

Wang, Rongbo, Zheru Chi, and Changle Zhou (2004) 'An English-to-Chinese Machine Translation Method Based on Combining and Mapping Rules', in *Proceedings of Asian Symposium on Natural Language Processing to Overcome Language Barriers*, Sanya, China, 97–102.

Wong, Billy T-M and Chunyu Kit (2008) 'Word Choice and Word Position for Automatic MT Evaluation', in *Proceedings of AMTA 2008 Workshop: MetricsMATR*, Waikiki, HI.

Wong, Billy T-M and Chunyu Kit (2009a) 'Meta-evaluation of Machine Translation on Legal Texts', in *Proceedings of the 22nd International Conference on the Computer Processing of Oriental Languages (ICCPOL-09)*, Hong Kong, China, 337–344.

Wong, Billy and Chunyu Kit (2009b) 'ATEC: Automatic Evaluation of Machine Translation via Word Choice and Word Order', *Machine Translation* 23(2): 141–155.

Wong, Billy and Chunyu Kit (2010) 'The Parameter-optimized ATEC Metric for MT Evaluation', in *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden, 360–364.

Wong, Billy and Chunyu Kit (2011) 'Comparative Evaluation of Term Informativeness Measures for Machine Translation Evaluation Metrics', in *Proceedings of Machine Translation Summit XII*, Xiamen, China, 537–544.

Wong, Billy T.M. and Chunyu Kit (2012) 'Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level', in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 12)*, Jeju Island, Korea, 1060–1068.

Wong, Billy T.M., Cecilia F.K. Pun, Chunyu Kit, and Jonathan J. Webster (2011) 'Lexical Cohesion for Evaluation of Machine Translation at Document Level', in *Proceedings of the 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2011)*, Tokushima, Japan, 238–242.

Wu, Andy and Randall Tan (2009) 'Tree-based Approaches to Biblical Text', BibleTech Conference.

Wu, Dekai (1995) 'Trainable Coarse Bilingual Grammars for Parallel Text Bracketing', in *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, Cambridge, MA, 69–82.

Wu, Dekai (1996) 'A Polynomial-time Algorithm for Statistical Machine Translation', in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL 1996)*, Santa Cruz, CA, USA, 152–158.

Wu, Dekai (1997) 'Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora', *Computational Linguistics* 23(3): 377–404.

Wu, Dekai and Pascale Fung (2009) 'Semantic Roles for SMT: A Hybrid Two-pass Model', in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2009)*, Boulder, CO, 13–16.

Wu, Dekai and Hongsing Wong (1998) 'Machine Translation with a Stochastic Grammatical Channel', in *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (ACL-COLING 1998)*, Montreal, Canada, 1408–1414.

Wu, Dekai, Pascale Fung, Marine Carpuat, Chi-kiu Lo, Yongsheng Yang, and Zhaojun Wu (2010) 'Lexical Semantics for Statistical Machine Translation', in Joseph Olive, Caitlin Christianson, and John McCary (eds) *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, New York: Springer, 236–252.

Wu, Yan and James Liu (1999) 'A Cantonese-English Machine Translation System PolyU-MT-99', in *Proceedings of the Machine Translation Summit VII: MT in the Great Translation Era*, 13–17 September 1999, Kent Ridge Digital Labs, Singapore, 481–486.

Zhang, Xiaoheng (1998) 'Dialect Words Processing in Cantonese-Putonghua Text Machine Translation' (粵-普書面語機器翻譯中的方言詞處理), in *Proceedings of International Conference on Chinese Information Processing* (*1998* 中文信息處理國際會議論文集), Beijing: Tsinghua University Press, 499–506.

Zhang, Xiaoheng (1999) 'Words Processing in Cantonese-Putonghua Machine Translation' (粵-普機器翻譯中的詞處理), *Journal of Chinese Information Processing* (中文信息學報) 13(3): 40–47.

Zhang, Xiaoheng (2005) '上下文索引程序 ABCD 及其在語言翻譯中的應用' (The ABCD Concordancer and Its Application to Language Translation), in Li Yashu 李亞舒, Zhao Wenli 趙文利, and An Qin 晏勤 (eds) *Informationization in Science and Technology Translation* (《科技翻譯信息化》), Beijing: Science Press 科學出版社, 241–245.

Zhang, Xiaoheng and Chun-fat Lau (1996) 'Chinese Inter-dialect Machine Translation on the Web', in *Proceedings of the Asia-Pacific World Wide Web Conference and the 2nd Hong Kong Web Symposium: Collaboration via The Virtual Orient Express*, Hong Kong, 419–429.

Zhang, Xiaoheng and Dingxu Shi (1997) 'Chinese Inter-dialect Machine Translation', in Chen Liwei 陳力為 and Yuan Qi 袁琦 (eds) *Language Engineering* (《語言工程》), Beijing: Tsinghua University Press, 225–231.

Zhou, Lina and James Liu (1997a) 'An Efficient Algorithm for Bilingual Word Translation Acquisition', in *Proceedings of the 2nd Workshop on Multilinguality in Software Industry: The AI Contribution (MULSAIC'97)*, Nagoya, Japan, 65–72.

Zhou, Lina and James Liu (1997b) 'Extracting More Word Translation Pairs from Small-sized Bilingual Parallel Corpus - Integrating Rule and Statistical-based Method', in *Proceedings of International Conference on Computer Processing of Oriental Languages (ICCPOL-97)*, Hong Kong, 520–525.

Zhu, Chunshen (2005) 'Machine-aided Teaching of Translation: A Demonstration', in *META 50*: *For a Proactive Translatology*, Montreal, Canada.

Zhu, Chunshen (2007a) 'ClinkNotes: A Corpus-based, Machine-aided Tool for Translation Teaching', in *Proceedings of the International Symposium on Applied English Education: Trends, Issues and Interconnections*, Kaohsiung, Taiwan.

Zhu, Chunshen (2007b) 'ClinkNotes: Possibility and Feasibility of a Corpus-based, Machine-aided Mode of Translation Teaching', *Conference and Workshop on Corpora and Translation Studies*, Shanghai, China.

18

TRANSLATION TECHNOLOGY IN JAPAN

Hitoshi Isahara

TOYOHASHI UNIVERSITY OF TECHNOLOGY, JAPAN

Introduction

Various services, such as information retrieval and information extraction, using natural language processing technologies trained by huge corpora have become available. In the field of machine translation (MT), corpus-based machine translations, such as statistical machine translation (SMT) (Brown *et al.* 1993: 263–311) and example-based machine translation (EBMT), are typical applications of using a large volume of data in real business situations. Thanks to the availability of megadata, current machine translation systems have the capacity to produce quality translations for some specific language pairs. Yet there are still people who doubt the usefulness of machine translation, especially when it applies to translation among different families of languages, such as those of Japanese and English. A study was conducted to examine the types of machine translation systems which are useful (Fuji *et al.* 2001), by simulating the retrieval and reading of web pages in a language different from one's mother tongue. Research on the technologies which make MT systems more useful in real world situations, however, is scanty.

The problems facing the developers of Japanese-to-English and English-to-Japanese machine translation systems are more serious than those encountered by developers for machine translation systems for say English-to-French translation. This is because Japanese is very different in syntax and semantics from English, so we often need some context to translate Japanese into English (and English into Japanese) accurately. English uses a subject-verb-object word order, while in Japanese, the verb comes at the end of the sentence, i.e. a subject-object-verb order. This means that we have to provide more example sentence pairs of Japanese and English compared to translating most European languages into English, as they also use a subject-verb-object order. The computational power required for Japanese to come up with accurate matches is enormous. And accuracy is particularly necessary for businesses selling their products overseas, which is the reason why it is needed to help Japanese companies provide better translated manuals for their products.

Faced with such obstacles, Japanese researchers conduct studies on quality improvement of MT engines, which include a 5-year national project on development of Japanese–Chinese machine translation systems (Isahara *et al.* 2007). In parallel with this kind of MT research, we can take a three-step approach to improve the MT quality in real life environment: simplifying the Japanese source text (controlled language), enriching lexicon, and enhancing the post-editing process (see Figure 18.1).

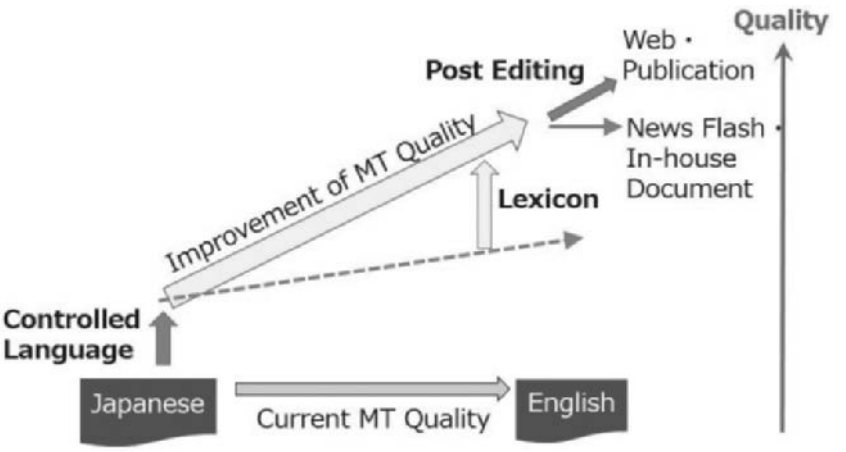


Figure 18.1 Quality improvement during translation procedure

In this chapter, a historical overview of research and development (R&D) of machine translation systems in Japan is given first. It goes on to discuss one of the latest government-funded MT projects, research activities related to pre- and post-editing, and the development of a linguistic resource which is utilized by MT systems, and describes research on the evaluation of MT systems.

History of MT R&D in Japan

The dawn of the MT age

Efforts to let a computer understand a natural language began almost at the same time as when computers were invented. In 1933, Petr Petrovich Smirnov-Troyanskii in Russia applied for a patent for a machine which selected and printed words during the process of translation. The first proposal for automatic translation using a computer was made by Warren Weaver in 1947 (Weaver 1955).

Modern MT research began in the 1950s. As a result of the so-called Sputnik crisis in 1957, R&D of machine translation became popular, especially in the United States. In Japan, Electrotechnical Laboratory developed the 'Yamato' English–Japanese machine translation system, which was presented at the first International Conference on Information Processing held in Paris in June 1959. As there was no electronic computer with a large memory capacity at that time, a dedicated machine which had a large storage for a dictionary for translation was fabricated.

In the past, the implementation of machine translation has adopted a range of approaches, including the transfer and the interlingua methods. For the transfer method, the input text in the original language is first analysed, and then the sentence structure is mapped out in accordance with the grammar of the original language. This sentence structure is then converted into that of the target language using transfer rules, to create a corresponding sentence. For the interlingua method (pivot method), the input sentence goes through a deeper analysis, and is converted into an expression described in an intermediate language that is independent of any specific language. The sentence in the target language is then created, based on the structure of the intermediate expression. Since the interlingua method generates a translation from the

identification of meaning, it allows for a more liberal translation and results in a more natural phrasing. However, this demands that processing provide a deeper understanding of meaning, while at the same time handling massive volumes of information. On the other hand, the transfer method requires the description of a great number of conversion rules, which results in a proportional increase in the number of required rules when multiple languages are involved. Both methods involve compilation from various sources (grammar rules, lexicons, thesaurus, etc.), which must be performed manually, and the establishment of a coherent approach to this task of compilation is extremely difficult. Recently, statistical machine translation (SMT) has been widely studied and shows its promising features. However, its capability to handle pairs of languages with very different grammatical and/or lexical structures is still questionable.

In contrast, it is believed that when a human performs translation, he or she is not strictly applying such knowledge, but is instead translating sentences through combinations of recollected phrases in the target language. Based on this hypothesis, Makoto Nagao of Kyoto University proposed an example-based machine translation (EBMT) in 1981 (Nagao 1984: 173–180). In an example-based machine translation system, translation is executed based on the similarity between the input sentence and an example contained in a huge parallel corpus. When EBMT was proposed, the capacity of the computer was insufficient to produce a practical system with this approach. In recent years, with rapid improvements in computer performance and the development of a method for judging similarity between examples (through reference to a database of syntactically analysed sentences accumulated in the system), the foundations for the establishment of a practical example-based machine translation system have been laid.

The golden age of MT in Japan

In the 1980s, machine translation studies in Japan became popular and Japan led the world in MT research. This had a lot to do with the Mu Project, which started in 1982 and ended in 1985. The Mu Project aimed to develop a Japanese–English and English–Japanese machine translation system for translating abstracts of scientific and technical papers. One of the factors that accounted for the creation of the Mu Project was the storage of scientific and technological information at the Japan Information Center of Science and Technology (JICST) in those days, and some of the items necessary for the development of a machine translation system, such as a documents database including the abstract, dictionaries and thesaurus, were adequate. In addition, as Japanese technology was as good as that of Europe and North America, the United States and other countries started to criticize Japan, insisting that Japan utilized a large number of technologies from overseas. Another factor that facilitated the development of machine translation systems was the availability of computer systems, such as Japanese word processors, which could handle the Japanese language with Kanji characters. What is more important was the leadership of distinguished scholars, such as Professor Makoto Nagao of Kyoto University, who proposed, launched and conducted national projects on natural language processing, especially machine translation.

Many Japanese companies participated in the Mu Project, and the knowledge gained from the project contributed to the study of machine translation among commercial companies, creating the golden age of machine translation study in Japan. Following the release of ATLAS-I, which was the first commercial MT system on a mainframe computer in Japan by Fujitsu in 1984, MT vendor companies, such as Bravice International, NEC, Toshiba, Hitachi Ltd., Mitsubishi Electric, Sharp, and Oki Electric Industry, released their machine translation systems. Among them, Toshiba started to sell its MT system working on a 'mini-computer'

and Bravice started to sell the system on a 'personal computer' in 1985. In the same year, some of these systems were demonstrated at the International Exposition, Tsukuba, Japan, 1985 (Tsukuba Expo '85). At major international conferences, such as COLING (International Conference on Computational Linguistics), a large number of research papers on machine translation were presented by researchers in Japanese companies and universities. The first Machine Translation Summit (MT Summit), which is one of the major biannual conferences related to MT, was held in 1987 at Hakone, Japan. These conferences have since been held in Asia, Europe, and North America.

In 1986, the Advanced Telecommunications Research Institute (ATR) was established and started research on speech translation. In 1987, the Japanese Ministry of International Trade and Industry (MITI) launched its multilingual machine translation project (MMT project) which aimed to develop machine translation systems for translation among Japanese, Chinese, Thai, Malay and Indonesian, and Japanese MT companies, such as NEC and Fujitsu, joined this project. The MMT project ended in 1996.

The 1990s and beyond

In the 1990s, there was considerable improvement in the performance of machine translation systems with their output quality reaching the practical use level if the domain of input text was properly restricted.

The Asia-Pacific Association for Machine Translation was established in 1991, initially with the name of the Japan Association for Machine Translation. To expand its operations, the name was subsequently changed to the Asia-Pacific Association for Machine Translation (AAMT). AAMT is one of three regional associations of the European Association of Machine Translation (EAMT) and the Association of Machine Translation in the Americas (AMTA), both members of the International Association of Machine Translation (IAMT) which organizes MT Summits.

For translation services, browsers interlocked with machine translation systems, such as PENSEE for Internet by Oki Electric, became commercially available in 1995. This shows that the development of MT systems was happening at the same pace as the expansion of the internet and WWW, which boosted the MT industries. Japan Patent Office (JPO) opened the Industrial Property Digital Library (IPDL) in 1999. At the beginning, abstracts of the patents were translated manually. In 2000, IPDL started to use MT systems to translate Japanese patents into English in full. As the first free translation site on the internet (Excite) was launched in 2000, information acquisition via the World Wide Web and utilization of MT became popular.

With regard to the price of software, MT software priced at less than 10,000 JPY was released in 1994. The software market then became very competitive. In 1996, personal computers with bundled MT software merged, and software practically became cost free for customers. MT software at less than 5,000 JPY appeared in 2002, and of less than 2,000 JPY in 2003. And as free online translation services became popular, the survival of companies selling package MT software became critical.

During this period, people began to have a better understanding of machine translation, and a number of MT systems received awards from various organizations, such as the 'Good Design Award' in 2001.

As for research projects, there were a few large-scale projects on text translation since the Mu Project and the MMT project. In 2006, a 5-year project on Japanese–Chinese machine translation using the example-based approach received funding from the Japan Science and Technology Agency.

Government-funded projects for developing a Chinese–Japanese and Japanese–Chinese machine translation system

In 2006, the National Institute of Information and Communications Technology (NICT) of Japan, the Japan Science and Technology Agency (JST), the University of Tokyo, Shizuoka University, and Kyoto University launched a 5-year government funded project on the development of a Chinese–Japanese machine translation system for scientific documents.

This project conducted research on augmented example-based machine translation as a verification of high-performance resource-based NLP, which was based on cutting-edge research on computational linguistics and natural language processing. It is found that, unlike what is happening in the West, the distribution of information in English in Asia is difficult. It is necessary for Asian countries to develop machine translation systems for Asian languages. As the first step in this endeavor, Japan started to develop a machine translation system for scientific and technological materials in Chinese and Japanese so as to keep pace with the significant progress that has been made in various fields.

As mentioned above, the construction of such a system serves as the first step in building systems which cover a wide variety of Asian languages. As China has made remarkable progress in science and technology, this Japanese–Chinese translation system had several objectives:

- to make scientific and technological information in China and other Asian countries easily usable in Japan;
- to promote the distribution of documents containing Japan's cutting-edge science and technology to China and other countries; and
- to contribute to the development of science and technology in Asian countries with the help of the information available through machine translation.

The goal of this project was to develop, within a period of five years, a practical machine translation system for translation between the Japanese and Chinese languages, focusing on scientific and technological materials. In this endeavor, research adopted the example-based approach, which provided a better reflection of the linguistic structures and syntactic information used in a number of parts in the translation engines.

Figure 18.2 presents an outline of the system under development. Its target domains were information science, biological science, and environmental science.

EBMT requires the accumulation of a large number of examples; accordingly, researchers planned to develop a parallel corpus of around 10 million sentences. They extracted parallel sentences from existing comparable texts semi-automatically and aligned words and phrases semi-automatically. They also planned to make the best use of existing linguistic resources and language processing technology owned by them.

In this five-year project, there were goals for the third year and the fifth year.

- Goal for the third year
  Evaluate the Japanese–Chinese machine translation prototype system for specific target domains.
- Goal for the fifth year
  Improve the Chinese analysis performance, and complete demonstration experiments on the Japanese–Chinese and Chinese–Japanese machine translation prototype system.

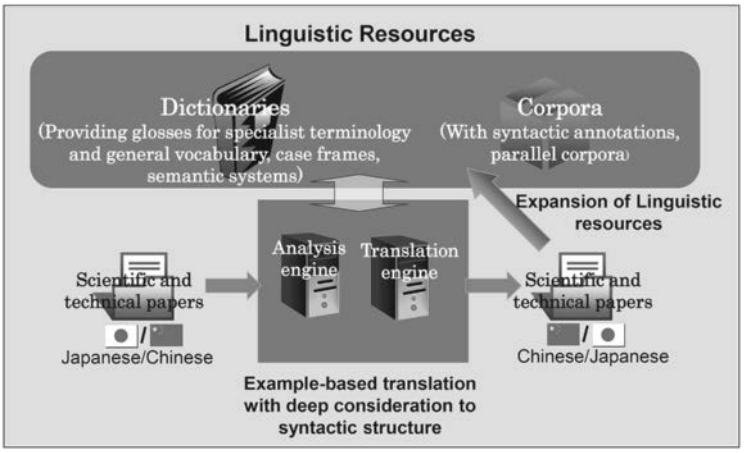


Figure 18.2 System overview

Even during the course of the project, researchers publicized the language resources (such as the corpus) to the fullest extent for research purposes. They also publicized the contents and results of their research widely, as part of their outreach activities.

EBMT systems basically generate sentences in the target language by extracting and combining examples in parallel text databases whose source language sentence is similar to the input sentence. A specific feature of the above system was the utilization of deep syntactic analysis. Parallel texts in the example database were analysed syntactically and aligned with words using syntactic information. In the translation phase, the system analysed the input sentence syntactically, extracted parts of sentences from the example database and combined them to generate sentences in the target language by considering their syntactic structures. At the last stage, the ordering of words in a sentence was made by using the information extracted from the monolingual corpora.

Practical use of machine translation – pre- and post-editing for multilingual information outbound

Some of major international companies use machine translation to meet their daily practical needs. These companies started to use English–Japanese MT at their branches in Japan to translate documents originally written in English into Japanese. To achieve their purposes, the performance of the MT engine and the support of the translation process are crucial. As the documents are frequently revised and reused, it is important to develop a translation environment for text input, the retrieval and display of parallel text (translation memory), dictionary lookup, MT, and formatting documents.

Japanese–English translation in Japan is mainly used for information dispatch, and high-quality translation is needed for such information dispatch from industries. Due to the linguistic features of Japanese, such as the omission of the subject, computational treatment of the Japanese text is more difficult than that of other languages, such as English. The performance of Japanese–English MT is therefore not as good as that of English–Japanese MT. The editing of the output from an MT system is difficult task for a non-native speaker of English.

To overcome this difficulty, control language and crowdsourcing post-editing have been proposed.

Control language

The output quality of MT depends heavily on the quality of the analysis of input sentences. Long and complex sentences in syntax and semantics are mostly very difficult for automatic analysers to output in proper structures. Restricting the structures of the input text is therefore beneficial to MT systems to achieve high-quality translation. Research has been carried out to address this issue by investigating the feasibility of developing a 'controlled Japanese' with explicit restrictions on the vocabulary, syntax, and style when authoring technical documentation. An example was the research project which was being conducted in collaboration with an automobile related company, an MT vendor and a university in Japan.

This project aimed to arouse translation awareness within a global Japanese company where non-professional authors are called upon to write 'global job manuals' for internal dissemination. Following an analysis of the current practice, researchers devised a document template and simple writing rules which were tested experimentally with MT systems. Sentences violating the rules were extracted from the original data and rewritten in accordance with respective rules. The original and rewritten sentences were then translated by MT systems, and the input and output were submitted to human evaluation. Overall, native-speaker judges found that the quality of the Japanese was maintained or improved, while the impact on the raw English translations varied according to MT systems. Researchers explained their template and rules to employees of the company and asked them to write their manuals articulating the know-how using the template and rules. They investigated their documents to identify the most promising avenues for further development (Tatsumi *et al.* 2012: 53–56; Hartley *et al.* 2012: 237–244). Table 18.1 lists the 20 problem features from the corpus with which they experimented. These gave rise to 28 pre-editing rules formulated as 'Omit …', 'Replace with …' or 'Add …'.

An alternative to controlled (or simplified) language is the translation between two languages both of which are properly controlled. If we train SMT with a parallel controlled language corpus, it can translate controlled input into controlled output with high quality. Some multilingual MT systems are combinations of MT engines for two languages and translations between non-English languages are performed via English. Such cascade translation usually amplifies errors during translation. Using controlled English as a pivot would be a promising solution of this problem.

There are several activities relating to controlled languages in Japan. The Technical Japanese Project, which was funded by the Japan Patent Information Organization (JAPIO), has conducted several activities in technical Japanese, a restricted language for business purposes. Its activities are divided into two parts, i.e. technical Japanese for general purposes and technical Japanese for patent documents. As for technical Japanese for patent documents, its committee comprises specialists on intellectual property, patents, natural language processing, machine translation, and patent translation. Their output includes a patent writing manual, guideline for human writers, format for patent ontology, and a patent writing support system. The Japan Technical Communicators Association (JCTA) published a book on writing guidelines for technical communicators who are mainly writing business documents. The Association of System Documentation Quality (ASDoQ) has created a list of terms and technologies related to system documentation, and has collected example sentences, both the good ones and the bad ones.

Table18.1 'Avoid' features of UM guidelines

F1	Long sentences (> 50 characters)
F2	Sentences of 3 or more clauses
F3	Negative expressions
F4	Verb + nominaliser こと
F5	Nominaliser もの
F6	Verb + ように ('it is suggested that')
F7	Topicalizing particle は
F8	Coordinating conjunction または ('or')
F9	Modal れる・られる ('can')
F10	Verb 見える ('can be seen')
F11	Compound noun strings
F12	Particle など ('and so on')
F13	Single use of conjunctionたり ('either')
F14	Katakana verbs
F15	Suffix 感 ('sense of')
F16	Verb かかる ('start')
F17	Verb 成る ('become')
F18	Verb 行う ('perform')
F19	Case-marking particle で ('with', 'by')
F20	Verb ある・あります ('exist')

Crowdsourcing post-editing

With the use of properly controlled input sentences and substantial dictionaries, the current MT systems are useful, for example, for quick translations, such as news flashes, and in-house translations (Figure 18.1).

For documents which need higher quality, post-editing is required. Post-editing, however, can be costly and time-consuming, and is not affordable to everybody. There is a preliminary investigation by Toyohashi University of Technology (TUT) in Japan on the impact of crowdsourcing post-editing through the so-called 'Collaborative Translation Framework' (CTF) developed by the Machine Translation team at Microsoft Research (Aikawa *et al.* 2012: 1–10). Crowdsourcing translation has become increasingly popular in the MT community, and it is hoped that this approach can shed new light on the research direction of translation.

To study this issue, researchers used foreign students at TUT and asked them to post-edit the MT output of TUT's websites (http://www.tut.ac.jp/english/introduction) via Microsoft Translator into their own languages using the CTF functionalities. Though they did not expect the students to have the same degree of accuracy from the professionals, they did note that students had a better understanding of the context, and this kind of collaboration could improve and reduce the cost of the post-editing process.

TUT completed the first experiment with its foreign students attending our university to post-edit the MT output of the English version of the university's website into their own languages. TUT also conducted an experiment with Japanese students with more precise settings, such as the ordering of post-editing. The experimental results show that it was possible to reduce the cost of post-editing drastically.

The development of linguistic resources

As the current mainstream of MT research is the resource-based MT system, the development of linguistic resources is obviously one of the main considerations in MT technology.

Parallel or multilingual corpora

A parallel corpus is a collection of articles, paragraphs, or sentences in two different languages. Since a parallel corpus contains translation correspondences between the source text and its translations at a different level of constituents, it is a critical resource for extracting translation knowledge in machine translation. In the development of MT systems, the example-based and the statistics-based approaches have been widely researched and applied. Parallel corpora are essential for the growth of translation studies and development of practical systems. The raw text of a parallel corpus contains implicit knowledge. If we annotate its information, we can get explicit knowledge from the corpus. The more information that is annotated on a parallel corpus, the more knowledge we can get from the corpus.

NICT started a project to build multilingual parallel corpora in 2002. This project focuses on Asian language pairs and the annotation of detailed information, including syntactic structures and alignment at the word and phrase levels. The corpus is known as the NICT Multilingual Corpora. A Japanese–English parallel corpus and a Japanese–Chinese parallel corpus were completed following systematic specifications. Details of the current version of the NICT Multilingual Corpora are listed in Table 18.2.

Table 18.2 Details of the current version of NICT Multilingual Corpora

Corpora	*Total*	*Original*	*Translation*
Japanese–English Parallel Corpus	38,383 sentence pairs; (English 900,000 words)	Japanese (38,383 sentences, Mainichi Newspaper)	English Translation
		English (18,318 Sentences, *Wall Street Journal*)	Japanese Translation
Japanese–Chinese Parallel Corpus	38,383 sentence pairs; (Chinese 1,410,892 Characters, 926,838 words)	Japanese (38,383 sentences, Mainichi Newspaper)	Chinese Translation

EDR Lexicon

Though current research on NLP and MT utilizes the machine-learning mechanism based on a huge amount of linguistic data, human-coded lexical resources are still very important.

The EDR Electronic Dictionary was developed for advanced processing of natural language by computers, and has eleven sub-dictionaries, which include, among others, a concept dictionary, word dictionaries, and bilingual dictionaries. The EDR Electronic Dictionary is the result of a 9-year project (from 1986 to 1994), aiming at establishing an infrastructure for knowledge information processing. The project was funded by the Japan Key Technology Center and eight computer manufacturers.

The EDR Electronic Dictionary is a machine-tractable dictionary that catalogues the lexical knowledge of English and Chinese (the Word Dictionary, the Bilingual Dictionary, and the Co-occurrence Dictionary), and has unified thesaurus-like concept classifications (the Concept Dictionary) with corpus databases (the EDR Corpus). The Concept Classification Dictionary,

a sub-dictionary of the Concept Dictionary, describes the similarity relation among concepts listed in the Word Dictionary. The EDR Corpus is the source for the information described in each of the sub-dictionaries. The basic approach taken during the development of the dictionary was to avoid a particular linguistic theory and to allow for adoptability to various applications.

The EDR Electronic Dictionary, thus developed, is believed to be useful in the R&D of natural language processing and the next generation of knowledge processing systems. In addition, it will become part of an infrastructure that provides new types of activities in information services.

A universal format for the user dictionary

The development of a lexicon is normally very cost-consuming. Sharing lexicons among groups and/or reusing lexicons between the previous system and the current system are one of the key technologies for the development of efficient MT systems and translation procedure.

As there was no widely used standard for user dictionaries in the Japanese/English MT market, AAMT developed a common format for lexicons for machine translation, and opened it as UPF (Universal PlatForm) in 1997. Currently its new format is available as Universal Terminology Exchange (UTX) (http://aamt.info/english/utx/index.htm).

UTX (Universal Terminology eXchange) is a common format for the user dictionary. In 2009, AAMT established the UTX-Simple (later renamed as 'UTX'), which was an open format in a tab-delimited text. UTX greatly improves the accuracy of translation software by sharing the knowledge of terminology through dictionaries in a bilingual format. The goal of UTX is to create a simple, easy-to-make, easy-to-use dictionary from a user's perspective, not from that of a developer. A user can easily convert a UTX dictionary into various formats. With or without such conversion, the content of the same UTX dictionary can be used with various translation software or computer-aided translation (CAT) tools. In addition, a UTX dictionary can also be used as a glossary without involving translation software.

An example of UTX is shown in Table 18.3.

Table 18.3 Example of an English-to-Japanese dictionary in UTX

#UTX-S 1.00; en-US/ja-JP; 2008-03-15T10:00:00Z+09:00; copyright: AAMT, license: CC-by 3.0									
#src	**tgt**	**src:pos**	src:plural	src:3sp	src:past	src:pastp	src:presp	src:comp	src:super
new	新規の	adjective						newer	newest
fast	高速な	adjective						faster	fastest
# prosody should be uncountable									
prosody	韻律	noun	prosodies						
save	保存する	verb		saves	saved	saved	saving		
good evening	今晩は	sentence							

Evaluation

The Japan Electronic Industry Development Association (JEIDA) formulated three criteria for evaluating MT systems: (1) technical evaluations by users; (2) financial evaluations by users; and (3) technical evaluation by developers. JEIDA has since 1992 developed a method to evaluate quality for the developers of machine translation systems so that they can easily check the imperfections in their systems. In 1995, JEIDA's two test-sets (English-to-Japanese and Japanese-to-English) were completed and made publicly available. During the development of these test-sets, JEIDA laid down the following two types of objectivity, which included

1 objectivity in the evaluation process; and
2 objectivity in the judgment of the evaluation results.

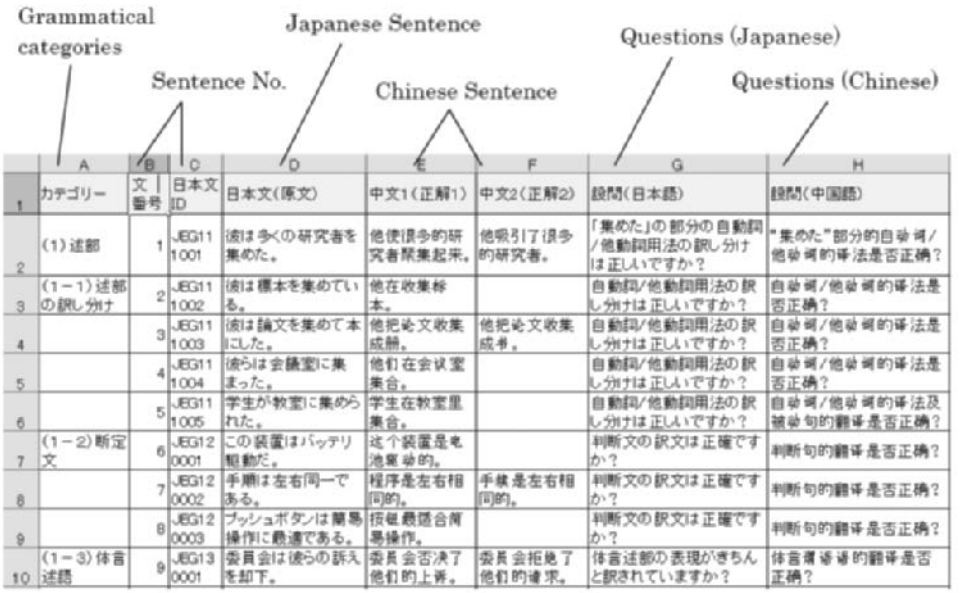
In an evaluation method such as the one proposed in the ALPAC report, 'fidelity' and 'intelligibility' are employed as evaluation measures, though they are dependent on subjective human judgment. Consequently, the results may differ according to who makes the evaluations, which means they do not satisfy the objectivity criterion (1). Theoretically, the evaluation method in the ALPAC report satisfies criterion (2) since the evaluation results are given as numbers. The system developers, however, fail to recognize which items cannot be handled in their own system. This is because the test example in question covers various kinds of grammatical items. Their interpretation of the evaluation result for further improvement of their system is therefore still subjective, which, for practical purposes, does not satisfy criterion (2).

On the other hand, JEIDA created test-sets that can satisfy both criteria. JEIDA explains how to evaluate individual examples by posing yes/no questions which enable the system developers to make an evaluation just by answering them. With this method, everyone can evaluate MT systems equally, as his/her answers require only a simple yes or no. Even for imperfect translation results, judgment will not vary widely among evaluators. In addition, JEIDA assigned to each example an explanation which gives the relationship of the translation mechanism to the linguistic phenomenon, thus enabling the system developer to know why the linguistic phenomenon in question was not analysed correctly. Consequently, with JEIDA's test-set method, the evaluation results can be utilized for improving MT systems.

In JEIDA's test-sets, we have systematically sampled the grammatical items that ought to be taken up, and listed some examples for each item. The test-sets clearly describe what linguistic phenomenon should be evaluated in each example so that the developers can easily understand the problems they need to solve in their systems. The system developer can then identify causes of translation failures.

Following JEIDA's test-set for MT evaluation, AAMT has continued its development of the MT evaluation method. Its aim is to establish a satisfactory evaluation method to provide an objective criterion, reduce man-hour costs, and identify weaknesses of MT systems. It followed the previous approach by JEIDA which was a binary classification evaluation in which judgment was conducted via Yes/No answers for grammatical questions. So far, AAMT has developed approximately 400 test sentences (46 grammatical items) for Japanese–English/Chinese MT (Figure 18.3).

After some experiments using these test-sets, AAMT found that the test-set-based evaluation needed less than half the time than the conventional method and its test-set-based evaluation has given a higher score to Japanese–English MT than Japanese–Chinese MT, which reflects the true state of MT technology.



	A	B	C	D	E	F	G	H
1	カテゴリー	文番号	日本文ID	日本文(原文)	中文1(正解1)	中文2(正解2)	設問(日本語)	設問(中国語)
2	(1)述部	1	JEG111001	彼は多くの研究者を集めた。	他使很多的研究者聚集起来。	他吸引了很多的研究者。	「集めた」の部分の自動詞/他動詞用法の訳し分けは正しいですか?	"集めた"部分的自动词/他动词的译法是否正确?
3	(1－1)述部の訳し分け	2	JEG111002	彼は標本を集めている。	他在收集标本。		自動詞/他動詞用法の訳し分けは正しいですか?	自动词/他动词的译法是否正确?
4		3	JEG111003	彼は論文を集めて本にした。	他把论文收集成册。	他把论文收集成书。	自動詞/他動詞用法の訳し分けは正しいですか?	自动词/他动词的译法是否正确?
5		4	JEG111004	彼らは会議室に集まった。	他们在会议室集合。		自動詞/他動詞用法の訳し分けは正しいですか?	自动词/他动词的译法是否正确?
6		5	JEG111005	学生が教室に集められた。	学生在教室里集合。		自動詞/他動詞用法の訳し分けは正しいですか?	自动词/他动词的译法及被动句的翻译是否正确?
7	(1－2)断定文	6	JEG120001	この装置はバッテリ駆動だ。	这个装置是电池驱动的。		判断文の訳文は正確ですか?	判断句的翻译是否正确?
8		7	JEG120002	手順は左右同一である。	程序是左右相同的。	手续是左右相同的。	判断文の訳文は正確ですか?	判断句的翻译是否正确?
9		8	JEG120003	プッシュボタンは簡易操作に最適である。	按钮最适合简易操作。		判断文の訳文は正確ですか?	判断句的翻译是否正确?
10	(1－3)体言述語	9	JEG130001	委員会は彼らの訴えを却下。	委员会否决了他们的上诉。	委员会拒绝了他们的请求。	体言述部の表現がきちんと訳されていますか?	体言谓语句的翻译是否正确?

Figure 18.3 Example of a Japanese–Chinese test set by AAMT

References

Aikawa, Takako, Kentaro Yamamoto, and Hitoshi Isahara (2012) 'The Impact of Crowdsourcing Post-editing with the Collaborative Translation Framework', in *Proceedings of the 8th International Conference on Natural Language Process (JapTAL2012)*, 22–24 October 2012, Kanazawa, Japan, 1–10.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational Linguistics* 19(2): 263–311.

Fuji, Masaru, N. Hatanaka, E. Ito, S. Kamei, H. Kumai, T. Sukehiro, T. Yoshimi, and Hitoshi Isahara (2001) 'Evaluation Method for Determining Groups of Users Who Find MT Useful', in *Proceedings of the Machine Translation Summit VIII: Machine Translation in the Information Age*, 18–22 September 2001, Santiago de Compostela, Spain.

Hartley, Anthony, Midori Tatsumi, Hitoshi Isahara, Kyo Kageura, and Rei Miyata (2012) 'Readability and Translatability Judgments for "Controlled Japanese"', in *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT-2012)*, 28–30 May 2012, Trento, Italy, 237–244.

http://aamt.info/english/utx/index.htm.

http://www.tut.ac.jp/english/introduction.

Isahara, Hitoshi, Sadao Kurohashi, Jun'ichi Tsujii, Kiyotaka Uchimoto, Hiroshi Nakagawa, Hiroyuki Kaji, and Shun'ichi Kikuchi (2007) 'Development of a Japanese-Chinese Machine Translation System', in *Proceedings of MT Summit XI*, 10–14 September 2007, Copenhagen, Denmark.

Nagao, Makoto (1984) 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle', in Alick Elithorn and Ranan Banerji (eds) *Artificial and Human Intelligence*, New York: Elsevier North-Holland Inc., 173–180.

Tatsumi, Matsumi, Anthony Hartley, Hitoshi Isahara, Kyo Kageura, Toshio Okamoto, and Katsumasa Shimizi (2012) 'Building Translation Awareness in Occasional Authors: A User Case from Japan', in *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT-2012)*, 28–30 May 2012, Trento, Italy, 53–56.

Weaver, Warren (1955) *'Translation (1949)': Machine Translation of Languages*, Cambridge, MA: MIT Press.

19

TRANSLATION TECHNOLOGY IN SOUTH AFRICA

Gerhard B. van Huyssteen

NORTH-WEST UNIVERSITY, SOUTH AFRICA

Marissa Griesel

NORTH-WEST UNIVERSITY, SOUTH AFRICA

Introduction

South Africa has a rich and diverse multilingual culture with eleven official languages, namely two Germanic languages (English and Afrikaans), four Nguni languages (Ndebele (isiNdebele), Swati (Siswati), Xhosa (isiXhosa), and Zulu (isiZulu)), three Sotho languages (Northern Sotho (Sesotho sa Lebo or Sepedi), Southern Sotho (Sesotho), and Tswana (Setswana)), and two other Bantu languages (Tsonga (Xitsonga) and Venda (Tshivenda)). These languages are granted official status in chapter one of the Constitution of the Republic of South Africa (6 of 1996), stating that 'the state must take practical and positive measures to elevate the status and advance the use of these languages'. To this effect, the Pan South African Language Board (PanSALB) was established in terms of the Pan South African Language Board Act (59 of 1995), with the goal to promote multilingualism in South Africa. Recently, the Use of Official Languages Act (12 of 2012) was promulgated, in which various conditions for the use of the official languages by government and other institutions are set in order to further a truly multilingual society. In addition to these acts, various other acts and industry regulations also contribute to create a progressive regulatory environment prescribing the use of multiple official languages. These include, inter alia, the Banking Association of South Africa (2004), and the National Consumer Protection Act (68 of 2008).

Despite the fact that English is only the sixth largest language in South Africa (with 9.6 per cent of speakers indicating English as their home language in the 2011 South African National Census; see Table 19.1), information in the business, health and government sectors is generally available only in English. Coupled with the fact that only a small portion of official South African government websites are available in all the South African languages (De Schryver and Prinsloo 2000: 89–106), it becomes clear that language practitioners and translators working with South African languages need all the help they can get to create texts in the South African languages as efficiently as possible.

Machine translation (MT) offers an attractive and viable option that is being explored only now on a more widely level in South Africa. However, as is well known, the quality of automated translation is not yet at a level, even internationally, to replace human translators for

Table 19.1 South African languages[1]

South African languages 2011		
Language	*Number of speakers*	*% of total*
Afrikaans	6 855 082	13.5%
English	4 892 623	9.6%
isiNdebele	1 090 223	2.1%
isiXhosa	8 154 258	16%
isiZulu	11 587 374	22.7%
Sepedi	4 618 576	9.1%
Sesotho	3 849 563	7.6%
Setswana	4 067 248	8%
Sign language	234 655	0.5%
SiSwati	1 297 046	2.5%
Tshivenda	1 209 388	2.4%
Xitsonga	2 277 148	4.5%
Other	828 258	1.6%
TOTAL	50 961 443	100%

translation of documents; human involvement in post-process editing of generated translations is still of the utmost importance. This is even more true in the South African context where MT is only available for a few select languages, with quality still reflecting the early days of such MT systems. However, using machine-aided human translation (where a human is responsible for the translation, but uses different technologies to ease and assist with the process) is already very useful and attainable in the South African context.

This article focuses on the history and state-of-the-art of MT research and development in South Africa for South African languages.[2] We will first provide an overview of the lead-up to MT development in South Africa, highlighting some related research, as well as the development of tools and data that could support MT in South Africa indirectly. Thereafter we give an overview of the first initiatives by the South African government to support the development of MT for South African languages. We then discuss individual research and development projects on MT for South African languages, before describing the Autshumato project, South Africa's first consolidated national MT project for South African languages, in more detail. We conclude with a look-ahead to post-Autshumato initiatives and possibilities for MT in South Africa.

Background: linguistics and language technology in South Africa

Linguistic research in all eleven South African languages has always been a rich field of study. Various aspects of the grammars of most of the languages have long since been described in various scholarly publications; for instance, as early as 1862, Bleek (1862) compared various aspects of the different Bantu languages. However, various political, socio-economic and socio-linguistic factors have slowed down processes of grammatical and lexical standardization, as well as the development of terminology in domains where higher functions are required (e.g. business, the judiciary, science and technology, mainstream media, etc.). Nonetheless, over the past twenty years more and more specialized dictionaries and terminology lists have

been developed through the establishment of government-supported national terminology units (not-for-profit companies) for each language. In addition, the national language bodies of PanSALB are responsible for language standardization and the development of orthographies for each of the eleven official languages.

The Bible Fellowship of South Africa has also contributed greatly, albeit unintentionally, to standardization of the South African languages. The Bible is available in all official languages, plus a few local variants like Fanagalo (a pidgin artificially created to support communication between English settlers and the local people, used extensively in the mines of South Africa, and incorporating words and structures from many different languages (Adendorff 2002: 179–198)). Professional language practitioners' forums, such as the South African Translators Institute or ProLingua, have become hubs of both knowledge in human translation practice, as well as sources for data such as personal wordlists and translation memories. These organizations have also become key partners in empowering freelance translators with tools to incorporate electronic resources such as translation memories, electronic dictionaries and term banks into translation practice.

In recent years, the human language technology (HLT) fraternity in South Africa has also become an important enabler, addressing some of the needs of translators and language practitioners. Since the South African HLT industry as a whole is still fairly young, only a few good quality core technologies exist for only some of the official languages. For example, automatic part-of-speech (POS) taggers utilizing different machine learning techniques have been developed for Afrikaans (Pilon 2005) and Northern Sotho (Heid *et al.* 2009: 1–19); lemmatisers for Afrikaans (Groenewald and van Huyssteen, 2008: 65–91) and Tswana (Brits *et al.* 2006: 37–47); morphological analysers for Zulu (Pretorius and Bosch 2003: 191–212; Spiegler *et al.* 2008: 9–12), etc. (see Sharma Grover *et al.* (2011: 271–288) for an overview of technologies and resources available for the South African languages). Most of these and other similar core technologies have yielded good results and can, for instance, be used in pre- and post-processing to improve machine translation output quality.

Spelling checkers, like those developed by the Centre for Text Technology (CTexT) at the North-West University (NWU) in South Africa, can also contribute greatly to the usefulness of an MT system by providing spelling variants, or checking the validity of generated constructions. Languages with conjunctive orthographies (like Afrikaans and Zulu) form new words (and even phrases) by combining words and morphemes; spelling and grammar checkers could play an important role in validating such combinations in the context of MT.

Another related development has been the creation and expansion of wordnets for five South African languages. A good quality wordnet could add valuable linguistic information to any MT system or for MT evaluation, as it includes various semantic relations, definitions and usage examples. The Afrikaans wordnet (Kotzé 2008: 163–184; Botha *et al.* 2013: 1–6) currently has more than 11,000 synsets, and is modelled to the standards set in the Princeton WordNet and the Balkanet project. A joint effort by the University of South Africa (UNISA) and the NWU, funded by the South African Department of Arts and Culture (DAC), also saw the development of wordnets for Northern Sotho, Tswana, Xhosa and Zulu, with more than 5,000 synsets in each of these wordnets. The project received renewed funding from UNISA to expand these wordnets even further, and to add other South African languages from 2012 to 2014.

With a view on automated speech translation in the future, the Meraka Institute at the Council for Scientific and Industrial Research (CSIR) and the speech research group at NWU have been the driving forces behind many projects to create core speech technologies and resources which could eventually be used for spoken MT. These include, inter alia, grapheme-to-phoneme conversion, speech recognition and speech synthesis, as well as a

large-scale data collection efforts in various projects. However, no spoken MT system is foreseen for the immediate future.

The South African government and HLT

The establishment of HLT as a viable industry in South Africa has a history extending back to 1988, with the publication of the LEXINET Report by the Human Sciences Research Council (Morris 1988). This report highlighted the importance of technological developments to foster communication in a multilingual society.

From the 1990s, South Africa was consumed with more pressing political matters, and the next government report to mention HLT explicitly only appeared in 1996. The final report by the Language Plan Task Force of the then Department of Arts, Culture, Science and Technology (DACST) included both short and long term action plans for language equality in South Africa (LANGTAG 1996). As a direct result of this report, a steering committee on translation and interpreting, as part of PanSALB, was established in 1998. A second steering committee, in collaboration with DACST, was formed in 1999, and was tasked to investigate and advise regarding HLTs in South Africa. The report by this joint steering committee was released in 2000, and a ministerial committee was established to develop a strategy for developing HLT in South Africa. The ministerial committee's report appeared in 2002, at which stage DACST split into two sections, viz. Department of Arts and Culture (DAC) and Department of Science and Technology (DST), with DAC retaining the primary responsibility for the development of HLT. (For an overview of the early history of HLT in South Africa, see Roux and du Plessis 2005: 24–38.)

Following the recommendations of a ministerial advisory panel on HLT in 2002, three major research and development projects were funded subsequently by DAC, including a speech project to foster information access via interactive voice response systems (the Lwazi project), a project to develop spelling checkers for the ten indigenous languages, and a project to develop MT systems for three language pairs (the Autshumato project); see the section on 'The Autshumato Project' below for details.

Based on a decision taken by the South African cabinet on 3 December 2008, the National Centre for Human Language Technology (NCHLT) was established in 2009. As one of its first large-scale projects, the NCHLT announced a call for proposals to create reusable text and speech resources that are to serve as the basis for HLT development, to stimulate national interest in the field of HLT, and to demonstrate its potential impact on the multilingual South African society. CTexT, in collaboration with the University of Pretoria (UP) and language experts across the country, was designated as the agency responsible for the development of various text resources. The following resources have been completed by 2013:

- corpora (one million words for each language);
- aligned corpora (fifty thousand words for each language, aligned on sentence level);
- wordlists for all eleven languages; and
- part-of-speech taggers, morphological analysers and lemmatizers for the ten indigenous languages.

Given all these projects funded by government, it soon became clear that a central repository should be established to manage multilingual digital text and speech resources for all official languages in a sustainable manner, in order to ensure the availability and reusability of this data for educational, research and development purposes. In 2011, CTexT was appointed to set up

the so-called Resource Management Agency (RMA[3]); the RMA works in close co-operation with the Dutch TST-Centrale. Data hosted by the RMA include broad categories such as text, speech, language related video and multimodal resources (including sign language), as well as pathological and forensic language data. It is now required that all past, current and future HLT projects funded by government have to deliver project data to the RMA, in order to prevent loss of data and to promote reusability of the data. The RMA also aims to position South Africa strategically through collaboration with other similar agencies worldwide, with the long-term vision of becoming the hub for language resource management in Africa.

Early MT projects in South Africa

Since the beginning of this century, when the South African government made it clear that it would be investing in and supporting initiatives for developing HLTs for South African languages, numerous research projects with smaller goals have begun exploring the possibilities that MT could hold for the South African community. One of the earliest projects (established in 2002 at the University of Stellenbosch) concerned the development of an experimental South African Sign Language MT system (van Zijl and Barker 2003). We could unfortunately not find any details on the performance of the system – from the latest publication from the project it seems as if it might still be under development (van Zijl and Olivrin 2008: 7–12).

As Afrikaans has the most available resources (data and core technologies) compared to the other indigenous languages (Sharma Grover *et al.* 2011: 271–288), most of the early developments in MT research for South African languages have had Afrikaans as either the source or target language. Ronald and Barnard (2006: 136–140) showed that, even with very limited amounts of data, a first MT system for translation from English to Afrikaans was indeed possible, using a statistical MT approach. They used a parallel corpus of only 40,000 sentences, and achieved a BLEU score (Papineni *et al.*, 2002: 311–318) of 0.3. Their study also included systems with even smaller datasets (3,800 sentences per language pair), translating from English to Tswana (BLEU = 0.32), Xhosa (BLEU = 0.23), and Zulu (BLEU = 0.29). This study set the scene for machine translation in South Africa, and made it very clear that data collection was a big part of the effort needed to improve the quality of translation output.

Another early project, established in 2003 at the University of the Free State (UFS), was the EtsaTrans project, which built on a rule-based legacy system, Lexica. The EtsaTrans system used example-based MT for domain-specific purposes (i.e. for meeting administration at UFS). Initially, it provided only for English, Afrikaans and Southern Sotho, but later developments also aimed to include Xhosa and Zulu (Snyman *et al.* 2007: 225–238).

Another independent study is that of Pilon *et al.* (2010: 219–224), which investigated the possibility of recycling (port/transfer/re-engineer) existing technologies for Dutch to the benefit of Afrikaans, a language closely related to Dutch. They convert (i.e. as a basic form of translation) Afrikaans text to Dutch, so that the Afrikaans text resembles Dutch more closely. After conversion, they use Dutch technologies (e.g. part-of-speech taggers) to annotate or process the converted text, resulting in the fast-tracking of resources for Afrikaans. Their conversion approach is similar to grapheme-to-phoneme conversion, in the sense that transformations are only applied on the graphemic level, and not, for instance, changing word order, etc. Similarities and differences between these two languages are captured as rules and wordlists, and require very few other resources (such as large datasets and probability estimations usually used in statistical MT methods). Although their recycling approach holds much promise for resource development for closely related languages, as an MT approach it is, of course, inefficient, since it does not deal with translation units larger than words. Pilon *et al.* (2010:

219–224) reported a BLEU score of 0.22 for converting Dutch to Afrikaans (compared to Google Translate's 0.40), and a BLEU score of 0.16 for Afrikaans to Dutch (compared to Google Translate's 0.44).

The Autshumato Project

As mentioned earlier, the Autshumato project was the first investment of the South African government to make MT a reality for South African languages. The aim of the project was to develop three MT systems (English to Afrikaans, English to Northern Sotho, and English to Zulu), an integrated translation environment (incorporating the MT systems in a user-friendly editing environment), and an online terminology management system. It was explicated that all resources and systems should be released in the open-source domain.[4] The project was funded by DAC, and executed by CTexT, in collaboration with UP.

The biggest portion of the budget and time for the MT subproject was spent on a drive to gather high-quality parallel corpora for the three chosen language pairs. These efforts commenced in early 2008, and included web crawling (mostly the government domain (gov.za), as this was to be the primary application domain), as well as acquiring personal translation memories, glossaries and other parallel text data from freelance translators and translation companies. Data collection was an on-going effort for the entire duration of the project, and proved to be a more difficult task than anticipated. Web crawling was especially ineffective for Zulu and Northern Sotho, as there simply are not that many parallel texts in these languages available on the web. Translators were also sceptical about sharing their parallel corpora, because of privacy concerns related to their clients. Subsequently the project team at CTexT developed an anonymizer that replaces names of people, places, organizations, monetary amounts, percentages, etc., in order to ensure that confidential information is not included in the parallel corpora; this proved to be an effective measure to convince some translators and companies to make their data available to the project. As a last resort, the project team decided to commission translations and create a custom corpus. This method is by no means ideal and was costly, but delivered excellent quality data as it was translated professionally.

While data collection continued, development of the three MT systems commenced in 2009 with the English–Afrikaans system. Based on the research of Ronald and Barnard (2006), statistical MT has seemed to be a viable option, and it was decided that the Autshumato systems would be based on the Moses statistical MT toolkit.[5] Data resources for all three systems include aligned units (sentences), wordlists and translation memories.

Since Zulu is a morphologically rich language with a conjunctive orthography, it poses many challenges for the development of HLTs in general. The English–Zulu system incorporated a very basic, rule-based morphologic analyser, but as it was still in early stages of development, it hindered development more than it helped. Although Northern Sotho is to some degree easier to process morphologically, performance of the English–Northern Sotho system was only slightly better than the English–Zulu system; compare Table 19.2 for a comparison of the three systems.

Table 19.2 Comparison of three MT systems

Language pair	*No. of aligned units*	*BLEU score*
English–Afrikaans	470,000	0.66
English–Northern Sotho	250,000	0.29
English–Zulu	230,000	0.26

All three systems include a pre-processing module to improve performance (e.g. Griesel *et al*. 2010: 205–210). In later stages of the project, the pre-processing module was further successfully adapted to manipulate the syntactic structure of the English source sentences to be more similar to the Afrikaans target structure, thereby eliminating some of the translation divergences before automatic translation (Griesel 2011). Since data was such a precious commodity in this project, efforts by McKellar (2011) to manipulate available data and selecting the best possible candidate sentences for human translation, were invaluable.

To make these MT systems practically available to translators, an integrated translation environment (ITE) was developed in a second subproject, which commenced in 2010. This computer-assisted translation (CAT) tool supports the workflow by incorporating the MT systems, glossaries, translation memories, and spellcheckers in a single, easy-to-use editing application. The ITE is based on the OmegaT platform,[6] an internationally recognized base for CAT tools. The ITE was designed and developed with continual inputs and evaluations by translators working for government, ensuring that the application would fit well in an everyday working environment. Since one of the functionalities of the ITE is to update translation memories and glossaries as you translate, these valuable resources are also currently being used to continually improve the MT systems.

The third subproject in the Autshumato project was the development of a terminology management system (TMS). One of the important functions of translators working for government is to keep a log of terminology that they come across while translating school books, government documents and pamphlets. This log serves as a way to standardize terms and encourage their use. The TMS is used for the development and management of a database of terms, including their various translations (in the eleven official languages), definitions, usage examples, images, sounds, mathematical equations, and additional notes by terminologists. It is searchable[7] by anyone outside of government, but only DAC translators can add terms or edit information. The database is continually expanded, while quality checks are performed regularly to ensure that the term base remains of a high standard.

In the course of these three subprojects, needs also arose for the development of various other tools, either for use by developers or by translators. These include a pdf-to-text convertor, language identifiers for all eleven official languages, text anonymizers (described earlier), and a graphical user interface for alignment of parallel texts on sentence level. These tools were also released on the official project website (see note 4) under open-source licences.

In addition, the parallel corpora and evaluation sets are also available to download from the project website – also under open-data licences. It is the intention that the Autshumato website, plus the accompanying forum, should become the central hub for the development of MT systems and related tools for the South African languages.

The first phase of the Autshumato project was completed in 2011, and the lessons learned by the development team will serve future projects well. Except for the scientific and technology benefits of the project, one of the most important accomplishments of the project was the engagement of the translation community in the development and eventual uptake of this new technology as an essential part of their workflow. Further uptake is ensured through continual training workshops for government translators, as well as support and maintenance of the existing systems.

Conclusion: the future of MT in South Africa

A few independent research projects and the government-funded Autshumato project have marked the entry of South Africa in the global MT field. Since the conclusion of the first phase

of the Autshumato project in 2011, research and development of MT systems and tools for other language pairs gained momentum. For example, Wilken *et al.* (2012) reported on the development of a baseline English–Tswana MT system, using the same syntactic pre-processing techniques described earlier. Griesel and McKellar (2012) continued work on the improvement of the English–Northern Sotho system by utilizing data from the closely related language, Southern Sotho. Sentences from a Southern Sotho corpus that were classified by the language identification tool as Northern Sotho, were added to the training data. This method improved the translation output quality noticeably, and showed that closely related languages could indeed benefit from pooling available resources.

The fact that the tools available in the Autshumato ITE are available for free in the open-source domain, also led to the development of a community of language practitioners using more sophisticated computer-based translation aids. Training workshops played a vital role in this regard, and also served as a marketing mechanism to draw the attention of businesses and other government departments. Through these workshops it has also become apparent that one of the biggest needs is for customization of translation memories and glossaries.

Resource scarceness is certainly the most pressing drawback for HLT and specifically MT development for the South African languages. As HLT and MT hold the potential to facilitate human–human and human–machine interaction through natural language, the continued investment by government in this budding industry is of vital importance. The South African government's commitment in this regard is illustrated through DAC's funding of the development of an English–Tsonga (a minority language) MT system from 2013 onwards, and with the hope of including more language pairs in future. It is an important step by DAC to ensure the momentum created in the Autshumato project does not go to waste, and to further establish MT as an area of interest for researchers, developers, and end-users.

Notes

1 http://www.southafrica.info/about/people/language.htm#.Ugo-V5LTw6A.
2 We do not give an overview of machine translation aids developed internationally for South African languages. In this regard, suffice to mention that Google Translate included Afrikaans as one of its first fifty languages, and that performance has increased significantly during the first few years. In September 2013 Zulu was released in Google Translate as a potential language, depending on community feedback.
3 http://rma.nwu.ac.za.
4 http://autshumato.sourceforge.net.
5 http://www.statmt.org/moses.
6 http://omegatplus.sourceforge.net.
7 https://ctext-data1.puk.ac.za:8080/tms2.

References

Adendorff, Ralph (2002) 'Fanakalo – A Pidgin in South Africa', in Ralph Adendorff (ed.) *Language in South Africa*, Cambridge: Cambridge University Press, 179–198.

De Schryver, Gilles-Maurice and D.J. Prinsloo (2000) 'The Compilation of Electronic Corpora, with Special Reference to the African Languages', *Southern African Linguistics and Applied Language Studies* 18(1–4): 89–106.

Banking Association of South Africa (2004) *Code of Banking Practice*. Available at: www.banking.org.za.

Bleek, Wilhelm Heinrich Immanuel (1862) *A Comparative Grammar of South African Languages*, London: Trübner & Co.

Botha, Zandré, Roald Eiselen, and Gerhand B. van Huyssteen (2013) 'Automatic Compound Semantic Analysis Using Wordnets', in *Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa*, 3 December 2013, University of Johannesburg, Johannesburg, South Africa, 1–6.

Brits, J.C., Rigardt Pretorius, and Gerhard B. van Huyssteen (2006) 'Automatic Lemmatisation in Setswana: Towards a Prototype', *South African Journal of African Languages* 25: −47.

Griesel, Marissa (2011) 'Sintaktiese herrangskikking as voorprosessering in die ontwikkeling van Engels na Afrikaanse statistiese masjienvertaalsisteem' (Syntactic Reordering as Pre-processing in the Development of English to Afrikaans Statistic Machine Translation), Unpublished MA dissertation, Potchefstroom: North-West University.

Griesel, Marissa and Cindy McKellar (2012) 'Sharing Corpora Effectively between Closely Related Languages: A Pilot Study for Improving the Quality of Sepedi-English Machine Translation Output', Poster presentation at the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), 29−30 November 2012, Pretoria, South Africa.

Griesel, Marissa, Cindy McKellar, and Danie Prinsloo (2010) 'Syntactic Reordering as Pre-processing Step in Statistical Machine Translation from English to Sesotho sa Leboa and Afrikaans', in Fred Nicolls (ed.) *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*−23 November 2010, Stellenbosch, South Africa, 205−210.

Groenewald, Handre J. and Gerhard B. van Huyssteen (2008) 'Outomatiese Lemma-identifisering vir Afrikaans' (Automatic Lemmatisation for Afrikaans), *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies: Special Issue on Human language technology for South African Languages* 29(1): 65−91.

Heid, Ulrich, Danie J. Prinsloo, Gertrud Faasz, and Elsabé Taljard (2009) 'Designing a Noun Guesser for Part of Speech Tagging in Northern Sotho', *South African Journal of African Languages* 29(1): 1−19.

http://autshumato.sourceforge.net.

http://omegatplus.sourceforge.net.

http://rma.nwu.ac.za.

http://www.southafrica.info/about/people/language.htm#.Ugo-V5LTw6A.

http://www.statmt.org/moses.

https://ctext-data1.puk.ac.za:8080/tms2.

Kotzé, Gideon (2008) 'Development of an Afrikaans Wordnet: Methodology and integration', *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies: Special Issue on Human language technology for South African Languages* 29(1): 163−184.

LANGTAG (1996) *Towards a National Language Plan for South Africa: Final Report of LANGTAG*, Pretoria: Department of Arts, Culture, Science and Technology.

McKellar, Cindy (2011) 'Dataselektering en –manipulering vir statistiese Engels–Afrikaanse masjienvertaling' (Data Selection and Manipulation for Statistical English-Afrikaans Machine Translation), Unpublished MA dissertation. Potchefstroom: North-West University.

Morris, Robin (1988) *LEXINET and the Computer Processing of Language: Main Report of the LEXINET Programme, LEXI-3*, Pretoria: Human Sciences Research Council.

Papineni, Kishore A., Salim Roukos, Todd Ward, and Zhu Wei-Jing (2002) 'BLEU: A Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002,* 7−12 July 2002, University of Pennsylvania, PA, 311−318.

Pilon, Suléne (2005) 'Outomatiese Afrikaanse Woordsoortetikettering' (Automatic Afrikaans Part-of-speech Tagging), Unpublished MA dissertation, Potchefstroom: North-West University.

Pilon, Suléne, Gerhard van Huyssteen, and Liesbeth Augustinus (2010) 'Converting Afrikaans to Dutch for Technology Recycling', in *Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 22−23 November 2010, Stellenbosch, South Africa, 219−224.

Pretorius, Laurette and Sonja E. Bosch (2003) 'Finite-state Computational Morphology: An Analyzer Prototype For Zulu', *Machine Translation* 18: 191−212.

Ronald, Kato and Etienne Barnard (2006) 'Statistical Translation with Scarce Resources: A South African Case Study', *SAIEE Africa Research Journal* 98(4): 136−140.

Roux, Justus and Theo du Plessis (2005) 'The Development of Human Language Technology Policy in South Africa', in Walter Daelemans, Theo du Plessis, Cobus Snyman, and Lut Teck (eds) *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium*, 22−23 September 2003, Bloemfontein, South Africa, Pretoria: Van Schaik, 24−38.

Sharma Grover, Aditi, Gerhard B. van Huyssteen, and Marthinus Pretorius (2011) 'The South African Human Language Technology Audit', *Language Resources and Evaluation* 45(3): 271−288.

Snyman, Cobus, Leandra Ehlers, and Jacobus A. Naudé (2007) 'Development of the EtsaTrans Translation System Prototype and Its Integration into the Parnassus Meeting Administration System', *Southern African Linguistics and Applied Language Studies* 25(2): 225−238.

Spiegler, Sebastian, Bruno Golenia, Ksenia Shalonova, Peter Flach, and Roger Tucker (2008) 'Learning the Morphology of Zulu with Different Degrees of Supervision', in Sinivas Bangalore (ed.) *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology (SLT 2008)*, 15–18 December 2008, Goa, India, 9–12.

van Zijl, Lynette and Dean Barker (2003) 'A Machine Translation System for South African Sign Language', in *Proceedings of the 2nd International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa, Afrigraph 2003*, 3–5 February 2003, Cape Town, South Africa, 49–52.

van Zijl, Lynette and Guillaume Olivrin (2008) 'South African Sign Language Assistive Translation', in Ronald Merrell (ed.) *Proceedings of the 4th Annual IASTED International Conference on Telehealth / Assistive Technologies*, 16–19 April 2008, Baltimore, MD, 7–12.

Wilken, Ilana, Marissa Griesel, and Cindy McKellar (2012) 'Developing and Improving a Statistical Machine Translation System for English to Setswana: A Linguistically-motivated Approach', in Alta de Waal (ed.) *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 29–30 November 2012, Pretoria, South Africa.

20

TRANSLATION TECHNOLOGY IN TAIWAN

Track and trend

Shih Chung-ling

NATIONAL KAOHSIUNG FIRST UNIVERSITY OF SCIENCE AND TECHNOLOGY, TAIWAN

Introduction

Living in a technological world, a growing number of people conduct their daily work using the technological tools afforded by computers, the Internet, and advanced information technologies in diverse sectors of life. These persons are closely connected to cloud mining, cloud rating and cloud information exchange and knowledge share in the Internet context on the daily basis. However, in creating a user-friendly, highly communicative Internet environment, translation plays a key role because it helps people break through language barriers and boost their bi/multi-lateral understandings and dialogues across borders. More importantly, diverse translation technologies (TT) can be used to help enlarge the translation scope, and increase the efficiency and cost-effectiveness of language services. All the modern professional translators are well aware of the business profits of TT in the translation industry, but their attitude towards it is different in different countries in different times. Still, translation scholars and instructors have learned the crucial part or/and function of TT in the contemporary translation workflow, but their reception of it and intention to incorporate it into their researches and teachings also vary in different countries and in different times. For this reason, beginning with Taiwan, a research report will be given on the application and teaching of TT in both theory and practice through an empirical investigation. However, before this, the landscape of TT development in Taiwan will be introduced to show its differences from other countries, and a general TT picture can be delineated as the background framework to support the findings in this research.

Today, MT development with technological advances has rekindled worldwide users' interest in it and the development of the integrated MT-TM system, such as Trados, has boosted professional translators' confidence in translation technology application. However, Taiwan started later and also made slower progress in TT development and application than did the United States and Europe.[1] A multilingual MT system, Systran, has been used by the Commission of the European Communities (CEC) since 1976, and 'TAUM METEO system [has been] implemented since 1978 by the Canadian Weather Service for routine translation of weather forecasts from English into French' (O'Hagan 1996: 30). Furthermore, many companies, such as Boeing, BMW, General Motors, and Caterpillar Inc., have developed

controlled English checkers to help author technical texts for effective MT application or/and for efficient post-MT editing (Torrejón and Rico 2002: 108). Also, the cost economics of MT and TM has been supported by some statistical reports from WTCC (World Translation Company of Canada), Systran Institute and Lynn E. Webb (1998–1999).[2] However, when European or/and American companies are enjoying the profits of using MT and MT tools, the majority of translation agencies and technological companies and governmental institutions in Taiwan still rely on human translators to perform daily translation tasks although the scenario has slowly undergone some transformations.

Scantier use of translation technologies in Taiwan can be attributed to the great linguistic differences between Chinese and Indo-European languages. MTs between Chinese and Indo-European languages are poorer than those involving some Indo-European languages. Furthermore, Taiwan's translation industry is less robust or/and less prosperous than Europe's, and there is no urgent need for using technological tools to increase translation efficiency and productivity. It wasn't until 1985 that an English–Chinese MT project started through joint efforts between the Department of Electrical Engineering of National Tsing Hua University and Behavior Tech Computer Corporation. This research resulted in the development of Behavior Tran[3] which is now exclusively used to aid in the translation service offered to the clients by Behavior Design Corporation (BDC). Subsequently, some graduate institutes of computer science and information engineering in Taiwan's universities engaged in MT research and computational linguistics, and the R.O.C. Computational Linguistics Society was formally established in March 1990, assuming the responsibility for organizing some events or conferences related to information technology subjects.

Taiwan's translation software market did not gain public attention until the late 1990s, nearly 30 years behind their counterparts in the US, Europe and Russia (Shih 2002). A chain of MT software emerged in the local market, such as Dr Eye[4] in 1996, JinXlat 1.0 and JinXlat 3.0 in 1998, TransWhiz in 2000, TransBridge in 2001 and others. Dr Eye's affordable price propelled sales to the 200,000 unit level in just one year (Shih 2002), but the poor quality of its automatic translation limited its product lifespan only to two years in the TT market. In 2001, TransWhiz debuted as Taiwan's first MT+TM system developed by Otek Company. However, after the debut of the corpus/statistics-driven MT system, such as Google Translate, and the import of Trados (a renowned TM system) from Germany, TransWhiz immediately gave way. Otek offered free online MT service [Yi-Yan-Tang], but its Chinese–English translations are not as accurate as Google Translate's, so most MT users in Taiwan prefer Google Translate. Furthermore, the companies who use TM tools favor SDL Trados.[5]

To understand the current status and role of MT and TM applications, this research has conducted some surveys to investigate the differences before and after 2000 in Taiwan's translation industry, university language education and academic research. Some of these investigations aim to detect the gap among translator training in Taiwan's universities, translation research concerns and professional translation in domestic translation agencies and technological companies. These three aspects embody interactive relations, because translator training would affect their reception of TT application. Whether professional translators do or do not use translation technologies is connected to the adequacy of their training in TT at school. Inadequate learning of TT at school often makes professional translators turn their backs on TT. Furthermore, research subjects often pertain to the instructor/scholar's teaching, and TT application in industry also affects the content of TT teaching at school. The interactive relations among school, research and industry in TT are inevitable and so this investigation focuses on the three areas. Three research questions (RQ) are raised as follows.

RQ1. What are the differences in TT and human translation (HT) courses offered in both MA (Master of Arts) and BA (Bachelor of Arts) programmes before and after 2000 in Taiwan's universities?
RQ2. What are the differences in TT research in terms of subjects, quantity and publication media before and after 2000 in Taiwan, and what has caused these differences?
RQ3. What are the differences in TT application in Taiwan's translation agencies and technological companies before and after 2000, and what accounts for these differences?

RQ1 asks how far translator training has been modified by integrating TT components, such as MT and TM, into conventional translation teaching at Taiwan's universities. RQ2 examines the evolution of translation research on TT in research subjects over the past several decades. RQ3 explores the ways in which Taiwan's translation agencies and technological companies have integrated TT into their daily workflow.

Translator training in TT

To arrive at a sense of MT/TM (TT) teaching in institutions of higher language education in Taiwan, a 2012 survey of curriculums online[6] has been conducted by examining TT teaching at both MA and BA levels. This survey can be measured against a website survey of the pre-2000 curriculum conducted by Shih (2002) in her research. The subjects investigated include the Graduate Institutes and Departments of Translation and Interpretation (T & I), English or/and Applied Linguistics (E/AL), Foreign Languages and Literature (FLL), and Applied Foreign Languages (AFL). The number of graduate institutes in the post-2000 investigation has increased to thirty because Taiwan's government approved of the establishment of many new MA programmes after 2000 under the policy of promoting higher education in Taiwan.[7] The findings showed that in the pre-2000 BA programmes, MT/TM took up 2.4 percent and HT, 88.1 percent, but in the MA programmes, MT/TM had a slight rise, 7.7 percent, and HT had a slight fall, 69.2 percent. In contrast, in the post-2000 BA programmes, MT/TM doubled the former one (14.3 percent) and HT became essential, reaching (100 percent), but in the MA programme, HT fell to 73.3 percent and MT/TM rose slightly to 16.7 percent. Figure 20.1 shows different percentages of TT and HT courses among the Departments of T & I, E/AI, FLL and AFL at both BA and MA levels before and after 2000.

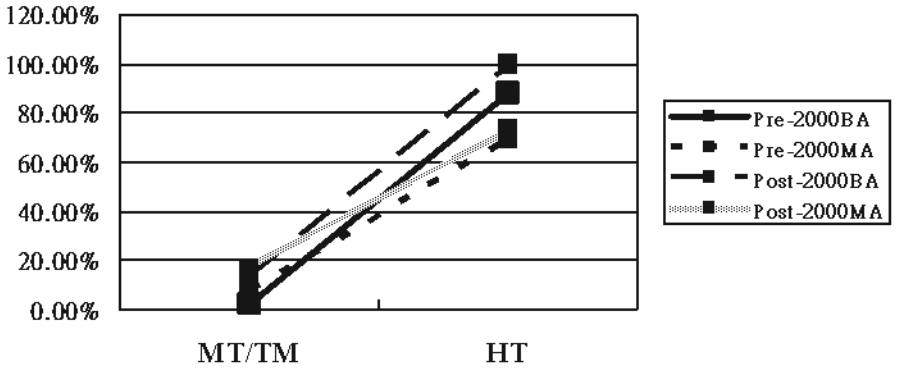


Figure 20.1 A survey of TT and HT courses before and after 2000

As Figure 20.1 indicates, TT (MT/TM) in MA and BA programmes in the pre-2000 survey shows a lower average percentage (3.6 percent) than HT (83.6 percent). This case holds true in the post-2000 survey in which only 15.3 percent of MA and BA programmes together offer TT courses and 87.5 percent offer HT courses. However, there is a higher percentage of TT teaching in both MA and BA programmes in the post-2000 survey than what it was in the pre-2000 survey. This phenomenon suggests that although some language instructors recognize the importance of translation, and view translation as the fifth language skill in addition to the four skills of speaking, listening, reading and writing in foreign language education, they still do not accept technology-enabled translation, nor do they identify computer-aided translation as a specialized subject in translation that is not only fit for translation specialists but also good for language majors. Most language instructors in Taiwan's universities do not learn TT when they receive education for their PhD, so they find it easier to teach HT than to teach CAT. Furthermore, many language or translation instructors in Taiwan do not have a clear notion of TT or adequate knowledge about CAT, and therefore decline incorporating it into their teaching. One more important reason is that translation instructors are TT-phobic. Miss Hwang, Taiwan's exclusive agent of Trados in early times, told me that she regretted seeing that many of Taiwan's translation instructors were lazy about learning Trados or other TT although they knew these tools were useful aids to professional translators.

Academic research on TT

In addition to identifying the trend of TT education, there is a need to map out the evolutionary line in TT-related academic research by conducting an online survey of TT research. The collection includes 8 books, 42 journal papers, 20 conference papers and 42 theses. The survey results serve as an index of the growing profession-oriented concerns with TT either at school or in industry. The subjects under investigation fall into three categories:

1 TT system design and language engineering
2 MT/TM use, MT error analysis and pre/post-MT editing, and
3 TT teaching.

For each category, there are three sub-categories such as MT, TM and MT plus TM. The finding showed that TT system design and language engineering held the highest frequency (53.56 percent) with 49.10 percent of MT researches and 4.46 percent of TM researches. TT application held the second highest (28.56 percent) with 14.28 percent of MT researches, 13.39 percent of TM researches and 0.89 percent of MT plus TM researches. TT teaching showed the lowest frequency (17.8 percent) with 7.14 percent of MT researches, 5.35 percent of TM researches and 5.35 percent of MT plus TM researches. Figure 20.2 shows the results of a website survey of TT-related researches published in books, journals, and presented in conferences and theses. SDLE represents System Design and Language Engineering; APP means application TT and TEA, Teaching of TT.

Generally viewed, the most frequently studied area is system design and language engineering, doubling that of the other two. The main reason is that a distinctly high percentage of theses discuss the issue of MT system design and language engineering. In Taiwan, there are more graduate institutes of computer science and information engineering than those of interpreting and translation, and thus among 42 theses investigated online, a total of 32 address the subject of technological design and computational linguistics. Those graduate students who major in translation and interpretation are not masters of computer programming and cannot deal with

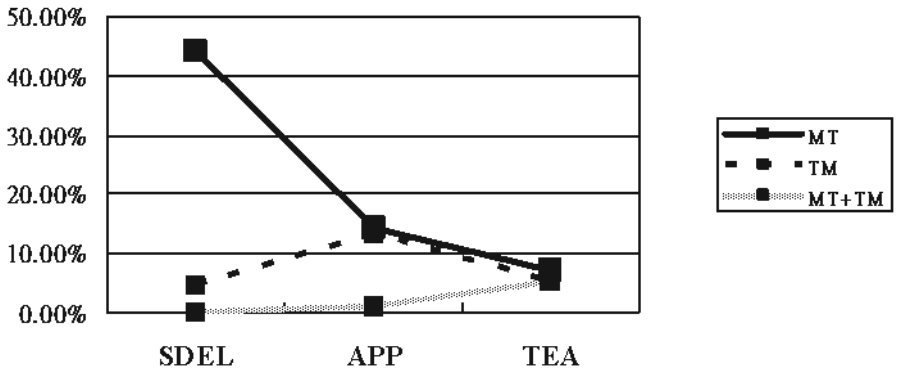


Figure 20.2 A subject-oriented investigation of TT-specific research

technological problems and MT/TM system development. They can only handle pedagogical and practical issues pertaining to TT under the supervision of the instructors who are not masters of language engineering. Thus, there is a clear division between two camps, with one focusing on information retrieval from the corpus or MT system design with quality improvement solutions, and the other emphasizing the identification of pre/post-MT editing rules for better MT performance, and relevance of MT errors or/and TM alignments to text types and linguistic problems. Only one thesis deals with an online survey of how freelance translators plus few in-house translators apply TM systems and term bases.

In addition, the highest frequency of SDLE can be attributed to the majority of journal papers dealing with the development of MT systems and computing linguistics. It is found that journal papers are evenly split with 21 papers discussing MT system design and engineering, and the other 21 papers handling MT editing or error analysis and teaching. Interestingly, the papers addressing MT/TM teaching are slightly higher than those on editing and error analysis. My inference is that after more masters programmes in translation and interpretation were offered in universities after 2000 in Taiwan, more translation instructors noticed the importance of TT and so started to propose TT-aided translation teaching. The result of their teaching research wass presented at some conferences and wass finally published in journal papers. Scanning the papers on TT application, we find that a wide range of topics cover contrastive analysis of MT systems, development of knowledge-based MT systems, sentence-based statistical MT models, production and consumption of MT, impacts of the technological turn, teaching text types with MT errors analysis and post-MT editing, teaching the concept of equivalence using TM tools, the shift in controlled English norms for different MT systems, and corpus-based study of differences in explicitation between literature translations for children and for adults, a teaching challenge to TM, the constructivist educational effectiveness of TM-aided specialized translation and others.

Another reason for the highest ranking of SDLE is that 50 percent of books in the survey address MT or TM system design and engineering, and the other half pertain to editing, error analysis and teaching. Four books in the area of SDLE provide a historical sketch of MT system design, approaches and developments with an introduction to the basic functions, practical problems, strengths and weakness of TT application. One of them is the translated book by W. John Hutchins, *Machine Translation: The Past, Present and Future* (1993), published by BDC. It is translated by the MT system, Behavior Trans, and post-edited by some in-house translators. This case suggests that although nonfiction is the right text type for MT application, the MT

output still requires editing prior to publication. The books in the areas of application and teaching include *Computer-aided Translation* (Shih 2004) and *Helpful Assistance to Translators: MT and TM* (Shih 2006). These two books shift the main focus from technical issues to pedagogical ones although they also give a brief overview of MT/TM functions and operational procedures. *Computer-aided Translation* elaborates the three-stage MT editing in register, discourse and context areas with supportive examples, and provides some exercises for student practice. The other one supplements some theoretical discussions, and reports some case studies by applying MT and TM in a translation class, such as 'Using Trados TagEditor in the teaching of web translation', 'Using the tool of Trados WinAlign to teach the translation equivalence concept', 'The use of TM as the scaffold in translation teaching' and others.

The two books, *Real-time Communication through Machine-enabled Translation: Taiwan's Oracle Poetry* (Shih 2011) and *New Web Textual Writing: Fast Communication across Borders* (Shih 2013), discuss cost-effective benefits of editing source texts in controlled Chinese for multilingual machine translations, with the former using Taiwan's oracle poetry as examples, and the latter, the web texts on Taiwan's festivals, folk culture and company profile as examples. A set of pre-editing rules is designed based on the linguistic differences between Chinese and English, such as clarifying the grammatical features of words by using *-de* before an adjective and *-di* before an adverb, using an article or quantifier, using more passive voice than active voice, and others. Idiomatic expressions must be adapted in controlled Chinese, and Chengyu or/and fixed four-character phrases must be paraphrased. The finding shows that oracle poetry and allusive stories after controlled editing have dramatically improved semantic clarity, grammatical accuracy and pragmatic appropriateness of their multilingual machine translations. In the two books, Shih (2011) emphasized that controlled Chinese was a new concept in the Chinese community and its use could meet some opposition, but this new language was designed for machine-friendly application and MT-enabled communication, not for daily writing. Just as we can have multiple choices for daily necessities, we can also be allowed to choose one language customized to optimize the effectiveness of MT application on the Internet.

In addition to a subject-oriented survey, the statistical results of varied channels of publicizing TT research need to be reported. In all the publications before and after 2000, journal papers and theses that address TT showed the highest percentage (37.5 percent); conference papers or/and presentations ranked second (17.8 percent), and books, third (7.1 percent). Figure 20.3 shows the statistical results of TT-related researches published in different media from the past to the present in Taiwan.

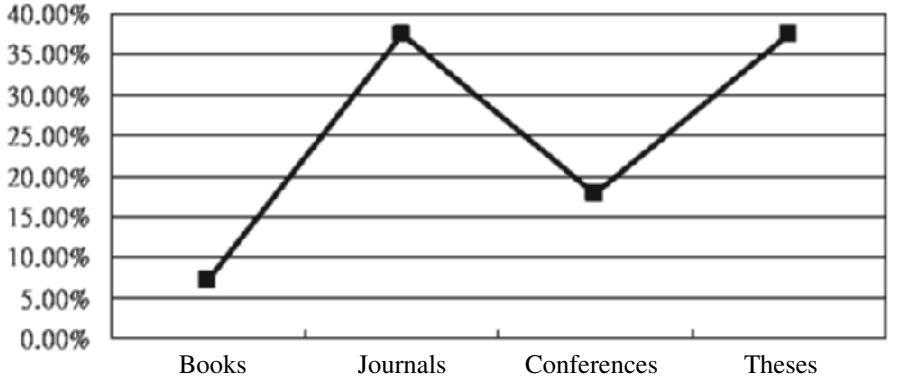


Figure 20.3 Media-oriented investigation of research on TT

The high percentage of journal papers is attributed to the result of a website survey. Many authors of journal articles in Taiwan are asked by publishers to endorse an agreement to have papers digitalized online for public access and information sharing. In contrast, conference papers are only collected in the proceedings and are not published with a copyright of ISBN. If conference papers are not uploaded by conference organizers, they cannot be accessed by this online survey. Furthermore, theses on TT hold the same high percentage as that of journal papers because Taiwan's Ministry of Education (MOE) requires all theses in Taiwan to be uploaded onto the website of National Digital Library of Theses and Dissertations. In short, theses and journals are two main sources of TT data in Taiwan. However, the contents of the theses cannot be accessed without the permission of authors.

Regardless of varied ways of publicizing TT research results, the publications showed a frequency difference according to their type and at different times and thereby a chronological survey was needed. The finding showed that in the post-2000 period, the number of journal papers on TT had risen significantly from 8 to 34; conference presentations and papers, from 1 to 19, and theses from 16 to 26. In the pre-2000 period, books on TT took up 13.79 percent (4/29); journals, 27.58 percent (8/29); conferences, 3.44 percent (1/29), and theses, 55.2 percent (16/29). In contrast, books on TT took up 4.8 percent (4/83); journals, 40.96 percent (34/83); conferences, 22.89 percent (19/83), and theses, 31.32 percent (26/83) after 2000**.** The distinctive difference was that theses on TT ranked first before 2000, but journal papers on TT showed the highest percentage after 2000. Furthermore, conferences on TT showed the lowest percentage before 2000, but books on TT showed the lowest one after 2000. Figure 20.4 shows the result of a chronological survey of the TT researches before and after 2000.

In spite of the different TT publications before and after 2000, the average percentage in the post-2000 period remains higher than that in the pre-2000 period. Apparently, the concept of TT and its application before 2000 were not widespread and research issues on TT were limited to MT or relevant ones, but after 2000 the TM issue was supplemented and thereby the amount of research doubled the previous amount. More importantly, more satisfying MT performance due to technical improvements has increased the users' faith and rekindled their interest. This reason accounts for an increase in the number of TT researches and publications, particularly on the subject of controlled language and effective MT editing for the creation of comprehensible multilingual machine translations. Overall, this phenomenon suggests that many translators, scholars and translation instructors in Taiwan have started to realize some benefits of technology-enabled translation in recent years.

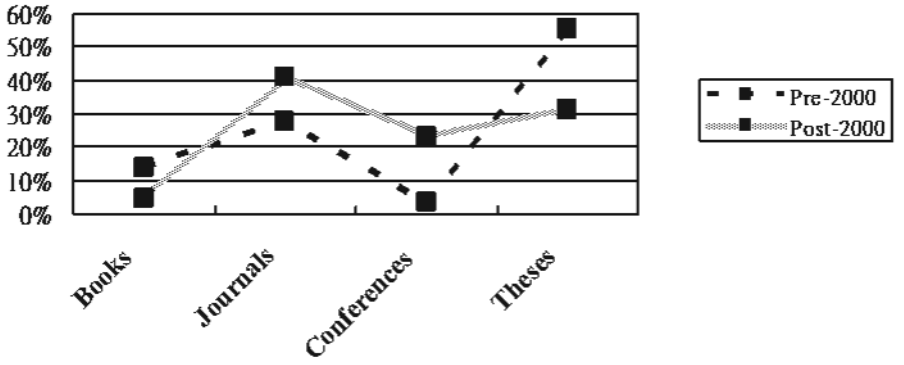


Figure 20.4 A chronological investigation of TT publications before and after 2000

Professional application of TT

An investigation of the professional application of TT after 2000 targets 19 translation agencies and four technological translation companies, such as Otek International Incorporation, Fohigh Technological Translation Company, Shinewave International Incorporation and Syzygy Information Services Company in Taiwan. The finding shows that there were 14 users of MT tools and 11 users of TM tools. These companies used MT tools for different purposes. For example, BDC used Google Translate for quality assessment against Behavior Trans system; Syzygy Company used Google Translate for accuracy tests, and Ests Company, Otek International Incorporation and Ya-Hsin Company used Google Translate or TransWhiz for gist translation and post-MT editing. Nine users viewed MT systems, such as Google Translate, as an online dictionary for specialized term lookup and did not rely on the quality of MT outputs. Figure 20.5 shows different purposes of using MT tools.

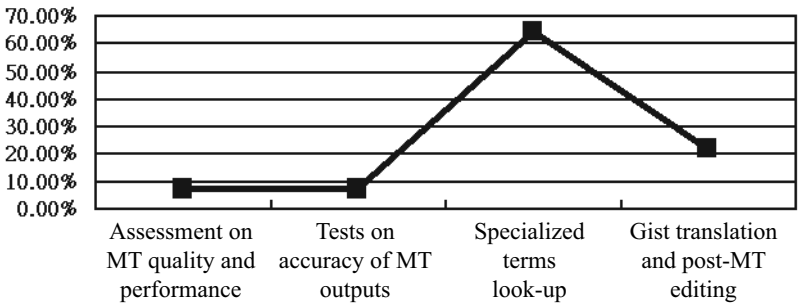


Figure 20.5 Different purposes of using MT in Taiwan's translation agencies and companies in the 2012 survey

The figure above shows that over half of MT users view MT tools as an alternative to the dictionary and do not use MT outputs as the scripts for post-MT editing. Their opinion is that current Chinese–English MT performance is not good enough for gist translation and post-MT editing. In light of the limitation, Shih (2011) has proposed pre-MT editing for effective MT application, particularly for the creation of multilingual translations.

With respect to TM application, the most frequently used tools are Trados, TM/Win and some localization software programs such as Catalyst, Passolo, RCWin Trans, Microsoft Helium, Microsoft LosStudio, Logoport and others. Since quality assurance is a key part in the project management of localization industry, ApSIC Xbench 2.9 serves as a favorable and helpful tool, free and accessible on the Internet. When TM users were asked to evaluate the performance of SDL Trados in my telephone interviews, six users gave 80 points; two users, 85; two users, 70; and one user, 90. The average score is 81. This statistical result suggests that the majority of TM users in Taiwan are satisfied with TM performance, but they also expect some technical improvements and reduction in price. One user complained that the speed was slow when the processed file was extremely big, and negative responses involved occasional breakdown, inadequate fuzzy matches, high price, no entire textual translation and complex operating procedures. Lower capital investment and friendly hands-on experience are users' primary concerns.

Another finding showed that among the 11 agencies and companies that did not use TM tools, six (6/11=54.5 percent) held that artificial intelligence could never compete with the

human brain, so human translation would always be more reliable. Two of them (2/11=18.2 percent) claimed that they handled only small amounts of translations and thereby did not need TM or corpus. Four (4/11=36.4 percent) maintained that their translations were not highly repetitive in content or/and sentence patterns, so human translation was faster. It is noted that most of the subjects in the present investigation are small translation agencies, and for this reason they think it is not worth an investment in the costly TM software.

A comparison between the 2012 survey and the 2001 survey of TT application (Shih 2002) shows that the percentage of MT and TM users about 11 years ago (28.6 percent) is lower than what it is in 2012 (54.34 percent). In the 2001 survey on the pre-2000 MT and TM use, there are 6 MT users (6/14=42.9 percent) and 2 TM users (2/14=14.3 percent). In contrast, the 2012 survey on the post-2000 MT and TM use shows that MT users rise from 6 to 14, and TM users, from 2 to 11. Furthermore, the gap between MT and TM users in the 2001 survey (28.6 percent) is higher than that in the 2012 survey (13.1 percent), giving evidence of an increase in the number of professional MT and TM users in recent years. Figure 20.6 shows the difference in the results of the two surveys.

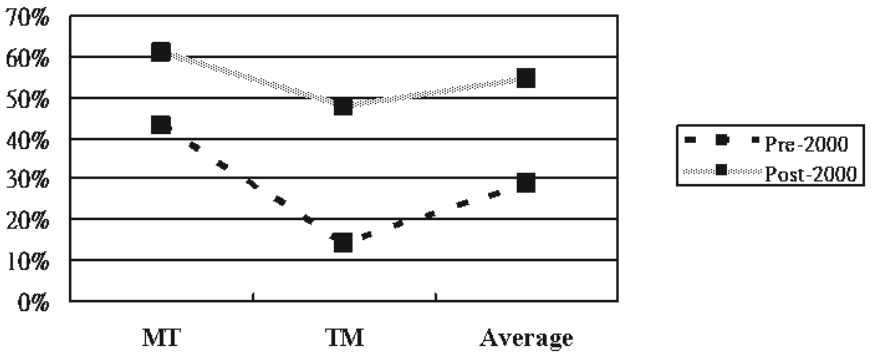


Figure 20.6 The differences in MT and TM use between 2001 and 2012 surveys

Although the average percentage of TT users after 2001 is higher than before, the percentage of MT users in both surveys is higher than that of TM users. One possible reason is that MT tools are much less expensive than TM tools, and Google Translate offers a free automatic translation service. In contrast, sophisticated TM tools such as SDL Trados are costly and post-sale training is also needed because of their complicated operational procedures.

In Shih W-M's (2007) survey, the percentage of TM tools used by Taiwan-based translators is lower (16 percent) than Shih's 2012 survey (47.8 percent) but higher than her 2001 survey (14.3 percent). This implies that from 2007 to 2012, there are a growing number of freelance translators or translation agencies willing to invest in costly TM tools even though they handle only small amounts of translation. However, TT application remains inadequate in Taiwan's translation industry, suggesting that the disconnection from the international translation world has blinded professional translators to the fast-changing international translation market. The narrow scope of text type and limited language pairs of translations they handle on a regular basis is a key barrier to the use of TM tools in daily translation work. In addition, many of Taiwan's local enterprises are not internationally marketed and their user manuals, product instructions and relevant documents do not have to be multilingual. For this reason, they do not resort to the localization company for the service of multilingual translations. Some renowned international companies such as Asus have their branch localization company (e.g.,

Shinewave International Incorporation) to help in handling multilingual translations, and they also try to customize their own TM for their clients. Their TM is confidential and is not released on the market.

Conclusion and suggestions

A combination of TT surveys in the areas of application, research and teaching has shown some tracks and trends, including:

1 optional TT training, not required
2 research preferences for technical TT, not TT-aided training
3 misconception and ignorance of MT and TM education, and
4 underdeveloped localization industry.

These weaknesses can be diagnosed and improved with some possible solutions as follows.

Joint-lecture and on-the-job TT training

In answer to RQ1 about TT courses offered at Taiwan's universities, the result indicates that translator training generally uses the conventional method of HT, and most MT and TM courses are electives, not required. This suggests that most T & I instructors still view technology-enabled translation as a supplementary course, not as a necessity for translation and language majors. This is partly due to the instructors' negligence of the international trend of technology-enabled translation in the real working scenario, and partly to their identification of translation as a linguistic subject or an art rather than as a practical, professional science that requires market-oriented training.

Seeing the low percentage of TT training in language and translation education in universities, some solutions are proposed and prepared to take action for a change of the status quo. Taiwan's MOE sponsors Excellence-in-Teaching projects at universities and encourages the joint-lecture practice. When the grant-funded projects are in force, TT professionals in the translation industry can be invited to lecture in class as the translation instructor's partner. Shih (2006: 357) proposed that sales managers or technicians 'could be invited from software companies to teach translation [students] how to operate MT and TM tools'. As a follow-up cooperation, instructors and professional translators in the localization industry can work together to design technology-enabled translation software. In the translation classes at Lunghwa University of Science and Technology, National Sun Yat-sen University and National Kaohsiung First University of Science and Technology, project managers and professional translators were invited from the localization companies to give lectures and all students were positive about the joint lectures, agreeing that this teaching method provided them with a window on the real translation world. Understanding the employment requirements for translation professionals is a catalyst to motivate them to learn the practical TT tools at school and push them to receive an internship outside the class. Training in TT gives students market-oriented expertise and enhances their employment prospects in the international translation market.

Additionally, 'local governments could consider funding and sponsoring some in-service training activities on MT and TM' and 'the school or technological companies could organize seminars and conferences on the issues of MT and TM to disseminate the knowledge of technology-enabled translation' (Shih 2002: 358). The Taiwan Association of Translation and

Interpretation has provided training in SDL Trados to the public in October 2012, and many translation instructors and freelancers participated in the event with much ardor. Nevertheless, the training time was too short to give the participants adequate practice and some teachers complained about the complicated procedures of operating the TM tool. Since this is the first time free training in TT has been provided to the public audiences in Taiwan, its flaws can be corrected and future training sessions could be more rewarding and have a positive reception. Furthermore, 'internship or on-the-job training programmes' can be provided by collaborating with software or technological companies (Shih 2002: 358). However, many technological or localization companies do not want to take interns because of the confidentiality of company documents, and some clients also forbid them to do so. Finally, the publication of more books on TT-related pedagogy is encouraged. Sufficient teaching resources on TT would encourage more teachers to teach TT in their translation or/and language classes.

Regular MT/TM conferences with the help of government's fund grants

In response to RQ2 concerning differences in translation research before and after 2000, it is gratifying to see that the quantity and scope of TT-related research has increased after 2000. This means that an increasing number of translation scholars and/or instructors have shown greater interest in TT and are more devoted to TT research than before. However, the percentage of research on language engineering and technical solution remains overwhelmingly higher than that of TT teaching. Many language and translation instructors continue to assume that TT is immature and unacceptable, and they cannot trust the accuracy and quality of the computer-aided translation. Since many language instructors still view translation as art, they devalue the translation created with the aid of technology as unreadable.

To strengthen the confidence in language and translation instructors and change their bias about TT, some conferences about MT/TM technology can be regularly organized in Taiwan's universities with the help of the government's fund grants. Adequate information input about TT can change the language and translation instructors' concept and this can urge them to introduce TT to their students. Finally, instructors will be more willing to study technology-enabled translation and teach it in class. Academic research needs a connection to the global translation world, but most scholars lack the momentum to act upon the concept. One more important point that translation scholars must know is that translation pedagogy with the help of TT is not a rejection of conventional translation teaching; rather it enriches its teaching context through integration of old and new.

TT education as one of the criteria for university evaluation

With respect to the evolution of TT application in the translation industry, the percentage of MT and TM use in the 2012 survey is higher than the corresponding percentages in the 2005 and 2001 surveys, and the rate of TT application remains inadequate in Taiwan's current translation industry. This phenomenon can be attributed to the lesser internalization or globalization of Taiwan's local businesses, which are also less likely to consider using TT to help handle their translation. Furthermore, far less use of MT tools than TM tools results from the role of a single language vendor (SLV) that Taiwan's localization companies have been playing, not that of a multiple language vendor (MLV). Since their service scope is limited to Chinese to/from English translation, they do not think of using controlled language (CL) and MT application to promote translation efficiency. Many European companies act as MLVs, but Taiwan's localization companies do not because they find it very hard to get competent

multilingual translators in Taiwan. Several languages spoken by different regions in Europe share similar linguistic features as they belong to the same Indo-European linguistic system and the European multilingual environments easily develop native speakers into those who are able to simultaneously use two or three languages. In contrast, the limited language environment plus the inadequate training in the translation of non-English languages at universities results in the severe shortage of multilingual translators in Taiwan. Various factors concur that the development and application of MT tools causes concern among translation professionals in Europe, but the majority of translators in Taiwan have been mistrustful of MT application and inclined to overlook it.

In Taiwan, one reason for the severe shortage of MT education is that many instructors have a misconception of MT and think MT training is useless because MT tools cannot be used to translate literary works. Thus, MT education should be first given to instructors, making them understand that no dish can be washed using the washing machine. Just as a dish can only be washed using the dishwasher, the MT tool can only be used to translate informative texts, such as user's manuals, product instructions and relevant others, for effective communication, not for aesthetic appreciation or replication of the author's creative style. MT education should emphasize the use of controlled language for improved MT performance, and the use of customized post-MT editing skills to meet the diverse functions of translations. In this respect, it is important to educate the translation and language instructors about MT and TT-relevant knowledge. One of the effective ways of doing this is the government's intervention by including TT education as one of the criteria for university evaluation. This instrumental purpose will motivate schools to stress TT education and instructors will teach students the functions, strengths and weaknesses of MT, TM and others. After TT education gains popularity, the governmental policy can be modified and its intervention can be reduced or eliminated.

Government incentives to boost the growth of localization industry

With regard to the use of TM tools, cost, complicated operational procedures and lack of training are common factors to hinder wide application. To raise the application rate, TT can be technically improved on one hand, and users must also learn how to use it appropriately on the other. If employees in industrial companies or translation agencies have not already received TT training in school, they must receive in-service or on-the-job training. After they have learnt the genuine benefits of using TT, professional translators would cease to resist it.[8] However, many companies in Taiwan would not invest in this costly training and therefore they cannot employ qualified TT experts for their work. The localization industry is shrinking in Taiwan and the number of localization companies (fewer than 10) is fewer than China (about twenty) and European countries.

To boost Taiwan's localization industry, the government can offer some incentives such as tax reduction, favorable loan interest rates and others. Also the government can intervene to set up an official committee for the localization industry and therefore the translation professionals can have a venue for exchanging their TT experiences and relevant information. Currently, localization companies in Taiwan maintain a competitive relationship without any interaction or, dialogue, not to mention any cooperation. Domestic competition only makes their business decline. The government should do something to help the local translation industry develop into an international-scale one, and the promotion of their business status would compel them to take notice of TT application. I would contend that if more of Taiwan's localization businesses, at the request of international clients, had to create product instructions

in multiple languages, the use of MT and TM tools would gain increasing attention, and more professional translators would apply TT tools to their daily work.

In fact, MT and TM tools are not a panacea, but their use can help to cope with an increasing quantity of translations. The economic profits of TT application are doubtless acknowledged. Perkins Engines has 'saved around 40000 pounds on each diesel engine manual translation using the Weidner system to translate technical documentation' (Hatim and Munday 2004: 216), and as of 1990, Mėtėo 'was regularly translating around 45,000 words of weather bulletins every day' (ibid.). Use of MT/TM systems is really a cost-effective gateway to the professional world of localization. Insofar as the technological trend has been swept through the international translation market, I think Taiwan's local professional translators, even without the help of the government, must find a way to integrate TT into their workflow to enhance their future competitive edge under the mantle of globalization.

In the conference organized by Taiwan's National Academy for Educational Research, Lin and others (2012) presented a report entitled 'A study of translation development strategies in Taiwan' and spoke of the development of diverse bilingual corpora as one of the important and practical strategies for boosting Taiwan's translation industry. This proposal in the governmental blueprint for Taiwan's future translation industry shows upper management's increasing attention to the importance of TT application. It is expected that translation practitioners and professionals can benefit from the corpus use to enhance their service quality and strengthen their image of professionalization. Lin's proposal concurs with Chen's (2012) finding in his lecture on '2012 Taiwan translation and interpretation industry survey report' when he claimed that more than 80 percent of Taiwan's respondents in his survey expected Taiwan's government to provide them with diverse bilingual corpora for free use. This case suggests that since no specialized English to/from Chinese bilingual corporas are released on the local market, professional translators only turn to seek help and support from the government. Governmental intervention is also expected to help solve the problem of textual copyrights when noncommercial huge corpora are developed for public use.

Actually, no approaches or options, however sophisticated, can provide a once-for-all solution to all problems arising from the shortage of TT professionals, and TT-relevant teaching and research in Taiwan. Remodeling and modification are confronting and challenging translators, instructors and scholars. According to J. Abaitua (2002), the information technology and localization industries are evolving rapidly and translators need to evolve with them. Dale Bostad claimed that 'if you could not beat [translation technologies], join them' and 'if you cannot strike [translation technologies], connect them' (quoted in Budiansky 1998: 84). Thus, a greater concern for and more dedication to TT application and development is urgently required in Taiwan.

Notes

1 European countries, the United States and Canada are far ahead of Taiwan in developing MT, TM systems and integrating them into the real translation working setting for time and cost benefits. A litany of MT success stories include TAUM METEO system in the Canadian Weather Service, Systran in France, U.S. Air Force and Xerox in America, and Logos by Lexi-tech in France and others (O'Hagan 1996; qtd. in Shih 2002).

2 WTCC (World Translation Company of Canada) released the results of the English–French Systran II's application in 1980, [claiming] that "the HT cost for 100 words was US$ 16.50, but the MT cost for the same 100 words was US$ 8.56" (Chen and Li 1991; qtd. in Shih 2002: 214). Systran Institute Gmbh's estimate maintained that 'the HT cost of 100 words was US$9.53 whereas the MT cost of the same 100 words was US$3.39' (Chen and Li 1991; qtd. in Shih 2002: 214). As Lynn E. Webb (1998–1999: 32 and 35) put it, 'company savings after using TM tools' were $3,360 for 'the

translation of 40,000 words' and 'translation agency savings' were '$3,503 to $5,215' for the same number of words.

3 Behavior Tran, not commercially released on Taiwan's market, has been used by BDC in translating computer manuals, user guidelines and books or articles on electrical engineering, mechanical engineering, aviation and psychology (Zhang and Chen 2001; qtd. in Shih 2002: 49).

4 Dr Eye was sold at prices ranging from NT$399 to NT$900 and thereby caused a big sale, roughly one unit for every hundred Taiwan residents around 1997–1998. The price of TransWhiz was more costly, NT$ 2990 without TM and NT$78000 with TM embedded in the MT system (Shih 2002).

5 Trados users take up more than 70 per cent of companies worldwide according to Lisa's report (2002–2004).

6 Among various forms of survey, a survey of curriculum design online is the easiest method, but it is hard to identify some courses in which TT training is only a part of the content, and is not shown as the course title suggests. Thus, the course whose title does not give any clue is deemed as only using the HT method without MT or TM teaching.

7 Before 2000, there were only two MA programmes in translation and interpretation offered by universities such as Fu-Jen Catholic University and Taiwan Normal University. After 2000, five MA programmes in T and I were offered by the universities such as Taiwan University, Kaohsiung First University of Science and Technology, National Changhua University of Education, Chang Jung Christian University and Wenzao Ursuline College of Languages. MA programmes in Applied Foreign Language and Literature have also dramatically increased in Taiwan after 2000.

8 It needs to be clarified that if translation achieves the goal of aesthetic appreciation, it serves as art. In contrast, if large numbers of translations need to be processed for information communication, it can be viewed as science, and the use of technological tools would shorten the turnaround time of translation, and boost its productivity. Other benefits include terminological consistency and no need for translating similar or the same sentences again.

Bibliography

Abaitua, Joseba (2002) 'Is It Worth Learning Translation Technology?' in *Proceedings of the 3rd Forum on Translation in Vic. Training Translators and Interpreter: New Directions for the Millennium*. Available at: http://sivio.deustro.es/abaitua/konzeptu/ta/vic.htm.

Behavior Design Corporation, Centre for Translation and Publication 致遠科技股份有限公司翻譯出版中心 (tr.) (1993) 《機器翻譯：過去、現在、未來》 (*Machine Translation: The Past, Present and Future*), Hsinchu: Behavior Design Corporation.

Budiansky, Stephen (1998) 'Lost in Translation', *The Atlantic Monthly* 282(6): 80–84.

Cervantes, Miguel de (2005) *Don Quixote*, Edith Grossman (trans.), New York: Harper Perennial.

Chen, Han-bin (2009) 'Learning Bilingual Linguistic Reordering Model for Statistical Machine Translation', unpublished Master's thesis, Department of Computer Science, National Tsing Hua University, Taiwan.

Chen, Pin-chi (2006) 'A Study of the Fuzzy Match Function in CAT Software Used in Technical Translation in Taiwan', unpublished Master's thesis, Department of Applied Foreign Language, National Taiwan University of Science and Technologies, Taiwan.

Chen, Tze-wei (2012)〈2012臺灣翻譯產業調查分析〉(2012 Taiwan Translation and Interpretation Industry Survey Report), Paper presented at 2012 International Conference on Translation and Interpretation: Quality Enhancement and Professionalization (2012 臺灣翻譯研討會－翻譯專業發展與品質提升), 23 November 2012, Development Centre for Compilation and Translation, National Academy for Educational Research.

Chen, Zi-ang 陳子昂 and Li Wei-quan 黎偉權 (1991)〈機器翻譯系統現況與展望〉(Current Status and Prospects of Machine Translation Systems), 《CIO 資訊傳真周刊》(*CW Infopro Weekly*) 143: 166–174.

Hara, Hiroyuki (2001) 'The Difference within the Sameness: A Comparative Study of MT Translatability across Medical Genres', unpublished Master's thesis, Graduate Institute of Translation and Interpreting, National Kaohisung First University of Science and Technology, Taiwan.

Hatim, Basil and Jeremy Munday (2004) *Translation: An Advanced Resource Book*, London and New York: Routledge.

Hsieh, Hung-chen 謝紅貞 (2008)〈電腦輔助翻譯軟體之翻譯記憶及匹配功能應用在西班牙文與中文新聞翻譯之可能性〉(The Possibility of Using Translation Memory and Alignment as

Computer-assisted Translation Tools for Spanish and Chinese News Text Translation), unpublished Master's thesis, Department of Spanish Language and literature, Providence University, Taiwan.

Hutchins, W. John (1986) *Machine Translation: The Past, Present and Future*, West Sussex, England: Ellis Horwood Limited.

Lee, Jason 李家璿 (2009)〈全自動機器翻譯加後編輯與人工翻譯之比較〉(A Comparative Study of Fully Automatic Machine Translation with Post-editing and Human Translation), unpublished Master's thesis, Graduate Institute of Interpreting and Translation, National Taiwan Normal University, Taiwan.

Lin, Ching-lung 林慶隆 Chen, Yun-xuan 陳昀萱 & Lin, Sinn-cheng 林信成 (2012)〈臺灣翻譯發展策略之探討〉(A Study of Translation Development Strategies in Taiwan) in *Proceedings of 2012 International Conference on Translation and Interpretation: Quality Enhancement and Professionalization*, Development Centre for Compilation and Translation, National Academy for Educational Research, Taipei, Taiwan, 1–21.

Lin, Chuan-jie 林川傑 (1997)〈國語-閩南語機器翻譯系統之研究〉(The Study of a Mandarin-Taiwanese Machine Translation System), unpublished Master's thesis, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan.

Liu, Sheng-liang 劉聖良 (1996)〈機器翻譯中多字動詞問題之研究〉(A Study on the Problems of Multi-word Verbs in Machine Translation), unpublished Master's thesis, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan.

O'Hagan, Minako (1996) *The Coming Industry of Teletranslation: Overcoming Communication Barriers through Telecommunication*, Bristol: Multilingual Matters Ltd.

Shih, Chung-ling (2002) *Theory and Application of MT/MAHT Pedagogy*, Taipei: Crane Publishing Co., Ltd.

Shih, Chung-ling 史宗玲 (2004)〈電腦輔助翻譯: MT & TM〉*Computer-aided Translation MT and TM*, Taipei: Bookman Books Ltd.

Shih, Chung-ling (2006) *Helpful Assistance to Translators: MT and TM*, Taipei: Bookman Books Ltd.

Shih, Chung-ling 史宗玲 (2011)《機器翻譯即時通：臺灣籤詩嘛ㄟ通》(*Real-time Communication through Machine-enabled Translation: Taiwan's Oracle Poetry*), Taipei: Bookman Books Ltd.

Shih, Chung-ling 史宗玲 (2013) 《網頁書寫新文體：跨界交流「快譯通」》(*New Web Textual Writing: Fast Communication across Borders*), Taichung: White Elephant Ltd., Company.

Shih, Wei-ming 施偉銘 (2007)〈台灣地區筆譯工作者運用翻譯工具之現況〉(Translation Tools: A Survey of Their Adoption by Taiwan-based Translators), unpublished Master's thesis, The Graduate Institute of Translation and Interpretation in National Taiwan Normal University, Taiwan.

Torrejón, Enrique and Celia Rico (2002) 'Controlled Translation: A New Teaching Scenario Tailor-made for the Translation Industry', in *Proceedings of the 6th EAMT Workshop—Teaching Machine Translation*, Centre for Computational Linguistics, UMIST, Manchester, UK, 107−116.

Webb, Lynn E. (1998−1999) 'Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis', unpublished thesis, German Graduate Division of Monterey Institute of International Studies.

Yu, Jian-heng 余鍵亨 (2005)〈機器翻譯系統為本之雙語網頁對應〉(Automatic Alignment of Bilingual Web Pages Using Machine Translation Systems), unpublished Master's thesis, Department of Computer Science, National Tsing Hua University, Taiwan.

Zhang, Jing-xin 張景新 and Chen Shu-juan 陳淑娟 (2001)〈機器翻譯的最新發展趨勢〉(The Latest Trend in Machine Translation). Available at: http://nlp.csie.ncnu.edu.tw/~shin/bdc/doc/INTRO.201.

21

TRANSLATION TECHNOLOGY IN THE NETHERLANDS AND BELGIUM

Leonoor van der Beek

INDEPENDENT SCHOLAR

Antal van den Bosch

RADBOUD UNIVERSITY NIJMEGEN, THE NETHERLANDS

Introduction

The Netherlands and Belgium share a history in translation technology research and an interest in Dutch, the language spoken by most of the inhabitants of the Netherlands and by about 60 percent of Belgium's population, mostly concentrated in the Dutch-speaking region of Flanders. Translation technology research and development came, went, and came back again in a span of three decades. Early Dutch and Belgian computational linguistics research was boosted significantly by roles that researchers and teams took in national and international knowledge-based machine translation system development programmes in the 1980s. Industrial spin-offs were created, remainders of which can still be found in present-day industrial translation technology providers. Currently, research follows the typical trends in translation technologies: statistical machine translation and hybridizations with linguistic knowledge, and business-oriented translation process automation.

In this overview we begin with a historical listing of the most visible machine translation projects in the two countries: the international projects Eurotra and METAL, and the Dutch industrial projects DLT and Rosetta. We then sketch the current state of affairs in academic research in the Netherlands and Belgium. Much of the current work focuses on statistical machine translation systems and translating between closely related languages such as Frisian and Afrikaans, but there is also work on building translation memories from monolingual corpora and bilingual lexica, and usability of translation tools such as post-editing software. We look at industry, and observe that most current industrial activity is located in Belgium, offering translation, terminology, and localization services.

We end our overview with a note on the position of the Dutch language in present-day translation technologies. It occurs as a language in about 15 percent of currently commercially available machine translation systems listed in the EAMT software compendium. Being one of the official languages of the European Union, there is a considerable amount of public parallel data available. With several research teams working on machine translation and with a host of

active SMEs, translation technology for Dutch is in good shape given its approximate 23 million speakers and its ranking in the low top-30 of numbers of native speakers of languages worldwide.

Early days

Dutch and Belgian academics were not among the global pioneers in translation technology. Mathematics Professor Adriaan van Wijngaarden did advocate a collaboration between mathematicians and linguists in order to jointly develop automated translation systems as early as 1952, just months after the very first workshop on machine translation at the Massachusetts Institute of Technology, organized by Yehoshua Bar-Hillel. '"Cinderella Calculation" should try to charm "Prince Linguistics" to establish this new field,' he claimed (Wijngaarden 1952), but his words fell on barren ground.

The required computing power was not available: there was not one working computer in the Netherlands. The first Dutch computer, the ARRA I, had been demoed a couple of months earlier, but had never been able to do another calculation since. At the same time the field of linguistics was dominated by structuralism. Formal linguistics was banned from the linguistic departments, publications, and conferences. On top of that, Van Wijngaarden's PhD student Hugo Brandt Corstius, who is considered by many to be the founder of computational linguistics in the Netherlands and to whom the professor had given the responsibility to investigate the feasibility of machine translation, developed into a profound and well-spoken opponent. Brandt Corstius did develop a procedure for automated translation of numbers in the 1960s, but he shared Bar-Hillel's opinion that in order to perform high-quality open domain machine translation, extensive encyclopedic knowledge was required (Battus 1973; Brandt Corstius 1978) – a problem which he did not believe could be solved any time soon. Meanwhile in Leuven, Flanders, professor Flip Droste had also reached the conclusion that fully automated high-quality machine translation was not feasible in the near future (Droste 1969). These fundamental issues added to the argument that there was no economic ground for machine translation, as formulated in the American ALPAC report (ALPAC 1966).

The adverse conditions and negative opinions meant that very little happened in the field of machine translation in the Netherlands and Belgium until the 1980s. Between 1963 and 1965, another Amsterdam-based mathematician, Evert Willem Beth, secured some of the Euratom funds for machine translation research, but this was mostly spent on allowing linguists to develop their interest in formal linguistics (van der Beek 2001: 1–60). In 1980 the sentiment towards MT changed. Professor Bondi Sciarone of Delft Technical University was asked to represent the Netherlands in the European MT project Eurotra, and after some investigation, he stated in his inaugural lecture that the conditions that led to the pessimistic view of the ALPAC report no longer applied to the then-current situation in Europe, and that the time was ripe for a large-scale investigation of machine translation (Sciarone 1980): computing power was increasing rapidly, and while the cost of human labour was increasing, the cost of computing was going down. In the same period the European Union recognized an increase in the need for translation technology. With Eurotra the Netherlands and Belgium saw their first large-scale MT project.

Eurotra

Three groups represented the Dutch language within the Eurotra consortium: the Belgian group in the Flemish city of Leuven, and Dutch groups in Delft and Utrecht. The University of Leuven was the first to get involved: computational linguist Dirk Geens had already

participated in the steering group that prepared the official start of the Eurotra project in the late 1970s. In 1980, Bondi Sciarone became the first representative of the Netherlands, soon followed by Steven Krauwer and Louis des Tombe from the University of Utrecht. The Flemish and the Dutch divided the money and the workload in a ratio of 2:1 (two thirds for the Netherlands, one third for Flanders). As Flanders was the first to join, they had the first pick of the foreign languages to work with. They chose English and German, leaving the teams from the Netherlands with French, Danish, Spanish, Portuguese, Greek, and Italian.

The real impact of the Dutch teams on the Eurotra project, however, was not in the language-specific work packages, but in the thematic central committees. Geens and his students Frank Van Eynde and Lieven Jaspaerts, as well as Krauwer and Des Tombe, gained positions in various central teams, which had their own funding. Sciarone on the other hand mostly worked on language-specific study contracts. By 1984 most countries had signed their official participation contracts with Eurotra, and the first wave of language-specific contracts stopped. It took the Netherlands until 1986 before they secured their contracts, but Sciarone, not eligible for additional funding from central committees, had no budget for Eurotra work anymore, and decided to withdraw from the project, leaving it to the larger group in Utrecht to complete the Dutch work.

As members of the Eurotra committee for linguistic specifications, Des Tombe and Jaspaerts advocated a more solid linguistic base. They argued that more linguistic research was required and an agreed-upon, well-founded linguistic basis was needed before any implementation could be done. Krauwer meanwhile pleaded for more solid system specifications. Both Krauwer and Des Tombes were strong advocates of the <<C,A>,T> or CAT framework (Debille 1986). This system was Eurotra's response to the rise of unification-based grammars. Eurotra had originally been based on transformational grammar, which at the start of the project was considered a novel and state-of-the-art approach. However, during the programme unification-based grammars such as GSPG, HPSG, and LFG gained ground and proved well suited for computational implementations. Rather than adopting one of these frameworks, the committee for linguistic specifications created a new one, which incorporated some of the insights from unification-based grammars. The group in Utrecht actively collaborated with researchers in Essex on this topic, and even built a working pilot system. It was called MiMo, as Eurotra leader Serge Perschke had derogatively called it a 'Micky Mouse system'. Neither the system, nor the ideas behind it were implemented in the larger Eurotra project, but the group was quietly allowed to continue its development. This eventually led to MiMo II, a working translation system for a subset of Dutch, English, and Spanish – although the final system was completely unification-based and had very little to do with the original Eurotra design (Noord *et al.* 1990).

Just as the Dutch-speaking region of Belgium – Flanders – contributed to the development of the Dutch components of Eurotra, the French-speaking region – Wallonia – contributed to the French components: the University of Liège was contracted to work on the monolingual French components, totaling 8 percent of the French work. The bilingual components were covered by the groups in Paris (southern languages) and Nancy (northern languages). The main focus of the group in Liège, headed by professor Jacques Noël, was in fact on computational lexicography and terminology. Noël, a professor of English Linguistics, had access to a digital version of the Longman Dictionary of Contemporary English since the mid 1970s, from which the group developed a more general interest in the reusability of resources for MT. The team did not gain much influence in the Eurotra organization, and their most important proposal for the integration of terminology in the Eurotra framework was never accepted.

The Eurotra project was unprecedented in its scale and funding. The impact of this enormous project on MT in Belgium and the Netherlands differs per group. For French, there was an

official Eurotra demo. For Dutch, the best demo was the unofficial MiMo system. Yet, the Dutch groups both in Leuven and Utrecht benefited greatly from the project as they were able to acquire hardware for future research and set up programmes for teaching MT (and, more broadly, computational linguistics) to a new generation of researchers. Perhaps most importantly, the project helped them establish strong ties with other MT researchers in Europe. This network had a long-lasting effect on MT and NLP research in Flanders and the Netherlands. Liège on the other hand never fully integrated in the project, and did not benefit from it in the same way.

METAL

In Leuven, Flanders, a second group was working on MT in the second half of the 1980s: Herman Caeyers set up a team for the French–Dutch and Dutch–French translation pairs of the American–German METAL project. METAL originated in Austin, Texas, and was based on the work of Jonathan Slocum and Winfield Bennett (Bennett and Slocum 1985: 111–121). The METAL research at the University of Austin was heavily sponsored by Siemens, the German company, and focused on translation between English and German. Siemens wanted to move all research and development to Munich, but when the Belgian government made a big deal with Siemens they required Siemens to invest in Research and Development in Belgium. Hence the one type of research which could not easily be done in Germany moved to Belgium: dictionary development (Mons) and grammar writing (Leuven) for machine translation between Dutch and French.

In contrast to Eurotra, which focused on full translation, METAL aimed at translation support tools. It did not have the ambition to cover complex infrequent linguistic structures. Rather, it focused on building a working prototype that would cover as many as possible of the most common and frequent constructions. The group in Leuven, which included computational linguists Rudi Gebruers and Geert Adriaens, made important contributions to the base system, which had not been designed with Romance languages like French in mind. Caeyers reports apologies from the US development team for having assumed that every 'foreign' language had a case system (van der Beek 2011). Among other things, the group built a valence system that recognized the syntactic role of phrases without relying on case marking. Building on their expertise Adriaens was able to secure European funding later on in 1995 with the SECC (Simplified English Grammar Checker and Corrector) project, in which a grammar front-end for MT was built.

Eventually, Siemens sold the rights to METAL to GSM for the C++ version for the consumer market, and the Lisp version of the English–Dutch and French–Dutch translation pairs to Caeyers' company Lant, which later merged into Xplanation, which still runs a translation support tool including the original METAL product.

DLT

Besides the two international projects in which the Dutch and Flemish participated, there were also two Dutch MT projects: Distributed Language Translation (DLT) and Rosetta. DLT was an industrial project that ran at the Dutch company BSO (Buro voor Systeem Ontwikkeling – System Development Office) from 1980 until 1990. In contrast to Eurotra and METAL, DLT was not transfer-based, but instead worked with a remarkable interlingua: Esperanto. The project was initiated by Toon Witkam, an aeronaut by training who worked on automation projects for BSO before dedicating himself to DLT.

The main argument for using an interlingua is well known: it reduces the number of translation pairs drastically. This comes at the cost of having to specify in the interlingua every distinction made in any of the languages translated to or from. The arguments for using Esperanto as interlingua, according to Witkam, were that Esperanto could be encoded compactly due to its regularity, that its degree of lexical ambiguity is supposed to be a lot lower than in other languages, and that it is a fully understandable and accessible language independent of any other language (van der Beek 2011).

The efficient encoding was important because of the envisaged application: in contrast to Eurotra, which focused on batch processing of documents, resulting in acceptable (though imperfect) translations which would then be post-edited, DLT aimed at an interactive translation system, which required the translator to disambiguate any ambiguities in the input. The disambiguated representation in the interlingua would then be distributed to work stations, where translation into the target language was to take place. The output would be clean translations that would not require post-editing.

DLT started out as a personal project of Toon Witkam. He presented the outline of his plan to the heads of the company in 1980, but even though the reactions were 'very positive' according to Witkam, no budget was allotted. An application for Dutch funding was also refused. Witkam then recruited an intern, and reduced his paid work week to four days in order to work on his plan. In 1982 he turned lucky: the European Union sponsored an investigation into the feasibility of his plan with 250,000 Dutch guilders. The final report of this feasibility study was well received and led to substantial follow-up funding of 8 million guilders from the Dutch Ministry of Economic Affairs and 8 million guilders from BSO in 1984.

Witkam proceeded to appoint Esperanto specialists Klaus Schubert, who focused on syntax, and Victor Sadler, who focused on semantics. A supervisory board of three was appointed, referred to as ABK, after the last names of the members: Bernhard Al (Van Dale lexicography), Harry Bunt (professor of Computational Linguistics at Tilburg University), and Gerard Kempen (professor of Psycholinguistics at the University of Nijmegen). DLT adopted the dependency grammar of Tesnière as the syntactic framework. The system was originally designed to be knowledge-based, with dependency grammars for source and target languages, and an enriched version of Esperanto in the middle, which would allow for an unambiguous representation of the input.

In order to disambiguate the source language input, external knowledge sources such as taxonomies were used. However, after the COLING conference of 1988, where Witkam was first introduced to statistical machine translation (Brown *et al.* 1988: 71–76), the design of DLT changed significantly. The team switched from a knowledge-based system to a corpus-based design, where disambiguation was achieved through bilingual knowledge banks (BKBs). Although DLT continued to use Esperanto as an interlingua in order to reduce the number of BKBs necessary, it was no longer considered a key element (Witkam 2006). A second important development in the project was the switch of focus from general, informative texts to simplified English as was used in maintenance manuals and technical documentation of the (now defunct) Dutch aerospace company Fokker.

Although there were some superficial contacts between DLT and other MT projects in the Netherlands and Flanders, the project was generally met with scepticism by peers. This was mainly caused by two factors: the choice for Esperanto as an interlingua, which was considered eccentric, and the bold claims made in the press. In 1990, when the project was running out of funds, BSO launched a media campaign in order to find new external investors for the project. 'Computer speaks every language' it said in one newspaper, and 'Computer translates any language in any language' in another.

Witkam estimated at the end of the project that it would require ten times the earlier funds to build the actual translation product. BSO could not or was not willing to supply those funds, and new investors were never found. At the end of the project, in 1990, the results of the programme were a demo from 1987 (from before the introduction of the corpus-based approach) and an estimated 1,800 pages of documentation, mostly in a series of books published by Foris Publications (Dordrecht, the Netherlands).

Rosetta

Rosetta was a Machine Translation project that ran throughout the 1980s at Natlab, a research institute at the Technical University of Eindhoven and a subsidiary of Philips. The design of the system was due to Jan Landsbergen, who had previously worked on the Question Answering system PHLIQA. He had come up with a new grammar formalism for PHLIQA which was called M-grammar. PHLIQA was discontinued before he could implement it, but Landsbergen already envisaged another application of M-grammar: Machine Translation.

M-grammar is based on Montague Grammar (Montague 1973: 221–242), a generative grammar that regards all sentences as compositionally built from basic expressions of intensional formal logic. Landsbergen applied a number of changes to this basic setup to avoid overgeneration, and to allow for the reverse process, parsing, in addition to generation. M-grammar generally allows for transformations in addition to the concatenations of traditional Montague Grammar. These transformations were crucially reversible: fit both for parsing and for generation. The powerful rules with transformations would overgenerate, but Landsbergen prevented this by applying them to constituent structures instead of unstructured sentences – an extension already suggested by Barbara Partee (Partee 1976: 51–76). A context-free grammar was written to provide the constituent structures to which M-grammar was applied.

The key idea behind Rosetta is that for each language, an M-grammar is developed that can parse (in conjunction with the context-free grammar) a sentence in the source language and output some expression in intensional logic that captures the semantics of it, but that can also generate a sentence (or multiple sentences) in the target language from the expression in intensional logic. It is crucial that the M-grammars are isomorphic: for each lexical entry, phrase or rule in one language, there is a corresponding entry, phrase or rule in all of the other languages. A successful parse then guarantees a successful translation. This setup means that most work for developing the system is in developing the M-grammars for all languages. It also means that although the logical expressions can be viewed as an interlingua, the system is not a pure interlingual system, as it is not possible to develop the modules for each language independent of the other languages.

Landsbergen proposed his ideas for a Machine Translation system to the Philips management in 1979, and got approval to spend one year on the project, together with engineer Joep Rous. One year later they were able to demo a pilot system for Dutch, English and Italian: Rosetta I, named after the Rosetta stone. The demo was received with enthusiasm. Although Natlab was not willing to employ any linguists, Landsbergen did manage to get some extra hands on board through his contacts with the Eurotra group in Utrecht: Natlab was willing to pay the University of Utrecht to employ linguists to work at Natlab. An elaborate project proposal was put together for funding from the Dutch Ministry of Economic Affairs. The proposal talked of a collaboration between Natlab, the University of Utrecht, and Bondi Sciarone's group in Delft. Each group would contribute five participants. Utrecht would focus on grammar writing, Delft on lexicon development. However, the subsidy was granted to DLT instead of Rosetta. Philips then decided to step in and fund the project, although not quite to the full extent of the proposal and on the condition that all work was to take place at Natlab. Due to

those two constraints, Delft stepped out, and Natlab and Utrecht continued together. Italian was replaced by Spanish as a target language.

The original plan was to develop Rosetta II, a version with greatly extended lexicons and grammars and a much larger coverage. The experience accumulated during the development and testing of this version would then lead to a third version, in which fundamental changes could be applied to the framework and formalism. Yet lexicon extension was delayed as a result of Delft leaving the consortium, and the lack of consistent electronic databases of lexical information. The team could make use of the tapes of leading dictionary publisher Van Dale, but they turned out to contain large numbers of inconsistencies. It also took more time than anticipated to develop the supporting software needed for an efficient large-scale system. In the meantime the linguists developed ideas to treat more complex constructions. The planning was adapted, and instead of generating grammar rules and dictionaries, the team developed a new version of the system, Rosetta III, which was more advanced, but also more complex. A new type of transformation was introduced that did not change the semantics of a phrase and that did not need to have a counterpart in other languages. This transformation allowed for the treatment of many new and more complex syntactic phenomena. As a result, Rosetta is famous for being able to correctly translate complex sentences with the notoriously complex Dutch pronoun *er*, but it was unable to handle most sentences of newspaper text or a corpus of hotel reservations.

The project was funded until 1991, but as early as 1987 Philips started pressing for concrete applications. Although various options were researched – from integration in electronic typewriters to Philips' interactive CD (CD-i) – no well-suited application was found. The most vexing bottleneck remained the high costs associated with the development of large-scale dictionaries. When the project ended, it was not renewed. Rosetta IV, the version that would have been made for a specific application, was never built. No working version of the software remains, but the project did result in a number of PhD theses and the publication of Compositional translation (Rosetta 1994), in which Rosetta is explained in detail.

Statistical and hybrid machine translation research

When IBM presented their corpus-based MT methods in the late 1980s, the Dutch MT community's reactions were sceptical. Steven Krauwer refused to report on IBM's Peter Brown's session at the TMI conference in 1989, because he thought the proposal ridiculous. DLT did embrace the new approach, but the project came to an end in 1990, before the Statistical MT (SMT) revolution really took off. When the large knowledge-based projects also ended in the early 1990s, they were not replaced by SMT projects. Instead, MT research in the Netherlands fell silent, while in Belgium it reduced to a trickle. The one Dutch MT project in the 1990s was still knowledge-based. In Nijmegen, Albert Stoop built a system based on professor Jan van Bakel's AMAZON-parser. The name of his thesis, completed in 1995, illustrates the predominant sentiment regarding MT in the Netherlands in the 1990s: TRANSIT: A linguistically motivated MT system.

The turn to SMT was only made ten years later, when a new generation of computational linguists got public funding for research proposals in MT. Partly encouraged by the success of the open source SMT software package Moses (Koehn *et al*. 2007: 177–180), and the increasing availability of parallel corpora, Dutch researchers developed their own brands of SMT systems. In Amsterdam, research projects headed by Khalil Sima'an, Rens Bod, and Christof Monz followed the probabilistic trail. A common thread in the works of Sima'an and Bod is the inclusion of linguistically motivated features, such as tree structure (Bod 2007: 51–57, Mylonakis and Sima'an 2011: 642–652). Monz has been an active co-organizer of the

Workshop on Statistical Machine Translation (WMT) series.[1] With his colleagues he has also been active in the IWSLT shared task (Martzoukos and Monz 2010: 205–208; Yahyaei and Monz 2010: 157–162), a joint and open benchmarking effort that has pushed international MT research forward, and that has lowered the bar of entering the field of MT research, together with the advent of more public parallel corpora, open source machine translation tools, and ever faster computers equipped with ever more memory.

The increased availability of parallel corpora can to some extent be attributed personally to Jörg Tiedemann, who, in cooperation with Lars Nygaard, initiated the Opus Corpus.[2] Tiedemann expanded the corpus while working at the University of Groningen, the Netherlands, in the second half of the 2000s. The Opus Corpus gathers publicly and freely available parallel corpora such as the proceedings of the European Parliament and documents of the European Medicines Agency (EMEA), and offers automated preprocessing and alignment at the sentence level (Tiedemann 2012: 2214–2218). While in Groningen, Tiedemann also published on transliteration and translation of closely-related languages (Tiedemann 2009: 12–19).

In the same period a group of researchers in Tilburg University developed memory-based machine translation (Van den Bosch and Berck 2009: 17–26; Canisius and Van den Bosch 2009: 182-189; van Gompel *et al.* 2009: 17–26). MBMT is a hybrid of SMT with example-based machine translation (EBMT), a data-driven approach that predates SMT (Nagao 1984: 173-180; Carl and Way 2003). The Tilburg group furthermore adopted SMT for paraphrasing, by treating paraphrasing as monolingual translation, and using aligned headlines of articles covering the same news story as parallel data (Wubben *et al.* 2011: 27–33; Wubben *et al.* 2012: 1015–1024). Both the Tilburg group leader Van den Bosch and Amsterdam's Sima'an became active as international collaborators of the Irish Centre for Next-Generation Localization (CNGL),[3] advising CNGL PhD students (Hassan *et al.* 2008; Haque *et al.* 2011: 239–285).

The computational linguistics research group in Leuven that had been active in the METAL and SCC projects on grammar, syntax, and resource development in the late 1990s, returned to machine translation with the METIS (2001–2004) and METIS-II (2004–2007) EU projects.[4] The METIS projects were carried out with project coordinator ILSP in Athens, Greece, and project partners in Spain and Germany. The goal of METIS, full name 'Statistical Machine Translation Using Monolingual Corpora', was aimed at developing an SMT system without the typical but sometimes unrealistic starting condition of having a (large) parallel corpus. The METIS method involves the search for text subsequences (syntactic chunks, word n-grams) in monolingual corpora, the statistical alignment of these subsequences using bilingual lexica, and the use of these alignments in example-based MT, SMT, or directly as translation memory in TM systems (Dirix *et al.* 2005: 43–50). The Leuven group focused on hybridizing the statistical alignment of subsequences with linguistic knowledge, such as automatically computed part-of-speech tags. When computed on both sides of a language pair, part-of-speech tags are helpful in translating ambiguous high-frequency words such as the Dutch word *zijn*, which as a verb translates to *to be*, while as a pronoun to *his*.

In Belgium, MT also found a place in academia outside Leuven in the new LT^3 (Language Translation and Technology Team) at the University College Ghent. Besides NLP and text analytics, the group's expertise relevant for translation technologies is in terminology, usability of translation tools such as post-editing software, and machine translation. The group organized a shared task on cross-lingual word sense disambiguation at the SemEval 2010 workshop (Lefever and Hoste 2010: 15–20), raising the intriguing suggestion that word sense disambiguation, when seen as a subtask of translation, is more grounded than in the case when monolingual sense distinctions come from a lexical semantic resource (Lefever *et al.* 2011: 311–322). Another focus of the group is business-oriented multilingual terminology extraction;

it attracted public funding for the 2011–2012 project TExSIS ('Terminology Extraction for Semantic Interoperability and Standardization').

Recent and current translation technology industry

There has been relatively little activity in the Dutch translation technology industry since the high-ambition projects that wound down in the 1990s. In the Compendium of Translation Software: Directory of commercial machine translation systems and computer-aided translation support tools,[5] an updated reference guide to software compiled by W. John Hutchins and Declan Groves, two Dutch translation technology companies are listed: Syn-Tactic, a spin-off of the Dutch printing and copying hardware company Océ, specializing in localization of software and translation of technical manuals, and Lingvistica, now an OEM for products of LEC (Language Engineering Company LLC). A third company, Linguistic Systems BV, has been developing a multilingual thesaurus initially called PolyGlot, now called EuroGlot. Founded in 1985, the Nijmegen-based company chose to build a multilingual thesaurus from scratch, and re-implemented this over time to keep up with standard requirements. EuroGlot is concept-based; when the user selects one of the possible conceptual spaces of an ambiguous word, its concept-specific translations are shown. Domain-specific add-ons are available.

Belgium continues to host more translation technology industry than its northern neighbor country. The aforementioned Xplanation (which took over the METAL/Siemens spin-off Lant, later Lantmark and Lantworks) offers human translation services supported by Tstream, a suite of in-house developed tools for terminology extraction and resource management, translation memories, document processing formats and tools. The original METAL LISP software, ported to Linux, now part of Tstream, still receives occasional dictionary updates, but continues to use the same grammars. Occasionally the company employs SMT software to train MT systems on customer-specific translation memories.

Lernout and Hauspie, the former Flemish language and speech technology company, also developed activities in translation technology in the late 1990s, mostly through its acquisition of the GlobaLink company and its GTS MT system, which was renamed Power Translator, a software package that still exists and is now a product of LEC. Lernout and Hauspie's acquisition of the translation bureau Mendez led to the development of a hybrid translation division that became quite successful. An online free 'gist-quality' MT service gave the user the option to have the output post-edited by professional human translations for a fee (Sayer 2000).

LandC (Language and Computing) was another Belgian company offering translation services in the medical domain; LandC is now part of Nuance. The current Ghent-based company CrossLang, formerly Cross Language, specializes in optimizing business translation processes, employing both existing technology and developing custom SMT systems. Notably they coordinate the Bologna Translation Services project (2011–2013), which specializes in translation services in higher education. A second currently active Ghent company is Yamagata, offering QA Distiller, an automatic tool for the detection and correction of errors and inconsistencies in translations and TMs. Together with the LT^3 group, Ghent is the current capital of translation technology in the Low Countries.

Dutch in current translation technology

Dutch is the first language of an estimated 23 million people worldwide. It is spoken by the majority of the population in the Netherlands and an estimated 60 percent of the populations of Belgium (mostly in the Flanders region) and Surinam. It is the eighth language in the

European Union in terms of the number of speakers.[6] Dutch has been among the most frequently included languages in academic and industrial translation technology projects worldwide, but has been losing ground due to the global rise of interest in growing-economy languages such as Chinese, (Brazilian) Portuguese, and Korean. In the EAMT software compendium, Dutch is listed as a source or target language in about 15 percent (66 out of 447) of the mentioned commercially available translation technology products. In MT systems Dutch is most frequently paired with its geographical and historical linguistic neighbors English (39 systems), French (26 systems), and German (12 systems). Other frequent pairings are with Spanish (11), Italian (9), Russian (9), and Chinese (9).

As one of the official European languages, Dutch is present in the larger public European parallel corpora such as those in the Opus Corpus, or in the JRC Acquis corpus[7] (Steinberger *et al.* 2006). This allows SMT systems to easily include Dutch paired with other official European languages. Recent academic work in Ghent, Leuven, and Tilburg has used Dutch as one of the languages. When the Tilburg translation group moved to Radboud University, Nijmegen, the Netherlands in 2011 they developed and launched an online Dutch−Frisian SMT system called Oersetter.nl.[8] The West-Frisian variant of Frisian, another West-Germanic language that shares its origin with English and Dutch, is spoken mostly in the Dutch province of Friesland. A similar effort with a historically related language, Afrikaans, is the work by the CTEXT lab at Northwest University, Potchefstroom, South Africa, where rule-based methods for transliteration and word reordering are developed to directly convert Dutch to Afrikaans (Pilon and van Huyssteen 2009: 23–28).

Notes

1 http://www.statmt.org/wmt12.
2 http://opus.lingfil.uu.se.
3 http://www.cngl.ie.
4 http://www.ilsp.gr/metis2.
5 http://www.eamt.org/soft_comp.php - consulted in August 2012.
6 http://taalunieversum.org/taal/feiten_en_weetjes/#feitencijfers.
7 http://langtech.jrc.it/JRC-Acquis.html.
8 http://oersetter.nl.

References

ALPAC (1966) *Languages and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC: National Academy of Sciences, National Research Council, 1966.

Battus (1973) *De vertaalmachine*, Holland Maandblad.

Bédard, Claude (1991) 'What Do Translators Want?' *Language Industry Monitor* (4): 1–3.

Beek, Leonoor van der (2001) 'Van Beth tot Van Benthem: de opkomst van de Nederlandse semantiek', *Tabu* 31(1−2): 1–60.

Beek, Leonoor van der (2011) 'Van Rekenmachine to Taalautomaat'. Available at: http://linqd.nl/book.html.

Bennett, Winfield S. and Jonathan Slocum (1985) 'The LRC Machine Translation System', *Computational Linguistics* 11(2−3): 111–121.

Bod, Rens (2007) 'Unsupervised Syntax-based Machine Translation: The Contribution of Discontiguous Phrase', in Bente Maegaard (ed.) *Proceedings of the Machine Translation Summit XI*, 10-14 September 2007, Copenhagen Business School, Copenhagen, Denmark, 51–57.

Brandt Corstius, Hugo (1978) *Computer-taalkunde*, Muiderberg: Dirk Coutinho.

Brown, Peter, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, Robert L. Mercer, and Paul S. Roossin (1988) 'A Statistical Approach to Language Translation', in *Proceedings of the 12th Conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ: 1: 71–76.

Canisius, Sander and Antal van den Bosch (2009) 'A Constraint Satisfaction Approach to Machine Translation', in *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT-2009)*, 14-15 May 2009, Barcelona, Spain, 182–189.

Carl, Michael and Andy Way (eds) (2003) *Recent Advances in Example-based Machine Translation,* Dordrecht: Kluwer Academic Publishers.

Debille, L. (1986) *Het basismodel van het EUROTRA-vertaalsysteem*, Automatische vertaling aan de K.U. Leuven.

Dirix, Peter, Ineke Schuurman, and Vincent Vandeghinste (2005) 'Metis II: Example-based Machine Translation Using Monolingual Corpora – System Description', in *Proceedings of the Example-based Machine Translation Workshop Held in Conjunction with the 10th Machine Translation Summit*, 16 September 2005, Phuket, Thailand, 43-50.

Droste, Flip G. (1969) *Vertalen met de computer; mogelijkheden en moeilijkheden*, Groningen: Wolters-Noordhoff.

Haque, Rejwanul, Sudip Kumar Naskar, Antal van den Bosch, and Andy Way (2011) 'Integrating Source-language Context into Phrase-based Statistical Machine Translation', *Machine Translation* 25(3): 239–285.

Hassan, Hany, Khalil Sima'an, and Andy Way (2008) 'Syntactically Lexicalized Phrase-based Statistical Translation', *IEEE Transactions on Audio, Speech and Language Processing* 16(7).

Koehn, Philipp, Hieu Hoang, Alexandre Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst (2007) 'Moses: Open Source Toolkit for Statistical Machine Translation', in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume: Proceedings of the Demo and Poster Sessions*, Association for Computational Linguistics, 25-27 June 2007, Prague, Czech Republic, 177–180.

Lefever, Els and Veronique Hoste (2010) 'Semeval-2010 Task 3: Cross-lingual Word Sense Disambiguation', in *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, Sweden, 15–20.

Lefever, Els, Veronique Hoste, and Martine De Cock (2011) 'Parasense or How to Use Parallel Corpora for Word Sense Disambiguation', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Portland, OR, 317–322.

Martzoukos, Spyros and Christof Monz (2010) 'The UvA System Description for IWSLT 2010', in *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT-2010)*, 2-3 December 2010, Paris, France, 205–208.

Montague, Richard (1973) 'The Proper Treatment of Quantification in Ordinary English', in Patrick Suppes, Julius Moravcsik, and Jaakko Hintikka (eds) *Approaches to Natural Language*, Dordrecht: Springer Verlag, 221–242.

Mylonakis, Markos and Khalil Sima'an (2011) 'Learning Hierarchical Translation Structure with Linguistic Annotations', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 19-24 June 2011, Portland, OR, 642–652.

Nagao, Makoto (1984) 'A Framework of a Mechanical Translation between Japanese and English by Analogy Principle', in Alick Elithorn and Ranan Banerji (eds) *Artificial and Human Intelligence*, Amsterdam: North-Holland, 173–180.

Noord, Geertjan van, Joke Dorrepaal, Pim van der Eijk, Maria Florenza, and Louis des Tombe (1990) 'The MiMo2 Research System', in *Proceedings of the 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Austin, TX, USA, 213-233. Available at: http://www.let.rug.nl/~vannoord/papers.

Partee, Barbara H. (1976) 'Some Transformational Extensions of Montague Grammar', in Barbara H. Partee (ed.) *Montague Grammar*, New York: Academic Press, 51–76.

Pilon, Suléne and Gerhard B van Huyssteen (2009) 'Rule-based Conversion of Closely-related Languages: A Dutch-to-Afrikaans Convertor', in *Proceedings of the 20th Annual Symposium of the Pattern Recognition Association of South Africa* (PRASA), 30 November – 1 December 2009, Stellenbosch, South Africa, 23–28.

Rosetta, M.T. (1994) *Compositional Translation*, Dordrecht: Kluwer Academic Publishers.

Sayer, Peter (2000) 'Lernout and Hauspie Translates Free', *PCWorld* 2 October 2000.

Sciarone, A.C. (1980) *Over automatisch vertalen*, Inaugural address.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufiş (2006) 'The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages', in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 2142–2147.

Tiedemann, Jörg (2009) 'Character-based PSMT for Closely Related Languages', in Lluís Màrquès and Harold Somers (eds) *Proceedings of 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, 14-15 May 2009, Barcelona, Spain, 12–19.

Tiedemann, Jörg (2012) 'Parallel Data, Tools and Interfaces in OPUS', in Khalid Choukri, Thierry Declerck, Mehmet Uğur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds) *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)*, European Language Resources Association (ELRA), 21–27 May 2012, Istanbul, Turkey, 2214–2218.

van den Bosch, Antal and Peter Berck (2009) 'Memory-based Machine Translation and Language Modeling', *The Prague Bulletin of Mathematical Linguistics* 91: 17–26.

van Gompel, Maarten, Antal van den Bosch, and Peter Berck (2009) 'Extending Memory-based Machine Translation to Phrases', in Mikel L. Forcada and Andy Way (eds) *Proceedings of the 3rd International Workshop on Example-based Machine Translation*, 12–13 November 2009, Dublin City University, Dublin, Ireland, 79–86.

Wijngaarden, A. van (1952) *Rekenen en vertalen*, Delft: Uitgeverij Waltman.

Witkam, Toon (2006) 'History and Heritage of the DLT (Distributed Language Translation) Project'. Available at: http://www.mt-archive.info/Witkam-2006.pdf.

Wubben, Sander, Erwin Marsi, Antal van den Bosch, and Emiel Krahmer (2011) 'Comparing Phrase-based and Syntax-based Paraphrase Generation', in *Proceedings of the Workshop on Monolingual Text-to-text Generation*, Association for Computational Linguistics, 24 June 2011, Portland, OR, 27–33.

Wubben, Sander, Antal van den Bosch, and Emiel Krahmer (2012) 'Sentence Simplification by Monolingual Machine Translation', in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Association for Computational Linguistics, 8–14 July 2012, Jeju Island, Korea, 1015–1024.

Yahyaei, Sirvan and Christof Monz (2010) 'The QMUL System Description for IWSLT 2010', in *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT-2010)*, 2–3 December 2010, Paris, France, 157–162.

22

TRANSLATION TECHNOLOGY IN THE UNITED KINGDOM

Christophe Declercq

UNIVERSITY COLLEGE LONDON, THE UNITED KINGDOM

Introduction

The European Association for Machine Translation (EAMT)[1] might very well be registered in Switzerland,[2] the organization is an official supporter of 'Translating and the Computer' conferences, held by the Association for Information Management (ASLIB) in London each year. The former president of EAMT and noted machine translation (MT) authority W. John Hutchins himself resides in Norwich, England. His Compendium of Translation Software is close to being the Bible of what actually is covered by the concept 'translation software' and arguably one of the most sensible approaches to the concept 'translation technology'.

Translation software largely covers two subcomponents. Automatic translation systems, on the one hand, are machine translation systems of various kinds (rule-based, statistical, online/standalone …) and for various purposes (from enterprise over professional to website and mobile). Translation support systems, on the other hand, range from electronic dictionaries over localization support and alignment tools to translation workstations with a translation memory at their core. Basically, the two approaches distinguish between language technology and translation technology. However, language technology increasingly is paired with translation technology (see Chapter 30, 'Editing in Translation Technology'). Therefore, this contribution covers translation software use in the United Kingdom.

Special relations

This contribution in particular depicts the role and position of translation and especially translation technology in a society which has a special relationship with its neighbouring countries, the EU and the US. And each time the English language serves as a unique currency.[3]

With the widespread uptake of the English language in former British colonies and with British English acting as one of the three main working languages in the European Union (and *de facto* the ultimate working language?), the United Kingdom finds itself in a peculiar position where the role of translation, and technology in support of it, is concerned. In stark contrast to many of its fellow EU member states, British society has never had a historical urge to accommodate translation needs. On the contrary, English as a global lingua franca steers translation itself. English has a well-established tradition of acting as a source language, from which translation happens, or as a target language.[4] And even if English is not involved as such, then it most likely acts as a relay language. Moreover, as a member of the United Nations Security Council, the

European Union (EU), the Commonwealth of Nations, the Council of Europe, the G7, the G8, the G20, NATO, the Organization for Economic Co-operation and Development (OECD), and the World Trade Organization (WTO), the United Kingdom most certainly finds itself amidst intense political and economic communication. This not only triggers translation into or out of English,[5] but also sees excellent communication happen across language barriers.

Education

Sampled from a dozen people using Google Advanced within the space of a few days and performing the same search (*'translation technology' site:uk*), a few interesting results appeared. First and foremost, Imperial College London's MSc in Scientific, Medical and Technical Translation with Translation Technology accounted for half the top ten results.[6] Joined by university courses at Swansea University, University College London (UCL), the School of Oriental and African Studies (SOAS) and elsewhere, the straightforward online search provided an important overall insight into 'Translation Technology' and British web pages: according to Google translation technology is mainly an academic field. In order to maintain a sustainable translation turnover, the United Kingdom needs a substantial continued stream of translation students on the one hand and an awareness that actively mastering languages is beneficial to professional development and career progress on the other hand. Here, the United Kingdom confirms its status as the odd one out. Nowhere else in Europe is the presence of languages in education, both secondary and higher education, under threat as in the United Kingdom.

Despite continued efforts of government, institutions such as the British Academy and frequent attention by most of the quality newspapers, languages in the UK decline at secondary education level. This is evident from the headlines retained below:

- Languages in state schools 'decline further' (BBC News 27 January 2011 cited in Selgren 2011)
- GCSE results set records but spark row over decline in modern languages (*Huffington Post*, 25 August 2011: GCSE 2011)
- A-level foreign languages decline alarms examiners (*Guardian* online, 16 August 2012 cited in Vasagar 2012)
- Anti-European attitudes 'turning pupils off languages' (*Daily Telegraph* online, 20 March 2013 cited in Paton 2013)
- How to encourage students to pursue languages at GCSE and A-level (*Guardian* online, 17 May 2013 cited in Drabble 2013)
- How can schools encourage students to take languages further? (Drury online, 3 July 2013)
- UCAS stats reveal languages decline (*Times Higher Education* online, 23 July 2013 cited in Matthews 2013)

In contrast with falling language numbers in secondary education, language-learning summer courses or language evening classes continue to spike. But more often than not this hunger for languages comes from non-native English speakers. This most peculiar situation, which in many aspects is the opposite of other EU member states, is mirrored in the specific situation of translation technology in the United Kingdom too. Whereas in many other EU member states professional translators and translation students alike are served by a local subsidiary of a translation technology software developer, the UK, more specifically in London, attracts other-lingual translators and students.

In Higher Education, the situation does not seem to improve much either. Despite the substantial influx of non-UK students, translation and language departments have been curtailed at an ongoing basis. From Imperial College London over City University London to Salford University,[7] language sections of Humanities departments have been scrapped along with their translation units.

It is very difficult to count the number of Master's in Translation in the United Kingdom. Often, a new Master's is created by recreating most of another one, by changing core modules to optional modules, or by adding one or two new modules. Despite shifting relevance, it can be argued that translation software features heavily in virtually all translation studies courses. One of the features that sets particular Master's apart from others is the label granted by the European Commission's Directorate-General for Translation to higher-education institutions offering Master's level translation programmes: European Master's in Translation (EMT). The quality label is granted to 'university translation programmes that meet agreed professional standards and market demands' and that answer to an elaborate 'translator competence profile, drawn up by European experts' (EMT online 2013). Transferable skills in project management and using translation software feature prominently in the EMT competences.[8] The following UK Universities offer a Master's course in translation that has been granted the label.

Aston University	MA in Translation in a European Context
Durham University	MA in Translation Studies
University of Surrey	MA in Translation
Imperial College London	MSc in Scientific, Technical and Medical Translation with Translation Technology (MscTrans)
London Metropolitan University	MA Applied Translation Studies
Roehampton University	MA in Audiovisual Translation
University of Westminster	MA in Technical and Specialised Translation
University of Manchester	MA in Translation and Interpreting Studies
University of Portsmouth	MA in Translation Studies
University of Salford	MA in Translating
Swansea University	MA in Translation with Language Technology

Sampling from translation software module descriptions, differences between the various EMT universities become clear. Whereas the MA at Aston University clearly keeps an open perspective on translation studies by incorporating media translation, MT, dubbing and subtitling, the MA at Birmingham retains a corpus linguistics approach ('using the Internet to search for terminology, comparable and parallel texts; using translation forums and other specialized translation resources websites', Birmingham 2013). Teaching audio-visual translation technology has increased substantially in the last few years (Roehampton, Surrey, Imperial College …) and a convergence in text and speech technology can be expected even more in the near future. The courses at Imperial and Swansea have a clear and open link with technology, the latter even offering a Postgraduate Certificate in Translation Technology for freelance translators who need to step up their technological skills.

However, when it comes to Translation Programmes in the UK, at least three non-EMT ones come to mind. The University of Bristol runs a straightforward computer-aided translation module, in which students gain 'an understanding of and familiarity with translation software applications and develop a practical competence in the range of functionalities offered' (Bristol 2013). This module is supplemented by 'The Translation Industry', in which ethics and quality assurance issues are coupled with insight into the wider business context of translation technology. The business of translation, in which technology features heavily, is indeed often lacking in translation technology modules, which frequently focus on utilizing a variety of

tools. Heriot Watt University offers an MSc in Translation and Computer-Assisted Translation Tools, but arguably the mother of including technology in translation studies is the University of Leeds, which uses 'an unrivalled range of software tools that are widely used by leading translation companies – Déjà Vu X, LTC Worx, MemoQ, OmegaT, Passolo, SDL Trados, STAR Transit, and Wordfast' and whose module 'is driven by multilingual group projects, which provide valuable experience of translation project management' (Leeds 2013).

And yet, the United Kingdom has so much more to offer the world of translation technology than just university courses at a Master's level. Arguably the most striking element on the assumed fully English native soil of the UK is the presence – albeit limited – of other languages that are officially recognized.

Devolution and translation technology

The United Kingdom (which reads in full: the United Kingdom of Great Britain and Northern Ireland) consists of England, Wales, Scotland and Northern Ireland. Other than England, each of the constituent nations has devolved powers. When it comes to translation software and using technology to cross linguistic barriers, Wales arguably is the most special case.

Even though the percentage of the Welsh population able to speak, read and write Welsh decreased by 1.5 per cent in the period 2001–2011, there are still nearly half a million people[9] who master the other official language of Wales besides English. That bilingual nature of Wales drives the difference in translation technology usage between Wales and its fellow constituent UK nations. Driven by the Welsh Assembly and central authorities many associations, bodies and events have taken place in the past years that have not had an equivalent in England or Scotland.

In 2005, the Welsh Government's Translation Service was established, supporting the Assembly Government in the delivery of bilingual public services. The services were the prolongation of services that already had been running since 1999. Whereas the focus lied with terminology (Prys 2006: 50) and translation provision in the early to mid-2000s, this shifted more to translation technology later on. In 2009, the report 'Improved translation tools for the Translation Industry in Wales' was published, written by Delyth Prys, Gruffudd Prys and Dewi Jones. The report highlighted the significance of the translation industry for the Welsh economy and even stressed that the sector was an important employer of women and was located in any possible area of the country (urban, rural, semi-rural). The report focused on the two-fold provision of the translation industry: it served the bilingual services in Wales and aided 'other sectors of the Welsh economy market in the export of their goods and services in the global marketplace'. Also, 'translation technology tools, regardless of the languages translated' (Prys 2009: 3) were seen to be underused in Wales. In 2009, core benefits of using translation technology in Wales were considered the following:

- increasing capacity by 40 per cent, and saving 20 per cent in administrative time without any increase in staffing levels by appropriate use of translation technology;
- 50 per cent further growth in the sector through expanding capacity to meet domestic demand, and 300 per cent growth in attracting translation business from outside Wales;
- increasing export opportunities for customers and potential customers by 19 per cent by making appropriate use of translation and multilingual services. (Prys 2009: 3)

However, a clear threat was seen in competition from companies from other parts of the UK and EU, which could possibly 'only be countered by equipping the industry within Wales

with the means to become more technologically competent themselves'. In order to better face future threats from competitors 'a demonstrator centre for the translation industry' (Prys 2009: 3) was sought to be established and compiling relevant tools into a toolkit[10] for industry-wide use in Wales was advocated.

The Language Technologies Unit (LTU) and more specifically SALT Cymru (Speech and Language Technology) at Bangor University now assume that role of demonstrator centre.[11] With research project and resources such as a Welsh Basic Speech Recognition Project, CEG (an electronic corpus of the Welsh language) and Maes-T (a web interface for the online development of terminology databases), SALT Cymru covers a wide variety of translation software such as

- speech technology: speech recognition, speaker recognition, text-to-speech techniques, speech coding and enhancement, multilingual speech processing;
- written language input: optical character recognition, handwriting recognition;
- language analysis, understanding and generation: grammar, semantics, parsing, discourse and dialogue;
- document processing: text and term extraction, interpretation, summarization;
- machine translation: including computer-aided translation, multilingual information retrieval;
- multimodality: gesture and facial movement recognition, visualisation of text data;
- language resources: written and spoken corpora, lexica, terminology;
- evaluation: of all of the above. (SALT 2010)[12]

What was retained from Hutchins' Compendium of Translation Software might not have been as elaborate as the above, but that does not mean the Compendium did not go to similar lengths. It also means that (albeit independently) an MT expert from Norwich and a European-funded University in Wales could think along the same lines.

An outcome of the report was the creation of a national terminology portal for Wales. Another outcome of the report was an event in January 2011, organized by the Universities of Bangor and Swansea,[13] to showcase how translation technology could support companies in reaching new markets, but also to service further needs of Wales' bilingual communities. No surprise then that the Welsh Minister for Heritage, Alun Fred Jones, who opened the conference, reiterated that developing expertise in the translation industry was a win-win situation for Wales[14] (Bangor 2011).

The focus on translation technology and Welsh does not only come from inside the Welsh borders. Google Translate added Welsh to its languages in August 2009, Microsoft produces many Welsh User Interfaces for its applications and regularly updates its Welsh Style Guide for localization purposes. As early as 2004 Harold Somers at the Centre for Computational Linguists at the University of Manchester had produced a report for the Welsh Language Board about the possibilities of machine translation for Welsh. Aptly called 'The Way Forward' Somers focused on three items mainly: language technology provision for Welsh at that time, three types of machine translation (SMT, RBMT, EBMT[15]) and their possible contribution to support Welsh language provision, and a comparison with minority languages elsewhere in Europe (such as Irish, Basque, Catalan and Galician). The contrast with Northern Ireland and Scotland, where only a very small number of people speak a language different from English or Irish or Scottish English, let alone England, could not be bigger.

The relative success in Wales of translation technology in government, executive associations and companies can hardly be replicated in Scotland.[16] The number of speakers of Scottish

Gaelic simply is of a different proportion.[17] However, a study by Commun na Gaidhlig found that businesses that use Gaelic in their visual marketing stood out to consumers even though several groups of people strive for more prominence of Scottish Gaelic in the Western Isles, utilizing translation technology to help those people who do not speak that language (Language Insight 2011). The attempt seems to find difficulty in establishing momentum beyond its own fragmented geographical area, which can be seen in the fact that the translation memory service for Scottish Gaelic is confined to the University of the Highlands and Islands alone,[18] whereas language technology efforts in Wales are shared across various higher education institutions and governmental organizations.

Most definitely a much more striking Scottish presence in the world of translation software is provided by the University of Edinburgh and their machine translation research and development. Several years in the making and fully released to the world in 2007 and 2008, the Moses Open Source Toolkit for Statistical Machine Translation[19] provided for much of the seismic power of the shift in language technology use at the time (most of the other seismic shock of the time was attributed to Google Translate). It is not fair to lay credit for Moses solely with Edinburgh; half a dozen of other institutions such as MIT and Aachen also provided substantial R&D and financial input. However, Philippe Koehn, Hieu Hoang, Alexandra Birch and Chris Callison-Burch are names that still stand out in the field of MT.

Especially the case of translation technology use in Wales has seen a government-backed and EU-supported move to team up local authority, education and companies. It is therefore the purpose of the following section to provide an overview of British organizations and companies working the field of translation technology.

Translation technology companies

The United Kingdom is one of the key players in translation technology. One of the homes of the global lingua franca, the UK also has London, Europe's biggest city and one of the world's financial centres.

With widely respected British television channels, especially the BBC, and newspapers,[20] the UK is an important centre for printed and broadcast media. Television and technology go hand in hand. In order to further accessibility to the media, subtitling, live captioning and audio description are increasingly used.[21] It should therefore not be a surprise that the UK is also a hub for translation technology.

Among the companies that have their headquarters in the UK are Applied Language Solutions, ArabNet, ATA Software, ITR (International Translation Resources), LTC, Network Translation, Prolingua, Screen Systems, SDL, Software Partners, Translation Experts, Translution, Wizart, Wordbank and XTM. With SDL, based in Maidenhead (and offices in Sheffield and Bristol), one of the giants of the translation software industry, the UK is provided for very well. Among the notable translation software companies that have a main base elsewhere but an important UK hub are ABBYY, LionBridge, TEMIS and Worldlingo. Overall, these companies offer software provision that is shifting towards a more diverse platform or customizable applications that include language technology, translation technology, project management, quality assurance, collaborative aspects and the like.

Just how diverse translation software and management of translation projects have become, is clear from the above screenshot from SDLs Products homepage. In fact, some of their resources aren't even fully referenced in the list. BeGlobal Trainer, the functionality to customize SDL's BeGlobal MT component, is not overtly included here, as is Contenta, the customisable XML solution, nor is LiveContent, which is also XML related. Although the

days of a handful of translation memory tools are over and the concept of 'new kid on the block' no longer applies in an era of apps, crowd and cloud, XTM International and their XTM modules offers a range of tools, including XTM Cloud, which is SaaS (software as a service). They have provided the diagram that is represented below.

Web content management
- SDL Tridion

Automated translation
- SDL BeGlobal

Realtime targeting
- SDL Fredhopper

Analytics and optimization
- SDL Customer Analytics
- SDL Global AMS

Structured content
- SDL LiveContent

Media management
- SDL Media Manager

Multichannel delivery
- SDL Email Manager
- SDL Mobile
- SDL XPP

Social intelligence
- SDL Customer Commitment Dashboard
- SDL SM2

Translation management
- SDL TMS
- SDL WorldServer

Translation productivity
- SDL Studio GroupShare
- SDL Trados Studio 2011

Campaign management
- SDL Campaign Manager
- SDL Quatron

Figure 22.1 Overview of SDL products

Source: SDL online

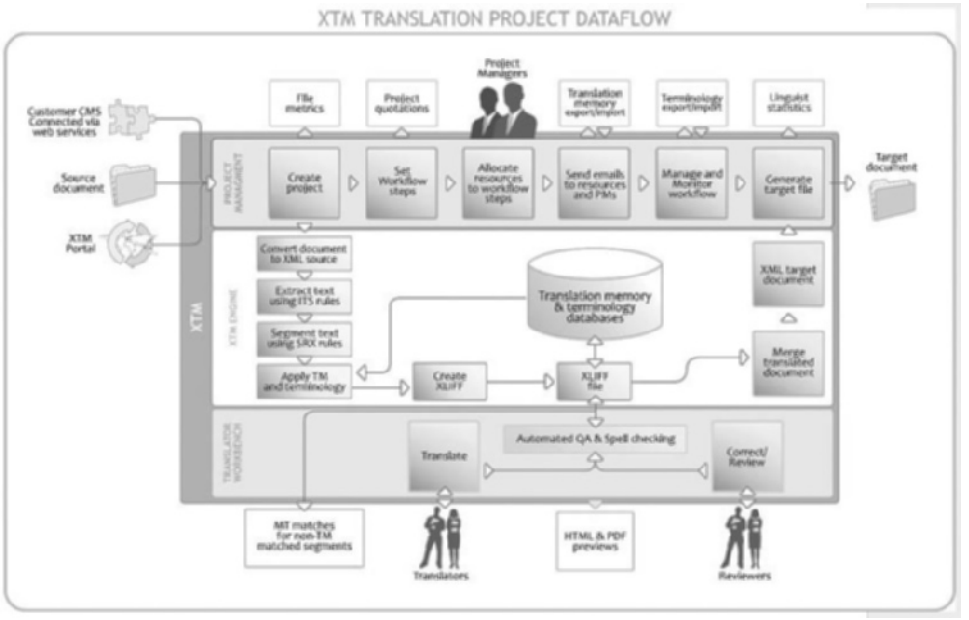


Figure 22.2 A typical XTM project dataflow

Source: XTM Online

Although the representation itself might be XTM International's, the workflow is very familiar to any company working with translation software. SDL and XTM differ in their approach in that SDL covers this workflow with some of their tools, whereas XTM's tools concern technology at various steps in this workflow only (a logical difference between an international corporation on the one hand and an emerging technology company on the other hand). But it can be argued that a workflow including project managers, translators and reviewers on the one hand and XLIFF, HTML and PDF files on the other hand, still is a more traditional workflow with translation technology involved.

Nataly Kelly, formerly of Common Sense Advisory fame, now Smartling, confirms that in the translation industry 'prevailing business models are exactly the same as they were two decades ago' and that change is necessary, an emergence of 'technological advances that impact the translation market at large, including the hundreds of thousands of professional translators out there'. Kelly advocates the position that MT and crowdsourcing 'do not begin to touch most of the market activities yet' and that they might not do so soon but that the translation industry is losing out on a major opportunity. This drive towards new opportunities is clear from SDL's approach, but not yet from XTM's. However, in the competitive and open market that is the translation software world in the UK, both models keep feeding into one another. This makes for British translation technology to remain solid at the base, without losing sight of future opportunities that are not catered for yet.

It is true that pervasive use of social media and MT, the emergence of the crowd and the crucial role of utilizing big data has confronted language technology with a new paradigm. If the current age is not the time for translation technology entities to open up to the wider Information Technology world, then when is? The greater London area remains the prime hub in Europe where teaching and learning experts as well as software developers come and work together.

Acknowledgements

Many thanks to several members of the former Translation Unit at Imperial College London (which has now transferred to UCL) and a dozen of selected LinkedIn contacts for providing me with screenshots of Google Advanced search results. Bettina Bajaj, Rocio Banos Pinero, Lindsay Bywood and Jorge Diaz-Cintas also provided input about what makes the United Kingdom stand out in the field of translation technology.

Notes

1 The current President is Andy Way, also UK. The EAMT oversees European R&D groups active in the field of machine translation. Of the 30 groups, 3 are located in the UK: the Statistical Machine Translation Group (University of Edinburgh), Machine Intelligence Laboratory (Cambridge University) and Information Retrieval Group (Queen Mary, University of London).

2 Similarly, the International Organization for Standardization (ISO) is registered in Switzerland as well (like EAMT in Geneva too). ISO was established in 1947, after 'delegates from 25 countries met at the Institute of Civil Engineers in London and decided to create a new international organization "to facilitate the international coordination and unification of industrial standards"' (ISO 2012).

3 It is an old adage among London cabbies that they do not need to learn a language because the world comes to London and speaks English there.

4 A noted exception to this in the field of translation software concerns Galician and Catalan, core to the many languages pairs covered by both Apertium, Lucy and Translendium, open source MT and online MT services.

5 The UK is also home to the following diverse organizations among others: the International Mobile Satellite Organisation, the International Maritime Organisation, Unicef UK, Amnesty International, PEN International, UN's Save the Children, European Bank for Reconstruction and Development, BP … the list is very long. Also, international assocations such as the International Cocoa Organisation, the International Grains Council and the International Sugar Organisation are based in London. Sadly, it is beyond the scope of this contribution to analyse translation needs and translation technology use among these institutions and associations.
6 From 2013/14 onwards this MSc is organized at University College London.
7 At the time of writing, the Translation Studies Unit at Imperial College will be discontinued for the subsequent academic year. The Unit was negotiating a transfer deal at the level of ongoing implementation meetings. At London City University, the MA in Audiovisual Translation no longer ran from September 2013 onwards. In June 2013, the University of Salford confirmed its plans to close all courses in modern languages, despite the fact that it leads the National Network for Translation.
8 Another European project, called OPTIMALE (Optimising Professional Translator Training in a Multilingual Europe), focuses on the training of trainers. 8 UK Translation Programmes are part of OPTIMALE and educators have been taking part in workshops that eyed best practice in training students language and translation technology.
9 Wales has a population of nearly 3.1 million people.
10 'The toolkit will comprise an illustrative integration of translation memories, terminologies, language proofing tools and workflow managers in an attractive translation environment. The toolkit will be generic and exemplary to avoid licensing issues with commercial software providers, and will include trial versions of new solutions under development at the LTU.' (TIKEP, no date)
11 In earlier different forms, the LTU has been active since the early 1990s.
12 In many ways both Hutchins and SALT honour Cole's 1996 *Survey of the State of the Art in Human Language Technology*, only they take it 15 years further.
13 Swansea University also meets the individual needs of freelance translators who remain undecided as to which translation memory to use or how to apply term recognition and to that end offers a Postgraduate Certificate in Translation Technology. Bangor and Swansea are not the only universities in Wales that deal with language technology. Aberystwyth University, for instance, holds the Centre for Welsh Language Services.
14 The European Regional Development Fund co-funds a lot of the research and development activities.
15 Statistical machine translation, rule-based machine translation, phrase-based machine translation.
16 However, increasingly minority languages are included in language technology conferences, such as 'Language in Minority/ised Language Media' held in July 2013 at the University of Aberystwyth.
17 In 2001, it was estimated that hardly 1.2 per cent of the Scottish population could only speak some Gaelic, compared with 10 per cent of all the Welsh people who speak, read and write Welsh.
18 On 29 May 2013, the University of Glasgow announced that among 20 funded collaborative PhD studentships with industry partners, one of their research projects concerned translation technology (Pittock 2013).
19 The article with the same title has a Google reference of nearly 2000 citations (date 1 July 2013). The core arguments of the paper concerned "(a) support for linguistically motivated factors, (b) confusion network decoding, and (c) efficient data formats for translation models and language models" (Koehn *et al.* 2007: 177).
20 In a list of global newspapers ranked according to their circulation, 8 British newspapers feature among the top 40 (source: IFABC)
21 In September 2012, London-based Red Bee Media landed an exclusive deal to be providing the BBC with subtitling, signing and audio description services for the subsequent seven years. (Laughlin 2012) Other Red Bee clients include Channel 4, Canal+ and Discovery Channel.

References

Apertium. Available at: www.apertium.org.
ASLIB's Annual Conference Translating and the Computer. Available at: www.aslib.co.uk/conferences/tcc/index.htm.

'Can Translation Technology Rescue Scottish Gaelic' (2011) *Language Insight*. Available at www.languageinsight.com/blog/2011/10/25/can-translation-technology-rescue-scottish-gaelic.

Cole, Ronald A. (1996) *Survey of the State of the Art in Human Language Technology*, Cambridge: Cambridge University Press. Available at: www.cslu.ogi.edu/HLTsurvey.

Drabble, Emily (2013) 'How to Encourage Students to Pursue Languages at GCSE and A-level', *The Guardian* online, 17 May 2013. Available at: www.guardian.co.uk/teacher-network/teacher-blog/2013/may/17/languages-schools-students-gcse-alevels-mfl.

Drury, Emma (2013) 'How Can Schools Encourage Students to Take Languages Further?' *Guardian* online, 3 July 2013. Available at: www.guardian.co.uk/teacher-network/teacher-blog/2013/jul/03/schools-encourage-students-languages-advanced-level.

EAMT: European Association for Machine Translation. Available at: www.eamt.org.

EMT: European Master's in Translation. Available at: http://ec.europa.eu/dgs/translation/programmes/emt/index_en.htm.

GCSE (2011) 'GCSE Results Set Records But Spark Row Over Decline', *Modern Languages*, Huffington Post online, 25 August 2011. Available at: www.huffingtonpost.co.uk/2011/08/25/gcse-results-spark-row-over-languages-decline_n_936006.html.

Hutchins, W. John (2010) 'Compendium of Translation Software'. Available at: www.hutchinsweb.me.uk/Compendium-16.pdf.

ISO (2012). Available at: www.iso.org/iso/home/about.htm

IFABC: International Federation of Audit Bureaux of Circulations (2011) 'National Newspapers Total Circulation'. Available at: www.ifabc.org/site/assets/media/National-Newspapers_total-circulation_IFABC_17-01-13.xls.

Kelly, Nataly (2013) 'Why I Joined Smartling'. Available at: www.smartling.com/blog/2013/04/04/joined-smartling.

Koehn, Philip, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst (2007) 'Moses: Open Source Toolkit for Statistical Machine Translation', in *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague, Czech Republic, 177-180. Available at: http://acl.ldc.upenn.edu/P/P07/P07-2045.pdf.

Kwik Translator. Available at: www.lucysoftware.com.

'Language in Minority/ised Language Media'. (2013) Available at: www.aber.ac.uk/en/tfts/latest-news/news-article/title-137696-en.html.

Laughlin, Andrew (2012) 'Red Bee Media Wins New Seven Year Subtitling Deal', in DigitalSpy. Available at: www.digitalspy.co.uk/media/news/a404168/red-bee-media-wins-new-seven-year-bbc-subtitling-deal.html.

Matthews, David (2013) 'UCAS Stats Reveal Languages Decline', *Times Higher Education* online, 23 July 2013. Available at: www.timeshighereducation.co.uk/news/ucas-stats-reveal-languages-decline/2005890.article.

Microsoft (2011) 'Welsh Style Guide'. Available at: http://goo.gl/m2k1EB.

OPTIMALE 'Optimising Professional Translator Training in a Multilingual Europe'. Available at: www.translator-training.eu.

Paton, Graeme (2013) 'Anti-European Attitudes "Turning Pupils off Languages"', *Daily Telegraph* online, 20 March 2013. Available at: www.telegraph.co.uk/education/educationnews/9943592/Anti-European-attitudes-turning-pupils-off-languages.html.

Pittock, Murray (2013) 'Agenda: Our Culture Can Boost the Economy', Herald Scotland online, 29 May 2013. Available at: www.heraldscotland.com/comment/columnists/agenda-our-culture-can-boost-the-economy.21177340.

Proficiency in Welsh (2012) '2011 Census: Key Statistics for Wales, March 2011', Office for National Statistics. Available at: http://ons.gov.uk/ons/dcp171778_290982.pdf.

Prys, Delyth (2006) 'Setting the Standards: Ten Years of Terminology Work', in Pius Ten Hacken (ed.) *Terminology, Computing and Translation*, Tübingen: Gunter Narr Verlag, 41–57.

Prys, Delyth, Gruffudd Prys, and Dewi Jones (2009) 'Improved Translation Tools for the Translation Industry in Wales: An Investigation'. Available at: www.catcymru.org/wordpress/wp-content/uploads/Final%20ReportHE06fspRevised.pdf.

Prys, Gruffud, Tegus Andrews, Dewi B. Jones, and Delyth Prys (2012) 'Distributing Terminology Resources Online: Multiple Outlet and Centralized Outlet Distribution Models in Wales', in *Proceedings of CHAT 2012: The 2nd Workshop on the Creation, Harmonization and Application of Terminology*

Resources, 22 June 2012, Madrid, Spain, Linkoping: Linkoping University Electronic Press, Linkopings universitet, 37-40. Available at: www.ep.liu.se/ecp/072/005/ecp12072005.pdf.
SALT Cymru (2010) SALT Definition. Available at: www.saltcymru.org/wordpress/?p=80&lang=en#038;lang=en.
SDL products (2013). Available at: http://www.sdl.com.
Selgren, Katherine (2011) 'Languages in State Schools "Decline Further,"' BBC News, 27 January 2011. Available at: www.bbc.co.uk/news/education-12288511.
Somers, Harold L. (2004) *Machine Translation and Welsh: The Way Forward*, A Report for the Welsh Language Board. Available at: http://mti.ugm.ac.id/~adji/courses/resources/doctor/MT_book/Machine%20Translation%20and%20Welsh%20%28PDF%29.pdf.
Translation Industry Knowledge Exchange Project (TIKEP, no date) Bangor University/Welsh Assembly.
Translation Service (2010) 'Welsh Government'. Available at: http://cymru.gov.uk/about/civilservice/directorates/ppcs/translationservice/?lang=en.
Translation Technology Helps Welsh Industry (2011) Bangor University. Available at: www.bangor.ac.uk/news/full.php.en?nid=3206&tnid=3206.
University of Birmingham (2013) 'Translation Studies MA Details'. Available at: www.birmingham.ac.uk/students/courses/postgraduate/taught/arts-law-inter/translation-studies.aspx#CourseDetailsTab.
University of Bristol (2013) 'MA in Translation'. Available at: www.bris.ac.uk/sml/courses/postgraduate/ma-translation.html.
University of Leeds (2013) 'MA in Applied Translation Studies'. Available at: www.leeds.ac.uk/arts/info/125053/centre_for_translation_studies/1803/taught_programmes/2.
University of Swansea (2013) 'Postgraduate Certificate in Translation Technology'. Available at: www.swansea.ac.uk/artsandhumanities/artsandhumanitiesadmissions/translationstudies/postgraduatedegrees/postgraduatecertificateintranslationtechnology.
Vasagar, Jonathan (2012) 'A-level Foreign Languages Decline Alarms Examiners', *Guardian* online, 16 August 2012. Available at: www.guardian.co.uk/education/2012/aug/16/alevel-foreign-languages-decline.
XTM Workflow Diagram (2013). Available at: www.xtm-intl.com/files/content/xtm/images/XTM_Workflow_diagram.png.

23

A HISTORY OF TRANSLATION TECHNOLOGY IN THE UNITED STATES

Jennifer DeCamp

THE MITRE CORPORATION, MCLEAN, VIRGINIA, THE UNITED STATES

Jost Zetzsche

INTERNATIONAL WRITERS' GROUP

The history of translation technology has been highly international. Research in one country spurs research funding and ideas in another part of the world. Researchers and their research efforts move around the globe, often following business opportunities. Collaborations occur across international borders for studies and products. Professional and standards organizations bring people together for common pursuits. This global story has been chronicled at length in publications and websites by John Hutchins (1986, 2000, 2005), the Translation Automation User Society (TAUS 2010), and others. This chapter focuses only on the history of translation technology in the United States (U.S.) and should be read as a complement to other articles in the *Routledge Encyclopedia of Translation Technology* (2014) exploring translation technology from the perspective of other countries.

Complex influences and environmental factors impact the history of translation technology not only in the U.S. but also in other parts of the world. Most notable among these influences has been the exponential increase in machine-readable data in the last several decades that has required translation and has fed translation memories and Statistical Machine Translation (SMT). Other common factors include the variety of data that people want translated, including interfaces to electronic systems, social media, news broadcasts, and many new languages and dialects. Additional factors include tightening economies, which have increased cost consciousness and created a need to show greater cost accountability and Return on Investment (ROI).

The history of translation technology in the U.S. is eventful, but little of its development is unique. Differences emerge more in timing or degree. The U.S. can claim a number of firsts in translation, such as the first written discussion of machine translation, which was produced by a U.S. citizen, Warren Weaver, but likely influenced by correspondence with European colleagues. The U.S. also had the first public demonstration of Machine Translation (MT), the first SMT, the first systematic evaluation of MT, and the first integration of MT with a variety of other technologies. The U.S. can also claim the development of many supporting tools and standards, including the mainframe computer, the microcomputer, and the initial version of

the Unicode Standard, which provided a means of efficiently working with multiple languages in documents, databases, and other tools.

One unique feature of translation technology development in the U.S. has been the extensive investment by its military. Whereas European investment has focused on providing global access to government information and services, U.S. government investment has focused on improving analysis of foreign documents and communication in war environments. This focus has been highly visible in programs by the Advanced Research Projects Agency (ARPA) and the Defense Advanced Research Projects Agency (DARPA), which have funded the great majority of machine translation research in the U.S. These programs are discussed at greater length throughout this chapter.

Another unique feature has been the wide gulf between machine translation researchers and human translators until well into the first decade of the current millennium. In the U.S., MT research grew out of interest in cryptology and mathematics (a legacy of World War II), as well as the emerging field of Artificial Intelligence (AI). As Hutchins describes (2006), the field of linguistics was too immature in the 1950s to offer much value to MT researchers. In addition, in the 1950s and through the rest of the century, human translation (HT) was not a prestigious career or a field of research. Indeed, until 2000 the U.S. census classified translators alongside mimes and interpretive artists. There was no easy bridge from the study of MT to the study of human translation or vice versa.

Moreover, the drive in the U.S. to achieve the stellar research challenge of fully automated machine translation alienated many translators, who worried about job security. In addition, the U.S. had many government-sponsored research programs for machine translation but none for improving the productivity of human translators. Even in the field of enterprise-scale translation management technology, there was a focus on the tools and the business rather than on the translator.

The extensive involvement of religious groups in the development and use of translation technology is also an unusual historical characteristic. The technology for productivity tools such as translation management systems and terminology management systems grew out of work by two different groups at the Church of the Latter Day Saints (LDS Church, also known as the Mormon Church). Members of the Mormon Church and the associated Brigham Young University—most notably Alan Melby—were also active in the development of other tools and of standards that would increase the productivity of translators. Another religiously affiliated group within the U.S., the Summer Institute of Linguistics (SIL), also developed a number of software tools specific to Bible translation as well as a large number of Unicode-based fonts and three-letter codes for the world's languages that eventually became the internationally accepted code for the representation of names of languages (International Standards Organization (ISO) 639–3).

The history of translation technology in the U.S. is described below in a decade-by-decade account and in a timeline. Not all significant U.S. researchers and not all products could be described in the limited space of this chapter.

1940s: articulating a new concept

As described by Hutchins (2005), machine translation grew out of wartime experiences and research advances in cryptography. In 1949, Warren Weaver, Director of the Natural Sciences Division at the Rockefeller Center, published a 'Memorandum on Translation' describing an experience decrypting a Turkish message. He concluded: 'The most important point … is that the decoding was done by someone who did not know Turkish, and did not know that the

message was in Turkish.' MT thus started out with the implied requirements that the source language would be unknown and that the person doing the translation or decoding would not comprehend that source language anyway.

Weaver also addressed many of the issues involved in translation, particularly the diversity of meanings for a single term. His approach to looking at past translations formed the basis for the concept of translation memory (a technology described in greater detail later in this chapter). He also explored the idea of possible language universals, which helped shape research into creating an interlingua or common form into which all other languages could be converted. Finally, Weaver raised the area of technical writing as an application which might be usefully, even if not elegantly, handled this way. From the academic beginning of machine translation, the focus was primarily on the technology of MT and only secondarily on how it might be used.

The Cold War began the same year as Weaver's memo, and the Board of Directors of the Bulletin of Atomic Scientists created the highly publicized Doomsday Clock. The clock's hands were set at seven minutes to midnight, with midnight designating nuclear annihilation. In this high-threat environment, a high priority for the U.S. government was to monitor technical journals from the Soviet Union in order to identify developments related to weaponry development, among others (Vasconcellos, 1996).

1950s: demonstrating a new capability

The research field of machine translation developed quickly, first in the U.S. and then in other countries. Georgetown University in Washington, DC established the first full-time research position in 1951, filled by Yehoshua Bar-Hillel. The following year, they established the first MT research center and held the first MT conference. In 1954, Leon Dostert and Paul Garvin from Georgetown University joined International Business Machines (IBM) staff members Cuthbert Hurd and Peter Sheridan to develop the Georgetown-IBM Experiment. This experiment used 250 lexical items and six rules to translate Russian to English. Hutchins comments that the media coverage for this experiment 'was probably the most widespread and influential publicity that MT has ever received' (Hutchins 2006).

As Hutchins points out, the media represented the Georgetown-IBM Experiment not as a showcase or first effort but as a working prototype (ibid.). Project leader Dostert predicted that 'it is not yet possible to insert a Russian book at one end and come out with an English book at the other … five, perhaps three years hence, interlingual meaning conversion by electronic process in important functional areas of several languages may well be an accomplished fact' (IBM press release 8 January 1954). This assertion created expectations for high-quality, near-term automated translation that could not be met. Moreover, the media hype annoyed many influential researchers, who became highly critical of the experiment (Hutchins 2006). At the same time, there was a similar disillusionment about the related and somewhat overlapping field of AI.

However, the focus on machine translation helped to secure U.S. government funding (Hutchins 2006). The research and the research community rapidly expanded in the U.S. and internationally through the late 1950s and early 1960s. In 1958, software developed at Georgetown University was installed by the Air Force to translate technical materials from Russian to English.

MT was soon starring in movies. In 1956 Robby the Robot in the movie *The Forbidden Planet* offered visitors flawless translation assistance in '187 other languages along with their various dialects and sub-tongues' (Internet Movie Database, *Forbidden Planet* Quotes). He also supplied meals, transportation, emeralds, couture gowns, and high-quality rum, setting a public expectation not only for highly accurate automated translation but also for translation integrated

with the many other benefits available from artificial intelligence and particularly from robots. This tradition continued through dozens of movies, television shows, and books, most notably C-3PO in *Star Wars* and the Universal Translator in *Star Trek*, and still today reinforces unrealistic expectations for machine translation.

In 1959 the American Translators Association (ATA) was founded to enhance professionalism among translators and interpreters. ATA eventually not only introduced translators to technology but also was instrumental in developing standards, particularly for exchanging data across different tools.

1960s: switching from research to operations

This first wave of enthusiasm about MT was followed by a period of disillusionment. In 1960, Bar-Hillel published a report on 'The Present Status of Automated Translation of Language,' questioning the feasibility of providing Fully Automated High-Quality Translation (FAHQT) and advocating the use of post-editing. Bar-Hillel was also critical of the idea that an interlingua (a common pivot point between languages being translated) would be a simpler approach to MT.

Also and possibly even more influential was the establishment of the Automatic Language Processing Advisory Committee (ALPAC) by the National Academy of Sciences in 1958 to review the state of MT research. The resulting ALPAC report (1966), 'A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation,' concluded that 'there is no immediate or predictable prospect of useful machine translation.' It recommended expenditures in computational linguistics as basic science that 'should not be judged by any immediate or foreseeable contribution to practical translation.' It also recommended investments in the improvement of translation, specifically:

1 practical methods for evaluation of translations;
2 means for speeding up the translation process;
3 evaluation of quality and cost of various sources of translations;
4 investigation of the utilization of translation, to guard against production of translations that are never read;
5 study of delays in the overall translation process, and means for eliminating them, both in journals and in individual items;
6 evaluation of the relative speed and cost of various sorts of machine-aided translation;
7 adaptation of existing mechanized editing and production processes in translation;
8 analysis of the overall translation process;
9 production of adequate reference works for the translator, including the adaptation of glossaries that now exist primarily for automatic dictionary lookup in machine translation.

However, the ALPAC Report not only curtailed most MT research but also failed to stimulate funding and/or interest in pursuing its own recommendations in the U.S. It was not until nearly 40 years later that the U.S. human translation community gained sufficient professional and academic standing to pursue these highly productive avenues of investigation. Cuts in research funding were also due to similar disillusionment with the broader field of AI. In addition, the Mansfield Amendment was passed in 1969, requiring DARPA to fund 'mission-oriented direct research, rather than basic undirected research' (National Research Council, 1999).

While research funding was greatly diminished, machine translation began to find a commercial footing. In 1961 the Linguistics Research Center was established at the University of Texas at Austin. The following year, the Association for Machine Translation and

Computational Linguistics was established and the Air Force adopted the Mark II system developed by IBM and Washington University. In 1963 Georgetown University systems for Chinese-English MT were installed in Euratom and at Oak Ridge National Laboratory.

One of the Georgetown MT researchers, Peter Toma, left in 1962 to establish a private computer company, Computer Concepts. The software, AUTOTRAN, and the marketing arm, Language Automated System and Electronic Communications (LATSEC), became Systran (System Translation) in 1968. Computer Concepts and then Systran continued to develop rule-based MT. They also expanded their customer base to industry, which presented new requirements and challenges, as well as a more stable if less research-oriented funding base. The Air Force adopted Systran software in 1970. In 1986 Systran was sold to a French company and in 2014 to a Korean company but maintained a San Diego-based U.S. presence.

1970s: using elementary translator productivity tools

In 1969 Bernard Scott founded the Logos Development Corporation for research on an English–Vietnamese MT system, and in the 1970s Logos obtained contracts from the U.S. government to develop MT to translate weapons documentation into Vietnamese and Persian. Logos provided the capability for users to interact with the machine translation to clarify the meaning of a source sentence (i.e., the sentence that is to be translated). However, this process-interactive MT in the U.S. was designed primarily to support users without expertise in the target language (the language into which the document is to be translated).

In 1971, Bar-Hillel defined the practical roles of MT as '(1) machine-aided human translation; (2) man-aided machine translation, and (3) low-quality autonomous machine translation,' a taxonomy that still stands today with minor variations. He also noted that the concept of translation quality depended on the particular user in a particular situation: 'A translation which is of good quality for a certain user in a certain situation might be of lesser quality for the same user in a different situation or for a different user, whether in the same or in a different situation. What is satisfactory for one need not be satisfactory for another' (Bar-Hillel 1971: 75f.).

In 1972 the first word processors were introduced to the U.S. by Wang Laboratories. Word processors enabled translators to produce professional print-like documents. These documents did not have the quality of type-setting but were acceptable under most circumstances and usually a great improvement over material produced on a typewriter. Word processors enabled editing without retyping of the entire page or document. Later, word processing combined with the Kermit protocol and, still later, email or File Transfer Protocol (FTP) enabled the transmission of documents electronically.

Computer tools for translators first appeared in the 1970s. In 1970 a group around Eldon Lytle, Daryl Gibb, and Alan Melby formed at the Brigham Young University (BYU) Translation Sciences Institute. Their Interactive Translation System (ITS) was based on Lytle's Junction Grammar. As its name suggests, it was not an MT system that worked independently of the translator; instead, it consisted of a number of dictionary facilities along with a suggested translation that was to be corrected by the human translator. The tool was intended for the translation of materials for the LDS church, which eventually decided not to use the system and essentially abandoned the development. After the group's shutdown in 1980, five members of the group incorporated as Automatic Language Processing Systems (ALP Systems or ALPS) and took over ITS. ALPS was later renamed ALPNet and began providing translation services as well as technology. This company was purchased by the British translation and technology provider SDL International in 2001 and incorporated into its products (see Slocum 1985; Melby and Warner 1995).

In a parallel but independent development, Bruce Weidner (also spelled Wydner) and his brother Stephen Weidner in 1977 at BYU formed Weidner Communication to develop and market the Weidner Multi-Lingual Word Processing System. This machine translation system translated English into various Western European languages in a less interactive fashion and following a different linguistic theory (developed by B. Weidner) rather than ITS, including the automatic transfer of formatting within documents. In 1981 the company was bought by the Japanese company Bravice and eventually was acquired by SDL via the companies Intergraph and Transparent Language Solutions.

1980s: improving tools and tool applications

In 1980 Muriel Vasconcellos and her team began developing and using MT at the Pan American Health Organization (PAHO). The system was rule-based and focused on combinations of English, Spanish, and Portuguese in the health domain. This team made important contributions to the development of post-editing procedures. The PAHO Machine Translation System (PAHOMTS) is still in use and has enabled significant productivity gains when compared to manual translation (Aymerich and Camelo 2009).

Melby, who after the demise of the ITS project stayed at Brigham Young University, formed the company Linguatech and worked on developing a standalone terminology tool for translators. Independently of that, translator and consultant Leland Wright had been working on a Terminate and Stay Resident (TSR) program to support translation since the 1970s. The two combined their efforts and made their terminology tool MTX (the product was initially called Mercury, then Termex and due to copyright complications later just MTX) commercially available in 1984. It enabled translators to compile their own glossaries as a separate task or while working in documents. This tool employed a Machine-Readable Terminology Interchange Format (MARTIF) which was based on Standard Generalized Markup Language (SGML) that formed the basis of today's Extensible Markup Language (XML)-based TermBase eXchange (TBX) standard.

Industry began experimenting with MT in the early 1980s to handle the translation or localization of its software and product documentation into foreign languages. As Weaver pointed out in 1949, technical manuals were a good application for MT because of the consistent and narrowly constrained use of terminology and because of the lack of need for very high-quality translation. Unlike much of the translation being conducted by the U.S. government, which was into English, localization efforts in the U.S. were mainly out of English into the languages where the companies wanted to sell their products. Moreover, a driving force for using MT was to decrease time-to-market by reducing the time from when the English source text was completed to when the translation was completed. Much of the work, particularly development of the translation lexicons, could be completed while the source text was being finalized.

To help bridge the gap between MT output and an acceptable level of translation quality, companies began developing means of limiting the input to what the MT system could handle and providing post-editing. Maria Russo and her team at Xerox Corporation developed tools that would check the original English documentation for terminology, punctuation, and sentence construction. The tools would alert the writers and/or pre-editors to change the text in order to provide better input. In addition, Xerox developed tools that highlighted sections of the output where there was low confidence in the translation (e.g., which contained not-found words or complex sentence constructions), thus focusing the efforts of translators in post-editing. In the late 1980s and early 1990s, Xerox marketed a combination of its pre-

editing and post-editing tools, Systran and Mechanical Translation and Analysis of Languages (METAL) MT, and an IBM mainframe, together known as DocuTrans.

As the field of localization matured throughout the 1980s, focus was placed on writing software and documentation so that it could be more easily localized, a process known as 'internationalization.' As the field emerged still further, practitioners actively worked with the end-to-end processes of developing, delivering, and maintaining products, a process known as 'globalization.' In 1989 the Localisation Industry Standards Association (LISA) was founded in Switzerland to take on challenges of providing standards for localization, internationalization, and globalization, particularly the standardization of the format for exchanging translation memories and terminology databases.

Research efforts for MT in this decade were funded mainly by operational efforts. Developers, including those at Carnegie Mellon University (CMU), found ways to make components more modular and thus reusable, with rules defining the source language, the target language, and/or the transfer from the source to the target. Experiments were conducted with data-driven MT, a concept first described by Weaver in 1949. In 1986 Xerox developed TOVNA, a pilot of an example-based data-driven MT system designed for production (TAUS 2010).

The use of statistical frequency also emerged within rule-based MT in order to better adapt translation for specific subject areas. For instance, 'bank' was statistically more likely to be a river edge in agricultural documents and a building in financial ones. CMU also experimented with indirect translation through intermediary representations and SMT. In 1986, Peter Brown and his colleagues at IBM presented their experiments on SMT, sparking widespread interest in this approach.

The availability of cheaper microprocessors made MT and Machine Assisted Human Translation (MAHT) more readily available to a wider number of translation organizations. Personal computers were often difficult to use for translation, particularly of languages with non-Latin scripts. For many languages (e.g., Chinese and Arabic), software was available only from foreign companies or foreign branches and was not well supported in the U.S.

1990s: adding in computer-assisted translation and SMT

The decade of the 1990s widely exposed translators to MAHT, also known as Computer-Assisted Translation (CAT) or more recently as Translation Environment Tools (TEnTs). CAT tools include translation memory, terminology management, and project management software. The translation memory component allows the translator to build up databases of translated material on a segment (typically a sentence) level on-the-fly during translation or alignment of existing translation, and it then enables the translator to leverage that segment against newly translatable content. The terminology management system makes it possible to enter terms along with documentation about the term. This process complements and extends the functionality of the translation memories by further controlling the usage of relevant terminology. The project management components enable translators to analyze the translatable texts and productivity as well as quality assurance of the translations.

IBM released Translation Manager (TM/2) in 1992. It was discontinued as a commercial offering in 2002 and revived once again in 2010 as the open-source product OpenTM2.

Trados, today's market-leading CAT tool, was originally developed by the German translation company of that name. In 1990 the company released its first commercial product, MultiTerm (Trados' terminology management component), and in 1992 Workbench (Trados' translation memory application) for the Disk Operating System (DOS). In 1994 Trados

released a Windows version with a Microsoft Word interface. Between 2002 and 2005, Trados was headquartered in the U.S., and in 2005 it was purchased by SDL International.

In 1991 the first version of the Unicode Standard was published, based on the work of Joe Becker from Xerox Corporation and Lee Collins and Mark Davis from Apple Computing. This widely implemented standard—a subset of which became ISO/International European Commission (IEC) 10646—provided a single standard for encoding most of the world's languages and writing systems. This common standard was critical for manipulating large quantities of corpora that could be used as the basis both for statistical machine translation and translation memory systems.

Translation also became more oriented toward conveying information in a format appropriate and appealing to specific cultures, a process aided by the Unicode Consortium's Common Locale Date Repository (CLDR). The CLDR provided programming-accessible information on locale or country-specific conventions. Such conventions included formatting of numbers, calendar information, and telephone codes, as well as a broad array of other information, including languages, character sets, and sorting rules.

In 1991, due in part to the efforts of Vasconcellos, the Association for Machine Translation in the Americas (AMTA) and the International Association for Machine Translation (IAMT) were founded, along with the European Association for Machine Translation (EAMT) and the Asian Association for Machine Translation (AAMT). With meetings alternating between local organizations and the IAMT, the exchange of information regarding MT development and use rapidly increased.

In 1991, DARPA started a Machine Translation initiative under Charles Wayne that was carried forward by George Doddington and Tom Crystal. In 1992, DARPA started a multifaceted Human Language Technology (HLT) program under Doddington. DARPA's Machine Translation initiative (1991-1994) funded three competing approaches. The first was a rule-based, interlingual approach called Pangloss, which was a joint effort of Carnegie Mellon University (CMU), University of Southern California (USC) Information Sciences Institute (ISI), and New Mexico State University (NMSU). The second was a statistical approach by IBM called Candide. The third was a combination of statistical and linguistic techniques from Dragon Systems called LINGSTAT. In 1992 Caterpillar and CMU launched the Caterpillar Machine Translation System based on a controlled language known as Caterpillar Technical English (CTE). CTE was based on Caterpillar Fundamental English (CFE), which was developed in 1972 to better match Caterpillar's technical manuals to their readers' capabilities (Kamprath, Adolphson, Mitaruma, and Nyberg, 1998). Also in 1992, CMU and Microsoft developed statistical methods that were particularly helpful for improving text alignment and extracting lexical and terminological data.

The mid-1990s also saw the development of several Translation Management Systems (TMS) in the U.S. While TMSs typically have applications for translation memory and terminology management, the emphasis is on the management or translation processes within large corporations. This process also includes sending out projects to outside translators who have to work on the translation in the environment that is provided by the tool.

The earliest of these companies was Uniscape, which was founded in 1996. Uniscape also built the first translation portal in 1999. (Translation portals are websites where translation jobs are posted and translators can bid on them.) As of this writing the largest portal is the U.S.-based ProZ.com. Uniscape was purchased by Trados in 2002, but the software did not continue as a separate or distinguishable product. In 1997 another large-scale TMS product, Ambassador Suite, was launched by GlobalSight Corporation. It was sold to the Irish translation company

Transware, Inc. in 2005. In 2008 the U.S. language service provider Welocalize purchased the company and renamed the software GlobalSight.

Idiom's WorldServer was launched in 1998. Like its competitors, Idiom experienced some quick growth but encountered difficulties as the landscape changed. In 2005 it started a program to give away fully functional free licenses to qualifying language service providers, causing a lot of excitement in the translation industry. This program lasted until 2008 when Idiom was bought by SDL; SDL did not continue the language service provider (LSP) program, originally planning to completely integrate WorldServer in its own product offering and shut it down as a separate product. This caused concern among some of the large enterprises that were still using the WorldServer product, however, so SDL reversed its decision and now (2014) continues to sell SDL WorldServer as a separate product.

Additional tools that support translators include systems such as Highlight from Basis Technology, appearing 1998, which provides automatic transliteration (i.e., phonetic representation of a term) in any of a set of different standards. The tool also enables users to type Arabic in Latin script roughly the way it sounds, with the system resolving the spellings into Arabic text. This capability enables translators who are not familiar with Arabic keyboards to efficiently type in that language.

Another development in the late 1990s was the use of the Internet to provide MT. In 1988 Globalink had already provided the first MT web service by email. In 1996 Systran began offering free web-based translation, the first available service of that kind in the U.S., though it had offered a paid service in the 1980s in France. The next year, Systran integrated its software with the then widely used search engine AltaVista to create the highly popular Babelfish.

Web-based MT provided easy access to MT with no systems administration and often with no licensing. It enabled people to experiment with MT and thus to learn about the technology. It also enabled increasingly sophisticated access of the Internet, enabling users to search across languages and to rapidly translate the materials they found. Web-based MT also helped change many environments from production-centric to user-centric. Translation had traditionally been production-based. An organization would produce or receive documents in one language, translate them, review them, and make them available to a set of customers. It was an expensive and time-consuming process. MT, however, has enabled a different process where end users find their own materials and use MT to get a rough or gist translation. These users may check a different online machine translation system, ask friends who may have limited translation ability, or look into chat rooms online rather than contacting certified translators. The downside is that decisions about translation quality and the need for retranslation are often being made by people without appropriate skills and training.

Web-based MT systems have also greatly influenced the general public's image of MT. Since these systems are usually not user-definable, their output is often less accurate than the results of customizable MT.

2000s: going online

In 2000, DARPA started the Translingual Information Detection, Extraction and Summarization (TIDES) program under Gary Strong. In 2001, Wayne strengthened and refocused the TIDES program and added the Effective, Affordable, Reusable Speech-to Text (EARS) program. All of these research programs included strong MT evaluation components.

In 2001, the Society of Automotive Engineers (SAE) committee published the SAE Standard J2450. This standard was the first effort to provide detailed guidelines for translations in a

specific domain into any target language. Also in 2001, terrorists attacked the Twin Towers in New York and the Pentagon in Virginia, starting the War on Terrorism and unprecedented spending on improving translation and related language technologies. That year, DARPA established the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) under Mari Meida.

That year, the U.S. National Institute for Science and Technology (NIST) began its annual MT benchmarking, a process that spurred the development of better MT and better MT evaluation methods. Most prominent among these methods was the Bilingual Evaluation Understudy (BLEU), a method of automatically scoring machine translation output that was fast and—for the end user—free. It was devised by IBM and was widely adopted. BLEU enabled researchers to iterate quickly, propelling steady advances in MT accuracy (similar to what the automated calculation of Word Error Rate [WER] had been doing for speech-to-text technology since 1987.)

BLEU compared the output from the MT system with multiple translations from professional translators (i.e., with a 'gold standard'). Scores were dependent on the distance of the words in the MT output from the words in the gold standard. The scores correlated—although not always reliably—with assessments by humans of the translation quality. However, the system was only usable for assessment of MT engines rather than assessments of translations, since the source text needed to be identical to what was provided for the gold standard. BLEU also resulted in various anomalies: with BLEU, the sentences 'John hit Mary' and 'Mary hit John' are rated as similar and the translation thus received a high score; but the sentences 'She participated in the Olympic Games' and 'She participated in the Olympic Competition' were rated as distant and the translation thus receives a low score. Alternative systems evolved, including a NIST program based on BLEU but with weighted scores, Metric for Evaluation of Translation with Explicit Ordering (METEOR), and the Language-independent Model for Machine Translation Evaluation with Reinforced Factors (LEPOR) and enhanced LEPOR (hLEPOR). Efforts were made to provide measures that better reflected human assessment.

In 2005, DARPA established the Global Autonomous Language Exploitation (GALE) program under Joseph Olive, which continued and substantially expanded the research started in TIDES. They established the Multilingual Automatic Document Classification Analysis and Translation (MADCAT) program the following year, also under Olive. These programs focused on improving translation, as well as other language technologies, for Arabic and other Middle Eastern Languages as well as for Chinese. In 2009, DARPA established the Robust Automatic Transcription of Speech (RATS), which improved transcription and thus translation of speech.

ATA continued its work on text-based evaluation in order to provide certification in translation for its members. A detailed method of hand scoring that was based on a numerical metric was developed. Each text was scored by multiple raters who received significant training. However, there was little effort to bring together work being done in the arenas of human translation and machine translation. Doug Oard worked late into the decade to tie MT quality to language proficiency scores (i.e., specifically to Defense Language Proficiency Test scores). This approach took into account the issues of different requirements for different users and situations, as described by Bar-Hillel in 1971. However, it did not produce the common measurement that a text-based evaluation system could have afforded.

Throughout this decade, the quantity of material that needed translation was exponentially growing, due to many reasons, including the continued concern about international terrorist attacks. Needs included translation of a broad range of media, from yellow paper sticky notes

to hard drives, in many languages such as Dari where MT, other language tools, and sometimes even data with which to train the tools were scarce or non-existent. Needs also included face-to-face negotiations and information exchange, areas covered traditionally by interpretation rather than translation. The U.S. National Virtual Translation Center (NVTC) was established to handle and/or coordinate surge translation needs from across the government.

Expanding translation corpora and particularly expanding Internet corpora produced significant gains with SMT. In 2002 Kevin Knight and Daniel Marcu founded Language Weaver, based on SMT. In 2003 Franz-Josef Och won the DARPA competition for speed of MT with an SMT engine. He later became head of machine translation at Google. In the late 2000s various companies, including Apptek and Systran, released hybrid MT systems using rule-based MT and SMT.

While the translation products of more than a dozen European companies appeared in the U.S. market, the only standalone tools to directly support human translators developed in the U.S. were Lingotek in 2006 and Fluency in 2010. The developers of both of these tools were based in Utah and came from within the LDS Church. In fact, at the time of writing, the LDS Church uses Lingotek as its preferred tool for its crowdsourcing translation. While Lingotek was originally marketed to government entities, translation companies, and freelance translators, the current marketing effort is focused on larger corporations with translation needs. Another tool, MadCap Lingo, was first released in 2010 as an add-on to the documentation authoring system of Madcap Software. In 2009 Systran released a version of its own server-based software using post-editing capabilities.

In addition, the Internet has also made crowdsourcing translation possible, enabling hundreds or thousands of contributors to translate and evaluate each other's work. This process has been particularly successful for large companies that invested in the development of sophisticated collaboration platforms where their often passionate users could work with each other. One of the first efforts in this area was pioneered by Google in 2001 with its now discontinued Google in Your Language (GIYL) campaign, which translated Google Search into 118 languages. In 2007 Facebook started to roll out its translation crowdsourcing application, which has now been translated into more than 70 languages. Many other organizations have employed the same model of building crowdsourcing platforms and engaging their users, including versions of WordNet and Wikipedia in many languages. There are a large number of readily available applications for a variety of crowdsourcing purposes, from video subtitling (e.g., Amara) to language learning and website translation (e.g., Duolingo).

While this kind of integration involves incorporating translation capacities for the product into the actual product (and binding users by fostering greater loyalty), the first two decades of the twenty-first century have also seen another kind of translation and translation technology integration. Early pioneers IBM and Xerox were examples of companies that invested heavily in translation technology and machine translation as a separate endeavor for product documentation and interfaces, but now companies such as Microsoft and Google view translation as integral to each of their primary business ventures. Translation and the technology necessary to carry it out have reached a level of commercial relevance rarely—if ever—seen before, from the need to translate social media formats with non-traditional means to integrating translation into development tools and being able to control applications with language commands.

In 2007 Google began replacing a version of Systran with its own statistical machine translation engine. Today Google Translate supports more than 70 languages (by pivoting through a common third language, this results in more than 5,000 language combinations). This service is integrated into many different Google services and products and a large number

of third-party products—including virtually all CAT tools—through its application programming interface (API). In 2009 Google also released its own CAT tool, Google Translator Toolkit, which relies heavily on the machine translation backbone but also uses features such as translation memory and terminology databases. The goal of this tool—like any of the other Google Translate features that encourage users to correct machine translation output—is to refine the data that the machine translation engine relies on.

Microsoft has followed a similar path with its Microsoft Bing Translator. In 2012 it released the Microsoft Translator Hub, which enables users to build their own SMT engine by uploading their translation memory data and refining existing or building completely new machine translation engines. The stated goal for this project is also to collect high-quality data to continue the optimization of existing machine translation capabilities. One organization that uses the Microsoft Translator Hub system today is the LDS church. Its key proponent, Steve Richardson, was one of the senior developers for Microsoft's MT efforts and before that a member of the development at BYU for ITS in the 1970s. He now (2013) spearheads the implementation of Microsoft's MT system at the LDS church.

2010s: facing the Big Data challenge

When the earthquake hit Haiti in January of 2010, both Google and Microsoft were able to use materials collected by a team at CMU and other sources to release a version of a Haitian Creole MT engine within days, an achievement due to the extensive research already conducted on SMT. In 2010, IBM launched nFluent, which included post-editing.

In 2010 SDL acquired Language Weaver, thus adding MT as a rough and usually unvetted backup to Trados translation memory. SDL soon added additional MT engines, as have virtually all other providers of CAT tools and translation management systems. In addition to translation memory and machine translation features, many tools also provide access to very large online corpora and translation memories (such as the TAUS Data Human Language Project or MyMemory). However, with the basic framework of the CAT tool, the translator can choose to view any or all matches of text with found translations and approve it or make changes.

In 2011, DARPA established the Broad Operational Language Translation (BOLT) program and in 2012, the Deep Exploration and Filtering (DEFT) program, building on the work of GALE. These programs—both under Bonnie Dorr—addressed challenges such as the diversity of languages, dialects, and registers. Social media was a particular concern, with problems such as lack of complete sentences, a lack of punctuation, missing characters, inconsistent spelling, non-standardized transliterations, a substitution of numbers and special characters for some letters (e.g., 'l8r' for 'later'), and short text segments (a tweet, for instance, can be only 140 characters). To address these challenges, DARPA began to incorporate conversational analysis (Dorr 2012).

In 2010 and 2012 ATA and AMTA co-located their annual conferences, arranging special events to educate MT researchers about what translators do and to educate translators on how MT can help. This blending of cultures stimulated more rigorous research in computer-assisted human translation.

The quantity of data for SMT and translation memory continues to grow exponentially, and researchers now focus on specialized domain-specific subsets of data to improve quality. While SMT has long been solely driven by mathematicians, linguists have now been asked to come back to help improve the systems based on their expertise. MT researchers are also using

advances in interoperability to pass data between tools in order to get the benefits of each to improve the MT.

A third characteristic of this time period is an increased focus on ROI. While assessments on the ROI of tools have been conducted for decades, there has been deeper analysis of the measurement of translation productivity. These studies further segment the types of translation, translators, and tools, and provide more rigorous standards for evaluation.

Future

There are many challenges still to address in translation technology, both within the U.S. and worldwide. The quantity of data and thus the need for translation continues to grow exponentially, affecting translation technology worldwide. More people are gaining access to computer communication devices and are using those devices with their own languages as well as with languages of wider communication. Language is changing at an unprecedented rate, particularly for social media. More research is being conducted concerning the extensive translation processes. More standards are becoming available for improving translation and interoperability of translation memory and terminology data. There is extensive international sharing and cooperation.

Researchers and practitioners are just beginning to address the following challenges:

1. **Development of automated evaluation for new translations.** A reliable automated evaluation system would provide a more flexible common measurement between output from various systems, including MT and human translators. Such an evaluation system would also enable modeling of the right level of accuracy for particular applications. Today, automated evaluation is done with a limited selection of source texts that have already been translated by human translators.
2. **Study of the translation process.** While efforts are underway to study the translation process, particularly by cognitive linguists through eye-tracking, think-aloud protocols, keyboard logging, and electroencephalograms, there is still much that is unknown about how people produce effective translations.
3. **Combination of TM, MT, and Terminology Management in a more integrated fashion.** This recommendation, provided by Melby and Wright in their article in this encyclopedia on translation memory, would enable translators to more efficiently prioritize terminology.
4. **Provision of more accurate indicators and/or assessments of translator proficiency.** More study is also needed of how to assess the ability of a translator to provide a particular kind and quality of translation.
5. **Systematic coverage of large numbers of new language pairs.** Of the more than 6,000 languages in the world, there are MT systems for only a few hundred language pairs at most, and those are of varying quality.
6. **Means of increasing translation accuracy and efficiency.** Researchers are experimenting with means to increase accuracy and efficiency for MT and CAT. Among the many issues is the difficulty of identifying high-quality translation on the Internet.
7. **Resolution of issues related to privacy.** Use of Internet data in translation memories for machine translation and for TMS raises issues of protecting the privacy of the authors of that data. Practices and policies to protect privacy are evolving independent of translation technology, but will impact on the use of such tools.

8 **Addressing the rapidly changing field of social media.** As discussed, social media presents a significant number of factors that once would have made it a poor candidate for the use of MT.

9 **Provision of improved speech translation.** Despite extensive progress made via the DARPA Babylon, TRANSTAC, and BABLE programs and extensive research in the private sector, speech translation still presents many challenges. Improved speech translation is making possible new systems for communication between people and between people and their computer devices.

10 **Education of the U.S. public on the need for and processes of human and machine translation.** While the War on Terror increased the military's understanding of the need for translation, there is still a broad section of the U.S. population that believes that fluency is an indicator of the fidelity of a translation. That population also believes that quality translation can be accomplished by anyone who knows the language or even by anyone who has studied the language. Greater understanding of the need for languages and the work involved in developing human and machine resources will better prepare the U.S. for an increasingly international future.

TIMELINE

1949 Warren Weaver, Director of the Natural Sciences Division at the Rockefeller Center, sends his 'Memorandum on Translation,' outlining issues and directions for machine translation
The Cold War begins

1951 Yehoshua Bar-Hillel at Georgetown University becomes the first full-time MT researcher

1952 Georgetown MT Research Center is established
First conference on MT is held at Georgetown University

1954 The Georgetown-IBM Experiment generates significant media coverage
Journal of Machine Translation is launched

1956 *The Forbidden Planet* popularizes MT with Robby the Robot

1957 Noam Chomsky publishes 'Syntactic Structures,' enabling the design of rule-based MT systems
U.S. Air Force starts using Russian-to-English MT for scientific work
American Translators Association is established
Bar-Hillel publishes a report on 'The Status of Automated Translation of Language'

1960 Bar-Hillel publishes 'Demonstration of the Nonfeasibility of Fully Automatic High-Quality Translation,' concluding that 'there is no immediate or predictable prospect of useful machine translation' and recommending investments in improving human translation

1961 Linguistics Research Center is founded at the University of Texas at Austin

1962 Association for Machine Translation and Computational Linguistics is established
U.S. Air Force adopts the Mark II system developed by IBM and Washington University
Peter Toma establishes Computer Concepts, with the marketing arm of Language Automated System and Electronic Communications (LATSEC)

1963 Georgetown University systems for Chinese–English MT are established in Euratom and at Oak Ridge National Laboratory

1965 U.S. troops are sent to Vietnam

1966 The National Academy of Sciences publishes the Automatic Language Processing Advisory Committee (ALPAC) report, recommending that research for MT be discontinued

1968 The first MT company, Language Automated Translation, System and Electronic Communications (LATSAC), is founded by Peter Toma; it later becomes Systran

1969 Logos is founded to work on MT for the U.S. government

1970 U.S. Air Force adopts Systran software
Group around Eldon Lytle, Daryl Gibb, and Alan Melby start to develop Interactive Translation System (ITS) at Brigham Young University (BYU)

1971 Bar-Hillel defines roles of MT and makes the distinction that translation quality depends on the particular users and their situations

1972 Wang launches the first office word processor in the U.S.
Caterpillar introduces Caterpillar Fundamental English

1975 The U.S. Air Force develops the QUINCE Chinese-to-English MT system and work on a German-to-English model

1976 Logos starts developing English-to-Persian MT to support sales of military systems to the Shah of Persia

1977 Weidner Communications Corporation is founded to produce computer-assisted translation for computer systems
Smart Al Inc. is established for controlled language

1978 Xerox starts using Systran to translate technical manuals

1980 ITS is shut down and technology is taken over by ALPS

1982 Linguatech develops MTX, a standalone terminology tool
Caterpillar launches Caterpillar Technical English
NEC demonstrates speech translation

1984 Trados is established in Germany

1985 The Pan American Health Organization (PAHO) implements ENSPAN for English-to-Spanish translation

1986 A Center for Machine Translation Systems is established at Carnegie Mellon University, focused on knowledge-based MT
Xerox develops the first data-driven MT system, TOVNA
Systran is sold to a private French company

1987 First Machine Translation Summit
Beginning of Unicode project

1988 Globalink is founded, providing the first personal MT software and the first MT web services
Peter Brown reports on experiments with statistical MT
Joe Becker publishes a proposal for Unicode
Xerox starts marketing its pre- and post-editing capabilities, tied to Systran and METAL MT on an IBM mainframe

1989 Localisation Industry Standards Association (LISA) is founded
IBM runs an R&D project in statistical MT

1990 IBM launches the first PC
Trados, a German company, releases its first commercial product, MultiTerm

1991 First version of the Unicode Standard is released
International Association for Machine Translation (IAMT) and Association for Machine Translation in the Americas (AMTA) are founded
DARPA starts an MT initiative

1992 DARPA establishes the Human Language Technology Program, including the Machine Translation Evaluation Program.
Language Data Consortium is founded with a grant from ARPA.
DARPA establishes the Translingual Information Detection, Extraction and Summarization (TIDES) program, Effective Affordable Reusable Speech-to-text, and Babylon.
IBM releases CAT tool Translation Manager (TM/2)
Trados releases its first translation memory application, Workbench
Caterpillar and Carnegie Mellon University launch Caterpillar's Automated Machine Translation
1994 Association for Machine Translation in the Americas (AMTA) holds its first conference
1995 Lernout & Hauspie acquire translation and other language technology, including METAL, with investments from Microsoft
1996 Systran offers free translation on the Internet in the U.S. (having offered a similar service in France in the 1980s)
Uniscape is founded
1997 AltaVista Babelfish is launched using Systran software
GlobalSight launches Ambassador Suite
Robert Palmquist develops first translation system for continuous speech
1998 Idiom launches WorldServer
2000 The U.S. Department of Labor recognizes translation as its own field with a category in the 2000 census
DARPA establishes the Translingual Information Detection, Extraction, and Summarization (TIDES) program
2001 DARPA establishes the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program and the Effective Affordable Reusable Speech-to-text (EARS) program
Lernout & Hauspie declare bankruptcy
SDL purchases ALPNet and Transparent Language software and incorporates the technology into its products
IBM launches WebSphere
Phraselator speech-to-speech translation system is tested in Afghanistan
National Institute for Science and Technology (NIST) launches its first MT competition
Terrorists attack New York and Virginia in what becomes known as 9/11
2002 Globalization and Localization Industry Standards Association (GALA) is established
Language Weaver is founded
An SMT entry from an ISI team headed by Franz-Josef Och wins the DARPA/NIST MT competition
Trados establishes headquarters in the U.S.
Trados acquires Uniscape
2003 Yahoo! acquires AltaVista Babelfish
2004 The Translation Association Users Society (TAUS) is founded and holds its first forum
2005 DARPA establishes the Global Autonomous Language Exploitation (GALE) program
2006 First U.S.-based CAT tool Lingotek is launched
2008 Welocalize acquires Transware
SDL acquires Idiom
DARPA establishes the Multilingual Automatic Document Classification Analysis and Translation (MADCAT) program

2009 Welocalize releases open-source product based on GlobalSight
Apptek launches a hybrid MT system using statistical MT (SMT) and rule-based MT (RBMT)
Systran releases version 7, which provides a hybrid version of RBMT and SMT as well as a post-editing module
Google releases CAT tool Google Translator Toolkit

2010 ATA and AMTA hold co-located conferences
Asia Online launches Language Studio, including MT and post-editing
Language Weaver launches quality confidence measure
Language Weaver is acquired by SDL
TM/2—developed by IBM—is released as open source in Open TM2
CAT tool Fluency is launched
DARPA establishes the Robust Automatic Transcription of Speech (RATS) program

2011 Committee in ASTM is founded to focus on translation standards
DARPA establishes the Broad Operational Language Translation (BOLT) program

2012 Yahoo! replaces Babelfish with Microsoft Bing Translator
Microsoft launches Microsoft Translator Hub
DARPA establishes the Deep Exploration and Filtering of Text (DEFT) program
Intelligence Advanced Research Projects Activity (IARPA) establishes BABEL program for speech translation and generation

2013 SDL establishes a wholly owned U.S. subsidiary

References

Aymerich, Julia and Hermes Camelo (2009) 'The MT Maturity Model at PAHO', in *Proceedings of the Conference for the International Association for Machine Translation* (IAMT), 26 August 2009, Ontario, Canada.

Automatic Language Processing Advisory Committee (ALPAC) (1966) *Language and Machines: Computers in Translation and Linguistics*, Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington DC, National Academy of Sciences.

Bar-Hillel, Yehoshua (1960) 'The Present Status of Automatic Translation of Languages', *Advances in Computers* 1: 91–163.

Bar-Hillel, Yehoshua (1971) 'Some Reflections on the Present Outlook for High-quality Machine Translation', Linguistics Research Center: The University of Texas at Austin. Available at: http://www.mt-archive.info/LRC-1971-Bar-Hillel.pdf. Retrieved 9/16/2013.

Dorr, Bonnie J. (2012) 'Language Research at DARPA: Machine Translation and Beyond', in *Proceedings of the 10th Biennial Conference for the Association for Machine Translation in the Americas* (AMTA), 28 October –1 November 2012, San Diego, CA.

Chomsky, Noam (1957) *Syntactic Structures*, The Hague and Paris: Mouton.

Hutchins, W. John (1986) *Machine Translation: Past, Present, Future*, Chichester: Ellis Horwood.

Hutchins, W. John (2000) *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*, Amsterdam and Philadelphia: John Benjamins.

Hutchins, W. John (2005) 'The History of Machine Translation in a Nutshell'. Available at: http://hutchinsweb.me.uk/Nutshell-2005.pdf. Retrieved 9/15/2013.

Hutchins, W. John (2006) 'Machine Translation: History', in Keith Brown (ed.) *Encyclopedia of Languages and Linguistics*, 2nd edition, Oxford: Elsevier, 7: 375–383.

IBM Press Release (1954) '701 Translator'. Available at: http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html, retrieved 9/16/2013.

Internet Movie Database 'Forbidden Planet Quotes'. Available at: http://www.imdb.com/title/tt0049223/quotes.

Kamprath, Christine, Eric Adolphson, Teruko Mitamura, and Eric Nyberg (1998) 'Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English', in *Proceedings*

of the 2nd International Workshop in Controlled Language Applications, Carnegie-Mellon University, Pittsburgh, PA.

Melby, Alan and Terry Warner (1995) The Possibility of Language: A Discussion of the Nature of Language, with Implications for Human and Machine Translation, Amsterdam and Philadelphia: John Benjamins.

National Research Council (1999) 'Developments in Artificial Intelligence', in *Funding a Revolution: Government Support for Computing Research*, Washington, D.C.: National Academy Press. Available at: http://web.archive.org/web/20080112001018/http://www.nap.edu/readingroom/books/far/ch9.html. Retrieved 9/16/2013.

Slocum, Jonathan (1985) 'A Survey of Machine Translation: Its History, Current Status, and Future Prospects', *Computational Linguistics* 11(1): 1–17.

Translation Automation User Society (TAUS) (2010) 'A Translation Automation Timeline'. Available at: https://www.taus.net/timeline/a-translation-automation-timeline. Retrieved 8 August 2013.

Vasconcellos, Muriel (1996) 'Trends in Machine Translation and the Forces that Shaped Them', in *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, 2-5 October 1996, Montreal, Canada, 1–10.

Weaver, Warren (1949) 'Memorandum on Translation'. Available at: http://www.mt-archive.info/Weaver-1949.pdf.

PART III

Specific topics in translation technology

This page intentionally left blank

24
ALIGNMENT

Lars Ahrenberg

LINKÖPING UNIVERSITY, SWEDEN

The alignment concept

In the context of translation, alignment refers to a process of relating parts of a source text to parts of a target text. As with the term translation itself, alignment may also refer to a product, the outcome of an alignment process.

The purpose of alignment is to capture relations of equivalence or correspondence in a translation. As these notions have no generally agreed definitions, and can be interpreted in different ways, it must be recognized that there often cannot exist a single, correct alignment. Instead, as with translations themselves, different alignments can be judged as more or less appropriate, given some relevant criteria for their intended use.

Historically, the notion of alignment in translation technology is intimately bound up with the interest in parallel corpora, or *bitexts*, that emerged in the last half of the 1980s, as a way to deal with the shortcomings of the then existing technologies for machine translation and translation aids. Isabelle (1992: 76–89) attributes the idea of alignment, or 'methods for reconstructing correspondences in pre-existing translations' to Martin Kay who used the term already in a 1988 precursor to the article 'Text-translation alignment' (Kay and Röscheisen 1993: 121–142).

In some works (e.g., Simard *et al.* 1992: 67–81; Melamed 2001), a distinction is upheld between alignment and correspondence. Alignment is then restricted to relations that are monotonic, so that if $<s_i, t_j>$ is a pair of aligned units, then a source unit $s_k<s_i$ can only be aligned to a target unit t_l if $t_l \leq t_j$. In this article the term alignment is used in the wide sense, as is currently normal, and modifiers such as 'monotonic' or 'functional' are used for alignments that are restricted in the relevant sense.

Alignment processes are classified according to the size of the units that are to be related. We talk of *sentence alignment* when the minimal unit is a sentence, or some text unit of equivalent status, and of *word alignment* when the minimal unit is a word. Aligning units in between words and sentences is called *sub-sentential alignment* or *phrase alignment*. The latter term, alongside *tree alignment*, is also used when source and target sentences have been assigned syntactic analyses in the form of trees, and the alignment relates nodes of the source tree to nodes of the target tree. This topic is not covered here.

Definitions and notation

It is common in the literature to call the two sides of a parallel corpus as the Foreign and English side, respectively, using the letters f and e, in various incarnations, to denote their parts. Here this convention will be followed. It is also common to regard one of the sides as the source side, usually the Foreign side, and the other, English side, as the target side. The following notational conventions will be used:

$P = <F, E>$ is a parallel corpus with the two halves F and E.

$F = <F_1, F_2, \ldots F_K>$ is the Foreign half divided into K sentence-level text units. Similarly, E is divided into L units $<E_1, E_2, \ldots E_L>$.

$A = <A_1, A_2, \ldots A_S>$ is an alignment of P into S pairs, and we can write $P^{(s)}=<f^{(s)},e^{(s)}>$. We will refer to this alignment as a sentence alignment and the pair itself as a *sentence pair*.

$\mathbf{f} = f_1, f_2, \ldots, f_J$ is a Foreign sentence with J words.

$\mathbf{e} = e_1, e_2, \ldots, e_I$ is an English sentence with I words.

An alignment **a** of **f** and **e** is a subset of the Cartesian product of their word positions: $\mathbf{a} \subseteq \{<j,i>: j=1,\ldots, J; I = 1,\ldots, I\}$. A pair <j,i> is called a *link*. A matrix with J rows and I columns where each element can be either 1, to indicate a link, or 0, to indicate no link, is called an alignment matrix for the pair <**f**, **e**>. The alignment problem can then be defined as a search for the best alignment(s) from the space of 2^{JI} possible alignments.

The alignment task

Alignment algorithms have to solve several problems. A first problem concerns how alignment characteristics are to be modeled. This is usually done by some sort of statistical model. More often than not several models are used and then one must also decide how to combine them, for instance, by assigning a weight to each model that indicates its importance for solving the problem.

A second problem is determining values for the model parameters, and, if weights are used, their values. This can be done by fiat, e.g., giving equal weights to all models, by direct computation on available data, or by a statistical learning process.

Third, there is the search problem of how to find the best alignment according to the models. This problem is intimately bound up with the other two. Without restrictions on the models the candidate alignments are just too many and an exhaustive search is out of the question. Restrictions on the models can be set as hard constraints, for example, by only considering functional or monotonic alignments. Even without such restrictions, learning the parameters and finding the best alignments will often be done iteratively, by using learning schemes such as Expectation-Maximization (EM) (Dempster *et al.* 1977: 1–38). Approximative search regimes, such as beam search, are also commonly used. In those regimes, partial alignments are compared based on their likelihood, or scores, and those that have too low scores are not considered further. This may happen if they are not among the N best alternatives, or if their relative score, compared to the best alternative, falls below a given threshold.

Evaluation

Alignment is a prerequisite for many tasks relating to translation technologies, including statistical machine translation, terminology extraction, population of bilingual lexicons and search in translation corpora. For this reason extrinsic evaluation is important for alignment systems. However, not least for the sake of system comparisons, intrinsic evaluation is also motivated and much used.

Precision (Pr) and recall (Rc) are basic metrics for intrinsic evaluation. These require comparisons with a gold standard. With L the set of links produced by a system, G the set of links in the gold standard, and $|X|$ indicating the cardinality of set X, we have

$$P = \frac{|L \cap G|}{|L|}$$

$$R = \frac{|L \cap G|}{|G|}$$

As usual these two metrics can be combined using F-measures, where the parameter α, $0 \leq \alpha \leq 1$, determines the relative weights of precision and recall ($\alpha=0$ gives the precision and $\alpha=1$ the recall).

$$F_{\pm} = \frac{P \star R}{\alpha P + (1-\alpha) R}$$

Another combination is the Alignment Error Rate (AER) and defined as follows:

$$AER = 1 - \frac{2 \star |L \cap G|}{|L| + |G|}$$

Given the subjective and use-dependent nature of alignment, it has been argued that systems should not be punished for proposing links that human evaluators disagree on. Thus, if such links are classed as Possible, while those that humans agree on are termed Sure, a metric that considers the difference may be useful. Och and Ney (2003) suggested that precision should be computed on the basis of recovered Possible links, while recall could be based on Sure links. This affects AER as follows, where S stands for Sure links and Possible P for possible links:

$$AER = 1 - \frac{|L \cap S| + |L \cap P|}{|L| + |S|}$$

Note that Sure links are considered a subset of the possible links, and that this definition coincides with the previous one, if S and P are identical.

While the revised AER has been a popular metric, it has been criticized for being a weak predictor of extrinsic measures, for instance in case of translation performance as measured by BLEU (Fraser and Marcu 2007: 293–303). Instead, they argue that a suitably weighted F-measure is preferable.

The AER can be particularly misleading when the number of Possible links is high compared to the number of Sure links. This will contribute to a low error rate but, when the Possible links emerge out of human disagreement, it is uncertain what is really measured.

Precision and recall at the level of links are problematic as well, when multiword units are present. Aligning the English word *gold ring* with its German translation *Goldring* should arguably result in two links. A system proposing only one of them, say <*gold*, *Goldring*>, will still be credited with a point which is valid both for precision and recall. This is reasonable for statistical translation but not if the task is bilingual lexicon generation or terminology extraction. In those cases it is arguable that alignments should have transitive closure (Goutte *et al*. 2004: 502–509), which means that if <j,i>, <j,i'> and <j',i> are non-null links, then <j',i'> is also a link. Søgaard and Kuhn (2009) calls such clusters translation units and defines the translation unit error rate, TUER, as

$$\mathrm{TUER} = 1 - \frac{2 * |U \cap G|}{|U| + |G|}$$

Here, U are the translation units produced by the system and G the translation units of the gold standard. The TUER is usually higher than the AER by several points.

Sentence alignment

Nowadays, if a bitext is included in a parallel corpus collected for research and/or distribution, we can expect it to be sentence-aligned and the sentence alignment to have high accuracy.

The performance of a sentence alignment system is clearly dependent on properties of the corpus, such as the presence of unambiguous sentence boundaries and the nature of the translation, in particular the frequency of non-translated, or extra sentences, and the occurrence of sentences that have been reordered, aggregated or split during translation. Sometimes large sections may even be missing from one of the texts, or moved from their original position, a situation which can be handled by reorganizing the data into smaller parts.

If the bitext is reasonably well-behaved, however, it can be sentence aligned with quite simple methods. We consider a bitext well-behaved if it is monotonic, or almost so, and sentence unit boundaries can be detected with high levels of accuracy. If so, the elements of a sentence pair <**f**, **e**>, could be

- single units
- up to *n* contiguous text units, where n is often set in advance, sometimes as low as 2
- A symbol such as 0 or ε, indicating absence of a corresponding unit.

The type of a pair is the number of units that it covers. Types are indicated as 1–1, 1–2, 1–0, 2–2, and so on.

All sentence alignment algorithms exploit statistical tendencies in well-behaved bitexts. These are of four basic kinds:

- *Distribution of matches on types*. 1–1 sentence pairs tend to account for some 90 per cent or more of all matches, and 1–2 or 2–1 types for most of the remainder.
- *Monotonicity*. A bitext may be represented as a matrix with rows and columns representing tokens or characters (Melamed 2001). Matches, and token associations, tend to occur near the diagonal of that matrix with only local deviations from strict monotonicity.

- *Length*, measured as number of characters (Gale and Church 1993) or number of words (Brown *et al.* 1993: 263–311). A short sentence tends to yield a short translation; a long sentence a long translation.
- *Token associations*, obtained by some association measure (Kay and Röscheisen 1993), from a dictionary (Varga *et al.* 2005: 590–596) or from string comparisons (Simard *et al.* 1992: 67–81). A token association signals the occurrence of a non-null pair of text units related under translation, and pairs that correspond under translation tend to contain more token associations than pairs that are not related under translation.

The first two tendencies are illustrated in Figure 24.1. Almost all sentences are part of a link and the links follow one another monotonically. The large majority are 1–1 matches, here interleaved with isolated occurrences of other types (2–1, 1–0, 1–5).

The tendencies can be exploited in different ways. A common approach is to assign a score to each pair based on measures of one or more of the statistical tendencies. By assuming independence the score for any complete or partial alignment can then be computed as a product of the scores for its individual matches:

$$score(A) = \prod_{i=1}^{N} score(A_i)$$

The best alignment, $\hat{A}$, is then taken to be the one with the highest score:

$$\hat{A} = \text{argmax}_A \; score(A)$$

It is usual to apply (negative) logarithms to both sides of the score equation so that computations can be performed more efficiently and the score becomes a cost measure. The best alignment is then the one with the lowest cost.

Gale and Church (1993) showed character length to be a very powerful feature for the language pairs English–French, and English–German, and used no token associations at all. Exploiting the independence assumption, they used dynamic programming to find the best alignment. Simard *et al.* (1992) showed that the length-based algorithm could be improved upon by applying token associations, in their case cognate-based, in a second pass for cases where Gale and Church's algorithm had about equal scores for the best and second-best alternatives.

Similarly, Melamed's algorithm (Melamed 2001), implemented in the **GMA system**, which has token associations as a base, asks Gale and Church's algorithm for a second opinion on any match which is not 1–1.

Other approaches that combine length-based metrics and token associations first attempt to establish a subset of 1–1 matches that have very high scores, and then work from there. This approach is more robust in the face of sentences and paragraphs that only appear on one side of the bitext. Moore (2002) finds such matches near the diagonal using a length-based statistic, while the **hunalign** system (Varga *et al.* 2005: 590–596) finds them based on a score combining length similarity and dictionary information. Moore's algorithm then creates a dictionary from the 1–1 matches found in the first pass, and use it to find more matches, while hunalign has this as an option, in the absence of a dictionary. Hunalign, unlike Moore's algorithm, has a final pass in which initial matches are expanded to 1-n matches, for arbitrary *n*, if certain conditions are met.

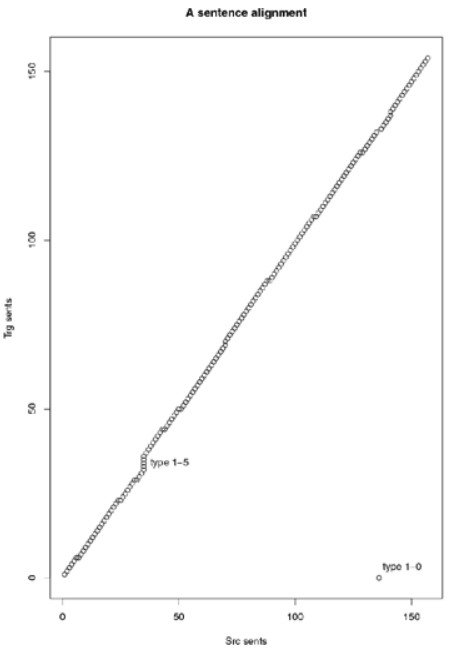


Figure 24.1 A sentence alignment from a Swedish–English novel translation with 157 source sentences (note the occurrence of a null match (lower right corner) and match of type 1-5)

Word alignment

Word alignment is usually performed on sentence-aligned bitexts, although good results may be reached also on bitexts that have been chopped up into equal-sized arbitrary parts (Fung and Church 1994: 1096–1102). On large bitexts, a limit is often set on sentence length, say, to 20 or 100 words, to reduce the size of the search space.

Word alignment as a computational problem is harder than sentence alignment. Many-to-many matches are more abundant, and matches may involve sets of words that are not adjacent. Moreover, null matches are generally more frequent, as is reordering. An example is shown in Figure 24.2.

Word alignments are also harder to establish for humans than sentence alignments. One reason is that structure and meaning differ between languages. One language may employ prepositions to express what another uses case-endings for, as in (1), or require an extra word to express some function in comparison with another language, as in (2).

(1)
EN: and they came to Bethlehem.
FI: ja he saapuivat Beetlehemiin.
GLOSS: and they came Bethlehem+Case

(2)
EN: they did not come
SE: de kom inte.
GLOSS: they came not.

In these cases one can argue either for not aligning the prepositions, or the extra verb form 'did', since they have no non-bound counterparts. On the other hand, one can argue that they should be aligned to the word that carries the function, although it does so morphologically.

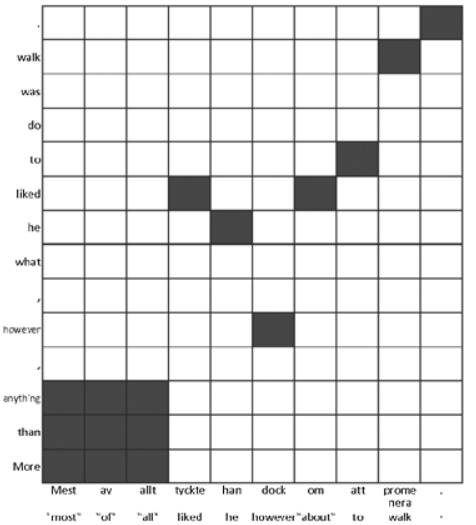


Figure 24.2 Word alignment of an English–Swedish sentence pair with null links, many-to-many links, reordered and discontinuous translation units

In addition, translation is not always meaning-preserving and the semantic difference is often a matter of degree. Thus, the decision whether two words or phrases should be aligned or not is essentially subjective. Still, there are many sentence pairs for which human experts can agree on the best alignment, and agreement can be improved if training and guidelines are used to harmonize judgements (Melamed 2001). Remaining disagreements can be solved by voting, or, as was described above for sentence alignment, be registered explicitly.

Statistical tendencies

While word alignment is harder than sentence alignment, similar statistical tendencies can be seen in the data.

- *Distribution of matches on types*. Often, a majority of matches, at least 50 percent, will be of type 1–1, and figures around 90 per cent have been reported for some corpora. Other types will however be more common than for sentence alignments and, in particular, include more null matches.
- *Monotonicity*. For many languages, especially if they are genetically related or of the same typological kind, the word order of a translation tends to follow that of the source. Thus, for the majority of links <j,i,>, i tends to be close to j, and the difference |i-j| tends to have a frequency maximum near or close to 0.
- *Token associations*. If the words associated with position <j,i> often co-occur in the sentence pairs of the bitext, this increases the chances that <j,i> in the specific sentence pair is a link. Similarly, the chances are increased, if the words can be found in a bilingual dictionary.

- *String similarity*. If the strings occupying position <j,i> are identical or similar, this increases the probability that the element <j,i> is a link. For languages using different alphabets simple string comparisons are not helpful, but comparisons can be made by conversion to phonetic strings.
- *Class-based associations*. Given that the two sentences have been tagged or parsed, comparisons can be made on the basis of syntactic similarity. In general, if the words at <j,i> have the same part of speech, or the same grammatical relation, the chances are higher that they form a link. Classes can also be learnt automatically with clustering methods (Och and Ney 2003: 19–51)

These tendencies have been modeled in different ways, and in different combinations, and there is a rich literature of alternative proposals. Here, only a small subset of them can be covered. For a comprehensive overview, see Tiedemann (2011).

Methods without learning

Since 1–1 matches are the most common, one idea is to find as many of these as possible and then continue from there. This restricts the number of possible alignments considerably, and thus simplifies search.

A particularly efficient instance of this idea is the **competitive linking algorithm** (Melamed 1997). Association metrics for the words of the bitext are computed and are used to score the positions of the alignment matrix. Positions of the matrix are then selected, starting with the highest scoring ones, and eliminating all positions belonging to rows or columns that have had a position selected. The greedy search comes to a halt when a threshold value is reached. The process may be iterated or supplemented with post-processes to extend the coverage to one–many or many–many matches. Melamed (2001) provides an extensive account of such algorithms.

Another efficient method is presented by Lardilleux and Lepage (2009). They observe that what they call 'perfect alignments', pairs of words or phrases having the same frequency *n*, and occurring in n sentence pairs of a bitext, are good link candidates, and this also when *n* is as low as 1. Based on this observation they devise a method for principled generation of small subcorpora of a given bitext, and extract the pairs that are perfect alignments in those subcorpora. The sampling process is fast and the same links are generated several times. Also, different links for the same word are derived, enabling computation of translation probabilities. Moreover, by arranging data, not in pairs, but in sequences, they can generate links for three or more languages at once, in all directions. This method is implemented in a system called **anymalign**.

Generative alignment models

From a given alignment, as in Figure 24.2, we can imagine the target words having been generated from the associated source words and then rearranged. Conversely, given a sentence pair from a bitext, we can look for an alignment as an explanation for how one sentence was translated from the other.

In this framework the task of translation is modeled as follows:

$$\hat{\mathbf{e}} = \underset{e}{\operatorname{argmax}}\, p(\mathbf{e}|\mathbf{f}) = \underset{e}{\operatorname{argmax}}\, \frac{p(\mathbf{f}|\mathbf{e})\, p(\mathbf{e})}{p(\mathbf{f})}$$

The best translation, ê, is the English string that has the highest probability given the foreign string, **f**. Since p(**f**) does not depend on **e**, it can be removed. Alignments are introduced as hidden variables in the equation, and there may be many alignments that produce the same end result. With A representing the set of all possible alignments, we have

$$\underset{\mathbf{e}}{\operatorname{argmax}} \sum_{\mathbf{a}\in\mathbf{A}} \mathrm{p}(\mathbf{f},\mathbf{a}\mid\mathbf{e})\mathrm{p}(\mathbf{e})$$

The alignment we are interested in is the one with the highest probability, i.e., the one that can give us the most likely explanation for f as an encoding of e. Thus,

$$\hat{\mathbf{a}} = \underset{\mathbf{a}\in\mathbf{A}}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{f},\mathbf{a}\mid\mathbf{e})$$

As the alignment of individual sentence pairs can be considered independent, taking the union of the best alignments for all pairs will give us the best alignment for the whole bitext.

A simple account of how a Foreign sentence is derived from an English one is the following:

1 Decide on the number of positions of the Foreign string.
2 Associate each Foreign position with at most one English position. This means that the alignment will be functional and that there may exist words in the Foreign strings that have no match in the English string.
3 Pick a Foreign word for each Foreign position.

To turn this into a stochastic model we need some probability distributions. A simple case, referred to as IBM Model 2, is the following:

1 length probabilities, lgth(J|I), which are usually regarded as uniform and not necessary to estimate.
2 alignment probabilities, a(i|j,I,J), where i=a(j) is a function of the Foreign positions and i=0 represents the case where a Foreign word has no correspondent in the English sentence.
3 translation probabilities, $t(f_j|e_{a_j})$ that express the dependence of a Foreign word on the English word in the aligned position.

To solve the search problem (Brown *et al.* 1993) proposed to start by learning a simple word-based alignment model and then introduce increasingly more complex models. This first model, IBM Model 1 is the special case of Model 2 where the alignment probabilities a(.) are considered uniform. Thus, sentences are essentially treated as sets. As the alignment is functional in the direction from Foreign to English, a foreign word f_j is associated with at most one English word, $e_{a(j)}$ and an alignment can be represented as an assignment of English positions to Foreign positions. A specific null token e_0 is used to represent Foreign words that are not aligned with any English word.

Starting from a uniform assignment of probabilities to the parameters, the probability of any alignment can be computed. This is the first E-step of the EM algorithm. Then, in the M-step, weighted counts are collected for new estimates of the probabilities, where the weights are based on the probabilities of the alignments in which the pair occurs. These steps are iterated a few times yielding new estimates for word translation probabilities t(f|e) and alignment

weights. In the next phase Model 2 alignment probabilities a(i|j,I,J) conditioned on the position of the foreign word and the lengths of the two sentences are also estimated. An alternative to Model 2 using relative rather than absolute position is the so-called HMM model introduced by Vogel *et al.* (1996). Here the probability of a position in the English string is conditioned on the position of its predecessor. Often a uniform distance-based probability is used so that $a(a_j|a_{j-1})$ is the same for all i=a_j, and i'=a_{j-1} with the same distance |i-i'|.

While the alignment is functional in the direction from Foreign to English in Models 1 and 2, there is nothing to prevent two (or more) Foreign words, to be matched with the same English word. The Models 3, 4 and 5, handles the tendency of English words to be associated with one, more or no Foreign words. This property of a word is termed its fertility.

In Model 3 fertility probabilities n(k|e) for k = 0,1,2, … are computed for English words. The case e=e_0 is treated separately by a single parameter, p_1, as the number of Foreign words that have no English correspondent are assumed to depend on the length of the input. While keeping to the initial assumption that alignment is functional in the direction from Foreign to English, the generative story for the alignment is now reversed. Starting from the English side, each word (and position) is assigned a fertility. In some cases the fertility will be 2 or more and extra positions are introduced. Extra Foreign words are introduced according to probability p_1. Then Foreign words are introduced to fill positions according to their translation probabilities with the associated English words. Finally, the Foreign string is reordered according to absolute position probabilities. However, in this case, with the reversed orientation, the probabilities of positions are modeled in the direction from English to Foreign and termed distortion probabilities.

Model 4 adds more parameters for the distribution of words within and between phrases. Similarly to the HMM model these probabilities are based on relative positioning, and also on classes.

Models 3 and 4 are deficient in the sense that they allow impossible alignments. This is because the fertility and positioning of one word is assumed to be independent of the fertilities and positioning of other words. Model 5 is basically a non-deficient version of Model 4. However, it is computationally more complex and for this reason, Model 4, although not giving quite as good results, is used more often in practice.

A freely available implementation of the IBM models is the system **Giza++** (Och and Ney 2003). In addition it includes the HMM-model. For almost a decade Giza++ has been the most heavily used alignment system in practice. A multi-threaded reimplementation, **MGIZA++** (Gao and Vogel 2008: 49–57), is also in wide use.

While IBM-style generative models have dominated the field, they are not without drawbacks. They are inherently asymmetric, as the Foreign and English halves have different roles, and cannot produce many-to-many translation units. They are also hard to extend with additional models as the generative framework must explain how one half can be generated from the other. Moreover, they are prone to overfitting to the training data and often propose many incorrect links for rare words such as numbers or proper names, a phenomenon called 'garbage collection' (Moore 2004: 518–525) .

Symmetrization

The asymmetry and functional character of the IBM models make it hard to derive many-to-many links. Also, if the roles of the two sides are reversed the associations found to have positive probabilities may be very different, especially for low-frequency words.

An obvious solution to this problem proposed already in Och and Ney (2003: 19–51) is to perform two alignments, reversing the roles of the two halves in a second round. From these two alignments it is possible to take the intersection as well as the union. The intersection will only contain 1–1 links, whereas the union can have many different types. Naturally, the intersection will have a high precision, but a low recall, while the situation for the union is the opposite. It has been found empirically that adding matches from the union to those of the intersection in a principled manner can increase recall substantially without sacrificing precision too much. Growing the intersection by adding neighbouring matches from the rows and columns is usually a good strategy, as long as the additions are not in conflict with existing matches. Growing along the diagonals may improve performance further (Och and Ney 2003: 19–51; Koehn *et al.* 2003).

Liang *et al.* (2006) describes extensions to Model 2 and the HMM model that perform joint estimation of the parameters for both directions. This means that symmetrization is performed on the go. The algorithm is implemented in the *Berkeley Aligner*. In another system, *SyMGiza++* (Junczys-Dowmunt and Szal 2012), the M-steps are modified by weighing the alignments on an average of the parameters for both directions, and the heuristic symmetrization steps after model training are incorporated in the general system flow.

A proposal for handling many-to-many relations as first class objects in a generative model was presented by Marcu and Wong (2002). They view the corpus as the result of simultaneous generation of a Foreign and English string, where the primitive pairs are not just word pairs, but possibly phrase pairs. In this model there is no need for fertilities, but on the other hand it is difficult to train. They could show, however, that the model produced better word alignments than IBM Model 4 on 100,000 sentence pairs from the French−English Hansard corpus.

Discriminative models

Discriminative models combine an arbitrary number of information-giving functions, h_n, usually called feature functions. Each function is supplied with a weight, w_n, that indicates its importance. The combination is in most cases linear, which gives the following decision rule:

$$\hat{\mathbf{a}} = \operatorname{argmax}_{\mathbf{a}\in\mathbf{A}} \sum_{n=1}^{N} w_n h_n(\mathbf{f}, \mathbf{e}, \mathbf{a})$$

To learn values for the weights it is necessary to have access to reliable reference data, which makes them semi-supervised. It has been shown that the required amounts need not be very large; a common figure is a few hundred sentence pairs. Generative models are in principle unsupervised, which means that they optimize their parameters without recourse to any information about human perspectives on alignments. This is mostly considered an advantage, but, on the other hand, if a system can make use of existing manual alignments, performed according to some standards it would be in a better position to produce alignments that agree with those standards. Such data, however, will never exist in large quantities, as they take large efforts to produce. For this reason, unaligned data are still required in large quantities.

Discriminative models have the advantage that any tendency observed in parallel corpora can be taken into account. Not only that, they can also use alignments produced by a generative model.

Central to the success of a discriminative system is the selection of feature functions. All systems use one or more features that capture token associations. Such feature functions can be based on association statistics such as the Dice coefficient, the log-likelihood-ratio, the χ^2-

statistic or translation probabilities from one of the IBM models. These are **local features** in the sense that their values depend only on the pair of words. Other local features may concern string similarity, parts-of-speech and properties of neighbouring words.

Global features, on the other hand, consider the alignment as a whole. For instance, the sum of all association scores for an alignment is a global feature. Other global features would relate to monotonicity or distortion. There is nothing that prevents using several features that relate to the same aspect. Moore (2005: 81–88) used both the number of backward jumps in an alignment, and the sum of their magnitudes. Liu *et al.* (2010: 303–339) used Neighbor count, the number of links for which both j-j' and i-i' equals one, and Cross count, the number of link pairs <j,i> and <j',i'> for which the product (j-j')*(i-i') is negative. For a monotonic alignment, this feature would have the value zero, while the value would be larger the more reorderings there are. Other features relating to the topology of the alignment can be concerned with the number of non-linked words, and the number of words that are linked to one, or more than one word. The value of these features can be taken as a sign of the normality of the alignment. Yet other features can refer to external resources; an indicator feature could register whether the words of a link can be found in a bilingual dictionary, and there may be features indicating whether other aligners have proposed a given link (Ayan and Dorr 2006: 96–103).

As with generative frameworks, a discriminative framework may perform search in several steps, using different features in different steps. For instance, Moore (2005: 81–88) used a two-step approach. In the second step, the global translation probability feature based on token associations computed with the log-likelihood-ratio was replaced with conditional link probabilities. These were estimated as the ratio of co-occurring word pairs that were actually linked in the first step.

The use of global features has the drawback that one has to resort to approximate methods in training and search. Alternative methods that have been used are average perceptron learning (Moore 2005: 81–88), SVM training (Moore *et al.* 2006), and Minimum Error Rate Training (Liu *et al.* 2010: 303–339). With only local features or first-order dependencies, globally optimal parameters can be obtained. Ayan and Dorr (2006) used Generalized Iterative Scaling, and Blunsom and Cohn (2006: 65–72) used forward-backward inference on two linear-chain Conditional Random Fields (CRF), one for each language direction.

Improved statistical learning

While discriminative systems have performed better on many corpora, the difference to Giza++ symmetrized alignments is not extreme. More recent work on generative modeling has approached the problem of overfitting by using regularization, i.e., by adding restrictions or penalties that favour smooth distributions. Graça *et al.* (2010: 481–504) used posterior regularization to enforce constraints of bijectivity and symmetry on alignments in the expectation step of an HMM model and show that it works well on six language pairs where manual alignments have a high percentage (over 90 percent) of 1–1 links. The main advantage of their method is that learning is tractable. Dyer *et al.* (2011: 409–419) devised a general model where only two parameters, the regularization strength and the learning rate, were learned from manual alignments, whereas features and weights were learned from the full unannotated bitext. They demonstrated clear improvements on Czech–English data compared with symmetrized Model 4 alignments. Vaswani *et al.* (2012: 311–319) proposed using a prior in the M-step to optimize translation probabilities with the effect that garbage collection behavior is much reduced. The set of translation parameters are optimized for each target word separately.

Using syntax

It has been debated whether linguistic structure is helpful for alignment. However, for some applications it clearly can be. Macken *et al.* (2008) describes a system for terminology extraction which aligns so called anchor chunks on the basis of lexical association and part-of-speech patterns for chunking. Given a sentence pair partially aligned with anchor chunks, further chunk pairs can be found in gaps between anchor chunks. There is also evidence that parsing one of the sides may help. Riesa and Marcu (2010) showed, for an Arabic–English corpus within a discriminative framework, that a search regime guided by a single phrase structure parse for the English sentences led to better performance than symmetrized IBM 4 alignments. Fossum *et al.* (2007: 44–52) could demonstrate improvements for Chinese–English data.

References

Ayan, Necip Fazil and Bonnie J. Dorr (2006) 'A Maximum Entropy Approach to Combining Word Alignments', in *Proceedings of the Human Language Technology Conference of the NAACL,* New York, 96–103.

Blunsom, Phil and Trevor Cohn (2006) 'Discriminative Word Alignment with Conditional Random Fields', in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 20 July 2006, Sydney, Australia, 65–72.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer (1993) 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational Linguistics* 19(2): 263–311.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin (1977) 'Maximum Likelihood from Incomplete Data via the EM Algorithm', *Journal of the Royal Statistical Society Series B* 39(1): 1–38.

Dyer, Chris, Jonathan Clark, Alon Lavie, and Noah A. Smith (2011) 'Unsupervised Word Alignment with Arbitrary Features', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 19–24 June 2011, Portland, OR, 409–419.

Fossum, Victoria, Kevin Knight, and Steven Abney (2008) 'Using Syntax to Improve Word Alignment Precision for Syntax-based Machine Translation', in *Proceedings of the 3rd Workshop on Statistical Machine Translation*, 19 June 2008, Ohio State University, Columbus, OH, 44–52.

Fraser, Alexander and Daniel Marcu (2007) 'Measuring Word Alignment Quality for Statistical Machine Translation', *Computational Linguistics* 33(3): 293–303.

Fung, Pascale and Kenneth W. Church (1994) 'K-vec: A New Approach for Aligning Parallel Texts', in *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, 5-9 August 1994, Kyoto, Japan, 1096–1102.

Gale, W. A. and Church, K. W. (1993) 'A Program for Aligning Sentences in Bilingual Corpora', *Computational Linguistics,* 19(1), 75–102.

Gao, Qin and Stephen Vogel (2008) 'Parallel Implementations of Word Alignment Tool', in *Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, June 2008, Columbia, OH, 49–57.

Goutte, Cyril, Kenji Yamada, and Eric Gaussier (2004) 'Aligning Words Using Matrix Factorization', in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 21–26 July 2004, Barcelona, Spain, 502–509.

Graça, João V., Kuzman Ganchev, and Ben Taskar (2010) 'Learning Tractable Word Alignment Models with Complex Constraints', *Computational Linguistics* 36(3): 481–504.

Isabelle, Pierre (1992) 'Bi-textual Aids for Translators', in *Proceedings of the 8th Annual Conference of the UW Centre for the New OED and Text Research,* University of Waterloo, Waterloo, Canada, 76–89.

Junczys-Dowmunt, M. and A. Szal (2012) 'SyMGiza++: Symmetrized Word Alignment Models for Statistical Machine Translation', *Security and Intelligent Information Systems, Lecture Notes in Computer Science*, Vol. 7053, pp. 379–390, Springer.

Kay, Martin and Martin Röscheisen (1993) 'Text-translation Alignment', *Computational Linguistics* 19(1): 121–142.

Koehn, P., R. J. Och and D. Marcu (2003) 'Statistical Phrase Based Translation', *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL),* pp. 48–54.

Lardilleux, Adrian and Yves Lepage (2009) 'Sampling-based Multilingual Alignment', in *Proceedings of Recent Advances in Natural language Processing (RANLP)*, Borovets, Bulgaria, 214–218.

Liang, P., B. Taskar and D. Klein (2006) 'Alignment by Agreement', *Proceedings of the Human Language Technology Conference of the North American Association for Computational Linguistics (NAACL)*, pp. 104–111.

Liu, Yang, Liu Qun, and Lin Shouxin (2010) 'Discriminative Word Alignment by Linear Modeling', *Computational Linguistics* 36(3): 303–339.

Macken, Lieve, Els Lefever, and Veronique Hoste (2008) 'Linguistically-based Sub-sentential Alignment for Terminology Extraction from a Bilingual Automotive Corpus', in Donia Scott and Hans Uszkoreit (eds) *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1,* 18-22 August 2008, Manchester, UK, 529–536.

Marcu, Daniel and William Wong (2002) 'A Phrase-based Joint Probability Model for Statistical Machine Translation', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6-7 July 2002, University of Pennsylvania, PA, 10: 133–139.

Melamed, I. Dan (2001) *Empirical Methods for Exploiting Parallel Texts*, Cambridge, MA: MIT Press.

Melamed, I. Dan (1997) 'A Word-to-word Model of Translational Equivalence', in *Proceedings of the 35th Conference of the Association for Computational Linguistics*, 7–10 July 1997, Madrid, Spain.

Moore, Robert C. (2004) 'Improving IBM Word Alignment Model 1', *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, 518–525.

Moore, Robert C. (2005) 'A Discriminative Framework for Bilingual Word Alignment', in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 6-8 October 2005, Vancouver, British Columbia, Canada, 81–88.

Moore, Robert C., Yih Wen-tau, and Andreas Bode (2006) 'Improved Discriminative Bilingual Word Alignment', in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, July 2006, Sydney, Australia, 513–520.

Och, Franz Josef and Hermann Ney (2003) 'A Systematic Comparison of Various Statistical Alignment Models', *Computational Linguistics* 29(1): 19–51.

Riesa, Jason and Daniel Marcu (2010) 'Hierarchical Search for Word Alignment', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 157–166.

Simard, Michel, George F. Foster, and Pierre Isabelle (1992) 'Using Cognates to Align Sentences in Bilingual Corpora', in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada, 67–81.

Søgaard, A. and Kuhn, J. (2009) 'Empirical lower bounds on alignment error rates in syntax-based machine translation', *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST '09),* pp. 19–27.

Tiedemann, Jörg (2011) *Bitext Alignment*, San Rafael, CA: Morgan and Claypool Publishers.

Varga, Daniel, Laszlo Németh, Peter Halácsy, Andras Kornai, Viktor Trón, and Viktor Nagy (2005) 'Parallel Corpora for Medium Density Languages', in *Proceedings of the International Conference RANLP 2005 (Recent Advances in Natural Language Processing)*, 21–23 September 2005, Borovets, Bulgaria, 590–596.

Vaswani, Ashish, Huang Liang, and Avid Chiang (2012) 'Smaller Alignment Models for Better Translations: Unsupervised Word Alignment with the 10-norm', in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 8–14 July 2012, Jeju Island, Korea, 311–319.

Vogel, Stephen, Hermann Ney, and Christoph Tillmann (1996) 'HMM-based Word Alignment in Statistical Translation', in *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, 5–9 August 1996, Centre for Sprogteknologi, Copenhagen, Denmark, 836–841.

25
BITEXT

Alan K. Melby

BRIGHAM YOUNG UNIVERSITY, THE UNITED STATES

Arle Lommel

GERMAN RESEARCH CENTER FOR ARTIFICIAL INTELLIGENCE (DFKI), GERMANY

Lucía Morado Vázquez

FACULTY OF TRANSLATION AND INTERPRETATION, UNIVERSITY OF GENEVA, SWITZERLAND

History of bitext

The term bitext was coined by Brian Harris in an article written in December 1987, while Harris was in Africa on leave from the University of Ottawa, and later published in *Language Monthly* (Harris 1988: 8–10). Harris described bitext (initially spelled with a hyphen, bi-text) as designating a new concept in translation theory, with its primary nature being psycholinguistic, even though it was seen to have applications in translation technology. According to Harris (ibid.), a bitext is a source text and its corresponding target text as they exist *in the mind of a translator*. As Harris points out, a human does not translate an entire text in one fell swoop but rather a segment at a time. Each segment of a source text is mentally linked to a corresponding segment of target text to form a cognitive translation unit. Segments can be phrases, clauses, or larger stretches of text. Together, the translation units of the bitext constitute the entire source and target texts 'laminated' to each other.

In an unpublished 1988 memo to some colleagues, Harris continued to discuss the notion of bitext and provided a concrete example:

> SAMPLE OF INTERLINEAR BITEXT [English to French]
> The Board of PAC unanimously confirms the PAC mandate and concept.
> *Le Conseil est unanime dans sa confirmation du mandat et du concept fondamental du PAC.*
>
> This includes support to long-term development through the strengthening of African NGOs;
> *Le concept comprend un appui au développement à long terme par le renforcement des ONG africaines;*
>
> supporting African awareness in Canada
> *un appui aux activités de sensibilisation du public canadien*
>
> which focusses on African abilities and strengths
> *qui accentuent les forces et habiletés africaines*

and the root causes of current problems;
et examinent les causes profondes de la crise;

encouraging partnerships based on a recognition and respect for mutual roles
l'encouragement de partenariats basés sur le respect mutuel

and confirmation of Africans as the agents of their own development;
et la reconnaissance que les Africain(e)s sont les premiers agents de leur développement;

supporting networking and linkage efforts both in Africa and in Canada.
l'appui à la création de liens et de réseaux en Afrique et au Canada.

PAC's emphasis is on African priorities
Le PAC met d'abord l'accent sur les priorités des Africain(e)s

and on activities which evolve out of the African context.
et les activités qui émanent du contexte africain.

Networking and linkages have been identified as priority areas for PAC
La promotion de liens et la formation de réseaux sont des domaines prioritaires pour le PAC

and are essential to developing true partnership relationships.
et sont essentiels pour la mise en oeuvre de relations de partenariat

(Note [from Harris]: [The vertical bar] marks translation unit boundaries. A search for any word, or combination of words, in the source text retrieves the segment containing it together with the corresponding translation segment (printed here in italics). This enables other translators, or the same translator at some future time, to perceive reusable translations like 'unanimously confirms / *est unanime dans sa confirmation du*' and 'awareness in Canada / *sensibilisation du public canadien*', which would not appear in the dictionaries or term banks because they are context specific, but which help get away from word-for-word equivalences.)

A bitext can be presented visually in various ways. Harris originally anticipated that the preferred presentation would be interlinear, with a segment of target text appearing directly beneath a segment of source text. However, a side-by-side display of source and target segments is currently more common.

As can be seen in this example, some segments of this bitext are entire sentences, others are independent clauses, and some are phrases, depending on what the creator of the bitext considered to be likely units of thought for a human translator.

Whatever its size and however it is identified, each segment of source text must be linked to its corresponding segment of target text. This segmentation and alignment process allows future reuse of a bitext.

Note that Harris uses 'translation unit' to refer to either a segment of source text or a segment of target text. However, in translation memory systems, generally in translation technology, and in the rest of this article, a translation unit is two segments, a source-text unit and its corresponding target-text segment, together with the link between them.

Interlinear translation has long been a part of literary studies but is not exactly a bitext. For example an interlinear translation of Chaucer into modern English is superficially similar to a bitext, but consists of a literal translation created specifically for the purpose of studying an important text. (See Interlinear 2013.)

A predecessor to bitext was part of a bilingual concordance system in Melby (1981), where segment pairs were identified by a human marking them in source texts and their corresponding translations. The term 'bitext' was not used at that time. After the translation units were marked by a human, software identified all the words in the source text and, for each word, located all the translation units containing that word. This allowed a report of how that word was translated in various contexts. For example, the entry for 'cease' in Melby (1981: 457–466) lists three occurrences in the bitext corpus:

> MRD0148 But when the echoes had fully <ceased>,
> *Mais quand l'écho s'était tout à fait évanoui*
>
> MRD0024 And then the music <ceased>, as I have told;
> *Alors, comme je l'ai dit, la musique s'arrêta;*
>
> CMO0321 when the motion of the hellish machine <ceased>,
> *que le mouvement de l'infernale machine cessa,*

The particular bitext from which this bilingual concordance entry was derived consisted mostly of Edgar Allen Poe stories and their translations into French by Baudelaire. The identifier at the beginning of the line indicated the translation unit. For example, MRD0148 was the 148th translation unit of the bitext of *The Masque of the Red Death*. For the reader who is not familiar with French, the three translation units retrieved for the word 'cease' show three different ways of translating it that are not fully interchangeable. In the first translation unit, an echo is ceasing and the French verb selected by the translator is typically used to translate 'to faint', with the image that the echoes faded away and eventually became inaudible. In the second and third translation units, 'cease' corresponds to different French verbs (*s'arrêter* and *cesser*) that are synonyms but are not used with the same frequency. Thus, this early bilingual concordance, derived from a bitext, fulfilled the hope that Harris expressed: 'to [help the translator] perceive reusable translations … which would not appear in the dictionaries or term banks because they are context specific, but which help get away from word-for-word equivalences'; but this bilingual concordance system was not further developed at the time, and the idea of a bitext remained dormant until Harris independently proposed it and coined the term seven years later.

Using current translation technology, a bitext such as the examples proposed by Harris and Melby could not be constructed automatically from the source and target texts in question. When a bitext is constructed automatically, one segment size is chosen in advance. Typically sentence-length or paragraph-length segments are used, since they can be automatically identified by segmentation software. Manual creation of large bitexts is too laborious. The notion of a bitext has strayed from its original conception by Harris as a reflection of the mental units used by a human translator, and has become the result of a mechanical process applied to source texts and their translations.

Note that source texts and their translations are often called parallel texts in the computational linguistics community. However, the term 'parallel texts' has a different meaning in translation studies, where it refers to texts in different languages and in the same domain that are not necessarily translations of each other. This additional sense of 'parallel texts' is linked to the term 'comparable texts' in computational linguistics. Thus, a bitext corpus can be automatically derived from parallel texts in the computational linguistics sense but not in the translation studies sense of comparable texts.

As another terminological note, in the article on translation memory in the present encyclopedia, the distinction between mental units and mechanical units is described as 'cognitive' vs. 'formal' units.

Bitext in current translation technology

In its application to translation technology, a bitext is ideally created segment by segment while a translator is in the act of translating. However, Harris also allows for the possibility of re-creating a bitext from a source text and a completed translation of it. Clearly, in the case of a bitext reconstructed after the fact, it is impossible for a third party to determine the segmentation that was performed in the mind of the translator. In addition, the initial translation may have been modified by a reviser or reviewer. Thus, there are two common cases: (1) a bitext created incrementally during the translation process; and (2) a bitext reconstructed from separate source and target texts, typically using automatic segmentation and alignment (often with additional manual correction of misalignments).

In current practice, most translation tools pre-segment the source text using a segmentation algorithm, primarily on sentence boundaries, and the translator is expected to translate one pre-defined segment at a time. Translation technology, by pre-defining segments and presenting them to the translator, may be changing the way humans think as they translate, but addressing this issue is beyond the scope of the present article. The original conception of a bitext as a reflection of the mental process lives on in translation studies, where eye tracking and keystroke capture studies are providing insight into how translators actually do their work (Christensen 2011: 137–160).

Once a bitext is available, it can be converted into a traditional translation memory, which usually involves eliminating duplicate translation units, some degree of normalization of the segments, and combining unordered sets of translation units from a number of source and target texts into an indexed database. Thus, the process of creating a traditional translation memory from a set of bitexts is a non-reversible process in the sense that the original source and target texts cannot be re-created solely from a classic translation memory database without access to the source text. Some translator tools blur the distinction between translation memory and bitext by retaining sufficient information in a translation memory database to reconstruct the original source and target texts.

As indicated in the historical section of the article on translation memory, the first commercial translation memory software system was released in 1986 by ALPS, about a year before Harris wrote his first article on bitext. However, Harris did not know about the ALPS system in 1987 (personal communication). Thus, bitext and translation memory can be considered to be concurrent, independent developments in the history of translation technology, each with a different original focus. The focus of the first translation memory systems was to retrieve entire sentences that had been previously translated, while the focus of a bitext corpus, as envisioned by Harris, was to assist a human translator in doing research on how other translators have dealt with particular words and phrases. With the recent rise of subsegment retrieval in translation memory systems, and the addition of more information in a translation memory database to indicate how a translation unit fits into the source and target texts from which it was derived, the distinction between translation memory and a bitext system is blurring.

Three standards that are highly relevant to bitext are SRX (Segmentation Rules eXchange), XLIFF (XML Localization Interchange File Format), and TMX (Translation Memory eXchange).

SRX provides a formal mechanism for describing how a text is to be segmented, and XLIFF provides a standard format for representing a bitext. See Appendix 1 and Appendix 2, respectively, for more information about SRX and XLIFF.

Monotonicity

TMX was developed in order to represent a translation memory database consisting of an unordered set of translation units, but sometimes TMX is used to represent a bitext by assuming that the order of the segments in the source text is identical to the order of the segments in the target text.

A strict segment-by-segment, typically sentence-by-sentence, correspondence between a source text and its translation is somewhat imposed on a translator using typical tools, sometimes called TEnTs (Translation Environment Tools) or Computer Assisted Translation (CAT) tools, where the source text is presented to the translator in a two-column table with a segment of source text on the left and a space for the corresponding segment of target text on the right. However, this segment-to-segment correspondence does not necessarily result in the most natural translation. It is based on the assumption that translations are monotonic, that is, segments of source and target text will progress in parallel, with no need for lines that link source and target segments to cross each other.

As pointed out by Quan *et al.*, this assumption is not necessarily valid:

> [M]ost existing approaches to sentence alignment follow the monotonicity assumption that coupled sentences in bitexts appear in a similar sequential order in two languages and crossings are not entertained in general (Langlais *et al.* 1998; Wu 2010). Consequently the task of sentence alignment becomes handily solvable by means of such basic techniques as dynamic programming. In many scenarios, however, this prerequisite monotonicity cannot be guaranteed. For example, bilingual clauses in legal bitexts are often coordinated in a way not to keep the same clause order, demanding fully or partially crossing pairings. … Such monotonicity seriously impairs the existing alignment approaches founded on the monotonicity assumption.
>
> (2013: 622–630, citations in original omitted)

Consider the following invented source text consisting of five sentences, designed to illustrate non-monotonicity in a simple fashion.

S1 – I was looking at dresses in the store.
S2 – I decided to buy the blue dress for several reasons.
S3 – I hate green, so the green dress was out.
S4 – The yellow dress was too expensive.
S5 – That was a week ago, and I am happy with the blue dress.

Suppose that the translation into some other language, viewed in English using back translation, consists of six sentences:

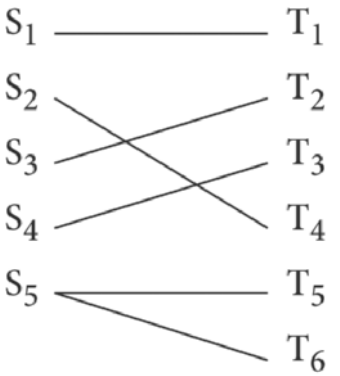
T1 – I was looking at dresses in the clothing store.
T2 – I hate green, so the green dress was not seriously considered.
T3 – The yellow dress was too expensive.
T4 – Therefore I decided to buy the blue dress.

T5 – That was a week ago.
T6 – I am happy with my purchase.

How could this non-monotonic translation be represented in a bitext without changing the sentence order of either the source text or the target text?

The fact that the fifth sentence of the source text becomes two sentences in the target text is not a problem. However, the rhetorical difference of introducing the conclusion early in the source text (in segment 2) but later in the target text (in segment 4), does cause difficulties for representation in a bitext.

The correlation of the translation units (source–target segment pairs) is as shown below:



This reordering and crossed sequencing is not a problem for a translation memory that consists of unordered translation units. However, a bitext is expected to represent the order of segments as found in the original source text and the original target text.

One way to deal with this problem is to make the segment unit a paragraph instead of a sentence and define S1–S5 and T1–T6 as one pair of paragraphs that correspond. However, there can be more dramatic ordering problems that would make this approach infeasible. Or there may be reasons to keep the segment size at the sentence level.

The most common format for representing a bitext outside any particular software application is XLIFF. Non-monotonic segment order is handled differently in XLIFF 1.2 and 2.0: in XLIFF 1.2, segments are represented by <mrk mtype="seg"> elements that are set within both the source and the target contents. Each of these markers has an ID, so the markers can be in different physical order in the source and target content while they are still linked by ID value. (See http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html#Struct_Segmentation for more information.)

Here is how the dress-buying example would be represented in XLIFF 1.2:

<trans-unit id="1">

<source>I was looking at dresses in the store. I decided to buy the blue dress for several reasons. I hate green, so the green dress was out. The yellow dress was too expensive. That was a week ago, and I am happy with the blue dress</source>

<seg-source>

<mrk mtype="seg" mid="1">I was looking at dresses in the store.</mrk>

<mrk mtype="seg" mid="2">I decided to buy the blue dress for several reasons.</mrk>

<mrk mtype="seg" mid="3">I hate green, so the green dress was out.</mrk>

<mrk mtype="seg" mid="4">The yellow dress was too expensive.</mrk>

<mrk mtype="seg" mid="5">That was a week ago, and I am happy with the blue dress.</mrk>

</seg-source>

<target>

<mrk mtype="seg" mid="1">I was looking at dresses in the store.</mrk>

<mrk mtype="seg" mid="3">I hate green, so the green dress was not seriously considered.</mrk>

<mrk mtype="seg" mid="4">The yellow dress was too expensive.</mrk>

<mrk mtype="seg" mid="2">Therefore I decided to buy the blue dress.</mrk>

<mrk mtype="seg" mid="5">That was a week ago. I am happy with my purchase.</mrk>

</target>

</trans-unit>

Many TeNTs have implemented XLIFF 1.2. However, few have implemented the use of IDs on segments in order to represent non-monotonic translations.

XLIFF 2.0 has not yet been approved as an OASIS standard. The current approved committee specification (http://docs.oasis-open.org/xliff/xliff-core/v2.0/cs01/xliff-core-v2.0-cs01.html) represents directly segments within the translation units. And each segment element includes its own source and target elements. A segment reordering mechanism has already been defined in this new version: the optional attribute 'order' indicates the order of the target segments, while the physical order of the elements tells us the order of the source. Here is how the dress-buying example would be represented in XLIFF 2.0:

<unit id="u1">

<source>I was looking at dresses in the store. </source>

<target order="1">I was looking at dresses in the clothing store. </target>

<source>I decided to buy the blue dress for several reasons. </source>

<target order="4">Therefore I decided to buy the blue dress. </target>

<source>I hate green, so the green dress was out. </source>

<target order="2">I hate green, so the green dress was not seriously considered. </target>

<source>The yellow dress was too expensive. </source>

<target order="3">The yellow dress was too expensive. </target>

<source>That was a week ago, and I am happy with the blue dress. </source>

<target order="5">That was a week ago. I am happy with my purchase. </target>

</unit>

See http://docs.oasis-open.org/xliff/xliff-core/v2.0/cs01/xliff-core-v2.0-cs01.html#segorder for more information.

One reason why it is important to consider non-monotonic translation in bitext, besides the fact that it occurs in real translations, is to avoid imposing a monotonic mindset on translators. Languages use a variety of rhetorical structures. See, for example, the seminal work of Kaplan (1966: 1–20). We thus come full circle back to the origin of bitext as a reflection of the mental process of a human translator, not an imposition on the mind of a translator intended to reduce diversity among languages.

Applications of bitext

Bitext has become highly influential in translation technology.

The major common applications of bitext are currently:

1 a method of keeping source and target texts aligned throughout the entire translation process, including quality assurance and quality control steps;
2 an intermediate stage toward the creation of a translation memory database; and
3 a source of training data for statistical machine translation systems.

However, there are other uses for a bitext. Among them are:

- The study of 'shifts' in human translation (Cyrus 2006: 1240–1245), such as:
  - passivization and depassivization;
  - number change (e.g. plural to singular);
  - explicitation and generalization.
- Terminology research:
  - TransSearch (Macklovitch *et al.* 2000 and www.terminotix.com);
  - Termight (see Dagan and Church 1994: 34–40);
  - Identifying concept relations (not just terms): (Marshman *et al.* 2012: 30–56).
- Word sense disambiguation (Diab and Resnik 2002: 255–262).
- Inducing transfer rules (Lavoie *et* al. 2001: 17-24 for rule-based machine translation, and Graham and van Genabith 2009: 1–10 for transfer-based statistical machine translation).

Conclusion

Although bitext is an idea from the 1980s that was originally intended to primarily assist human translators in retrieving instances of words and phrases as treated by other human translators, it has also turned out to be the basis for many other aspects of translation technology, from translation memory to machine translation. In a sense, it has evolved from a purely descriptive mechanism to a framework for translation that makes it difficult to break out of a sentence-by-sentence correspondence between source and target languages. One is led to wonder what effect bitext is having on language. Despite the undeniable benefits of bitext, has it reduced the richness of translation by imposing the sequence of source-language segments on the target language?

⋆ ⋆ ⋆

APPENDIX 1

Segmentation Rules eXchange (SRX) Format[1]

One significant problem that arises in building bitext (and multitext) corpora stems from segmentation, the division of texts into segments generally considered equivalent to sentences. Segmentation would pose little problem if there were an unambiguous character for marking sentence boundaries; but the full stop (.) character that indicates sentence boundaries for most Western languages is highly ambiguous. Besides ending sentences, it serves to mark abbreviations (e.g., 'etc.', 'Dr.', 'Mr.'), indicate decimals (in some languages) or serve as the thousands separator (in others), and is used for special purposes in certain areas (e.g., as the prefix for file-type extensions in many computer operating systems or to separate sections of numerical IP addresses). All of these uses mean that a full stop, by itself, is not a reliable indicator of sentence boundary.

At the same time, additional characters may serve as segment boundaries for Western languages. The following are some common examples:

- Carriage returns are frequently used to terminate list items, or after titles and headings that do not end in periods. At the same time, carriage returns are frequently used to force formatting line breaks that do not end sentences.
- Semicolons (in English at least) are frequently used to separate (grammatically) full sentences that have a closer logical relationship than would be implied if they were separated by a period. But semicolons are used for other purposes that do not end segments. (Semicolons are particularly problematic when aligning English texts with source or target texts in other languages where the other language uses two distinct segments in place of one in English.)
- Other punctuation marks, such as the exclamation point (!) and question mark (?) when followed by capital letters are, for most text types, generally unambiguous segment boundary markers when followed by a space and a capital letter, but they have special uses in some text domains (such as information technology and mathematics) that may render them ambiguous.
- Tab characters may be used to separate items in tabular data, but interpreting segment boundaries in tab-delimited data is frequently highly problematic since the tabs may be combined with carriage returns, spaces, and other characters in complex fashions.

All of these issues make accurate segmentation difficult from a machine-processing perspective. While parsing and data-driven approaches can help disambiguate text and identify correct segment boundaries, most commercial applications have tended instead to use regular-expression-based iterative processes that search a text for potential segment boundaries and then check them against exception lists to determine whether segmentation is appropriate.

Leaving Western languages, the situation may be better or worse, depending on the language. Chinese, Japanese, and Korean, for example, are much easier to segment accurately at the sentence level than Western languages because the sentence-terminating punctuation tends to be used exclusively for the purpose of terminating segments. The orthography of Thai, on the other hand, poses special problems for both word- and sentence-level segmentation because Thai generally lacks inter-word white space, but does use space characters to mark segment boundaries and in some other circumstances. Modern Hebrew and Arabic tend to be fairly simple to segment by comparison. (This chapter cannot address the specifics of world languages and focuses primarily on the orthographic challenges of Western languages written in Latin, Greek, and Cyrillic scripts.)

The Unicode Consortium in Unicode Standard Annex (UAX) 29 describes a process for segmenting text based on a standard algorithm. This approach, however, does not account for language-specific issues (and specifically notes that it cannot handle them). As a result, text segmented according to this specification is likely to contain errors. For example, UAX#29 rules would break this text:

> On Friday we saw Mr. Smith at the theater.

into two segments:

> On Friday we saw Mr.
> Smith at the theater.

Such problems are quite common and pose a particular challenge for segmenting and aligning text. Early research conducted by a group of IT companies found that they lost between 5 percent and 10 percent of translation memory matches in technical text due to incorrect or differing segmentation. The largest offender was abbreviations that end in full-stop characters, such as 'Mr.' and 'Prof.' (in English), 'Mme.' (French), and 'z.B' (German). Certain abbreviations that can be termed 'refixing abbreviations' are particularly likely to be followed by capital letters (thus triggering a simple segmentation boundary condition): these are abbreviations for titles and names. Other abbreviations, by contrast, such as 'etc.' are relatively less likely to be followed by capital letters outside of segment boundary conditions, but still may be followed by them in some cases (e.g., 'I saw the camels, donkeys, horses, etc. John had brought to market').

In response to these findings, the Localization Industry Standards Association developed the Segmentation Rules eXchange (SRX) format (LISA 2008, available at GALA 2012). SRX allows users to declare regular-expression-based sets of rules for segmentation. In particular, it allows them to create rules for breaking text and rules for preventing breaks in a standard XML format. SRX files specify a regular expression that defines the text that occurs before the possible break and a regular expression that defines the text after the possible break. For example, the following rule:

```
<rule break="no">
<beforebreak>\sMr\.</beforebreak>
<afterbreak>\s</afterbreak>
</rule>
```

indicates that no segment boundary should occur after the text 'Mr.', thus overriding the default UAX#29 algorithm. By contrast, the following rule:

```
<rule break="yes">
<beforebreak>[\.\?!]+</beforebreak>
<afterbreak>\s+[A-Z]</afterbreak>
</rule>
```

states that if a full stop, question mark, or exclamation point is followed by one or more whitespace characters and a capital letter, the text should be broken after the terminal punctuation.

SRX allows for the creation of general and language-specific sets of rules. In practice, these rule sets tend to consist of a list of rules that account for abbreviations that should not trigger a segmentation break, followed by a list of general break conditions. Each location in the file is evaluated against the rules sequentially. If a no-break condition is met, then the processing engine ceases to examine at that point and moves on to the next inter-character position in the file. If no no-break condition is met then breaking conditions are evaluated and, if one is met, the text is segmented. If no breaking condition is found, then the processor moves to the next position.

SRX rules are generally quite easy to interpret for individuals familiar with regular expressions. Complex regular expressions can be used. As mentioned above, most SRX files focus on exceptions to general rules triggered by abbreviations, since these account for most segmentation faults. However, they may also address specific conditions related to specific text domains or text types. For example, if an input file is 'hard wrapped' (that is, uses new-line characters at the end of lines within a paragraph), an SRX file can specify that new line characters, which would generally indicate a segment boundary, should be ignored unless they occur after a full stop or other trailing punctuation and are followed by a capital letter. (And this behavior, in turn, may be overridden by specific rules for abbreviations or other conditions.)

SRX serves a valuable function by allowing tools to declare how they segmented text to allow other tools to emulate or understand that behavior. Because of domain and text-type differences, there is no single segmentation algorithm that will suffice for all conditions. Furthermore, users of natural language processing (NLP) tools frequently 'tweak' segmentation engines to account for issues encountered in the text with which they work. SRX provides a way to document these modifications and specific functions to ensure interoperability between segmentation engines. In addition, by correcting for segmentation faults that might otherwise confuse NLP tools, SRX files can help improve automatic alignment results.

One usage scenario for SRX with benefit for working with bitexts involves using SRX to allow bitexts to be dynamically resegmented to allow interoperability between processes that work on bitexts. In this scenario a bitext could be segmented to match multiple existing translation memory databases to identify the best matches or a new text could be segmented to match an existing bitext for which the appropriate segmentation method has been declared in

SRX. Without SRX such dynamic analysis would be much more difficult and would require the creation of one-off segmentation routines that would offer little flexibility. Instead SRX permits researchers and commercial developers to implement effective segmentation rules to meet their needs without the worry that segmentation choices will negatively impact future activities due to incompatibilities.

One recent development of note with SRX is that the Unicode Consortium's Unicode Localization Interoperability (ULI) technical committee has started a project to document common abbreviations and other segmentation exceptions as part of the Common Locale Data Repository (CLDR) for many languages (Unicode Localization Interoperability Technical Committee 2013). These exceptions can be easily converted to SRX-format rule sets. While this resource is in its infancy, it will assist individuals building bitext corpora to ensure that their results are accurate. While the CLDR data cannot account for domain-specific or organization-specific exception lists, it will help improve general segmentation results.

SRX plays an important role in bitext applications, particularly in translation memory, by providing a formal mechanism to declare the specific segmentation rules used to generate a corpus. This transparency helps reduce incompatibility between bitext tools and can assist users in understanding how particular results were achieved.

SRX is recognized as the standard for segmentation rules and a number of TEnTs have implemented it (including XTM, CafeTran, MemoQ, and Swordfish[2]) and it is included the open-source Okapi, OmegaT, SRXEditor, and LanguageTool projects. A modified version of SRX is utilized by the widely used SDL Trados to permit export of segmentation information, but this version is not compliant with the SRX specification and Trados segmentation files require modification to be used with SRX processors.

Some basic sets of SRX rules for major languages have been made publicly available (see GALA 2012), but segmentation rules depend on domain and organization and it is not possible to generate universal rule sets since specific segmentation cases may directly conflict depending on specific instances. As mentioned above, Unicode Localization Interoperability Technical Committee has begun gathering information on prefixing abbreviations for various languages. Since CLDR is widely implemented by major corporations, the promotion of segmentation and SRX-related concerns within Unicode will be a driver for increased implementation of open segmentation algorithms that can be represented by SRX.

Given that segmentation into a TM is often a non-reversible process, SRX cannot address the problem of interoperability for unordered translation memories that were previously segmented using different rules. In other words, it cannot directly "fix" the segmentation of heterogeneous TM resources. However, if full texts are preserved as bitexts rather than reduced to TM databases, SRX can be used to adjust segmentation as needed.

In cases where segmentation differences impede interoperability, SRX may be more productively used to resegment complete texts to ensure compatibility moving forward while treating previously segmented resources as a fall-back for match purposes. This lack of the ability to directly address previously segmented legacy texts converted to TM databases has proved a barrier to greater adoption of SRX within the TM community. However, the ability to work with bitexts and SRX to allow dynamic resegmentation is an argument for greater use of bitexts instead of traditional TM resources.

SRX plays an important role in the open standards landscape, allowing easier and more reliable movement between translation tools for use of language resources with heterogenous segmentation processes. Such scenarios are of considerable importance for businesses with significant bitext resources and SRX can have a major business impact, especially in cases where organizations are required to merge language resources (e.g., in merger and acquisition scenarios).

APPENDIX 2

XLIFF

Introduction

The XML Localization Interchange File Format (XLIFF) is a tool-neutral data container that allows the interchange of localization data and metadata during the localization process. It is currently being developed by OASIS (Organization for the Advancement of Structured Information Standards). The standard was first created under the name of 'DataDefinition group' by members of three software companies: Novell, Oracle and Sun Microsystems, in Dublin in the year 2000 (Jewtushenko 2004). Their aim was to develop a single format that would allow the interchange of localizable data between tools during the localization process without loss or corruption of data. Two years later, XLIFF 1.0 was officially approved as a Committee Specification within OASIS (XLIFF TC 2003). Since then, many software companies, TEnT developers, localization companies and academicians have joined the OASIS XLIFF Technical Committee to work on its development and maintenance; two more versions (1.1 and 1.2) have already been approved, and a new one (XLIFF 2.0) is under review. Task-specific subcommittees have also been created to work on specific aspects of the standard; for example, the Promotion & Liaison subcommittee that maintains relationships with other related standardization bodies and carries out different promotional activities, such as the organization of yearly international symposia on XLIFF.

Extraction-merge principle in XLIFF

XLIFF is based on an extraction-merging concept (Savourel 2003) which can be explained as a three-step localization mechanism: in the first step it relies on the extraction of the localization-related data from an original format and its conversion to a valid XLIFF file. The second step involves the manipulation of that file by any TEnT that supports the standard. The manipulation would always depend on the specific localization project being converted, and could include some typical localization tasks such as translation, reviewing or QA checking. After finishing all the required processes, the XLIFF file can be declared as final. The third step involves merging the manipulated data into the original file to create a localized version in another language. As observed in this three-step mechanism, XLIFF was originally designed as a temporary format to be used during the localization process and discarded after the final merging process; however, it is now seen as a long-term representation of a text and its translation. Even after a translation or localization project has been completed, an XLIFF file can be used to generate a traditional translation memory, for reuse in XLIFF aware tools, and for various research tasks based on bitexts, as previously described in this article.

The current version of the standard (1.2) has been widely implemented in the TEnT ecosystem (Filip and Morado Vázquez, 2012) since its approval in 2008. The main criticism received during these years was the permissiveness of the specification in some points. This has resulted in some cases of different interpretations and tool-specific implementations of the standard, which could jeopardize the main feature of the standard: interoperability between tools. The XLIFF TC listened to the feedback and suggestions for improvement, and designed version 2.0 with them in mind.

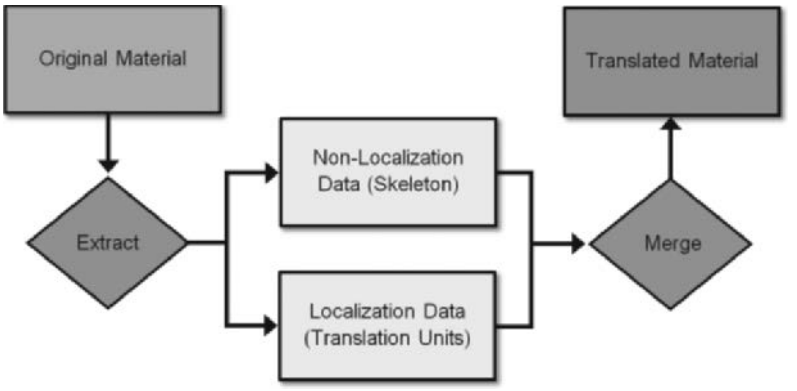


Figure 25.1 Extraction/merge principle of XLIFF

Source: Adapted from XLIFF TC (2003: 10)

XLIFF 2.0

The next version, 2.0, which is currently under review, differs substantially from the previous version (1.2), and introduces the core and module concepts for the first time. The core consists of

> the minimum set of XML elements and attributes required to (a) prepare a document that contains text extracted from one or more files for localization, (b) allow it to be completed with the translation of the extracted text, and (c) allow the generation of *Translated* versions of the original document
>
> (XLIFF TC 2013)

The core concept shares some similarities with the "minimal XLIFF" that was present in version 1.2. As well as the core, eight specific modules have been designed to store extra information about specific localization processes: Translation Candidates, Glossary, Format Style, Metadata, Resource Data, Change Tracking, Size Restriction and Validation. Each of them was designed with a specific process in mind; they have their own pre-defined XML elements and attributes in an individual XML schema and namespace (XLIFF TC 2013).

From this version on, if a TEnT wants to be declared as XLIFF compliant, it would need to support at least the XLIFF 2.0 core. Checking and certifying if a tool is truly XLIFF compliant is out of the scope of the XLIFF TC; however, this clear core-module distinction would help tool developers to concentrate their efforts on supporting at least the reduced set of XML elements and the attributes of the core. Depending on the nature of the specific TEnT and its needs, developers might also decide to implement some of the modules proposed in the 2.0 specification; for example, the Translation Candidates module where alternative translation proposals can be stored. The Translation Candidate module substitutes the <al-trans> element that was present in the previous version (1.2).

Below is an example of a basic (and valid) XLIFF 2.0 file which only contains core elements and attributes. The root element is <xliff>, which can contain one or more <file> elements. Please note that the structural elements <header> and <body> are no longer present in this version. Inside the file element we find the <unit> element where one or more

elements can be included. A compulsory <source> element was placed in the unit that stores the text to be translated, followed by an optional <target> element that stores the translated version.

```
<xliff version="2.0" srcLang="en" trgLang="es">
<file id="f1">
<unit id="1">
<segment>
<source> Hello World! </source>
<target> ¡Hola mundo! </target>
</segment>
</unit>
</file>
</xliff>
```

XLIFF 2.0 is under review at the time of this writing; therefore information included here is subject to change. The following months will be critical for the approval and implementation of the standard. Those implementations will be crucial to the future of XLIFF, as its use in TEnTs and mainstream localization industry processes represents the real success of the standard.

Notes

1 SRX is closely related to the article on segmentation (see Chapter 37).
2 This list of TEnTs and the following lists of implementations are not intended to be comprehensive.

References

Christensen, Tina Paulsen (2011) 'Studies on the Mental Processes in Translation Memory-assisted Translation – The State of the Art', *trans-kom. Zeitschrift für Translationswissenschaft und Fachkommunikation* 4(2): 137–160.

Cyrus, Lea (2006) 'Building a Resource for Studying Translation Shifts', in *LREC 2006: Proceedings of the International Conference on Language Resources and Evaluation*, 24–26 May 2006, Genoa, Italy, 1240–1245.

Dagan, Ido and Kenneth W. Church (1994) 'Termight: Identifying and Translating Technical Terminology', in *Proceedings of the 4th Conference on Applied Natural Language Processing*, 13-15 October 1994, Stuttgart, Germany/San Francisco, CA: Morgan Kaufmann, 34–40.

Diab, Mona and Philip Resnik (2002) 'An Unsupervised Method for Word Sense Tagging Using Parallel Corpora', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 7-12 July 2002, University of Pennsylvania, Philadelphia, PA/San Francisco, CA: Morgan Kaufmann, 255–262.

Filip, David and Lucía Morado Vázquez (2012) 'XLIFF Support in CAT Tools: Results of the Survey, January 2012', XLIFF Promotion and Liaison Subcommittee, OASIS.

GALA (2012) LISA OSCAR Standards. Available at: http://www.gala-global.org/lisa-oscar-standards.

Graham, Yvette and Josef van Genabith (2009) 'An Open Source Rule Induction Tool for Transfer-based SMT', *The Prague Bulletin of Mathematical Linguistics: Special Issue: Open Source Tools for Machine Translation* 91: 1–10.

Harris, Brian (1988) 'Bi-text, A New Concept in Translation Theory', *Language Monthly* 54: 8–10.

Interlinear (2013). Available at: http://sites.fas.harvard.edu/~chaucer/teachslf/tr-index.htm.

Jewtushenko, Tony (2004) 'An Introduction to XLIFF', in *IV International LRC Localisation Summer School 2004*, 2 June 2004, LRC, University of Limerick, Ireland.

Kaplan, Robert B. (1966) 'Cultural Thought Patterns in Inter-cultural Education', *Language Learning* 16(1–2): 1–20.

Lavoie, Benoit, Michael White, and Tanya Korelsky (2001) 'Inducing Lexico-structural Transfer Rules from Parsed Bi-texts', in Association for Computational Linguistics: 39th Annual Meeting and 10th Conference of the European Chapter: Workshop Proceedings: Data-driven Machine Translation, 6–11 July 2001, Toulouse, France, 17–24.

LISA (Localization Industry Standards Association) (2008) *Segmentation Rules eXchange (SRX)*, Féchy, Switzerland: Localization Industry Standards Association.

Macklovitch, Elliott, Michel Simard, and Philippe Langlais (2000) 'TransSearch: A Free Translation Memory on the World Wide Web', in *LREC 2000: Proceedings of the International Conference on Language Resources and Evaluation*, 31 May–2 June 2000, Athens, Greece, 1201–1208.

Marshman, Elizabeth, Julie L. Gariépy, and Charissa Harms (2012) 'Helping Language Professionals Relate to Terms: Terminological Relations and Termbases', *Journal of Specialized Translation* 18: 30–56.

Melby, Alan K. (1981) 'Linguistics and Machine Translation', in James Copeland and Philip W. Davis (eds) *The Seventh Lacus Forum*, Columbia, SC: Hornbeam Press, 457–466.

Quan, Xiaojun, Chunyu Kit, and Yan Song (2013) 'Non-monotonic Sentence Alignment via Semi-supervised Learning', in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 4–9 August 2013, Sofia, Bulgaria, 622–630. Available at: http://aclweb.org/anthology/P/P13/P13-1061.pdf.

Savourel, Yves (2003) 'An Introduction to Using XLIFF', *MultiLingual Computing & Technology* 14(2): 28–34.

TMX 1.4b Specification: *Translation Memory eXchange Format*, 2005-04-26. Available at: http://www.gala-global.org/oscarStandards/tmx/tmx14b.html (original: 1998).

Unicode Localization Interoperability Technical Committee (2013) 'ULI Segment Exceptions Posted in SVN and Demo Updated'. Available at: http://uli.unicode.org/home/announcement/ulisegmentexceptionspostedinsvnanddemoupdated.

XLIFF TC (2003) 'A white paper on version 1.1 of the XML Localisation Interchange File Format (XLIFF)'. Available at: http://www.oasis-open.org/committees/download.php/3110/XLIFF-core-whitepaper_1.1-cs.pdf.

XLIFF TC (2008) 'XLIFF Version 1.2, OASIS Standard, 1 February 2008'. Available at: http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html.

XLIFF TC (2013) 'XLIFF Version 2.0, Committee Specification Draft 02, Public Review Draft 02'. Available at: https://tools.oasis-open.org/version-control/browse/wsvn/xliff/trunk/xliff-20/xliff-core.pdf.

26

COMPUTATIONAL LEXICOGRAPHY

Zhang Yihua

GUANGDONG UNIVERSITY OF FOREIGN STUDIES, CHINA

Introduction

Lexicography is traditionally defined as the branch of applied linguistics concerned with the design and construction of lexicons for practical use. Or, as defined by Hartmann and James (2000), 'The professional activity and academic field concerned with dictionaries and other reference works'. Nowadays, lexicography has developed into a relatively independent cross-discipline involved with linguistics, language acquisition, cognition, cultural anthropology, terminography, translation studies, statistics, computer science, and technology, etc. The combination of lexicography and computer technology resulted in a specific area of study, computational lexicography.

The theoretical research on computational lexicography begins from as early as the 1960s, and its theoretical framework gradually took shape. The scope of research became more and more specifiable in the 1980s to 1990s. The annual journal *Lexicographica* published a special issue, *Computational Lexicography and Computational Linguistics* in 1988, and Boguraev and Briscoe put out a book entitled *Computational Lexicography for Natural Language Processing* in 1989. Atkins and Zampolli's *Computational Approaches to the Lexicon* (1994) and van Eynde and Gibbon's *Lexicon Development for Speech and Language Processing* (2000), both gave deep insights into issues concerning computational lexicography, and articles on computational lexicography are available in many journals and online media.

Meanwhile, writings on corpus lexicography coming off the press attracted lexicographers' attention. *Corpus, Concordance, Collocation* (Sinclair 1991) and *Computer Corpus Lexicography* (Ooi 1998) are two examples. In 2004, a somewhat systematic framework of computational lexicography was put forward by Zhang Yihua in his book *Computational Lexicography and New Dictionaries*, and *Computational Lexicography* (2013).

A general view of computational lexicography

From the perspective of lexical structure, the term 'computational lexicography' is a modifier-head construction. It can be interpreted as the study of lexicographical theory and practice by means of computational technology. Thus, computational lexicography should certainly function within the framework of traditional lexicography, but with the focus put on a new lexicographical methodology based on modern technology. Undoubtedly, the development of information technology and multi-media provide excellent tools for lexicographical study and practice.

In fact, the computer creates favourable conditions for such aspects of lexicography as data storage, extraction, analysis, transmission, and exchange, as well as corpus construction and dictionary compilation. Moreover, large databases or electronic dictionaries are needed to enhance data processing ability in natural language processing (NLP). All these external causes play a key role in the emergence and development of computational lexicography. In this sense, 'computational lexicography' refers both to the development of machine-readable dictionaries based on a printed version and a lexicographic database for computer use.

Many authors (Amsler 1982: 661; Ooi 1998: 1–2; Hartmamn and Gregory 2000; Bennett *et al.* 1986: 26) provide various definitions of computational lexicography (cf. Zhang Yihua 2004: 13–14). From the historical viewpoint of its evolution, computational lexicography should first deal with the electronization and machine-readability of the lexical knowledge of the printed dictionary, then study of automatic reading, recognition, conversion, and exchange of lexicographic data by computers.

Study must now be done on computer-aided compilation, editing, and revising of dictionaries, aiming at semi-automation in the near future, with the ultimate goal to realize automation throughout the whole process of dictionary-making and publishing. And last, attention should be given to the electronic adaptation of commercial dictionaries, or the design or compilation of electronic dictionaries or online dictionaries for human use.

With respect to lexicographic data processing and compiling digitalization, the analysis of authentic continuous texts, the index, and the extraction of lexical data should be taken into consideration, besides that of computer-aided dictionary compilation. The distribution of related lexical items in natural discourses can be investigated and analysed by the means of an index, so as to examine the functional attributes of various aspects such as grammar, semantics, and pragmatics, and acquire useful features concerning the function and usage of each lexical item. Simultaneously, the data of sense distribution and division can also be obtained through text analysis.

In general, computational lexicography deals with the study of the electronization of corpus processing, semi-automation or even total automation of dictionary compilation, formalization of microstructure arrangement, digitalization of dictionary media, intellectualization of the dictionary query, and integration of multimedia into lexical data representation. The major content of study includes language data collection and processing, sense-division support, comprehensive semantic analysis, illustrative example extraction, computer-aided dictionary compilation, lexicographic database construction, corpus and database management, (semi)-automatic dictionary generation or production, lexicographic data statistics, dictionary compilation management, and data export interface.

Computational lexicography and relevant subjects

Computational lexicography is technologically based on computational linguistics, which focuses mainly on computer-aided NLP, including the technology of information processing in various aspects or layers of both written and oral language. The achievements in computational linguistic research can only contribute to the practice of dictionary making when they are well integrated into lexicography. Corpus lexicography adopts or integrates the views and approaches of computational linguistics, computational lexicography, and corpus linguistics, dedicated essentially to corpus-based research in the principles and practice of dictionary making.

Computational lexicography is theoretically based on computational lexicology, which studies the application of computers to lexicon researches, especially the computational representation of the lexicon, methods of lexical data calculation, and the relation between the computerized lexicon and various parts of the system of NLP. Detailed study on computational

lexicology includes mental representation during cognition and the acquisition of natural vocabulary by means of computer simulation, the mechanism for forming lexical meanings, and the structural arrangement, storage model, extraction approach, and combinatory pattern of lexical information in the mental lexicon. Computational lexicology differs from computational lexicography in that the former emphasizes the analysis of grammatical function and semantic construction of the vocabulary or lexicon, whereas the latter pays more attention to the description of them. However, analysis and description are mutually complementary as an integral whole. They can never be set apart.

Main study field of computational lexicography

As a cross-disciplinary field of study, computational lexicography has developed into a relatively independent subject through serial researches over a rather long time, with a complete set of methodology and clear research objectives.

Corpus lexicography

The basic mechanism of corpus lexicography is corpus linguistics; it is the combination of the linguistic corpus and lexicography. Corpus linguistics proposes a new train of thought that linguistic research and NLP can be done based on computer corpora, which provides a new way for the lexicographer to compile contemporary learners' dictionaries and large-scale comprehensive dictionaries, and thus satisfy the requirements of current dictionary users. Corpus linguistics has its function and research focus as follows: (a) language performance, (b) language description, (c) quantitative and qualitative modelling of language, and (d) experimentalism (Leech 1992a: 107). Therefore, corpus-based theoretical research and practice of lexicography can justify being called corpus lexicography.

Property and characteristics of corpus lexicography

The application of corpora can be seen in almost every branch of linguistics, in which research can be done based upon corpora. As for corpus lexicography, the scope of study falls into three aspects:

1 corpus building, including the import, segmentation, lemmatization, tagging, arrangement, and storage of language materials;
2 the management of corpora, including the supplementation and updating of language materials, the statistics of corpus data, the generation of wordlists and frequencies, and the generation and management of illustrative examples; and
3 the use of corpora, including language data query, example extraction and use, and database building based on corpora for general or specialized dictionaries.

Sinclair (1985: 81–94) and Atkins (1991: 167–204) put forward a new methodology, which Atkins termed corpus lexicography, to evaluate instances of language performance by means of running texts in an attempt to build a more complete, coherent, and consistent set of language data compared to the traditional lexicon. Language data can be regarded as the representation of linguistic/lexical knowledge, which can be subdivided into two levels, the conceptual structure and the computational structure (Kim 1991: 129). The former is a format comprehensible by human beings and the latter is readable by machine. Computational

structure is characterized by its clear formulation and can directly reflect the conceptual structure, while the conceptual structure facilitates the compilation of theoretical universal or general language dictionaries.

Construction and processing of corpora

Since the construction of the *Brown Corpus* in 1964, the first representative computerized mega corpus, numerous corpora have been built around the world. Especially in Britain, a series of dictionary corpora have been built and put into use during the 1980s and 1990s, for example, *The Bank of English,* the *Longman Corpus Network,* the *British National Corpus* (*BNC*), the *BNC Spoken Corpus,* the *Longman Learner's Corpus,* the *Longman Written American Corpus,* the *Longman Spoken American Corpus,* the *Longman Lancaster English Language Corpus,* the *Cambridge International Corpus,* and the *International Corpus of English*.

Corpus building should take into consideration the following aspects: first, the basic features of the corpus, i.e., a corpus must be designed for a specific purpose, the language materials collected must be authentic and typical, and the lexical data encoding and decoding must be standardized and machine-readable. Second, the function of the corpus, i.e., a dictionary corpus should have the function of data management, indexing, statistics, tagging, speech analysis, and lexical data extraction. Third, the types of corpora, i.e., different types of corpora can be classified from the perspective of a specific purpose, languages involved, language forms, language use, text type distribution, processing degree treatment, and storage media (cf. Zhang 2004: 50–55).

Application or use of corpus

Since the late 1970s, the corpus first began to be used in the compilation of English learners' dictionaries, especially the five best-known learners' dictionaries: Longman, Oxford, Cambridge, Collins COBUILD, and Macmillan are all based on corpora.

The *Oxford Advanced Learner's Dictionary* pays special attention to the syntactic pattern of verbs and provides detailed collocational structures and abundant sentence/phrase examples to illustrate the usage of defined words. The *Longman English Dictionary* constructs macrostructure and microstructure in conformity with users' cognitive laws and practice, and uses special defining vocabularies to define words. All these are authenticated and controlled by computer programs. The first edition of the *Collins COBUILD Advanced Learners' English Dictionary* was not compiled, but rather generated, based on a huge database of 73 million words. The language data extracted from corpora is practical and reliable. It is noticeable that the above dictionaries make full use of the computer to complete the data processing stages that must be done manually in traditional lexicography: data collection, headword selection and establishment, and arrangement as well as entry-compilation.

Electronic dictionaries

The concept of the electronic dictionary came into being in the late 1940s when Americans began to study NLP or machine translation, and it attracted people's attention in the middle 1950s and 1960s. However, it came to a standstill mainly because no progress was made in machine readability. Then in the 1980s the exploitation of the electronic dictionary became active with the development of computer technology.

The electronic dictionary is so-called in contrast to the printed dictionary: the storage media ranges from the magnetic disk to the optical disk, magneto-optical disk, flash disk, and IC card

(chip), etc. and it can be queried and read through the microprocessor and related facilities. Hartmann and James (2000) define *electronic dictionary* as 'a type of reference work which utilizes computers and associated facilities to present information on-screen'. Electronic dictionaries can be classified into two types according to their function: (a) a non-coding natural language dictionary available for human users and (b) an encoded computer-language dictionary for machine translation and NLP. The two types can further be subdivided into monolingual, bilingual, and multilingual dictionaries according to the languages involved; they can also be subdivided into unidirectional, bidirectional, and multi-directional dictionaries according to the defining relation between source language and target language.

The non-coding dictionary is inputted, stored, displayed, and read through computers with natural language as its text form. The encoded dictionary stores and transmits natural language by means of computer language code, and is specially designed for machine translation or NLP. A machine translation system needs the support of various encoded dictionaries, including a monolingual dictionary, a bilingual dictionary, a collocation dictionary, a concept dictionary, and so on.

The electronic dictionary involves all types of dictionaries with a database stored in magneto-optical media, including online dictionaries or databases via Internet hyperlinks; even the spellchecker in word-processing platforms (e.g., Microsoft Word) can be considered as an electronic dictionary. In fact, the electronic dictionary is actually a hypertext language information framework composed of language data with related corpus and language processing technology. Here are some typical types of electronic dictionaries:

1 *CD-ROM Dictionaries* are dictionaries on compact disks, including DVDs. The storage capacity of a disk can embrace one or two large dictionaries without any difficulty. This very handy reference tool can be read by means of a computer or an electronic bookplayer. There are a large number of CD-ROM dictionaries on the market; mainstream printed dictionaries are usually sold with a CD-version, including the dictionary series published by Oxford, Longman, Cambridge, Collins COBUILD, Macmillan, Webster, Larousse, and Robert.
2 *Hand-held e-dictionaries* are composed of a micro CPU chip, data RAM, LCD module, keyboard module, and image DMA module. They are compact, lightweight, and portable, often containing lexicographic data of various printed dictionaries in one set, and suitable for school and college students. In recent years, they come with several new functions: handwriting input, downloading of upgraded versions from the Internet, and programming with GVBasic; some of them even provide various IC cards with built-in dictionaries.
3 *Online Dictionaries* can be divided into four categories according to their functionality:
   (a) single-unit versions usually can be downloaded and installed on the computer to translate Web pages and display data from different languages;
   (b) single online versions are usually attached to a website and can be consulted at any time. When users log on the website, they can use the dictionary to look up or translate new words;
   (c) a dictionary website puts together tens, hundreds, and even thousands of dictionaries in different languages and subjects on one home page or index page. This kind of website is often set up by a dictionary publishing house such as Oxford, Longman, Larousse, etc., or such independent sites as yourdictionary.com, onelook.com, vocabulary.com, 1000Dictionaries.com; thesaurus.com, etc.; and
   (d) a dictionary website that is in fact translation software based on bilingual dictionaries, e.g., babylon.com, translate.google.cn, iciba.net, netat.net, and chinafanyi.com.

The lexicon and the lexical database

The lexical database is a combination of the computational lexicon and lexicography, and the computational lexicon is usually a simulation of the human mental lexicon by means of information technology, designed to assist the interpretation and comprehension of natural language by machine. The relationship or difference between lexicon and dictionary is that the lexicon is an entity defined by linguistic theory, while the dictionary is the text representing the information of a particular aspect of lexicon in a certain format.

Since the 1970s, American scholars have established large-scale lexicons that could make semantic descriptions automatic, and began to put it into practice in the mid-1980s. Some scholars in China also made such an attempt in the 1990s. At present, the main lexicons and databases best known to us include WordNet, MindNet, FrameNet, HowNet, Integrated Linguistic Database, VerbNet, PropBank, and the Common-sense knowledge base of CYC, etc. Moreover, English and American lexicographers put out some lexical databases and interface software aimed at corpus datamation, e.g, Dante Database, Word Sketch Engine, Corpus Pattern Analysis, and the Wordlist and Frequency Dictionary of American English etc.

WordNet uses synonym sets (synsets) to represent the lexical concept and describe the lexical matrix, which builds a mapping between the form and meaning of words, and classified nouns, verbs, adjectives, and adverbs into sets of cognitive synonyms, each set representing a different concept.

MindNet uses Microsoft's broad-coverage parser to automatically analyse the dictionary definition and thus obtain linguistic knowledge. There are 24 relationships presented in MindNet, including the Attribute, Possessor, Co-Agent, Deep-Object, Deep-Subject, Domain, Material, Source, Goal, Cause, Purpose, Manner, Means, Subclass, and Synonym, etc.

FrameNet is a knowledge base built with the help of lexicographic definition and corpora within frame semantics. A frame is a basic way to describe the meanings of lexical units and organize lexical knowledge. Each frame has a number of frame elements which represent a precise semantic role.

WordNet, FrameNet, and MindNet are all characterized by describing mental representation through semantic frames, valences, and selected restrictions of frame elements, including semantic class and lexical aspect, or the relations within the language system, such as synonymy, antonymy, hyponymy, meronymy, and entailment, etc.

VerbNet doesn't define lexical units as precisely as FrameNet, but it relates them closer in terms of syntactic structures. PropBank is an annotated corpus which was developed with the idea of serving as training data for machine learning-based systems.

HowNet extracts all semantic relationships implied in the knowledge system of natural language, forms various relational tables, and then describes the intrinsic relationships among and between concepts and features, as well in the knowledge system, and eventually constructs a reticular knowledge and information structure system.

Researches on lexical databases mainly focus on the mental representation of language for NLP. In his 'Generative Lexicon', Pustejovsky (1991: 419) provides a kind of mechanism that roughly satisfies the requirements for this purpose. It consists of four levels of representations: Argument Structure, Event Structure, Qualia Structure, and Lexical Inheritance Structure.

The lexical database is a dictionary knowledge base that is built in light of the principle and method of data organization in the mental lexicon, and in conformity with dictionary macro- and microstructure. It is beneficial for dictionary compilation and revision, as well as the (semi-)automatic generation of dictionaries.

Computer-aided dictionary compilation

The most direct, typical, and revolutionary development of Computer Aided Dictionary Compilation (CADC) is the application of machine-readable corpora in dictionary making. Afterward comes the electronization or digitalization of lexicographical data for processing, arranging, storing, querying, and presenting, etc.

Computer-aided dictionary compilation tools

CADC tools are found in a special word processing platform and management system, and it is designed mainly for the compiling, editing, typesetting, and revising of dictionaries. The CADC system differs from general word processing tools (e.g., Microsoft Word) in that the input and display interface is designed especially to conform with dictionary microstructure and the user's needs for lexicographic data processing, including example extraction, corpus pattern analysis, semantic disambiguation, entry arrangement, text editing and typesetting.

The CADC system has incorporated the well-known Dictionary Writing System (DWS), which has been widely used across the world. Some representatives include: *Dictionary Production/Publishing System (DPS)* by IDM (in France), *ABBYY Lingvo Content* by ABBYY(in Russia), *TshwaneLex* by TshwaneDJe (in South Africa), *Lexique Pro* by SIL International, and so on.

These DWSs are designed for creating, updating, and managing lexicographical data for various types of monolingual and bilingual dictionaries. They may, to a certain degree, satisfy the requirements of dictionary authors and publishers, providing them with a multifunctional template for dictionary compilation.

Corpus and dictionary-compiling

The large-scale corpus has abundant authentic resources, sophisticated corpus processing and analysing tools, and a powerful indexing engine; all these provide advantages for dictionary compilation. The primary motivation for building a corpus is to extract examples from it. With the improvement of corpus managing and processing instruments, lexicographers find that word frequency analysis can be used as the basis for entry selection; classification of concordance lines can assist sense division; and corpus pattern analysis can provide collocational structure or construction of defined words. All these can contribute much to lexicographic definition, as well as the representation of grammatical, pragmatic, and cultural information for specific language aspects.

What's more, the corpus can evince the distribution and use of synonyms in the light of a corpus with sense relation tagging; relevant information about synonyms may be easily found and directly or indirectly used in definition and annotation; and the contextual selection restriction on synonyms may be specified concerning semantic valence, collocation, and usage domain according to the semantic distribution of the headword in concrete corpus samples.

Corpus extraction and application

In the dictionary compiling process, lexicographers should comb out and summarize separate word usage, abstract the word's different senses from its different distributions, and represent them in the dictionary. However, with the increasing expansion of the corpus the frequency of a word becomes increasingly higher. One single indexing of a common word would result

in numerous concordance lines, sometimes many thousands of them. The overloading of the concordance lines causes too much inconvenience to compilers to discover language regularities, and thus results in a rather low efficiency in dictionary compilation. Therefore, it is necessary to develop specialized software or tools to analyse and extract the right lexicographic data from a large-scale corpus.

Currently used techniques mainly include, first, an illustrative example generator that uses a key word and a specific syntactic pattern to generate or extract natural sentences with the same distribution structure from the corpus and then makes a contrastive analysis among them, which can considerably reduce 'noise information'. Second, the 'Word Sketch Engine' (Kilgarriff and Rundell 2002) uses NLP technology to realize the processing as tokenization of words and phrases, lemmatization of word variants, part-of-speech tagging, and grammatical parsing, and establishes a database on the basis of lexical collocation. In this way, a 'word sketch' based on grammatical and collocational features can be generated automatically, and thus greatly facilitate word sense disambiguation for lexicographers (cf. Kilgarriff *et al.* 2003). Third, data mining technology is used to search for useful data in a huge amount of linguistic material. More specifically, it helps to extract unknown knowledge and information that is potentially valuable to dictionary compilation from a mass of incomplete, noisy, vague, and random corpus items. The data mining technology can not only process structured data (like those in a relational database) or semi-structured data (like texts, graphs, or images), but also heterogeneous data distributed across the World Wide Web. These three technologies, though different in some ways, have one thing in common: by abstracting useful language rules or patterns from a large-scale corpus, they can all alleviate the load of lexicographers and improve their efficiency.

Dictionary generation system

Automatic generation of a dictionary requires rather complicated language processing. It requires not only language being processed by means of the NLP approach and conforming to artificial intelligence principles, but also a whole-hearted cooperation between lexicographers and computer professionals. It is predicable that there are at least two ways to generate a dictionary. One is corpus-based, and the other is database-based.

Dictionary generation based on a corpus

The ideal conception of computational lexicographical research is to generate automatically various types of dictionaries directly out of a corpus according to the lexicographer's intention and design. The following conditions must be created to meet the requirements of the dictionary generation process.

1 The corpora should be thoroughly processed in a detailed way: every lexical item in the corpus must be tagged at all levels and include aspects such as spelling, speech sound, inflection, morphology, syntax, semantic attribute, semantic feature, semantic valence structure, and pragmatic rules.
2 A systematic instrumental dictionary should be made for phonetic, morphological, syntactic, and semantic tagging, as well as the mapping and generation of various attributes and features of each lexical item.
3 A sophisticated computer program should be devised and dedicated to the control and management of the dictionary generation process.

These conditions constitute three essential factors for automatic generation of dictionaries. The implementation of the above conditions, however, requires a large amount of investment in manpower and material resources. Besides, there exist certain insurmountable technical barriers in the artificial intelligence which is needed for semantic tagging. In fact, the most time-consuming task is not the attribute-tagging itself, but the building of various systematic instrumental dictionaries. Considering the current technological situation and complexity of semantic tagging, the automatic generation of a dictionary out of a corpus is not practical yet.

But in recent years, people have tried to undertake research on the generation of smaller dictionaries from bigger ones or the generation of Chinese–English dictionaries based on English–Chinese dictionaries. Even though research achievements have been published, no dictionary of this kind has appeared.

Dictionary generation based on databases

Since there are still many technical problems to be solved in corpus-based dictionary generation, some researchers have turned to the development of database-based systems. The main principle of such a system is to create a dictionary database in conformity to lexicographic microstructure with the help of a computer-aided dictionary compilation system. Various dictionaries can be generated from the database.

The Center for lexicographic Studies at Guangdong University of Foreign Studies (in China) tried to undertake the project: *Bilingual Dictionary Generation System Based on Micro-data Structure*. The main features of this system are as follows:

1 The introduction of Wide Area Network technologies enables the system to make the best of all available human resources as well as information resources, thus considerably enhance the efficiency of dictionary making.
2 The editing module functions as the import platform of lexicographic data, while the Background Management Program puts each bit of micro-data in its right place and input data becomes automatically tagged during this process. The tagged lexicographical data can be accessed and reorganized to form as many new dictionaries as the editor wishes.
3 Each generated dictionary has its separate database, and new data can be added from the general database, which provides an effective solution to problems that traditional dictionaries face, such as the reuse of resources and the difficulties in revising and reprinting large-scale dictionaries
4 The system can greatly accelerate the speed of dictionary making, traditionally a time-consuming job, so as to hold an initiative in marketing. In this way, the dictionary making and dictionary generation can be digitalized, Internetized and paperless, and thus maximize the use of the existing resources, and they complete various editing work in a highly efficient way.

Dictionary data processing technology and standard

Data processing technology is a study that focuses on data entry, data tagging, data indexing, data transmission, and data recognizing as well as computer software navigation, which involves the formal structure of language and the rules that govern such a structure; while the standard can assure the uniform format or style of language organizing structure.

Dictionary data processing technology

Computer-aided dictionary making and dictionary generation are both realized by means of NLP technology. The automation of language information processing requires the formalization of language description, which means encoding natural language and computer information in a specially designed encoding meta-language. The encoding mode should be clear and understandable. Clarity is of great importance for computation, or else the information cannot be understood or processed by computer.

Naturally, different databases use different encoding modes. The most commonly used encoding languages or markup languages for electronic files are: Standard Generalized Markup Language (SGML), HyperText Markup Language (HTML), Extensible Markup Language (XML) and Document Type Definition (DTD). The grammar modules used for formal representation of languages are: Generalized Phrase Structure Grammar (GPSG), Lexical-Functional Grammar (LFG), Head Driven Phrase Structure Grammar (HDPSG), Categorical Grammar, etc. In addition, many linguistic theories can support computational lexicography, such as Logical and Mathematical Semantics, Conceptual Dependency, Case Grammar, Word Grammar, Montague Semantics, Meaning-Text theory, Frame Semantics and so on. These theories are widely discussed and studied by lexicographers, and the studies are of great significance for the digitalization of dictionary making, dictionary editing and dictionary publishing

The international standard for dictionary making

In setting standards for many products, the International Organization for Standardization (ISO 1951: 2007) deals with monolingual and multilingual, general and specialized dictionaries. It aims to establish a model for dictionary making, LEXml. It specifies a formal generic structure independent of publishing media and proposes a means to present entries in printed and electronic dictionaries in a digitalized way, including dictionary compiling, editing, publishing, and distributing. The objective of this standard is to facilitate the production, merging, comparison, extraction, exchange, dissemination, and retrieval of lexicographical data in dictionaries.

This processing model gives consideration not only to the methodical formal structure that is required for the automation of dictionary making, but also to its convenience for use; it not only introduces new ideas and new methods into lexicographic data processing but also takes into account the conventional stylistic layout and methods of the dictionary, which can be applied to both electronic/Internet dictionaries and printed dictionaries. The International Standard has the following main features:

1 Uniform framework and data items for microstructure: no matter what type of dictionary, work can be done on the basis of the same tree structure; the differences between large and small dictionaries, multilingual and monolingual dictionaries, or general and specialized dictionaries lie in the types and numbers of data items.
2 Separation of the compiling format from the display format: the format for the lexical data input is different from that for display; what compilers have to do is to input information according to the tree structure. All the punctuation marks and structure marks can be automatically generated during the phase of data display, and will be displayed on the preview interface.

3 Clear marking of the relations among microstructure information items: in the input process, imported data will be automatically tagged, and checked when they are accessed or put in use. In this way, all sorts of data can be retrieved according to the compiler's needs, which contribute to a 'smart search'.
4 Standardization of formats for all the information or data: the database is built in LEXml format, and the lexicographic data may be exported according to specific needs and be connected to the interface of special typesetting systems, or be applied to other language database, language-processing, or machine-translation systems.
5 Flexibility in data retrieval and display: the LEXml format is a universal model, and the data structure of subsets can meet specific needs as long as they are constructed according to the international standards of the XML format.
6 Good compatibility with current XML tools: if special modes and tools are used in language processing that will lead to incompatibility with other software tools. XML, along with its specifications, has become an industrial standard, and naturally the XML-based LEXml can be put into use as XML models are, and can even be edited with XML editors and XSL style sheets.

Conclusion

Computational lexicography has gone through decades of development, and has acquired a distinct theoretical framework. Many achievements have been made in theoretical and practical research concerning computational lexicography. Unquestionably, computer technology has contributed greatly to the development of lexicographical studies and dictionary making, and more than 30 years' experience has been accumulated for the building and use of corpora. The development of lexicographical databases, and the use of computer-aided compiling systems have achieved noticeable success. Electronic dictionaries and online dictionaries have become ubiquitous in many countries. However, in some other countries, the methodology of dictionary making is still restricted to the conventional operating mode, and print dictionary publishing houses seldom keep pace with the development of electronic dictionaries. Therefore, lexicographers still have much to learn and do in computer lexicography in order to modernize dictionary making and publishing.

References

Amsler, Robert A. (1982) 'Computational Lexicology: A Research Program', in *Proceedings of the National Computer Conference*, Houston, TX, May, AFIPS, 657–663.

Atkins, Sue T. (1991) 'Building a Lexicon: The Contribution of Lexicography', *International Journal of Lexicography* 3: 167–204.

Atkins, Sue T. and Antonio Zampolli (eds) (1994) *Computational Approaches to the Lexicon*, Oxford: Oxford University Press.

Bennett, Paul A., Rod L. Johnson, John McNaught, Jeanette Pugh, Juan C. Sager, and Harold L. Somers (1986) *Multilingual Aspects of Information Technology*, Brookfield: Ashgate Publishing Company.

Boguraev, Bran and Ted Briscoe (eds) (1989) *Computational Lexicography for Natural Language Processing*, London: Longman Science and Technology.

Hartmann, R.R.K. and Gregory James (2000) *Dictionary of Lexicography*, Beijing: Foreign Language Teaching and Research Press.

Kilgarriff, Adam and Michael Rundell (2002) 'Lexical Profiling Software and Its Lexicographic Applications: A Case Study', in *Proceedings of the 10th EURALEX International Conference*, 13-17 August 2002, Copenhagen, Denmark, 807–818.

Kilgarriff, Adam, Rob Koeling, David Tugwell, and Roger Evans (2003) 'An Evaluation of a Lexicographer's Workbench: Building Lexicons for Machine Translation', in *EAMT '03: Proceedings*

of the 7th International EAMT Workshop on MT and Other Language Techology Tools, Improving MT through Other Language Technology Tools: Resources and Tools for Building MT, April, Budapest, Hungary, 9–16.
Kim, Steven H. (1991) *Knowledge Systems through Prolog: An Introduction*, Oxford: Oxford University Press.
Leech, Geoffrey (1992a) 'Corpora and Theories of Linguistic Performance', in Jan Svartvik (ed.) *Directions in Corpus Linguistics*, Berlin: Mouton de Gruyter, 105–222.
Ooi, Vincent B.Y. (1998) *Computer Corpus Lexicography*, Edinburgh: Edinburgh University Press.
Pustejovsky, James (1991) 'The Generative Lexicon', *Computational Linguistics* 17(4): 409–441.
Sinclair, John (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.
Sinclair, John (1985) 'Lexicographic Evidence', in Robert Illson (ed.) *Dictionaries, Lexicography, and Language*, Oxford: Pergamon Press in association with the British Council, 81–94.
van Eynde, Frank and Dafydd Gibbon (eds) (2000) *Lexicon Development for Speech and Language Processing*, Dordrecht: Kluwer Academic Publishers.
Zhang, Yihua 章宜華 (2004)《詞算詞典學與新型詞典》(*Computational Lexicography and New Dictionaries*), Shanghai: Shanghai Lexicographical Publishing House 上海辭書出版社.
Zhang, Yihua 章宜華 (2013)《詞算詞典學》(*Computational Lexicography*), Shanghai: Shanghai Lexicographical Publishing House 上海辭書出版社.

27
CONCORDANCING

Federico Zanettin

UNIVERSITY OF PERUGIA, ITALY

What is a concordance?

A concordance is an index of all the contexts in which a word appears in a given text or corpus. Concordancing involves retrieving all the instances of a specific word or expression from the corpus and displaying them in such a way that they provide context-based information.

Long before the advent of computers, concordances were manually produced as lists of words arranged alphabetically with indications to enable the inquirer to find the passages of the text where the words occurred. The first concordances were produced in 1230 by Dominican friars from the Vulgate, the Bible in Latin used in the Middle Ages. It was simply an index to the positions of a word in the text, but was later expanded to include the complete quotations of the passages indicated. These concordances did not of course contain all the words in the Bible, but only those deemed to be most important, and they quoted enough from a passage for one familiar with it to recall it to memory. Bible concordancing continued during the centuries, and concordances were produced for the Hebrew and Greek Bible (the Septuagint), as well as for their translations into English and other languages. Near completion, sometimes with reference to different versions, was achieved at the price of considerable bulk and weight (Herbermann 1913). The production of concordances was time-consuming scholarly work, and was typically carried out only for important books such as sacred and literary texts, with the notable exception of Otto Käding, who based his 1897 frequency dictionary of German on a manually collected 11-million-word corpus of legal and commercial texts (Těšitelová 1992: 90).

The first computer-generated concordances were produced by Father Roberto Busa in 1951, after he had created a machine-readable version of the works of Saint Thomas Aquinas in order to carry out its lemmatization, known as the Index Thomisticus. Computers allowed the carrying out of consistent, quick indexing and retrieval of any text available in electronic format, and this prompted the first studies of language based on corpora, undertaken in the US and Europe in the 1960s (see e.g. Kučera and Francis 1967; Quirk *et al.* 1972). Corpus linguistics established itself as a fully fledged discipline and methodology starting from the late 1980s, when research focused on language as a social rather than a psychological phenomenon, and approaches to the study of language based on actual textual products rather than on abstract linguistic competence became mainstream. The development of concordancing and corpus linguistics techniques more generally has been consequential to progress in computational power and storage capacity. Starting from the 1990s 'second generation' large 'reference'

corpora of up to around 100 million words were created, while in the 2000s 'third generation' very large corpora consisting of billions of words have appeared. Corpora and concordancing, once the domain of a few linguists, have become a resource at hand for translators, terminologists, and other language services professionals alike.

McEnery and Hardie (2012: 36–48) distinguish between four generations of 'concordancers', which largely correspond to four different phases of ICT development. First-generation concordancers were programs running on large mainframe computers and which could generally produce Key Word In Context (KWIC) concordances, that is printouts of all the occurrences of a word in a corpus, displayed in the middle of the paper or screen and accompanied on each side by enough context to fill a line. This basic display format is still usually the default option for all concordancers (Figure 27.1).

Second-generation concordancers were the first such programs available when personal computers started to become a commodity for the corpus linguist. While these concordancers included features which previously had to be executed by external programs, such as the possibility to sort results according to the alphabetical order of the words surrounding the search word (for instance, the concordance in Figure 27.1 is sorted according to the first word to the left of the search word), generate a wordlist, and compute some basic descriptive statistics, they had less processing power than mainframe programs. However, they allowed interested researchers to undertake corpus-based studies without needing to be part of a dedicated team or possessing programming skills. Third-generation concordancers are stand-alone applications for corpus analysis, and are still currently in use. As opposed to those of the previous generation, these concordancers 'were able to deal with large datasets … had bundled in with them a wider range of tools … gave access to some meaningful statistical analyses [and] effectively supported a range of writing systems' (McEnery and Hardie (2012: 40). Finally, fourth-generation concordancers have become available as a result of 2.0 ICT developments. They are based on a client–server architecture, meaning that search processing is done by a server application, while the input for a search is received from and the output is displayed on

·amme of development for `reading,	translation	and the book trade', covering translati
ł its own translation of the Bible - a	translation	which Rome deemed unacceptable. In
self being written in Salzburg, but a	translation	into German was also undertaken there
sing emulation over binary-to-binary	translation	is that ISVs do not have to tinker with
up Echo Logic Inc's binary-to-binary	translation	of the 680x0 into native PowerPC code
·land and The Wanderer , an English	translation	of Alain-Fournier's Le Grand Meaulnes .
ı audit opinion. The English	translation	of the attestation, reproduced in the i
·ers in Scotland than the equivalent	translation	, stone parmelia. It dyes wool a reddish
d version is identical to the existing	translation	(the present tense is the correct one
eek (a Catalan sestina with a helpful	translation	- into French). But it also dares to tou
ılfilment of Tyndale's dream that his	translation	of the Bible would reach and transform
ish fiction, whether in original or in	translation	, often dominates sales across the wor
; of language development, machine	translation	, dictionary production and testing of
ed. He proffered a modern	translation	of Hamlet's `To be, or not to be'.
ıme of them; others had needed no	translation	. His caresses had urged her to a wild,
l that in antiquity big enterprises of	translation	were due to public, not private, initiat
tives might indicate the initiation of	translation	is at Met163 (nucleotide position 619),
dable. The concept of mediation or	translation	between presenting problems and avail
derivation from the methods of oral	translation	in the Synagogue - which makes it impr
d, but Biff frowned at the trooper's	translation	. His brow furrowed with the effort to

Figure 27.1 A KWIC concordance of the word 'translation' (from the Sketch Engine)

the client's application, using a common browser as an interface. Whereas stand-alone concordancers work with a corpus residing on the same machine or local network, online concordancing software and services interact with a corpus located on a remote machine. Such corpora, potentially available and searchable from any computer, are often linguistically annotated and indexed.

Annotation refers to the enrichment of running text with explicit linguistic labels, as regards for instance lemmatization and part-of-speech (pos) tagging. Labels for lemmas allow the researcher to include in the results of a search different forms of the same basic lemma, for instance singular and plural forms of nouns and inflected forms of verbs. Pos labels allow distinguishing between homographs belonging to different word classes, for instance between the word 'go' as a verb and the same word form used as a noun. Indexing refers to the fact that searches are not conducted on-the-fly, as usually happens with stand-alone concordancers, but on a database containing information about the position and frequency of each word in the corpus. Indexing allows more flexible and quicker retrieval of even very large and heavily annotated corpora, since information about word position and annotation is stored separately from the texts themselves.

What concordancing can do for translators

Translation practitioners and professionals have at their disposal a number of computational tools and resources to help them perform translation tasks and jobs, ranging from dictionary and reference sources, to dedicated forums and social networks, to specialized software tools and services. Corpora and corpus analysis software play an important role as they allow translators to tap into linguistic and textual knowledge in a way which no other resource can offer. By analyzing concordance lines translators can derive information about how words are used in actual texts, be they source language texts to better understand the language they are translating from, or target language texts to confirm candidate translations and find unforeseen solutions to translation problems.

Concordancing software is clearly of no use without corpora. Translators, as well as other language professionals and learners, have two options available; that is they can resort to existing corpora or create new ones to suit their needs. In the first case they can either access corpora available online through a Web-based interface, or download corpora already compiled and analyze them through a local concordancer. Online services which offer access to one or more corpora through a concordancing application include, for instance, Mark Davies' interface to a range of monolingual corpora at Brigham Young University (BYU corpora: http://corpus.byu.edu/), comprising among others the 450-million-word Corpus of Contemporary American English (COCA), the 100-million-word British National Corpus (BNC), and Spanish and Portuguese corpora; the Sketch Engine's pre-loaded corpora (http://www.sketchengine.co.uk/), which comprise very large Web corpora in many languages as well as the BNC and other smaller corpora; other 'national' corpora (e.g. German, Czech, etc.), accessible through dedicated Web sites.

The large and very large corpora available online for free or at a fee are very useful resources, but sometimes a translator may be better off, corpus-wise, with a smaller but more specialized corpus relating to a specific translation task to be performed. Some specialized corpora can be found at language resources repositories such as the European Language Resource Depository (ELRA: http://www.elra.info/) and the Linguistic Data Consortium (LDC: http://www.ldc.upenn.edu/) and can in principle be downloaded and used by translators, though the choice of genres, topics and text types offered by these archives is ultimately restricted. Furthermore,

many of these corpora were often created for use in machine translation or other automated technologies rather than for manual analysis through concordancing software, and may thus prove impractical and difficult to set up and exploit. Thus, translators may find it worthwhile to build their own corpora, turning to the Internet as a source of suitable texts. To build their own DIY corpus translators can use corpus creation software such as BootCaT (http://bootcat.sslmit.unibo.it), which allows the user to compile a corpus semi-automatically from a set of Internet texts meeting specific criteria. These corpora can then be analyzed using a concordancer of choice. Alternatively, translators can resort to an online service such as the Sketch Engine, which allows the user to acquire a corpus (using a version of BootCaT), as well as to annotate and analyze it using the system's standard interface.

Translators may avail themselves not only of monolingual but also of bilingual corpora, that is corpora comprising two components or subcorpora, one in the source and one in the target language, to compare lexical and grammatical features across two languages. A first type of bilingual corpus is the comparable bilingual corpus, created by putting together two sets of texts in different languages, paired on the basis of design similarity. In this sense, two general reference corpora with roughly the same composition can be used as a comparable corpus. Specialized comparable bilingual corpora are, however, not easily found for many language pairs, and this is where translators may have to create their own DIY corpora. When using a comparable bilingual corpus search techniques and display options are the same for the two (sub)corpora, though the user will have to consider differences in writing systems, text segmentation and structural linguistic features.

A more specific type of bilingual corpus is the parallel corpus, comprising a set of source texts in one language and their translations in the other, or two sets of texts in the two languages which are held to be 'equivalent', for instance the different language versions of EU legislation. Parallel corpora can be equally difficult to create, as they require a non-straightforward process of alignment, that is the pairing of source and target 'equivalent' segments, on a sequential basis. In order to take advantage of aligned parallel corpora specific search and display functions must be made available in addition to those found in monolingual concordancers (see below).

The usefulness of corpora and concordancing over more traditional tools may be assessed by comparing them with dictionaries. Both corpora and dictionaries can be consulted to help understand the source text and compose the target text. Large reference corpora can be seen as analogous to general language dictionaries, while smaller specialized corpora play a function similar to that of specialized monolingual dictionaries. Parallel corpora can instead be compared to bilingual dictionaries, as they both provide a direct link between lexical items in two languages.

Dictionaries, on paper and in electronic format, offer information about words which has already been distilled by lexicographers, often based on corpus evidence. While dictionaries favor a synthetic approach to lexical meaning via a definition, and by necessity condense and simplify the complexity of lexis, corpora allow for an analytic approach via multiple usage contexts. In selecting a target language equivalent from a target monolingual dictionary a translator has to appraise the appropriateness of the translation candidate to the new context by using a definition and a few examples. However, translators often need to understand precise senses of meaning and nuances of use, and corpus concordancing provides access to a range of examples of actual language use which no dictionary can offer. Clearly, the added comprehensiveness has a cost, which is that translators must make out by themselves the solution to their problems by exploring and interpreting a very large quantity of raw textual data. To this end, they must be able to take advantage as best as possible of corpus concordancing techniques, including concordancers' data search and display options.

Dictionaries are primarily accessed by looking up basic word forms (lemmas), though some electronic dictionaries allow for searching also within definitions and examples. Corpora instead allow searching for specific (groups of) word forms as well as (variable) phrases in the context of other words or expressions. Sinclair (1991) has argued that the meaning of a word is determined by the patterns in which it occurs, and that lexis and grammar cannot be treated separately. Rather, each word has its own specific grammar, which comprises the structures it appears in as well as its collocations. Close observation of corpus concordances and collocations can unveil syntactic and semantic patterns of lexis as occurring in natural language, as well as information relating to text type and textual organization. Specialized corpora can be especially useful in providing information on lexical, syntactic and rhetorical structures of a specific text type or genre. For instance, by concordancing even a very small corpus of medical articles translators can be helped to make well-informed choices on medical phraseology (Gavioli and Zanettin 2000), or find out whether a given expression is typically used in the first or in the last part of research articles (Aston 1997).

Bilingual dictionaries are repertories of lexical equivalents (general dictionaries) or terms (specialized dictionaries and term banks) established by dictionary makers which are offered as translation candidates. Parallel corpora are repertoires of translation equivalents as well as of strategies past translators have resorted to when confronted with problems similar to the ones that have prompted a search. Parallel corpora provide information that bilingual dictionaries do not usually contain, since while the former supply lexical equivalents, the latter also offer examples of lack of direct equivalence. A parallel corpus can, in fact, provide evidence of how actual translators have dealt with cases where there is no easy equivalent for words, terms or phrases across languages (Zanettin 2002).

Corpora and concordancing have also changed terminological practice, that is the way terminological entries are compiled and used. According to Bowker (2011), as personal terminology management systems have largely replaced large institutional data banks, terminological work has moved from an onomasiological to a semasiological approach. That is, rather than using a conceptual ontology to map the terms used in a specific domain, personal term banks are usually compiled from lists of words obtained from corpora. Furthermore, terminological entries do not necessarily fit into the traditional definition of terms as nominal constructs, as they may consist of frequent combinations of words belonging to different word classes. They are often recorded in their most frequent rather than in the base form, and synonyms may be registered as different entries. The entries will often contain basic information, i.e. target language equivalent(s) and selected concordance examples.

Translators can resort to monolingual and bilingual corpora and corpus analysis software to find information about terms, phraseology and textual patterns in both source and target languages, and to parallel corpora to find solutions to translation problems based on previous translational experience. Large, general monolingual corpora are now available for many languages, and translators can create their own small corpora from the Web by downloading and processing documents retrieved using search engines and compiled through semi-automatic routines implemented by *ad hoc* programs and online services. Proficiency in corpus concordancing skills and procedures has become an indispensable part of the translator's professional competence. Like the use of dictionaries, the use of corpus concordancing has to be learnt (Frankenberg-Garcia 2010). The usefulness of concordancing depends on the corpus, on the software and on the user. Users must be aware of what a corpus contains and to what extent observations derived from it are relevant, reliable, and applicable to a specific translation task. They must also be able to understand the potentialities and limitation of the software which is used to interrogate a corpus, and to interpret the results of a search appropriately.

Search and display options

Corpus software includes applications which are used to perform the various tasks associated with corpus construction and analysis, that is to acquire, process, manage, query and display corpus data. While translators may need to become involved in corpus compilation, this chapter focuses on corpus analysis, and on concordancing in particular. 'Concordancers', as corpus analysis tools are often referred to, can be either stand-alone or online applications which, for free or at a cost, allow the user access to corpus data. These applications offer different functionalities and features, each with advantages and disadvantages. While a comprehensive list of commercial and public domain software and services would be too long to include here (see e.g. Zanettin 2012 for a list of concordancing applications), some of the most commonly used are mentioned in the discussion, and bibliographical references are provided.

A search is typically carried out by typing a search string, i.e. a typographical word or series of words in a search box, much as it happens with a general purpose search engine. There are however decisive differences between general search engines like Google, Yahoo! or Bing and concordancers, both as regards search and display facilities. Furthermore, the Internet can only be considered a corpus to a certain extent, inasmuch as it is an open-ended repository and even though a search can be restricted to specific sectors of the WWW (e.g. a newspaper archive, a mailing list or a forum, Facebook, Twitter, Google Books), search and display functions of search engines are not fully within the control of the user. However, while first- and second-generation corpora were composed of printed texts typewritten or scanned in and OCRed, corpora of the 2000s are increasingly made up of 'native' electronic texts, often downloaded from the Internet through semi-automated procedures.

Concordancers may be assessed according to the options available for data search and display. These may vary depending on whether the concordancing application is the only or main tool of corpus analysis software, or a component of a different piece of software, for instance a translation memory management system. Both stand-alone programs and online services may include, besides the concordancing function proper, other options for displaying corpus data.

Search options

A simple search for a word or a phrase in a concordancer will retrieve all instances of that word or phrase in all the texts in the corpus. More advanced searches are generally based on regular expressions resembling those of programming languages, and allow for the retrieval of variable textual patterns. Thus, while in a simple search precision is ensured by retrieving all and only the citations containing the exact search string specified, in advanced searches different characters can be used to increase recall by allowing for variation. Non-alphabetic characters are attributed special meanings, some acting as wildcards. For instance, an asterisk * is used by some concordancing programs to represent one or more trailing characters, while the escape \ character can be used to invoke alternative interpretations of subsequent characters in a character sequence or to introduce a list of alternative characters, depending on the software used. For instance, in WordSmith Tools (Scott 2008) an asterisk can be used to retrieve all words beginning or ending with a specified sequence, so that 'go*' will retrieve 'go', 'going', 'Godzilla', etc., while '*go' will retrieve 'embargo,' 'forgo,' 'tango,' etc. The escape character can be used to provide alternatives, for instance the string 'go\goes\going\gone' can be used to retrieve all the forms of the verb.

Searches using regular expressions can be extremely powerful. For instance, a search for the regular expression

\bha[vs]e?\W\w{4,}e[nd]\b

in MonoConc Pro (Barlow 2004) will retrieve all instances of 'has' or 'have' followed by an *-en/-ed* form: The metacharacter \b is used to indicate word beginning and end, alternative characters are enclosed in square brackets. The question mark indicates an optional preceding character, while '[t]he part of the search query that we hope will match the participle is "padded" with alphanumeric characters (\w) to eliminate shorter words ending in *-en* or *–ed* such as ten and bed' (Barlow 2003a: 62). As Barlow explains, regular expression (regex) searches allow for very complex queries, though caution must be exerted:

> some good hits such as seen will also be omitted by this search query and in cases like this it is up to the user to formulate the search query in such a way as to get the right balance between a good retrieval rate and a high percentage of desired forms in the concordance results. The specification of a minimum of four letters in the participle in the search query above has the effect of increasing the percentage of good 'hits' in the results, at the cost of missing some instances of the present perfect that occur in the corpus.
>
> (ibid.: 70)

Annotated corpora can be used to both fine tune and simplify a search. In a corpus which has been lemmatized and pos-tagged a search can be carried out not only in the running text but also in the content of the labels attached to words. For instance, a search for verb forms in the present perfect in the English corpus which is part of the Leeds collection of Internet corpora (Sharoff 2006, http://corpus.leeds.ac.uk/internet.html), is formulated as follows:

[lemma='have'] [pos='V.*']

The search syntax specifies that the concordances returned must contain all forms of the lemma 'have' followed by any verb.

The Sketch Engine online concordancing service, which is similarly based on the IMS Corpus Workbench, a standard platform for corpus indexing and management, includes a more sophisticated interface. Figure 27.2 shows how the same search can be performed through a search input form, which allows the user to select the contents of the annotation (the tag attributes) from a predefined set of options contained in dropdown menus.

The interface to Mark Davies' annotated corpora allows the user access to variable phraseological units, for instance all instances of the construction 'VERB one's way PREP' (Figure 27.3).

A graphical user interface (GUI) makes queries more user friendly, though expert users can still run searches using the regular expression search syntax.

Additional annotation regarding text or discourse features can also be exploited if available. For instance, if the texts in a corpus include metalinguistic annotation concerning genre or text type, date, author, etc., these specifications can be used to filter out unwanted texts and create more precise subcorpora and therefore retrieve more accurate information from concordances.

A parallel concordancer allows the user to perform a search in either of the two subcorpora which make up the bilingual parallel corpus and retrieve, together with the lines or sentences containing the word, phrase or variable expression searched for in one language, the corresponding lines or sentences in the other language. In addition, some parallel concordancers allow the specification of search criteria in both languages, to return only those results from the

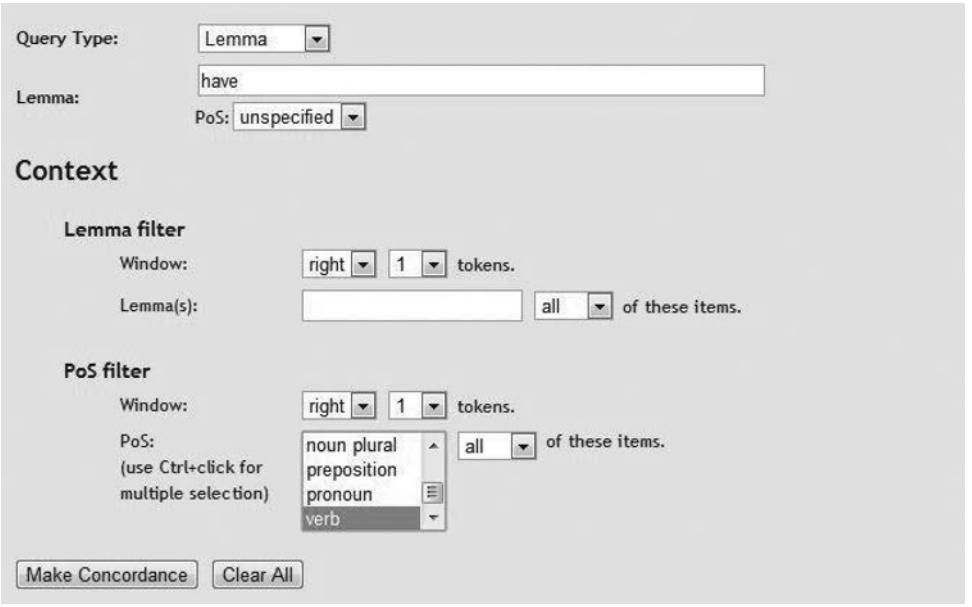


Figure 27.2 A search for the lemma 'have' immediately followed by a verb (from the Sketch Engine)

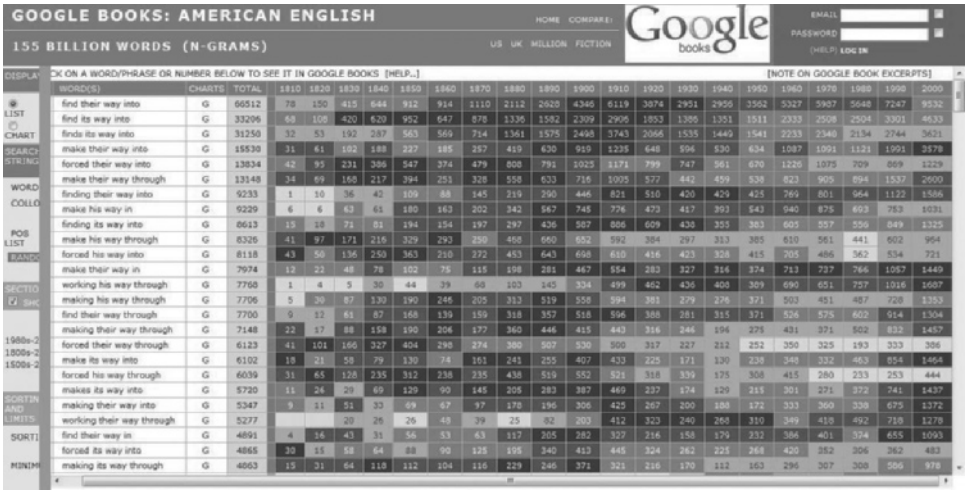


Figure 27.3 'verb + one's way + preposition' constructions in the 155-billion-word Google Books Corpus of American English

Source: from Davies' corpus.byu.edu

texts in one language for which the paired target segments contain the expression specified for the other language. For instance, a search in an English–Italian parallel corpus can be made to include in the results only those occurrences of the English word 'run' contained in sentences for which the corresponding aligned sentences in Italian contain a form of the verb 'correre'.

Statistical information about word frequency and position can be used to (semi)automatically retrieve the most likely translations for a given word or expression. For instance, the ParaConc

stand-alone parallel concordancer (Barlow 2005) has a 'hot words' function which allows the user to select from a list of 'possible translations and other associated words (collocates) … suggested by the program itself' (Barlow 2003b: 68) on the basis of how frequently these words appear in segments which translate the source text segments. Annotated corpora may provide additional information with which more precise bilingual (semi)automated searches can be performed.

A special type of parallel corpora is translation memories (TMs), which are created by translation memory management systems (TMMSs) as a by-product of the translation process. As translations are carried out source and target text segments are stored together in the system's 'memory'. In the TM database each Translation Unit (TU), that is, each aligned text segment pair, is archived together with administrative information and retrieved by the system in order to be considered again as a candidate for future translations. TMs are usually proprietary, that is they belong to individual translators or translation companies, but they can also be created from publicly available parallel corpora. For instance, some very large multilingual parallel corpora, including the Europarl corpus (the proceedings of the European Parliament from 1996 to the present, consisting of up to 50 million words for each official language of the European Union) and the Acquis Communautaire (the entire body of EU legislation, consisting of one billion words altogether) are available both as aligned parallel corpora and in standard TM format.

Most TMMSs offer a way to generate parallel concordances from a word or expression in the source text, that is a list of all the translation units in the memory in which that word or expression occurs, together with the target segments. Such applications are usually less sophisticated than most stand-alone bilingual concordancers. As opposed to the former, the latter allow more control over both what is searched by letting the user perform more flexible pattern searches, and over how results are displayed by letting the user sort the results and access the wider context of a given segment (Bowker and Barlow 2008). Though translation memory search engines allow for 'fuzzy' matches, that is to search for target segments which only partly match the new source text, the text retrieval system is often not geared to performing complex searches such as those described above, and results are not liable to be manipulated regarding the order and the format in which they are displayed (see below). Some hybrid tools, however, allow for the integration of the functions of TMMSs and parallel concordancers (ibid.: 20).

Data display options

General search engines typically return the hits of a search placing the search string in the middle of documents' extracts, accompanied by the documents' URL and title. Results are ordered according to non-linguistic, often commercial, reasons and they can be browsed in the order of retrieval but not further manipulated. Though the usefulness of general search engines and of the WWW as a corpus for linguistic purposes should not be underestimated, especially when it comes finding uncommon or long phrases (Zanettin 2009), concordancers and pre-defined corpora offer a number of advantages as regards display options, since they usually allow the user to view the results of a search in a number of visualization formats. The most common of these is the KWIC format illustrated in Figure 27.1, with the possibility of enlarging the context around the search word to more than one line. Most concordancers also allow the user to switch between line and sentence view, sentence boundaries being derived either from punctuation marks or explicit annotation. Concordances can be aligned around the search word in the middle of the screen, or by having sentences begin along the left margin. The

occurrences shown can be limited to a random selection in order to make the analysis more manageable. Results can be filtered according to contextual restrictions, that is by including only those where a given word of phrase occurs within a given right of left word span. In order to highlight linguistic patterns, concordance lines can be ordered alphabetically by sorting them according to the 'node' words or according to the words to the right or left of them, establishing different sorting levels if desired. Color coding and typeface can be used as further visual aids. Concordances may also be categorized according to user-decided criteria, by manually marking occurrences or by using existing mark-up such as lemma or pos tags if the corpora have been previously annotated. Sorting concordance results according to linguistic patterns and visualizing them in a clear and memorable way allows the user to acquire important information about how words and phrases are used in actual texts and contexts.

Concordancers sometimes include facilities which, on the basis of statistical analysis, provide information about the frequency of occurrence of words in relation to each other and in the corpus as a whole which would be otherwise impossible to recover. Thus, for instance, WordSmith Tools can not only generate concordances but also give access to information about *Collocates*, *Plots* (the dispersion of words in the texts in the corpus), *Patterns* and *Clusters*. The *Patterns* view shows collocates of a given word visually organized in terms of frequency. The *Clusters* view displays all recurrent groups of words which appear more frequently around the search word or expression, for example clusters of between three and six words, with a frequency of at least five occurrences. Warren's (2009) *ConcGram* was created specifically to 'identify all of the co-occurrences of two or more words irrespective of constituency and/or positional variation in order to account for phraseological variation and provide the raw data for identifying lexical items and other forms of phraseology' (Greaves 2009: 2). *ConcGram* displays concordance lines with all the words in a 'concgram' equally highlighted, and concordance output can be sorted and centered alternatively around any of these words.

Further options for the display of collocational data, which rely to an even greater extent on visual and graphical features, include 'collocate clouds', in which collocates are listed alphabetically, while frequency information is displayed as font size and collocational strength as text brightness, and 'concordance trees' (Luz 2011), in which collocational patterns can be made out using similar visual indicators (Figure 27.4).

Linguistic annotation can be used to sort collocational patterns on the basis of the relations between words, their grammatical class and their position, and to obtain automated summaries of such relations. Thus, the Sketch Engine provides 'word sketches', i.e. 'one-page, automatic, corpus-derived summary of a word's grammatical and collocational behavior'(http://www.sketchengine.co.uk) in which colligational relations (collocation between words and word classes rather than simply between words) are shown. Annotated corpora can also be used to generate automated thesauruses of words that tend to occur in similar contexts in terms of grammatical and collocational behavior.

Finally, bilingual concordances can be arranged on the screen either along the vertical or the horizontal axis: in the vertical display alignment units are arranged side by side; in the horizontal display concordance lines can either be presented as alternated sentences/segments in different languages, or the screen can be split into an upper and a lower window, one containing the output for the search expression in the source language and the other containing the aligned target segments in the same order. Within each presentation display results can be shown in KWIC format, with concordance lines centered around the search expression either or both in the source and target language, or in any other format usually available in monolingual concordancers.

Figure 27.5 shows a KWIC bilingual concordance in a horizontal split-screen display, in which concordances are sorted according to target language order. The link between segment

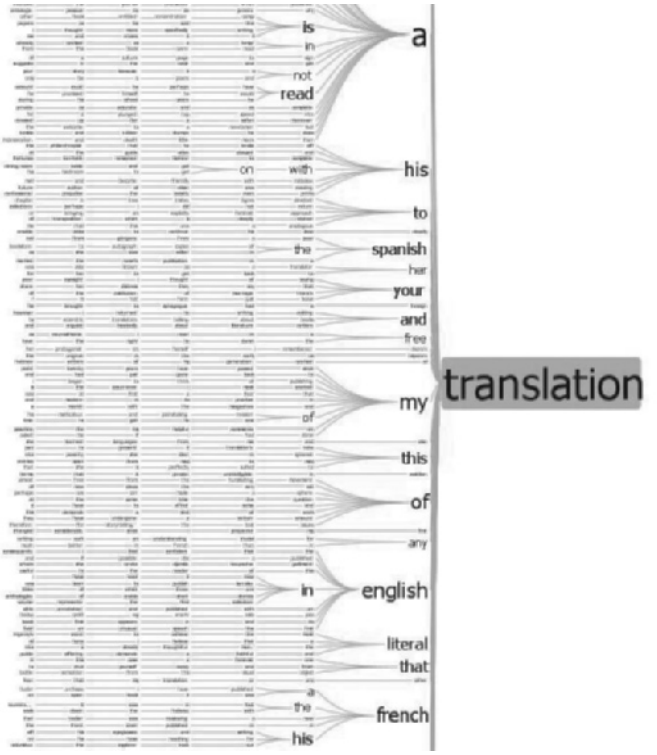


Figure 27.4 Left-side concordance tree of the word 'translation'

Source: Luz's TEC browser

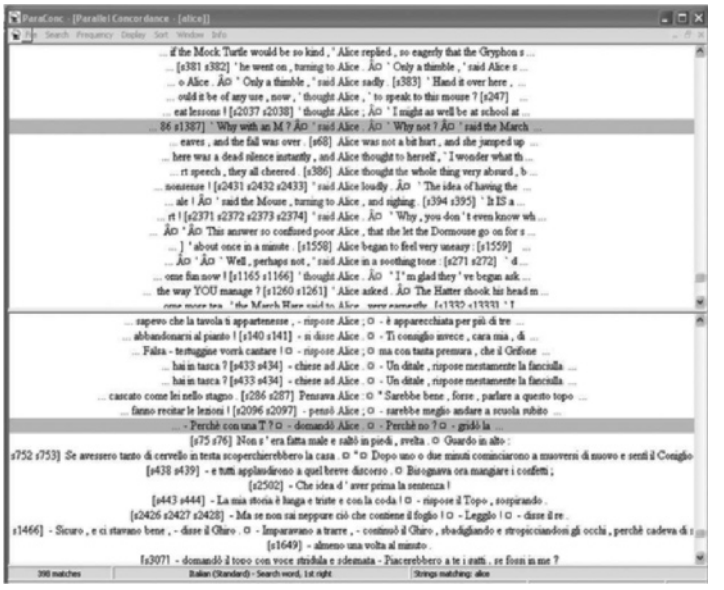


Figure 27.5 Parallel concordance ordered according to target language

Source: Barlow's ParaConc

pairs in different languages is not provided by spatial proximity, with segment pairs displayed in adjacent columns or lines, but by sorting order and highlighting. Source text segments and their translations are arranged in the same sequence, and concordance lines are highlighted in pairs. When the results in one language are re-sorted or otherwise manipulated, much as they are with a monolingual concordancer, aligned units in the other language are re-ordered accordingly.

Different display formats are useful for focusing on different aspects of bi-textual correspondences. The KWIC display in split-screen format allows a better visualization of linguistic patterns, whereas the sentence-by-sentence interlinear format is more efficient for comparing correspondences at sentence level.

References

Aston, Guy (1997) 'Small and Large Corpora in Language Learning', in Barbara Lewandowska-Tomaszczyk and Patrick James Melia (eds) *PALC 97: Practical Applications in Language Corpora*, Łodz: Łodz University Press, 51–62.

Barlow, Michael (2003a) *Concordancing and Corpus Analysis Using MP 2.2*, Houston, TX: Athelstan.

Barlow, Michael (2003b) *ParaConc: A Concordancer for Parallel Texts*, Houston, TX: Athelstan.

Barlow, Michael (2004) *MonoConc Pro 2.2*, Houston, TX: Athelstan. Available at: http://athel.com/product_info.php?products_id=80.

Barlow, Michael (2005) *ParaConc*, Houston, TX: Athelstan. Available at: http://www.athel.com/paraconc.pdf.

Barlow, Michael and Lynne Bowker (2008) 'A Comparative Evaluation of Bilingual Concordancers and Translation Memory Systems', in E. Yuste Rodrigo (ed.) *Topics in Language Resources for Translation and Localisation*, Amsterdam and Philadelphia: John Benjamins, 1–22.

Bowker, Lynne (2011) 'Off the Record and On the Fly: Examining the Impact of Corpora on Terminographic Practice in the Context of Translation', in Alet Kruger, Kim Wallmach, and Jeremy Munday (eds) *Corpus-based Translation Studies: Research and Applications*, London and New York: Continuum, 212–236.

Davies, Mark (2008) 'The Corpus of Contemporary American English: 450 Million Words, 1990–Present'. Available at: http://corpus.byu.edu/coca.

Frankenberg-Garcia, Ana (2010) 'Raising Teachers' Awareness of Corpora', *Language Teaching* 1(1): 1–15.

Gavioli, Laura and Federico Zanettin (2000) 'I corpora bilingui nell'apprendimento della traduzione. Riflessioni su un'esperienza pedagogica', in Silvia Bernardini and Federico Zanettin (eds) *I corpora nella didattica della traduzione* (*Corpus Use and Learning to Translate*), Bologna: Cooperativa Libraria Universitaria Editrice Bologna, 31–44.

Greaves, Chris (2009) *ConcGram 1.0*, Amsterdam and Philadelphia: John Benjamins. Available at: http://benjamins.com/#catalog/software/cls.1.

Herbermann, Charles (ed.) (1913) 'Concordances of the Bible', in *Catholic Encyclopedia*, New York: Robert Appleton Company, 4: 195–216. Available at: http://en.wikisource.org/w/index.php?title=Catholic_Encyclopedia_(1913)/Concordances_of_the_Bible&oldid=2168065.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell (2004) 'The Sketch Engine', in *Proceedings of the 11th Euralex International Congress*, Lorient, France: Université de Bretagne Sud, 105–116. Available at: http://www.sketchengine.co.uk.

Kučera, Henry and W. Nelson Francis (1967) *Computational Analysis of Present-day American English*, Providence, RI: Brown University Press.

Luz, Saturnino (2011) 'Web-based corpus software', in Alet Kruger, Kim Wallmach, and Jeremy Munday (eds) *Corpus-based Translation Studies – Research and Applications*, London and New York: Continuum, 124–149.

McEnery, Tony and Andrew Hardie (2012) *Corpus Linguistics: Method, Theory and Practice*, Cambridge: Cambridge University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1972) *A Grammar of Contemporary English*, London: Longman.

Scott, Mike (2008) *WordSmith Tools* 5.0, Oxford: Oxford University Press. Available at: http://www.lexically.net/wordsmith/version5.

Sharoff, Serge (2006) 'Creating General-purpose Corpora Using Automated Search Engine Queries', in Marco Baroni and Silvia Bernardini (eds) *WaCky! Working Papers on the Web as Corpus*, Bologna: GEDIT, 63-98. Available at: http://wackybook.sslmit.unibo.it.

Sinclair, John (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Těšitelová, Marie (1992) *Quantitative Linguistics*, Amsterdam and Philadelphia: John Benjamins.

Warren, Martin (2009) 'Introduction', in Chris Greaves *ConcGram 1.0*, Amsterdam and Philadelphia: John Benjamins.

Zanettin, Federico (2002) 'Corpora in Translation Practice', in Elia Yuste-Rodrigo (ed.) *Proceedings of the 1st International Workshop on Language Resources for Translation Work, Research and Training (LR4Trans-III)*, Paris: ELRA, 10-14. Available at: http://www.lrec-conf.org/proceedings/lrec2002/pdf/ws8.pdf.

Zanettin, Federico (2009) 'Corpus-based Translation Activities for Language Learners', *The Interpreter and Translator Trainer (ITT)* 3(2): 209–224.

Zanettin, Federico (2012) *Translation-driven Corpora Corpus Resources for Descriptive and Applied Translation Studies*, Manchester: St. Jerome Publishing.

28
CONTROLLED LANGUAGE

Rolf Schwitter

MACQUARIE UNIVERSITY, AUSTRALIA

Introduction

Natural languages are the primary mode of human communication. In their textual form they constitute the most widely used medium for storing human knowledge. Natural languages are also the most expressive knowledge representation languages that exist, far more expressive than any machine-processable formal language. While natural languages allow humans to deal with most aspects of everyday life, their expressive power can create problems for both humans and machines. Sometimes it is difficult for humans who have only limited knowledge of an official language that is used in a work environment to read and understand technical documents. It is also often difficult for machines to process full natural language for a given task because of its inherent ambiguity and complexity.

Controlled natural languages tackle these kinds of problems by restricting the size of the grammar and vocabulary in order to reduce or eliminate ambiguity and complexity inherent in natural languages. Controlled languages can be classified into four major groups according to the purpose they have been designed for. One group of controlled languages has been created to improve the communication between humans who do not share a common native language. A second group of controlled languages have been developed to make it easier for non-native speakers to read and understand technical documentation written in a foreign language. A third group of controlled languages aim to improve the quality of machine translation and reduce the post-editing effort for translated documents. And finally, a fourth group of controlled languages have been designed as high-level interface languages for semantic systems where different forms of automated reasoning are used to make inferences from knowledge expressed in controlled language, in particular to answer questions about this knowledge.

Controlled languages for human communication

Historically, the most famous controlled language for human communication is Basic English (Ogden 1930) which was created as an international auxiliary language to help non-native speakers to learn English as a second language and use the language for general and technical communication. Basic English derives its vocabulary and grammar from Standard English and eliminates those words (in particular verbs) that can be reconstructed by using simpler words and a number of prescribed grammar rules.

The core vocabulary of Basic English is very small. It consists of only 850 words: 600 of them are nouns (e.g., *act*, *hour*, *milk*, *town*, *tooth*), 150 are adjectives (e.g., *angry*, *conscious*, *loud*, *quiet*, *true*) including 50 opposites, and 100 words that are called operations (e.g., *come*, *enough*, *for*, *he*, *some*, *or*). This last category includes verbs, adverbs, prepositions, pronouns, quantifiers, and conjunctions; words that are used to put other words in statements into 'operation'. The size of the vocabulary of Basic English is kept small by using only 18 verbs (*come, get, give, go, keep, let, make, put, seem, take, be, do, have, say, see, send, may, will*). These verbs are also called operators and operator-auxiliaries in Basic English. The reduction to this subset is based on the assumption that irregular verb forms of English are difficult to learn for non-native speakers and that all important verbs can be expressed by more basic constructions. These alternative constructions use the 18 basic operators as a starting point and combine them with prepositions to express the intended meaning. For example, instead of the verb *approach* in *approach (a town)*, the operator *come* is used together with the preposition *to*; this results in the basic construction *come to (a town)*. Another example is the construction *take out (a tooth)* instead of *extract (a tooth)*.

To address the need of a particular work environment, the core vocabulary of Basic English is augmented by 100 words for any general environment (e.g., science or trade) and 50 words for any specific field in that environment. The resulting vocabulary of 1000 words is further enriched by a list of international words (e.g., *hotel*, *electricity*, *university*) that are presumed to be widely understood without additional instructions. According to the advocates of Basic English, this vocabulary is sufficient for any form of business communication or publication that is required for international use.

The grammar of Basic English follows the accepted rules of English but is subject to a number of restrictions. For example, compound nouns can be formed by combining two basic nouns (e.g., *footnote*) or an adjective and a noun (e.g., *blackberry*), and derivatives can only be constructed by using a specific group of suffixes (*-ed*, *-er*, *-ing*, *-ly*, *-s*) and one single prefix (*un-*).

Below is a short excerpt of the Atlantic Charter written in Standard English as well as in Basic English (Ogden 1968):

Standard English

> The President of the United States and the Prime Minister, Mr. Churchill, representing His Majesty's Government in the United Kingdom, being met together, deem it right to make known certain common principles in the national policies of their respective countries on which they base their hopes for a better future for the world.
>
> First, their countries seek no aggrandizement, territorial or other.
>
> Second, they desire to see no territorial changes that do not accord with the freely expressed wishes of the peoples concerned. …

Basic English

> The President of the United States and the Prime Minister, Mr. Churchill, acting for His Majesty's Government in the United Kingdom, being now together, are of the opinion that it is right to make public certain common ideas in the political outlook of their two countries, on which are based their hopes for a better future for all nations.
>
> First, their countries will do nothing to make themselves stronger by taking more land or increasing their power in any other way.

> Second, they have no desire for any land to be handed over from one nation to another without the freely voiced agreement of the men and women whose interests are in question. …

As this example illustrates, the simplification of the vocabulary is achieved at the expense of longer sentences that include sometimes lengthy paraphrases. Although experience proved that Basic English was easy to learn to read, it turned out that it was difficult to rewrite a given text in Basic English so as to preserve the original meaning.

A modified form of Basic English, known as Simple English, is nowadays used to write articles for Simple English Wikipedia,[1] an online encyclopaedia that uses fewer words and a simpler grammar than the ordinary English Wikipedia. This simplified encyclopaedia is designed for people who are trying to learn English or who have special needs (e.g., for children, students, and adults with learning difficulties). Most Simple English articles are not new Wikipedia articles, instead they have been taken from the ordinary English Wikipedia and rewritten in order to make them simpler and easier to understand. Some of these articles are written in Basic English; however, the writing guidelines of Simple English Wikipedia do not specify strict rules that prescribe which words or grammatical structures can be used, and alternative resources to Basic English are recommended. The writing guidelines suggest that about 2000 words are enough to write a normal Simple English Wikipedia article. Similar to Basic English, the use of Simple English does not result in shorter articles, although these articles often use shorter sentences than the original Wikipedia source. It is not uncommon that a Simple English article requires between 25 percent and 50 percent more words than the original Wikipedia article. However, we can observe an interesting shift between the early use of Basic English and Simple English: the guidelines of Simple English focus more on the use of simpler grammatical structures and shorter sentences compared to the guidelines of Basic English where the focus is on the control of the vocabulary. The guidelines of Simple English recommend, for example, the use of the following sentence patterns to reduce syntactic complexity:

1 Subject – Verb – Direct Object.
2 Subject – Verb – Indirect Object.
3 Subject – Verb – Direct Object – Indirect Object.
4 Subject – Verb – Direct Object – Subordinate Clause.
5 Subject – Verb – Direct Object – Indirect Object – Subordinate Clause.

The guidelines also discuss a number of techniques which illustrate how complex sentence structures can be simplified, for example by removing conjunction or by isolating multiple subordinate or dependent clauses.

Controlled languages for technical documentation

ASD Simplified Technical English (ASD-STE100) is a controlled natural language with a much narrower focus than Basic English or Simple English. STE, formerly known as AECMA Simplified English, was created for the aerospace industry to help readers easily understand maintenance documentation (ASD-STE100 2010). Most technical documentation in this industry is written in English and used in multi-national programmes. However, many readers of this documentation have only limited knowledge of English and are often overwhelmed by complex sentence structures and the number of meanings and synonyms of English words.

STE was developed to address these issues with the aim of improving the quality of procedural and descriptive texts in maintenance documentation so that human errors can be reduced during maintenance tasks in the aerospace industry.

The STE specification provides a set of about 60 writing rules and a basic dictionary with a vocabulary of about 870 approved words for writing technical documentation. If a word is not in the STE dictionary, then it is not approved and cannot be used in a technical document, unless it is a manufacturer-specific word that qualifies as a technical name or a technical verb and fits into one of the categories listed in the STE specification.

The words in the STE vocabulary were chosen for their simplicity and ease of recognition. In general, there exists only one part of speech for one word, and only one word for one meaning. For example, the word *test* is approved only as a noun but not as a verb, and the verb *follow* has only the approved meaning *come after* but not *obey*. The writing rules of STE cover aspects of grammar and style, and mainly regulate the use of word forms, grammatical voice, sentence lengths, and layout. Some of the writing rules are easy to check automatically, for example:

RULE: 5.1 Keep procedural sentences as short as possible (20 words maximum).

Other writing rules are difficult or even impossible to check automatically since they rely on domain-specific knowledge and on human experience:

RULE: 6.8 Present new and complex information slowly.

Writing correctly in STE is not an easy task since this requires a good command of English together with detailed knowledge of the domain and familiarity with the STE specification. There exist commercial word and rule checkers that support the writing process of STE, flag unapproved and unknown words, and violations of rules. However, these checkers are no replacement for training in STE authoring since these tools cannot do the hard work and transform a non-STE compliant text automatically into a compliant one.

Although STE was not intended for use as a general writing standard, it has been successfully adopted by other industries for their documentation needs, including the defence, construction, and medical industries. It turned out that STE is not only beneficial for those who do not have English as their first language, but also for native speakers since simple and unambiguous texts can improve the readability and comprehensibility of documents for all users, and as a consequence limit human factor risks, in particular in safety-critical domains.

Another benefit of writing in STE is that documents are easier to translate into other languages, although this is not the primary objective of STE. In some cases translation of safety-critical documentation is not even allowed by national regulations.

Controlled languages for machine translation

Machine translation (MT) is another interesting application area for controlled languages. Various controlled languages have been used in industry to improve the quality of MT output, in particular for multilingual translations in technical domains, when a document is authored in a source language and then translated into multiple target languages (Hutchins 2005: 5–38). The primary objective of using controlled language for MT is to limit lexical ambiguity in the source language and to rule out complex sentence structures in order to ease the processing and achieve better translation results. The overall quality of the translation depends on the rule set

that restricts the input language, the availability of tools that help authors comply with this rule set, and the MT system that is used in the translation process.

Traditionally, MT systems use either a rule-based or a corpus-based approach to translate a document. While rule-based machine translation systems use lexical and grammatical rules to govern the translation process, statistical machine translation systems use statistical models derived from bilingual text corpora to find the best translation. Nowadays, often hybrid machine translation systems are in practical use and combine the strengths of the first two approaches by post-processing the output of rule-based systems using statistical methods or by pre-processing the input or output of statistical systems with the help of rules.

Most commercial MT systems have been designed for processing full natural language but provide mechanisms for domain customization. This means that the input to the MT system is often restricted in a specific form by human intervention to improve the translation quality. Controlled languages can help to optimize this customization process in a systematic and linguistically motivated way for a particular application domain. The restrictions on the source language for MT are often stricter than those for writing technical documentation since the main goal is the reduction of ambiguity in input sentences for an MT system and not the improvement of readability for a human reader; however, these simplifications may work not only for machines but also for humans. However, caution is advised, if the controlled language that is used to write the source text for MT becomes too restrictive. This is because sentences that are not stylistically adequate will not be accepted by technical writers and this can lead to usability and productivity problems.

Many of these special requirements for controlled language processing for MT have been addressed in the KANT system (Mitamura 1999: 46-52). It is instructive to have a closer look at this rule-based system since it tightly integrates controlled language checking and multilingual translation. The input language to the KANT system, KANT Controlled English, specifies lexical and grammatical restrictions. It turned out that the most effective way to improve the translation accuracy of the KANT system is to limit lexical ambiguity. In most cases, the lexicon of the KANT system encodes only a single meaning for each word/part-of-speech pair, and alternative terms are used if a lexical item has more than one potential meaning in a domain. If a term must absolutely carry more than one meaning, then interactive lexical disambiguation is carried out during source language analysis. Other lexical restrictions concern the use of function words, modal verbs, participle forms, acronyms/abbreviations, and orthography.

The grammar of KANT Controlled English is based on two types of grammatical constraints: phrase-level constraints and sentence-level constraints. Phrase-level constraints govern the use of phrasal verbs; for example, particles of phrasal verbs are often ambiguous with prepositions, and these verbs should therefore be replaced by single-word verbs. Note that this is in contrast to Basic English where a small number of verbs that function as operators are combined with prepositions. Other phrasal-level constraints govern the use of coordinated verb phrases and conjoined prepositional phrases since these constructions can result in ambiguity. Sentence-level constraints ensure that two parts of a conjoined sentence are of the same type, that relative clauses are always introduced by a relative pronoun, and that the use of ellipses is ruled out whenever possible. To guarantee that a source text is compliant with the rules of KANT Controlled English, an interactive checker is used that performs vocabulary and grammar checking. The checker parses each sentence in the source text and supports interactive disambiguation of lexical and structural ambiguities. If no analysis can be found, then the sentence must be rewritten. The KANT system was successfully used in the heavy equipment industry. In particular the combination of constraining the domain lexicon and the grammar

together with interactive disambiguation by the authors resulted in a dramatic reduction in the number of parses per sentence (from 27.0 to 1.04) and improved the resulting translations.

Instead of using a specialized MT system such as KANT, researchers have tried to identify those controlled language rules that have a high impact on the translation quality of commercial MT systems (O'Brien and Roturier 2007: 345–352; Aikawa *et al.* 2007: 1–7). Implementing only rules that have a high impact on the resulting translation is an interesting idea since authoring in a controlled language with a large rule set can be time-consuming. A comparative study of two commercial rule-based MT systems found that a small set of high-impact rules can considerably reduce the post-editing effort and improve the comprehensibility of the MT output (O'Brien and Roturier 2007: 345–352). Interestingly, these rules are relatively simple to apply; they govern misspelling, misuse of punctuation, long sentences (more than 25 words), and personal pronouns without an immediate antecedent. The hypothesis that a small set of rules can reduce the post-editing effort and improve MT quality has also been confirmed for statistical MT (Aikawa *et al.* 2007: 1–7). In this study, it was found that in particular style restrictions on lexical and phrasal items, correct spelling and capitalization had the greatest cross-linguistic effects on four typologically different languages (Dutch, Chinese, Arabic, French). While the outcome of this research is promising, it is important to note that these high-impact rules depend on the capabilities of the MT system, and that language-specific rules are equally important to achieve good results. For example, in a multilingual MT scenario prepositional attachment ambiguity is a special problem for Chinese but not so much for French because this form of ambiguity can usually be preserved between English and French but not between English and Chinese.

Controlled languages for semantic systems

Another group of controlled languages have been designed and used as general-purpose knowledge representation languages, interface languages to knowledge systems, in particular to the Semantic Web, and as specification languages for business rules (Schwitter 2010: 1113–1121). These controlled languages can often be translated unambiguously into a formal target language and then be used for automated reasoning. Since these controlled languages correspond closely to a formal target language, their design is driven by theoretical considerations that require a careful balance between the expressive power of the language and computational complexity (Pratt-Hartmann 2010: 43–73).

Controlled languages for knowledge representation

While there are many proposals for representing knowledge in the context of automated reasoning, by far the most dominant approach is first-order logic or one of its variants. Unlike English, the language of first-order logic is completely formal. This enables us to write precise and unambiguous specifications in this notation. However, formal notations are difficult to understand by domain specialists who often do not have training in formal logic. This makes it difficult for them to check if a formal specification fulfils the intended purpose in a specific application domain. There exist a number of general-purpose controlled languages that can serve as high-level specification languages (Schwitter 2010: 1113–1121). These controlled languages can be translated into a version of first-order logic and then be used for automated reasoning tasks including question answering.

Attempto Controlled English (ACE)

ACE is a controlled natural language that has been designed as a specification and knowledge representation language (Fuchs and Schwitter 1996; Fuchs *et al.* 2008). It covers a well-defined subset of English and allows users to specify their knowledge about an application domain in the form of a text. ACE texts are computer-processable and can be unambiguously translated into discourse representation structures (DRSs) (van Eijck and Kamp 2011: 181–252). These DRSs are a variant of first-order logic and serve as an interlingua that can be translated into various other formal notations for the purpose of automated reasoning.

ACE is defined by a small number of construction rules that specify admissible sentence structures, and a small number of interpretation rules that disambiguate constructs that might appear ambiguous in full English. Simple ACE sentences have the following form:

> subject + verb + [complements] + { adjuncts }

Complements depend on the verb and are required to complete a sentence, while adjuncts are optional modifiers of the verb. Composite ACE sentences can be built recursively from simpler ACE sentences through coordination, subordination, quantification, and negation.

The vocabulary of ACE consists of predefined function words (e.g., determiners, conjunctions, and pronouns), some predefined fixed phrases (e.g., *there is*, *it is false that*), and approximately 100,000 content words (nouns, proper names, verbs, adjectives, and adverbs). Users can import additional content word lexicons, prefix unknown words in a sentence by their word class or let the ACE parser guess the word class.

ACE supports language constructs such as:

- active and passive verbs (incl. modal verbs);
- strong negation (e.g., no, does not) and weak negation (e.g., it is not provable that);
- subject and object relative clauses;
- declarative, conditional, interrogative and imperative sentences; and
- various forms of anaphoric references to noun phrases (e.g., he, himself, the man, X).

To make it easier to write in ACE, authors can use a predictive text editor that can help to construct a text in controlled language.

The following example shows Lewis Carroll's *Grocer puzzle* in ACE:

> (1) Every honest and industrious person is healthy. (2) No grocer is healthy. (3) Every industrious grocer is honest. (4) Every cyclist is industrious. (5) Every unhealthy cyclist is dishonest. (6) No healthy person is unhealthy. (7) No honest person is dishonest. (8) Every grocer is a person. (9) Every cyclist is a person.

Given this text, the ACE reasoner RACE (Fuchs 2012) can prove that a conclusion such as:

> No grocer is a cyclist.

can be derived from the premises expressed in the text. The reasoner proves this conclusion and gives a justification for the proof in ACE. For our example, RACE finds the following minimal subset of premises that entail the conclusion:

1: Every honest and industrious person is healthy.
2: No grocer is healthy.
3: Every industrious grocer is honest.
4: Every cyclist is industrious.
8: Every grocer is a person.

Variations of this reasoning process allow for consistency checking in an ACE text, as well as question answering. Some proofs require domain-independent linguistic and mathematical knowledge that is expressed in the form of additional auxiliary axioms.

It is important to note that ACE texts are not decidable and therefore the search for a proof might not terminate. RACE controls undecidability by a time limit for a proof. However, there are decidable fragments of ACE. One of these fragments can be translated into the web ontology language OWL2, and covers almost all of OWL2 (apart from some aspects of data properties).

ACE has been used for several applications, including software and hardware specifications, agent control, legal and medical regulations, and ontology construction.

Processable English (PENG)

PENG is a controlled language that is similar to ACE but adopts a more light-weight approach in that it covers a smaller subset of English (White and Schwitter 2009). The language processors of ACE and PENG are both implemented in Prolog and based on grammars that are written in a definite clause grammar (DCG) notation. These DCGs are enhanced with feature structures and are specifically designed to translate declarative, conditional, and interrogative sentences into a first-order logic notation via a discourse representation structure (DRS).

In contrast to the original version of ACE that uses the DCG directly and resolves anaphoric references only after a DRS has been constructed, the language processor of PENG transforms the DCG into a format that can be processed by a top-down chart parser and resolves anaphoric references during the parsing process. PENG was the first controlled language to be supported by a predictive editor (Schwitter *et al.* 2003: 141–150). This editor provides text- and menu-based writing support and partially removes the burden of learning and remembering the constraints of the controlled language. The editor enforces the grammatical restrictions of the controlled language via look-ahead information while a text is written, and displays a paraphrase that clarifies the interpretation of each sentence. For each word form that the user enters, look-ahead information is generated by the chart parser which informs the user how the current input can be completed. These restrictions ensure that the text follows the rules of the controlled language so that it can be translated unambiguously into a DRS and then be further transformed in order to be processed by an automated reasoner.

PENG has been used as a high-level interface language to specify dynamic scenarios and the relevant background knowledge required to reason about direct and indirect effects of events as well as about continuous change (Schwitter 2011: 12–21). Recently, PENG has been used as an interface language to Answer Set Programming (ASP). Instead of writing a problem specification in ASP, the specification can be expressed directly in PENG and then translated automatically into an ASP program to compute stable models for question answering (Schwitter 2012: 26–43).

Computer Processable Language (CPL)

CPL is a controlled language for knowledge representation which has been developed at Boeing Research and Technology (Clark *et al.* 2005). In contrast to ACE and PENG where

all syntactic constructions have a default interpretation, CPL allows for ambiguous constructions but to a lesser extent than full natural languages. In CPL multiple interpretations of a sentence are possible, and the task of the language processor is to find the best parse and interpretation using additional external knowledge sources. The CPL parser relies on a preference mechanism to resolve attachment ambiguities. During parsing a simplified logical form is generated by rules that are parallel to the grammar rules. This logical form does not contain explicit quantifier scoping. Additional disambiguation decisions are performed during the generation of the logical form while other decisions are deferred and handled during the translation of the logical form into the underlying frame-based knowledge representation language KM (Clark and Porter 1999). Each CPL sentence is interpreted interactively, and new sentences are added incrementally to the knowledge system. The interpretation of the system is then displayed to the user in paraphrased English. Furthermore, the KM system uses an inference mechanism that allows for reasoning about actions and dynamic worlds.

CPL accepts three types of sentences: ground facts, questions, and rules. In the case of ground facts, a basic CPL sentence takes one of the following three forms:

There is|are NP
NP verb [NP] [PP]*
NP is|are passive-verb [by NP] [PP]*

The nouns in noun phrases can be modified by other nouns, prepositional phrases, and adjectives. The verbs can include auxiliaries and particles. CPL accepts five forms of questions; the two main ones are:

What is NP?
Is it true that Sentence?

In the case of rules, CPL accepts sentence patterns of the form:

IF Sentence [**AND** Sentence]* **THEN** Sentence [**AND** Sentence]*

Recently, CPL has been used in the AURA system (Chaudhri *et al.* 2009) that is part of the project Halo (Gunning *et al.* 2010: 33-58). This project is an effort to develop a reasoning system that enables domain specialists in a broad range of scientific disciplines to author knowledge bases in controlled language and allow a different group of users to ask novel questions against the given knowledge bases. As the following example illustrates, the question answering process may involve a short scenario (1) and a question (2) against this scenario:

1 A car accelerates from 12 m/s to 25 m/s in 6.0 s.
2 How far did it travel in this time?

In order to answer the question, the user first reformulates the scenario and the question in CPL. This results in our case in the following specification:

A car is driving.
The initial speed of the car is 12 m/s.
The final speed of the car is 25 m/s.
The duration of the drive is 6.0 s.
What is the distance of the drive?

If a CPL guideline is violated during this reformulation process, the AURA system responds with a notification of the problem and gives advice about how to rephrase the input. In addition to this advice, the user has access to a vocabulary list that contains all words that the system understands and a searchable database of good CPL examples. If the input is a valid CPL sentence, then the AURA system displays its interpretation in graphical form so that the interpretation can be validated by the user.

Controlled languages for the Semantic Web

Over the last few years, a number of CNLs have been proposed as interface languages to the Semantic Web, such as Attempto Controlled English (Kaljurand and Fuchs 2007), Lite Natural Language (Bernardi *et al.* 2007), Sydney OWL Syntax (Cregan *et al.* 2007), Rabbit (Hart *et al.* 2008: 348-360), and OWL Simplified English (Power 2012: 44–60). Most of these CNLs have been used with the support of predictive authoring tools in systems for authoring and verbalizing ontologies of the description logic based OWL family (Krötzsch *et al.* 2012).

Let us have a closer look at OWL Simplified English since this language has a number of interesting features that have been introduced to simplify the learning and use of the language at the expense of its expressiveness (Power 2012: 44–60). This adjustment is supported by an empirical study of a large ontology corpus of about 500,000 axioms. This study revealed interesting details about the information structure and the semantic complexity of these axioms (Power and Third 2010: 1006–1013). Some logical patterns occur with high frequency in the axioms of these ontologies while others are very rare. For example, 99.8 percent of terms that occur as the first argument of an axiom (subject position) are atomic and only 0.2 percent consist of complex subject terms. Furthermore, names of individuals, classes, and properties have distinctive features. This makes it possible to define formation rules that allow a parser to determine with high accuracy where an entity name begins and ends, and then classify them accordingly. As a consequence, the authoring tool of OWL Simplified English requires only the specification of verbs; all other words can be automatically classified as long as they follow the formation rules. Another interesting finding of this study was that complex OWL expressions are invariably right-branching; this allows for verbalizations that are structurally unambiguous and can be described efficiently by a finite-state grammar.

The linguistics patterns that are used for expressing common axiom and class constructors in OWL Simplified English are similar to other CNLs (Schwitter *et al.* 2008). However, OWL Simplified English considerably restricts the structure of complex sentences. Only three strategies are allowed for constructing complex sentences: noun-phrase lists (1), verb-phrase lists (2), and verb-phrase chains (3):

1 London is a city and a capital and a tourist attraction.
2 London is capital of the UK and has as population 15000000.
3 London is capital of a country that is governed by a man that lives in Downing Street.

These constructions are free of ambiguity and can be combined in a systematic way to form more complex sentences, for example (4):

4 London is a city that has as population 15000000 and is capital of a country that is governed by a man that lives in Downing Street.

Note that OWL Simplified English does not allow the use of *and/that* and *or* in the same sentence since this would result in ambiguous structures. Furthermore, only three forms of negation are allowed in this language, and these forms can only occur in predicates: negating a simple class (*is not a* Class), negating a simple restriction (*does not* Property *a* Class), and negating the second term of a simple intersection (*is a* Class1 *that is not a* Class2).

Controlled languages for business rules

Another interesting application domain for controlled languages is the domain of business rules. Business rules are statements in natural language that describe how a person or a machine can perform a specific action in an organization. The process of writing useful business rules is a difficult task since these business rules need to be understandable by humans and processable by machines. Business rules fall into two major groups: (a) operative business rules that specify how things must be done in an organization, and (b) structural business rules that define how things must be understood in an organization. In contrast to operative business rules, structural business rules cannot be violated, but they can be ill-formed or inappropriate. Because of these characteristics, it is important that business rules are written in a precise and unambiguous manner so that they are easy to validate for business people and easy to verify automatically for consistency.

The Semantics of Business Vocabulary and Business Rules (SBVR) is a standard that allows business people to define business rules in a clear and unambiguous way so that these rules are translatable into other representations (OMG 2008). SBVR specifies a meta-model for describing the meaning of business vocabularies, facts and rules. The core idea of the SBVR approach is that rules build on facts, and facts build on concepts that are expressed by terms. SBVR's logic is founded upon typed first-order logic with some restricted extensions into modal logic and higher-order logic. SBVR does not standardize a particular surface notation, but SBVR meta-models can be rendered in graphical form, textual form or a combination of both.

SBVR Structured English is a controlled language that has a particular mapping to SBVR structures of meaning. SBVR uses deontic modalities (e.g., it is obligatory that …, it is permitted that …) for expressing operative business rules and alethic modalities (e.g., it is necessary that …, it is possible that …) for structural business rules.

The specification of a business rule in SBVR Structural English usually takes a vocabulary entry of a fact type (= verb concept) as a starting point, for example:

branch owns rental car

This is a binary fact type that uses two designations (*branch* and *rental car*) for the noun concepts and one designation (*owns*) for the verb concept. Vocabulary entries of fact types usually use singular, active forms of verbs; other forms of verbs are implicitly usable in business rules. For the specification of an operative business rule, a suitable deontic modal operator is selected and added to the representation of the fact type. Additionally, the grammatical voice is fixed, for example:

It is obligatory that rental car is owned by branch.

In the next step quantifiers are added to the designations of the noun concepts:

It is obligatory that each rental car is owned by exactly one branch.

The specification of a structural business rule works in a similar way but uses an alethic modal operator, for example:

It is necessary that each rental has exactly one requested car group.

This business rule is based on the following supporting fact type:

rental has requested car group

Fact types in SBVR cannot have properties, but they can be turned into an object by giving them a name. The name can then be used in other fact types. This process is called objectification and can be used to identify a state or an event, and relate this state of affairs to a time point, or to another state or event. For example, the designation *car assignment* in

It is obligatory that each car assignment of a rental occurs before the pick-up date of the rental.

represents the objectification of the following fact type:

car is assigned to rental

By additionally defining this objectification and using the following fact types:

car assignment is a state of affairs
state of affairs occurs before date/time

the state of affairs (*car assignment*) can be related to a time point (*pick-up date*) with respect to time (e.g., occurring before or after that time point).

The SBVR business rules introduced so far prefix a statement with key phrases which convey the intended modality. SBVR Structured English uses an alternative style to communicate the modality. This alternative embeds a keyword (in front of verbs) within rule statements, for example:

Each rental car must be owned by exactly one branch.
Each rental always has exactly one requested car group.

This embedded keyword style is the preferred style for expressing modalities in RuleSpeak® (Ross 2003), an existing business rule notation that builds on SBVR and has been used with business people in large-scale projects.

Despite the existence of formally grounded notations, SBVR lacks a logical formalization which would allow a reasoning tool to check automatically the consistency of a set of business rules. This is because SBVR uses very expressive constructs for which it is known that a sound and complete reasoner cannot be constructed. Research into logic-based reasoning support for subsets of SBVR using a specific first-order deontic-alethic logic has started only recently (Solomakhin *et al.* 2011: 311–325).

Conclusion

As we have seen, there exist a number of application areas that can benefit from restricting the expressivity of natural language in a systematic way in order to improve communication between humans; comprehensibility and processing of documents or interaction between humans and machines. Controlled natural languages achieve these improvements by carefully restricting the size of the grammar and vocabulary for a specific application area with the aim of reducing or eliminating ambiguity and complexity of full natural language. We have identified four main application areas in this article: controlled languages for human communication, controlled languages for technical documentation, controlled languages for machine translation, and controlled languages for semantic systems. Each of these areas has specific requirements for the design of a controlled language, and even within an area there often exist competing formally equivalent linguistic constructions, and it is not always clear which one works best. Without doubt there is a lot of room for comparative evaluations in this domain to determine which constructions work best for a particular user group.

Nevertheless, controlled languages have been used successfully in many industries over the last 20 years as a method to improve the readability of technical documents or to make these documents easier to translate into the language of their customers. Authoring support is absolutely essential for the production of texts and documents in controlled language since human writers need to be able to judge whether a sentence or a paragraph complies with the rules of the controlled language and whether an expression belongs to the approved vocabulary or not. In the future, we will see more companies using controlled languages for document production and producing their own controlled language standards. We expect to see many more semantic systems hereafter that will use controlled languages as high-level interface languages that allow humans and machines to communicate in a truly cooperative way without the need to formally encode the relevant knowledge.

Note

1 http://simple.wikipedia.org/wiki/Main_Page.

References

Aikawa, Takako, Lee Schwartz, Ronit King, Monica Corston-Oliver, and Carmen Lozano (2007) 'Impact of Controlled Language on Translation Quality and Post-editing in a Statistical Machine Translation Environment', in *Proceedings of MT Summit XI*, 10–14 September 2007, Copenhagen, Denmark, 1–7.

ASD (2010) *ASD Simplified Technical English*, Specification ASD-STE100, International Specification for the Preparation of Maintenance Documentation in a Controlled Language, Issue 5, April 2010.

Bernardi, Raffaella, Diego Calvanese, and Camilo Thorne (2007) 'Lite Natural Language', in *Proceedings of the 7th International Workshop on Computational Semantics (IWCS-7)*, 10–12 January 2007, Tilburg University, the Netherlands, 1–12.

Chaudhri, Vinay K., Peter E. Clark, Sunil Mishra, John Pacheco, Aaron Spaulding, and Jing Tien (2009) 'AURA: Capturing Knowledge and Answering Questions on Science Textbooks', *Technical Report*, SRI International.

Clark, Peter and Bruce Porter (1999) 'KM – The Knowledge Machine 2.0: Users Manual', *Technical Report*, AI Lab, University of Texas at Austin.

Clark, Peter, Phil Harrison, Tom Jenkins, John Thompson, and Rick Wojcik (2005) 'Acquiring and Using World Knowledge Using a Restricted Subset of English', in *Proceedings of the 18th International Florida Artificial Intelligence Research Society Conference (FLAIRS '05)*, 15–17 May 2005, Clearwater Beach, FL, 506–511.

Cregan, Anne, Rolf Schwitter, and Thomas Meyer (2007) 'Sydney OWL Syntax – Towards a Controlled Natural Language Syntax for OWL 1.1', in Christine Golbreich, Aditya Kalyanpur, and Bijan Parsia

(eds) *Proceedings of OWL: Experiences and Directions*, CEUR-WS, 6–7 June 2007, Innsbruck, Austria, 1–10.

Fuchs, Norbert E. and Rolf Schwitter (1996) 'Attempto Controlled English (ACE)', in *Proceedings of CLAW 96*, March 1996, University of Leuven, Belgium, 124–136.

Fuchs, Norbert E., Kaarel Kaljurand, and Tobias Kuhn (2008) 'Attempto Controlled English for Knowledge Representation', in Cristina Baroglio, Piero A. Bonatti, Jan Maluszynski, Massimo Marchiori, Axel Polleres, and Sebastian Schaffert (eds) *Reasoning Web, Fourth International Summer School 2008*, 7–11 September 2008, Venice, Italy/Berlin: Springer Verlag, 104–124.

Fuchs, Nobert E. (2012) 'First-order Reasoning for Attempto Controlled English', in *Proceedings of CNL 2010*, 13–15 September 2010, Marettimo Island, Sicily, Italy, 73–94.

Gunning, David, Vinay K. Chaudhri, Peter Clark, Ken Barker, Shaw-Yi Chaw, Mark Greaves, Benjamin Grosof, Alice Leung, David McDonald, Sunil Mishra, John Pacheco, Bruce Porter, Aaron Spaulding, Dan Tecuci, and Jing Tien (2010) 'Project Halo Update – Progress Toward Digital Aristotle', *AI Magazine* 31(3): 33–58.

Hart, Glen, Martina Johnson, and Catherine Dolbear (2008) 'Rabbit: Developing a Control Natural Language for Authoring Ontologies', in Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis (eds) *The Semantic Web: Research and Applications: 5th European Semantic Web Conference ESWC 2008*, 1–5 June 2008, Tenerife, Canary Islands, Spain, 348–360.

Hutchins, W. John (2005) 'Current Commercial Machine Translation Systems and Computer-based Translation Tools: System Types and Their Uses', *International Journal of Translation* 17(1–2): 5–38.

Kaljurand, Kaarel and Norbert E. Fuchs (2007) 'Verbalizing OWL in Attempto Controlled English', in *Proceedings of OWL: Experiences and Directions*, 6–7 June 2007, Innsbruck, Austria.

Krötzsch, Markus, Frantisek Simančik, and Ian Horrocks (2012) 'A Description Logic Primer', *CoRR* 19 January 2012.

Mitamura, Teruko (1999) 'Controlled Language for Multilingual Machine Translation', in *Proceedings of MT Summit VII*, 13–17 September 1999, Kent Ridge Digital Labs, Singapore, 46–52.

Object Management Group (OMG) (2008*) Semantics of Business Vocabulary and Business Rules (SBVR)*, v1.0., Specification, OMG (January).

O'Brien, Sharon and Johann Roturier (2007) 'How Portable Are Controlled Languages Rules? A Comparison of Two Empirical MT Studies', in *Proceedings of MT Summit XI*, 10–14 September 2007, Copenhagen, Denmark, 345–352.

Ogden, C.K. (1930) *Basic English: A General Introduction with Rules and Grammar*, London: Paul Treber and Co., Ltd.

Ogden, C.K. (1968) *Basic English: International Second Language*, New York: Harcourt, Brace and World.

Power, Richard (2012) 'OWL Simplified English: A Finite-state Language for Ontology Editing', in Tobias Kuhn and Norbert E. Fuchs (eds) *Controlled Natural Language: The 3rd International Workshop, CNL 2012*, 29–31 August 2012, Zurich, Switzerland/Berlin: Springer Verlag, 44–60.

Power, Richard and Allan Third (2010) 'Expressing OWL Axioms by English Sentences: Dubious in Theory, Feasible in Practice', in Chu-Ren Huang and Dan Jurafsky (eds) *Proceedings of the 23rd International Conference on Computational Linguistics*, 23–27 August 2010, Beijing International Convention Center, Beijing, China, 1006–1013.

Pratt-Hartmann, Ian (2010) 'Computational Complexity in Natural Language', in Alex Clark, Chris Fox, and Shalom Lappin (eds) *The Handbook of Computational Linguistics and Natural Language Processing*, Chichester, West Sussex: Wiley-Blackwell, 43–73.

Ross, Ronald G. (2003) *Principles of the Business Rule Approach*, Boston, MA: Addison-Wesley Professional.

Schwitter, Rolf, Anna Ljungberg, and David Hood (2003) 'ECOLE – A Look-ahead Editor for a Controlled Language', in *Proceedings of EAMT-CLAW03*, 15–17 May 2003, Dublin City University, Dublin, Ireland, 141–150.

Schwitter, Rolf (2011) 'Specifying Events and Their Effects in Controlled Natural Language', in Normaziah A. Aziz, Koiti Hasida, A.Wahab, A. Rahman, and Hiroaki Saito (eds) *Computational Linguistics and Related Fields, Procedia – Social and Behavioral Sciences*, Elsevier, 27: 12–21.

Schwitter, Rolf (2012) 'Answer Set Programming via Controlled Natural Language Processing', in Tobias Kuhn and Norbert E. Fuchs (eds) *Proceedings of the 3rd Workshop on Controlled Natural Language (CNL 2012)*, 29–31 August 2012, Zurich, Switzerland, Springer, 26–43.

Schwitter, Rolf, Anna Ljungberg, and David Hood (2003) 'ECOLE – A Look-ahead Editor for a Controlled Language', in *Proceedings of EAMT-CLAW03*, 15–17 May 2003, Dublin City University, Dublin, Ireland, 141–150.

Schwitter, Rolf, Kaarel Kaljurand, Anne Cregan, Catherine Dolbear, and Glen Hart (2008) 'A Comparison of Three Controlled Natural Languages for OWL 1.1.', in Kendall Clark and Peter F. Patel-Schneider (eds) *Proceedings of the 4th OWLED Workshop on OWL: Experiences and Directions*, 1–2 April 2008, Washington, DC.

Solomakhin, Dmitry, Enrico Franconi, and Alessandro Mosca (2011) 'Logic-based Reasoning Support for SBVR', in *Proceedings of the 26th Italian Conference on Computational Logics (CILC-2011)*, 31 August – 2 September 2011, Pescara, Italy, 311–325.

van Eijck, Jan and Hans Kamp (2011) 'Discourse Representation in Context', in Johan van Benthem and Alice ter Meulen (eds) *Handbook of Logic and Language*, 2nd edition, London/Amsterdam/ New York: Elsevier B.V., 181–252.

White, Colin and Rolf Schwitter (2009) 'An Update on PENG Light', in Luiz Pizzato and Rolf Schwitter (eds) *Proceedings of the Australasian Language Technology Association (ALTA 2009)*, 3–4 December 2009, Sydney, Australia, 80–88.

29
CORPUS

Li Lan

HONG KONG POLYTECHNIC UNIVERSITY, HONG KONG, CHINA

Introduction

The word *corpus* (plural *corpora*) originally came from Latin. According to the *Oxford English Dictionary* its sense of 'body of a person' started in the mid-fifteenth century and the sense of 'collection of facts or things' occurred later in 1727. The year 1956 saw an extension of the meaning to include 'the body of written or spoken material upon which a linguistic analysis is based'. A large number of index cards used by early dictionary compilers were in fact human-readable language corpora. As Leech (1992) observed, corpora of text collection had been used by linguists and grammarians for the study of language long before the invention of the computer; therefore he suggests that 'computer corpus linguistics' would be a more appropriate term for studies based on language database today. The corpus in linguistics is a large collection of machine-readable texts compiled with a specific purpose that can be retrieved with particular computer software for linguistic research.

Corpus-based translation study (CTS) is defined as the use of corpus linguistic technologies to inform and elucidate the translation process (Baker 1995: 223–243). In tandem with rapid developments in computational power and availability of electronic texts the corpus approach has become a truly empirical approach to language and translation studies (Granger 2003: 17–28). Applications of the corpus approach can bridge professional human translation to machine translation, and can enable descriptive linguistic and translation research in language teaching and translator training.

Development and typology of corpora

The landmark of modern corpora is generally attributed to the Brown Corpus of Standard American English. It consists of 500 text samples (2,000 words each) distributed in 15 categories forming a one-million-word selection of American English from a wide variety of sources. The corpus was compiled in the 1960s and was used in the analysis of linguistics, language teaching, psychology, statistics, and sociology (Kucera and Francis 1967), particularly as a foundation for the famous Survey of English Usage by Quirk *et al.* in the 1980s. One of its outcomes, *A Comprehensive Grammar of the English Language* (Quirk *et al.* 1985) is regarded as one of the most important English grammar books in the English language. In addition, the

Brown Corpus also provided support to the 1969 edition of *American Heritage Dictionary*. The AHD took the innovative step of combining prescriptive elements (how language *should* be used) with descriptive information (how it is actually used).

Corpus linguistics in North America, after the Brown Corpus project, seemed to enter a rather dormant phase throughout the 1980s and 1990s until fairly recently when a number of freely available online mega-corpora were introduced to the public domain. These mega-corpora include the 400-million-word *Corpus of Contemporary American English* (COCA), 200-million-word *Time Corpus,* COHA, Google Books and the latest release, 1.9-billion-word *GloWbE*, all by Mark Davis at Brigham Young University. These free mega-corpora have percolated different types of corpus research. As the compiler predicts, COHA, TIME, COCA and Google Books can be used for historical or diachronic studies of the English language; COCA and BYU-BNC are for genre studies and GloWbE, which consists of sub-corpora of English used in 20 countries, will contribute to the exploration of a variety of Englishes in the world.

The first corpus of British English was London-Lund, built in 1965, which also contributed to the Survey of English and *A Comprehensive Grammar of the English Language*. The late 1980s and 1990s saw corpus linguistics flourishing in Great Britain. A number of famous linguists entered the area of corpus linguistics and made a great contribution in terms of developing its theoretical premise and methodological application: John Sinclair and Geoffrey Leech making particularly notable contributions to the field. Under the leadership of Sinclair the *Bank of English*, or the COBUILD corpus, has served a large number of dictionaries, grammar books and ELT teaching materials. The project started in 1991 and has been growing, reaching a total of 650 million running words in 2012. The corpus is held both at HarperCollins Publishers and the University of Birmingham and is open only to paid academic institutions in Europe. Another influential standard corpus is the *British National Corpus* (BNC), a collection of standard British English. The compilation lasted from 1991 to1994. It is a balanced corpus with 100 million words of both spoken and written data with part of speech (POS) tagging. Since its publication, the BNC has been used as a reference corpus for many linguistic studies including general English versus specialized English, standard English versus a variety of English, native English versus non-native or learner English.

Although these monolingual English corpora are not specially for translation, they can be used in translation training, to strengthen students' knowledge of target language patterns and improve the quality of translation (Bowker 1999: 11–24; Kenny 2001), to help with terminology extraction (Pearson 1996: 85–95) and 'to allow patterns observed in a source or target text to be set off against what is known about the language in general' (Kenny 2001: 58).

The development of English corpora, together with the fast development of computer technology, has inspired corpora of different languages in many parts of the world. Up to now, more than 30 languages have built their own corpora, big or small, general or specific. Although a huge amount of data is available online today, it is important to notice the difference between corpora and archives of electronic texts. As observed by Granger, 'building a corpus requires not only a large quantity of data but also an information retrieval operation, in order to locate relevant and reliable documents, while an archive is only a repertory of electronic texts' (Granger 2003: 18).

Table 29.1 lists some influential monolingual non-English corpora from recent publications. Their websites can be easily googled on the internet. The application of these data in translation will be discussed later in this chapter.

Table 29.1 Monolingual non-English corpora

Language	*Title*	*Year of compilation*	*Size*	*Host institution*	*Website*
European languages					
Swedish	The Swedish Treebank	Not known	1.55 million?	Uppsala University; Växjö University	http://stp.lingfil.uu.se/~nivre/swedish_treebank/
Danish	Korpus 2000	Not known	28 million	Society for Danish Language and Literature	http://ordnet.dk/korpusdk_en
Spanish	Corpus de Referencia del Español Actual (CREA)	2000–2004?	3.5 million	REAL ACADEMIA ESPAÑOLA	http://corpus.rae.es/creanet.html
Dutch	The INL corpus	2002–2004	70 million	The Institute for Dutch Lexicology	http://www.inl.nl/pagina-niet-gevonden
French	French Treebank	2005–2007?	Not known	Laboratoire de Linguistique Formelle	http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php
German	Deutsches Referenzkorpus (DEREKO)	1999–2002	200 million	The Institut für deutsche Sprache (IDS) in Mannheim, the Seminar für Sprachwissenschaft (SfS) in Tübingen, and the Institut für Maschinelle Sprachverarbeitung (IMS) in Stuttgart	http://www.sfs.uni-tuebingen.de/dereko/
Scottish	The Scottish Corpus of Texts and Speech (SCOTS)	2004	4 million	The School of Critical Studies at Glasgow University	http://www.scottishcorpus.ac.uk/corpus/search/
Welsh	Cronfa Electroneg o Gymraeg	2001?	1 million	University of Wales, Bangor	http://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en
Irish	TOBAR NA GAEDHILGE	1995–2012?	3.5 million	The University of the Highlands and Islands	http://www.smo.uhi.ac.uk/~oduibhin/tobar/index.htm#history
Italian	Corpus of Italian Newspapers	1993–2010	500,000	Not known	http://ota.ahds.ac.uk/headers/1723.xml
Greek	The Corpus of Greek Texts (CGT)	2006?	Not known	The Universities of Athens and Cyprus	http://sek.edu.gr/index.php?en
Portuguese	The CETEMPúblico Corpus	2000	180 million	The Portuguese Ministry for Science and Technology (MCT)	http://www.linguateca.pt/cetempublico/
Czech	The Prague Dependency Treebank	1995?	2 million	The Institute of Formal and Applied Linguistics (ÚFAL)	http://ufal.mff.cuni.cz/pdt2.0/
Croatian	Croatian National Corpus	Not known	101.3 million	University of Zagreb	http://www.hnk.ffzg.hr/default_en.htm
Russian	BOKR (The Russian Reference Corpus)	2002	100 million	Leeds University?	http://bokrcorpora.narod.ru/index-en.html

Table 29.1 (continued)

Language	*Title*	*Year of compilation*	*Size*	*Host institution*	*Website*
Serbian	Corpus Of Serbian Language (CSL)	1996?	11 million	Institute for Experimental Phonetics and Speech Pathology, Belgrade; Laboratory for Experimental Psychology, University of Belgrade	http://www.serbian-corpus.edu.rs/ns/eindex.htm
Polish	The National Corpus of Polish	2008–2012	20 million	the Polish Ministry of Science and Higher Education	http://nkjp.pl/
Turkish	METU Turkish Corpus	Not known	2 million	Middle East Technical University	http://ii.metu.edu.tr/corpus
Asian languages					
Hebrew	Wiki-Segmentation Hebrew Corpus	Not known	523,599 words	Ben-Gurion University of the Negev	http://www.cs.bgu.ac.il/~nlpproj/wiki-seg-corpus/
Arabic	Buckwalter Arabic Corpus	1986–2003	2.5–3 million	Tim Buckwalter	http://www.qamus.org/wordlist.htm
Chinese	CLL (simplified Chinese)	2006?	477 million	Peking University	http://ccl.pku.edu.cn:8080/ccl_corpus/
	Spoken Chinese Corpus of Situated Discourse	2003?	Not known	The Chinese Academy of Social Science	http://ling.cass.cn/dangdai/corpus.htm
	Academia Sinica (traditional Chinese)	1996	5 million	Academia Sinica	http://app.sinica.edu.tw/kiwi/mkiwi/index.html
Japanese	Balanced Corpus of Contemporary Written Japanese	1999–2003	100 million	The National Institute for Japanese Language (NIJLA), the Communications Research Laboratory (CRL), and the Tokyo Institute of Technology (TITech)	http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/#3
Korean	Korean National Corpus	1998	57 million (as of 2002)	The Ministry of Culture and Tourism	http://www.sejong.or.kr/gopage.php?svc=intro.eintro
Malay	Malay Concordance Project	1991	5.8 million	Australian National University	http://mcp.anu.edu.au/
Mongolian	The Multi-dialectal speech corpus of Mongolia (MDSCM)	1998–2006	27 hours of speech	Waseda University and ATR of Japan	http://universal.elra.info/product_info.php?cPath=37_39&products_id=2222 http://www.isca-speech.org/archive_open/archive_papers/iscslp2006/B74.pdf

Table 29.1 (continued)

Language	*Title*	*Year of compilation*	*Size*	*Host institution*	*Website*
Thai	Thai National Corpus	Not known	80 million	Chulalongkorn university	http://ling.arts.chula.ac.th/TNC/category.php?id=32&lang=eng
South Asian languages	EMILLE (Enabling Minority Language Engineering)	2003?	97 million	Lancaster University	http://www.emille.lancs.ac.uk/
Nepali	Nepali Grammar Project	2004–2006?	Not known	Lancaster University	http://www.lancs.ac.uk/staff/hardiea/nepali/index.php
Bengali	SHRUTI Bengali Continuous ASR Speech Corpus	Not known	22,012	Society for Natural Language Technology Research	http://cse.iitkgp.ac.in/~pabitra/shruti_corpus.html
African languages					
Swahili	The Helsinki Corpus of Swahili (HCS)	Not known	12.5 million	The University of Helsinki	http://www.csc.fi/english/research/software/hcs
Zulu	Ukwabelana:	1995–2013	30,000 sentences	University of Bristol	http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/resources.jsp#corpus
Amharic	Amharic News Corpus	Not known	210,000 words	European Language Resources Association (ELRA)	http://aflat.org/content/tagging-and-verifying-amharic-news-corpus

Compared to the English mega-corpora, these linguistic databases may not be as standard and representative, but they have been used for national and international linguistic studies. It is obvious that the development of corpus study across languages is not balanced; 'less widespread language may not have any corpus resources at all or access to them may be severely limited' (Granger 2003: 22).

The role of corpus in language study

The role of corpus in language study offers new perspectives, allowing us 'to see phenomena that previously remained obscure because of the limitation of our vantage points' (Kenny 2001: xiii). John Sinclair (2003) believes that natural language use constitutes the best source of linguistic evidence. Such use can only be found in authentic communicative texts. He claims one of the main aims of creating the corpus *Bank of English* was to retrieve evidence in support of the learning of the English language (Sinclair 1991), and led the COBUILD team to compile one of earliest learner's dictionaries drawing on the data from the *Bank of English*.

Wallis and Nelson (2001) propose 3A perspectives for data processing:

> *Annotation* consists of the application of a scheme to texts. Annotations may include structural markup, part-of-speech (POS) tagging, parsing, and numerous other representations.

Abstraction consists of the translation (mapping) of terms in the scheme to terms in a theoretically motivated model or dataset. Abstraction typically includes linguist-directed search but may include e.g. rule-learning for parsers.

Analysis consists of statistically probing, manipulating and generalizing from the dataset. Analysis might include statistical evaluations, optimization of rule-bases or knowledge discovery methods.

Corpus approaches to language studies can be corpus-driven or corpus-based (Biber *et al.* 1998; Tognini-Bonelli 2001). A corpus-based approach is a top-down methodology 'that uses corpus evidence mainly as a repository of examples to expound, test or exemplify given theoretical statements' (Tognini-Bonelli 2001: 10). A corpus-driven approach involves a bottom-up methodology, beginning by selecting random examples from the corpus, identifying their shared and individual features, and then grouping them for different purposes. Researchers observe language facts from corpus data, formulate a hypothesis to account for these facts, make a generalization based on corpus evidence of the repeated patterns and then unify these observations in a theoretical statement (ibid.: 14–18). Given that translation studies are after all a linguistic study, both corpus-driven and corpus-based approaches can provide linguistic and cultural evidence and improve the quality of translation.

In language study as well as in translation study, dictionaries and corpora are indispensable tools. The difference between the two lies in the way they are used. Dictionaries provide word meanings or target language equivalents directly. Corpora, or rather the concordance lines from the data, require translators' judgement to choose proper information to meet the needs of translation.

Cross-linguistic corpora for translation

Cross-linguistic corpora are becoming increasingly available for a large number of languages and have been used for theoretical generalizations in a range of linguistic disciplines – from typology (van der Auwera *et al.* 2005: 201–217; Cysouw and Wälchli 2007: 95–99) to contrastive linguistics (Granger 2010: 14–21), to functional and cognitive linguistics (Croft 2010: 1–11) and dialectometry (Grieve *et al.* 2011: 193–221). However, the most important use of translation corpora is for translation (Baker 1993: 233–250 and 1995: 223–243). After two decades it is quite common now for translation researchers to use corpora to verify, refine or clarify theories that had little or no empirical support and to achieve a higher degree of descriptive adequacy. However, as in many new scientific fields, the terms of cross-lingual corpus have not received a general consensus which may lead to some confusion.

The field of translation turned to corpus when corpus linguistics began to thrive in the 1990s. Mona Baker initialized corpus-based translation studies (Baker 1993: 233–250) and started building the Translational English Corpus (TEC) for studying translated English at the University of Manchester (see Baker 1999: 281–298). The TEC corpus consists of written texts translated by native speakers of English in four genres: fiction, biography, newspaper articles and in-flight magazines. The collection constantly expanded with fresh materials from a range of source languages and reached a total of 10 million words in 2003. The TEC corpus is freely available on the internet and has stimulated a number of publications on translation patterning of translated text and non-translated text in the same language, and stylistic variation across individual translators.

As a comparatively new area, corpus-based translation study has not secured a chain of consistent terminology. Granger (2003: 17–28) regards the confusion of the terms as being caused by two different linguistic branches: translation studies (TS) and contrastive linguistics (CL). Translation researchers use the terms *translation corpus*, *parallel corpus* and *comparable corpus* to refer to various types of cross-linguistic texts. The terms are used interchangeably and can be confusing. Contrastive linguists, according to Johansson and Hasselgård (1999: 145–162), have different definitions:

1 *Translation corpora:* consisting of original texts in one language and their translations into one or more other languages.
2 *Comparable corpora:* consisting of original texts in two or more languages, matched by criteria such as the time of composition, text category, intended audience, etc.

The term *translation* (or *translational*) *corpus* refers to the corpus of translated texts (Baker 1999: 281–298), or bitexts in computer scientists' terms (Resnik and Smith 2003: 349). Translational data conveys the same semantic content therefore as an ideal resource for establishing equivalence, terminology and phraseology between languages. Translation corpora may bear many extra-linguistic features such as the translator's status or the direction of the translation process. The main drawback of translation corpora, however, is that they can hardly have balanced genres. Translation pairs of copyright free texts can be obtained from international organizations such as the UN and the EU. Bilingual documents are also common in bilingual societies such as Hong Kong and Francophone regions of Canada. There are translations of some older masterpieces of literature, film transcripts, standard company letters, most of which do not have machine readable source and target texts. Our experience with data collection has shown that collecting bilingual letters and email messages is virtually impossible; internal and external communications are not usually translated (Li and Bilbow 2001: 210). Translated news reports are also rare as news reporters mostly write in their own language even though thematically paralleled stories on the same event can be available in two or more languages. In addition, there are a few bi-direction translation corpora because the majority of translations are in one direction: from English to another language.

Comparable corpora, according to Granger (2003: 17–28), are not translational. They represent original texts in different languages. Produced by native speakers of the languages under comparison, the texts are in principle free from the influence of other linguistic systems. In the case of translation corpora, the original source text is in a different language and will naturally impose some influence over the translated text. The main drawback of comparable corpora lies in the difficulty of establishing comparability of texts. Johansson and Hasselgård (1999: 145–162) mentioned time, categories and audience in comparable corpora, but seemed to have overlooked an important item: theme. When compiling trilingual business corpora at Hong Kong Polytechnic University, we defined the three sub-corpora as thematically parallel, because the texts were collected at the same time, in the same genre and had similar topics, although they were not translated texts. In an early article, Baker used *comparable corpus* to mean translation corpus: 'the term *comparable corpus* is used to refer to two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages' (Baker 1995: 234). More recently, comparable and translation have been clearly distinguished by researchers. In short comparable corpora are thematically parallel non-translational corpora.

Table 29.2 Translation corpora

Title	*Language pair*	*Year of compilation*	*Host institution*	*URL*	*Size*	*Annotation*
Parallel Corpus Pro: The Bible	English and 14 other languages	1999	University of Maryland	http://www.umiacs.umd.edu/~resnik/parallel/bible.html	Not known	√
United Nations General Assembly Resolutions: A Six-Language Parallel Corpus		2009?	United Nations	http://www.uncorpora.org/	About 3 million per language (6 languages)	Not known
PELCRA Parallel Corpora	English–Polish	1997–2013?	University of Łódź Lancaster University	http://pelcra.pl/res/parallel/	225 million	Not known
MultiUN: Multilingual UN Parallel Text 2000–2009		2011	Language Technology Lab in DFKI GmbH (LT-DFKI), Germany	http://www.euromatrixplus.net/multi-un/	Not known	Not known
Hunglish Corpus Version 2.0	English–Hungarian	2005–2013?	The Budapest University of Technology and Economics; the Hungarian Academy of Sciences Institute of Linguistics	http://mokk.bme.hu/resources/hunglishcorpus/	120 million	Not known
The Kyoto Free Translation Task (KFTT)	English–Japanese	2011–2012	Nara Institute of Science and Technology (NAIST); Kyoto University	http://www.phontron.com/kftt/	Not known	√
IDENTIC	Indonesian–English	2011–2012	Charles University in Prague	http://ufal.mff.cuni.cz/~larasati/identic/#Introduction	Not known	√
Academia Sinica Balanced Corpus of Modern Chinese	English–Chinese	1991–1997	Academia Sinica, Taiwan	http://app.sinica.edu.tw/kiwi/mkiwi/	5 million	√
Babel Chinese–English Parallel Corpus	Chinese–English	2001–2004	Institute of Computational Linguistics, Peking University	http://www.icl.pku.edu.cn/icl_groups/parallel/default.htm	10 million Chinese characters	√

Table 29.2 (continued)

Title	*Language pair*	*Year of compilation*	*Host institution*	*URL*	*Size*	*Annotation*
Urdu–Nepali–English Parallel Corpus	Urdu–Nepali–English	2009	Center for Research in Urdu Language Processing (CRULP)	http://www.crulp.org/software/ling_resources/urdunepalienglishparallelcorpus.htm	100,000 words	√, POS
CLUVI Parallel Corpus	English–Galician	2012	Universida de Vigo. Grupo de investigación TALG	http://repositori.upf.edu/handle/10230/20051	23 million	Not known
An English–Inuktitut Parallel Corpus	English–Inuktitut	1999–2002	Institute for information technology, Canada	http://www.inuktitutcomputing.ca/NunavutHansard/en/	3 million English words; 1.5 million Inuktitut words	Not known
Dutch Parallel Corpus: a Balanced Parallel Corpus for Dutch–English and Dutch–French	Dutch–English; Dutch–French	2013	Hogeschool Gent	http://lt3.hogent.be/en/publications/dutch-parallel-corpus-a-balanced-parallel-corpus-for-dutch-e/	10 million	√
ELRA Multilingual and Parallel Corpora	English and 9 European languages	2008?	European Language Resources Association	http://catalog.elra.info/product_info.php?products_id=764	Not known	Not known
CzEng 1.0 (Czech–English Parallel Corpus, version 1.0)	Czech–English	2012?	Institute of Formal and Applied Linguistics (ÚFAL)	http://ufal.mff.cuni.cz/czeng/czeng10/	233 million English and 206 million Czech tokens	√

The term *parallel corpus* has been used to refer to different types of cross-linguistic corpora and seems to be the most confusing. Hartmann refers it to as a translation corpus (Hartmann 1980: 37). Aijmer (2008) calls parallel corpus 'comparable corpora'. The ParaConc software by Barlow (1999: 319–327) names translated texts as parallel corpora and has influenced many of its users; the users have to align the translated text at sentence level making it parallel before using the software. However, from the growing literature on corpus-based translation study, *parallel corpora* are also understood in the broadest possible sense as any collection of texts in different languages and language varieties conveying similar information produced under similar pragmatic conditions. They can include translation corpora, balanced samples of the same genres from different languages, as well as texts produced by different speakers of one language, therefore *parallel corpora* is an umbrella term for cross-lingual corpora.

To sum up, the term *translation corpora* has been explicitly defined, but *parallel* and *comparable* corpora may refer to any type of multilingual corpora, translational or non-translational. Granger (2003) attributes the difference to the two cross-linguistic approaches: comparative linguistics (CL) and translation studies (TS). Apart from the terminological difference she thinks

there is a more fundamental discrepancy. In the TS framework, translated texts are considered as texts in their own right, which are analysed in order to 'understand what translation is and how it works' (Baker 1993: 243). In the CL framework they are often presented as unreliable as the cross-linguistic similarities and differences that they help establish may be 'distorted' by the translation process, i.e. may be the result of interference from the source texts (Granger 2003: 34). Despite the discrepancies, recent years have witnessed linguists using different types of parallel corpora and employing new methodological approaches to cross-linguistic studies quantitatively and qualitatively. Corpus translation studies have been enriched with more empirical methods after a long reign of generative approaches to lexico-grammar and largely intuitive judgement on grammaticality and lexicality (Gries and Wulff 2012: 35).

Quantitative and qualitative research in corpus-based translation study

The application of the corpus approach has enabled the interplay of quantitative and qualitative methodologies used in translation studies, based on the concept that 'linguistic system as well as idiolectal uses is commensurable to a degree' (Lakoff 1987). Many different quantitative and qualitative methods have been used in translation studies, but they are largely tentative with limited construction and testing methods for theoretic models of translation, which in turn has hindered the expansion of the field (Oakes and Ji 2012: vii).

Things have changed in more recent years. Quantitative methods are preceded by the researcher's ideas and hypotheses about observed dimensions to calculable and measurable parameters. Frequency occurrence of a language form, its combinations with other items in discourse as well as patterns of semantic similarity, oppositeness and inclusion all contribute to a language-specific character of SL and TL forms. In her report on the interplay of quantitative and qualitative analysis of Polish and English cross-lingual corpora, Lewandowska-Tomaszczyk (2012) proposed general methods to conduct the explanatory analysis in translation study:

1 comparison of two or more translations of an original text (to study stylistic differences);
2 comparison of translation and monolingual corpora in the same languages as the translation (to study linguistic features of the translation as compared to the reference text in the same language as the translation).

(Lewandowska-Tomaszczyk 2012: 4)

Lewandowska-Tomaszczyk clearly described how lexical profiles of the TL and SL can be compared; how the keyness of certain grammatical patterns can be generated with a large reference corpus of the same language, and how collocation patterns of TL and SL can be presented statistically. The typology of translational quantitative criteria of resemblance in two languages can be shown by Lewandowska-Tomaszczyk's summary in Table 29.3.

Statistical analysis of translation texts has revealed some interesting findings. First the study of sentence length implicitly points to essential features of translated texts such as simplification and explicitation (Pym 2008: 128). In a study of specialized Italian–English translation by students, Laviosa (2008) found that sentence length may serve as an indication of the quality of translation; high-score translations tend to be associated with higher average sentence length and high level of lexical density, while low level translations often exhibit the reverse textual patterns. Ji and Oakes (2012: 177–208) compared three versions of early English translations of *Honglongmeng*, a masterpiece of Chinese literature, and demonstrated in detail that 'a set of bivariate statistics, commonly used for the comparison of corpora, can be applied in translation studies' (Ji and Oakes 2012: 176). The statistical analysis of sentence length, positive and

Table 29.3 The typology of translational quantitative criteria of resemblance

i.	Frequencies of occurrence of lexical units
ii.	Keyness
iii.	Frequencies of syntactic patterns (complex/simple constructions, sentence types and sentence patterns)
iv.	Frequencies of classes of lexical-semantic patterns
v.	Frequencies of types of figurative extensions (frequency of Source Domain and Target Domain patterns
vi.	Quantitative cross-correspondence of concepts from the same conceptual cluster
vii.	Distributional criteria

Source: Lewandowska-Tomaszczyk (2011: 32)

negative emotional words, value words, idioms and phrases used, presents stylistic differences of the translators, showing that 'textual and linguistic features can be demonstrated by the validity and productivity of statistics for the study of translation corpora' (ibid.: 206).

While quantitative research investigates relations between a few variables in larger samples, qualitative research deals with relations between many variables that can be investigated in smaller samples. Qualitative study is based on interpretations of resemblance between concepts presented in the original SL and TL translation from the experiences, actions and observations of individuals. The key skill in this new area, according to Sinclair is 'to be able to interrogate the corpus efficiently – to ask the right sort of questions, to refine the first responses and to control the retrieval process so as to reveal the way in which meaning and pattern interact in text' (Sinclair 2003: 3). These activities also need support by various types of computer software.

Multilingual computer tools can work on different corpora concurrently and generate bilingual or multilingual concordance lines at the same time. Two commonly used multilingual tools for non-computer scientists are *Multi-Concord* (Woolls 2002) and *ParaConc* (Barlow 1999). *ParaConc* is a bilingual or multilingual concordancer that can be used with translated texts in contrastive analyses, language learning, and translation studies. The *ParaConc* website shows that since its birth in 1990, it has been used at a variety of institutions for about 16 language pairs, such as English–Arabic, English–Chinese, English–French, English–Japanese, English–Korean, English–Russian, and so forth.

Computer software can generate paralleled concordance lines in both SL and TL, but it cannot explain them. Analysing and interpreting keywords in context cannot be conducted completely without human intelligence. The linguistic approaches to corpus study by John Sinclair may also apply to translation studies. The five co-selections are the core, collocation, colligation, semantic prosody and semantic preference. Collocations are word relations. They are 'the co-occurrence of words with no more than four intervening words' (Sinclair 2004: 34). Colligational patterns are lexicogrammatical realizations, and are relations between words and grammatical categories. Semantic preference is 'the restriction of regular co-occurrence to items which share a semantic feature, e.g. about sport or suffering' (ibid.: 141). Semantic prosody is the relation between words and lexical sets, and refers to a particular attitude or a particular point of view of a writer. Sinclair explains that 'the initial choice of semantic prosody is the functional choice which links meaning to purpose; all subsequent choices within the lexical item relate back to the prosody' (ibid.: 34). In other words, it is the semantic prosody selected by the speaker or writer that determines the semantic preference. The semantic preference then determines the collocational and colligational patterns. The five linguistic parameters can help establish textual profiles of both source language and target language, and further categorize the text's function and its communicative purposes.

Corpus analysis enables translators to compare the original text and the translated text with a number of criteria, namely perceptual, functional, emotional, axiological, ideological, logical, and associative (Lewandowska-Tomaszczyk 2012: 32). To realize these goals, the choice of qualitative and/or quantitative methods has to be taken in line with particular research questions. Both methods have advantages and limitations, but each can contribute to translation in a different manner. In practice, translators and researchers have to use a combined approach: qualitative data is in many cases also annotated and counted, and quantitative data is interpreted and explained. There is no universally 'best way' to combine methods.

Topics of corpus approach in translation studies

Translation memories and statistical machine translation have changed the way translated texts are generated. Improved alignment techniques at word level, phrase level and sentence level provide increasingly important resources for the proper use of both source language and target language. At the same time, theoretical and descriptive corpus-based research has investigated topics such as translation universals (Baker 1996; Laviosa 2002; Mauranen and Kujamäk 2004), translation ideology (Pérez 2003; Li *et al.* 2011), translator style (Baker 2000; Burrows 2002, 2007; Rybicki 2012), and translated/interpreted language (Granger 2003; Ji and Oakes 2012).

The corpus approach can investigate translation universals in that large amounts of data can better represent linguistic features of a particular language than individuals' intuitive judgements. Despite the argument on whether there is a translation universal across different languages, research has shown some universal features in translation such as simplification, convergency, explicitation, disambiguation, overrepresentation and conservatism. Baker defines simplification as 'a translator's attempt' to make things easier for the reader, but not necessarily more explicit (Baker 1996: 182). Laviosa-Braithwaite's study evidenced that simplification has at least three types: syntactic, stylistic and lexical, making translated texts simpler and easier to understand than non-translated texts (Laviosa-Braithwaite 1997: 533). In view of the convergence universal, translated texts are found to be more similar to each other than non-translated texts (Mauranen and Kujamak 2004). Explicitation indicates that some translated texts avoid extremes in translation and may generate a target language text with many more redundancies (Laviosa 2002; Moropa 2011: 259–281). Baker summarized that

> universal features can be seen as a product of constraints which are inherent in the translation process itself, and this accounts for the fact that they are universal. They do not vary across cultures. Other features have been observed to occur consistently in certain types of translation within a particular socio-cultural and historical context.
>
> (Baker 1993: 246)

The manifestation of ideology in translation has become an increasingly important issue in translation studies. According to Baker (2006), translation ideology aims to contribute to the broader discussion of a set of ideas, beliefs and codes of behaviour that govern a community. Some translation studies have compared ideological phenomena such as group interest, dominance, power relations in source language and target language (Pérez 2003; Petrescu 2009: 93–96). Others are more interested in the phenomenon of how translators implant their own viewpoints in translated texts. Not only questions of politics, but also reflections upon gender, sexuality, religion, secularity and technology provide a strong argument that such diversity of perspectives is highly desirable for good translation.

Translation style deals with a number of measurable features such as sentence length, vocabulary richness and various frequencies of words, word lengths and word forms, etc. There are numerous applications in authorship attribution research. Statistical analysis of conscious and unconscious elements of personal style can help detect the true author of an anonymous text, which means stylistic fingerprints can betray the plagiarist with more or less sophisticated statistical methods (Rybicki 2012: 231). Burrows (2002: 267–287) applied z-scores to establish a Delta system to evaluate the differences between the most frequent words in two corpora. Although some scholars challenged that 'it lacks any compelling theoretical justification' (Hoover 2004: 453–475), Delta has been used as a simple and intuitively reasonable method for traceable differences between texts. However, Rybicki's experience with English and Polish proves that Delta's precision in recognition of English texts is not matched by that in other languages (Rybicki 2012: 233), and might not adequately differentiate between individual translators' styles. For a better comparison, Rybicki used Delta measure plus Cluster Analysis to produce tree diagrams with a given set of parameters and concluded that discriminating between original authors is easier than discriminating between different translators of the same original.

Summary

The main benefits of corpora in translation are summarized by Granger (2003: 17–28) as: a great resource for content information, a great resource for terminology and phraseology, large quantity and good coverage of genres and texts, and improved operation of retrieving linguistic and contextual information. While the success of machine translation systems depends on automation and data quantity, descriptive applications rely on manual analysis and data quality. The availability of suitable tools and resources is crucial to corpus-based translation studies, but user-friendly tools and balanced translation corpora are yet to be produced. Given that the necessary steps to prepare a corpus may be technically complex or time-consuming in terms of manual labour required, it is necessary for the technical expertise, such as programmers and computational linguists, to team up with linguists and translation scholars who are willing to contribute their time and effort with corpus texts.

As Granger noticed, 'many corpus-based descriptive translation investigations suffer from a piecemeal, fragmentary and tentative approach; the variety of data sets, methods and tools used do not combine into a single overall framework and the results are often hardly commensurable' (Granger 2003: 22). In order to improve the quality of resources and make them available and accessible, and to realize the fuller potential of corpus linguistic methodologies, scholars of corpus-based translation studies need high-quality, easy-to-use linguistic resources and tools to bridge the gap between source language and target language.

References

Aijmer, Karin (2008) *English Discourse Particles: Evidence from a Corpus*, Amsterdam and Philadelphia: John Benjamins.

Baker, Mona (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications', in Mona Baker, Gill Francis, and Elena Tognini-Bonelli (eds) *Text and Technology*, Amsterdam and Philadelphia: John Benjamins, 233–250.

Baker, Mona (1995) 'Corpora in Translation Studies: An Overview and Some Suggestions for Future Research', *Target* 7(2): 223–243.

Baker, Mona (1996) 'Corpus-based Translation Studies: The Challenge That Lies Ahead', in Harold L. Somers (ed.) *Terminology, LSP, and Translation: Studies in Language Engineering*, Amsterdam and Philadelphia: John Benjamins, 175–188.

Baker Mona (1999) 'The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators', *International Journal of Corpus Linguistics* 4(2): 281–298.
Baker, Mona (2000) 'Towards a Methodology for Investigating the Style of a Literary Translator', *Target* 12(2): 241–266.
Baker, Mona (2006) *Translation and Conflict: A Narrative Account*, London and New York: Routledge.
Barlow, Michael (1999) 'MonoCon 1.5 and ParaConc', *International Journal of Corpus Linguistics* 4(1): 319–327.
Biber, Douglas, Susan Conrad, and Randi Reppen (1998) *Corpus Linguistics: Investigating Language Structure and Use,* Cambridge: Cambridge University Press.
Bowker, Lynne (1999) 'The Design and Development of a Corpus-based Aid for Assessing Translations', *Teanga* 18: 11–24.
Burrows, John (2002) '"Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship', *Literary and Linguistic Computing* 17(3): 267–287.
Burrows, John (2007) 'All the Way Through: Testing for Authorship in Difference Frequency Strata', *Literary and Linguistic Computing* 22(1): 27–48.
Croft, William (2010) 'Language Structure in Its Human Context: New Directions for the Language Sciences in the Twenty-first Century', in Patrick Hogan (ed.) *Cambridge Encyclopedia of the Language Sciences*, Cambridge: Cambridge University Press, 1–11.
Cysouw, Michael and Bernard Wälchli (2007) 'Parallel Texts: Using Translational Equivalents in Linguistic Typolog', in Michael Cysouw and Bernard Wälchli (eds) *Parallel Texts: Using Translational Equivalents in Linguistic Typology*, Sprachtypologie und Universalienforschung STUF, 95–99.
Davis, Mark (2008) *The Corpus of Contemporary American English (COCA): 410+ Million Words, 1990–Present*, Brigham Young University. Available at: http://www.americancorpus.org.
Granger, Sylviane (2003) **'**The Corpus Approach: A Common Way Forward for Contrastive Linguistics and Translation Studies?' in Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Amsterdam and Atlanta: Rodopi, 17–28.
Granger, Sylviane (2010) 'Comparable and Translation Corpora in Cross-linguistic Research: Design, Analysis and Applications', *Journal of Shanghai Jiaotong University* 2: 14–21.
Gries, Stefan and Stefanie Wulff (2012) 'Regression analysis in translation studies', in Michael Oakes and Ji Meng (eds) *Quantitative Methods in Corpus–based Translation Studies: A practical guide to descriptive translation research*, Amsterdam/Philadelphia: John Benjamins, 35–52.
Grieve, Jack, Dirk Speelman, and Dirk Geeraerts (2011) 'A Statistical Method for the Identification and Aggregation of Regional Linguistic Variation', *Language Variation and Change* 23: 193–221.
Hartmann, Reinhard (1980) *Contrastive Textology*, Heidelberg: Julius Groos Verlag.
Hoover, David (2004) 'Testing Burrows's Delta', *Literary and Linguistic Computing* 19 (4): 453–475.
Ji, Meng and Michael P. Oakes (2012) 'A Corpus Study of Early English Translations of Cao Xueqin's *Hongloumeng*', in Michael P. Oakes and Meng Ji (eds) *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research*, Amsterdam and Philadelphia: John Benjamins, 177–208.
Johansson, Stig and Hilde Hasselgård (1999) 'Corpora and Cross-linguistic Research in the Nordic Countries', in Sylviane Granger, Jacques Lerot, and Stephanie Petch-Tyson (eds) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Amsterdam and Atlanta: Rodopi, 145–162.
Kenny, Dorothy (2001) *Lexis and Creativity in Translation: A Corpus-based Study*, Manchester: St. Jerome Publishing.
Kučera, Henry and W. Nelson Francis (1967) *Computational Analysis of Present-day American English*, Providence, RI: Brown University Press.
Lakoff, George (1987) Women, Fire and Dangerous Things: What Categories Reveal about the Mind, Chicago: Chicago University Press.
Laviosa, Sara (2002) Corpus-based Translation Studies: Theory, Findings, Applications, Amsterdam and New York: Rodopi.
Laviosa, Sara (2008) 'Description in the Translation Classroom: Universals as a Case in Point', in Anthony Pym, Miriam Schlesinger, and Daniel Simeoni (eds) *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*, Amsterdam and Philadelphia: John Benjamins, 191–230.
Laviosa-Braithwaite, Sara (1997) 'Investing Simplification in an English Comparable Corpus of Newspaper Articles', in Kinga Klaudy and János Kohn (eds) *Transferre Necesse Est: Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting*, 5–7 September 1996, Budapest, Hungary, 531–540.

Leech, Geoffrey (1992) '100 Million Words of English: The British National Corpus (BNC)', *Language Research* 28(1): 1–13.

Lewandowska-Tomaszczyk, Barbara (2012) 'Explicit and tacit: An interplay of the Quantitative and Qualitative Approaches to Translation', in Michael P. Oakes and Meng Ji (eds) *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research*, Amsterdam and Philadelphia: John Benjamins, 3–34.

Li, Defeng, Zhang Chunling, and Liu Kanglong (2011) 'Translation Style and Ideology: A Corpus-assisted Analysis of Two English Translations of *Hongloumeng*', *Literary and Linguistic Computing* 26(2): 153–166.

Li, Lan and Graham Bilbow (2001) 'From a Business Corpus to a Business Lexicon', *Lexikos* 11: 209–221.

Mauranen, Anna and Pekka Kujamäk (eds) (2004) *Translation Universals: Do They Exist?* Amsterdam and Philadelphia: John Benjamins.

Moropa, Koliswa (2011) 'A Link between Simplification and Explicitation in English-Xhosa Parallel Texts: Do the Morphological Complexities of Xhosa Have an Influence?' in Alet Kruger, Kim Wallmach, and Jeremy Munday (eds) *Corpus-based Translation Studies: Research and Applications*, London: Continuum, 259–281.

Oakes, Michael P. and Meng Ji (eds) (2012) *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research*, Amsterdam and Philadelphia: John Benjamins.

Pearson, Jennifer (1996) 'Electronic Texts and Concordances in the Translation Classroom', *Teanga* 16: 85–95.

Pérez, Maria Calzada (ed.) (2003) *Apropos of Ideology: Translation of Ideologies: Ideologies in Translation Studies*, Manchester: St. Jerome Publishing.

Petrescu, Camelia (2009) 'Translation and Ideology', *Professional Communication and Translation Studies* 2(1–2): 93–96.

Pym, Anthony (2009) 'Using Process Studies in Translator Training. Self-discovery through lousy experiments', in Susanne Göpferich, Fabio Alves and Inger Mees (eds) *Methodology, Technology and Innovation in Translation Process Research*, Copenhagen: Samfundslitteratur, 135–156.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*, London: Longman.

Resnik, Philip and Noah Smith (2003) 'The Web as a Parallel Corpus', *Computational Linguistics*, 29 (3): 349–380.

Rybicki, Jan (2012) 'The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation', in Michael P. Oakes and Meng Ji (eds) *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research*, Amsterdam and Philadelphia: John Benjamins Publishing Company, 231–248.

Sinclar, John (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Sinclair, John (2003) *Reading Concordances: An Introduction*, Harlow: Longman.

Sinclair, John (2004) *Trust the Text: Language Corpus and Discourse,* London: Routledge.

Tognini-Bonelli, Elena (2001) *Corpus Linguistics at Work*, Amsterdam and Philadelphia: John Benjamins.

van der Auwera, Johan, Ewa Schalley, and Jan Nuyts (2005) 'Epistemic Possibility in a Slavonic Parallel Corpus – A Pilot Study', in Björn Hansen and Petr Karlik (eds) *Modality in Slavonic Languages, New Perspectives*, Munich: Sagner, 201–217.

Wallis, Sean and Nelson, Gerard (2001) 'Knowledge Discovery in Grammatically Analysed Corpora', *Data Mining and Knowledge Discovery* 5(4): 305–336.

Woolls, David (2002) 'Multiconcord: The Lingua Multilingual Parallel Concordancer for Windows'. Available at: http://artsweb.bham.ac.uk/pking/multiconc/l_text.htm.

30

EDITING IN TRANSLATION TECHNOLOGY

Christophe Declercq

UNIVERSITY COLLEGE LONDON, THE UNITED KINGDOM

Language and translation technology

With the emergence of translation memory technology in the early to mid-1990s,[1] the translation profession underwent a true technological turn that had been eagerly awaited by those working on machine translation systems since the 1950s. At the core of the translation memory systems (TMS) was a database of human translations, aided by the machine: machine-aided human translation, MAHT, or computer-aided translation, CAT. With the segment-based approach of re-use of previously translated material, traditional concepts and workflows changed dramatically too. Other than language skills and writing abilities, translation of texts included an increasing use of computer technology. Processes such as editing, revision and proof-reading should follow suit, but to date translators are struggling to cope with the speed of translation technology uptake.[2]

In the last few years, that uptake has assumed the shape of machine translation (MT), especially statistical machine translation (SMT). Yet, despite the fact that MT is into its seventh decade, deeply rooted reservations about quality output of automated translation engines remain commonplace. Peculiarly, that scepticism among many translators is overcome by millions of users of Google Translate or Microsoft's Bing Translator and by thousands of customized translation engines the world over.

An in-depth analysis of 'editing and translation today' therefore not only looks into workflows involving a translation memory (TM) and their innate verification processes of editing and proof-reading, but also into the emerging convergence of translation technology, in particular translation memories, and language technology, especially machine translation.[3] As no clear delineation can be established between translation technology (TT, the applied use of any computer application that supports the translation process as performed by a human translator) and language technology (LT, human language generated automatically by a computer system),[4] this contribution therefore aims to include any automated means of facilitating productivity and/or quality of human translation.

'Traditional' translation technology and editing

Traditional translation technology has at its core the translation memory system (TMS) and is likely to be supported further with a terminology database.[5] The TM application which re-uses segments that match previously translated material has also been described as machine-aided human translation (MAHT), computer-aided translation (CAT) or translation environment tools (TEnTs). Whatever the acronym and whatever the definition of a TMS, besides re-use of previously translated segments (or translation units), a key feature is often overlooked by translators themselves: TM systems allow translators to deal with complex file formats they do not necessarily master themselves.[6] They receive and deliver files in their native formats without interfering with the underlying code such as cross-references or mark-up language.[7] Typical material concerns FrameMaker or InDesign, but also DITA XML or Microsoft .NET files. As such a main benefit of any TMS is that it allows translators to be translating and editing much more material than in any typical word processing environments. With the translation interface, the user interface of any TM environment in which translators visually see the text on screen as they edit it (Biau Gil 2007), translating and editing converge also.

Not all types of matches from a TM occur in just any translation project. In the screenshot shown in Figure 30.1, the fourth and last segment still need translating from scratch and no source was copied across there. All the 100 per cent matches are re-used from the TM. Whether or not the perfect matches need editing should depend on the quality of the TM results, the formatting and quality assurance settings, the project requirements and the experience of the translator. However, this is often limited to contractual obligations which urge the translator not to alter any perfect match. Note that the named entities make up about 35 per cent of the overall word count. Copying across the source segments with added short cut expertise to be jumping across words in the target segment most certainly constitutes an increase in productivity (especially for this text type, i.e. sports).

However, beyond those stipulations, each degree of matching requires different cognitive processes of the translator. Whereas often minor brief additions or alterations might improve a segment to the level that is acceptable for the purpose for which it is used,[8] research on how translators maintain their awareness of possible flaws while re-using translation units from the TM might be relevant to analyses of editing MT output too.

Figure 30.1 Detail of the Editor Environment of SDL Trados Studio 2011 (SP1), with 3+1+4+1 units

Source: text by sport.be

Cognitive processes and editing

Lagoudaki (2006) was a reference work about the translators' perception and use of technology, but translation environments have moved on.[9] Among others, the pervasive use of SMT has effectuated a new paradigm in that perception of language and translation technology. More importantly, in the last few years, translation memory systems have broken away from the –

admittedly often preferred by translators – environment of word processors and moved to standalone applications and online software as a service (SaaS). However, what has remained ever since the increased uptake of TM systems in the 1990s, are widespread concerns about the effect of translating and editing in a TMS. Based on an empirical study, Dragsted 2008 proved that any TM's segmentation into units, usually sentences, creates a strong focus on those segments, which affects the overall quality of the translation as a final product.

With a text that is presented in a TMS in various segments or units, a sentiment of alienation lies in the balance between a steady pace and a structured approach. In fact, with translation technology as a form of human–computer interaction, it is very difficult to differentiate formal benefits/disadvantages from holistic ones.

Whether segmentation leads to an increased tendency towards more literal translation or not, remains a matter for scholars to discuss and for further empirical studies. In the debate about the consequences of segmentation, experience and maturity are often overlooked, along with the need for increased productivity. In fact, in his pilot study Biau Gil proves that subject-matter knowledge is more relevant than visual information (2007: 7). Taking this finding across the TM/MT threshold already, this is a further argument that post-editors should above all be knowledgeable about the subject topic.

Table 30.1 Benefits and disadvantages of segmentation in Translation Memory Systems

Benefits of (segmentation in) a TM	*Disadvantages (of segmentation in) a TM*
A sense of control on the segment level	The layout of the source text is lost
Similar pace	No feeling of overall view and alienation from the context
Close reading, no interference of non-verbal elements	Lack of non-verbal elements affects quality and productivity (Biau Gil 2007)
Added value of term recognition	Lack of control
No formatting issues	Formatting sometimes still requires editing
Increased accuracy and consistency	A tendency to more literal translation
Being able to monitor progress	
Auto-propagation	
Possible copying across of the source segment	

Forms of editing, other than translating

Editing in projects that involves translation technology run along two axes. A first axis ranges from TM to MT. A second axis then concerns editing, ranging from pre-editing to post-editing. As pre-editing and controlled language are discussed elsewhere in this encyclopaedia, post-editing is broken down into more sub-concepts. Editing, revision and proof-reading are fundamental elements in translation projects and as a consequence their validity in MAHT projects is equally important.

Translation Service Providers (TSPs, sometimes also referred to as LSPs, Language Service Providers) adhere to the TEP model (translation/editing/proof-reading). However, in marketing their services the added value, especially of proof-reading, is often sold as a separate service. In the next section, the differentiation between the various forms of going over a text other than translating is effectuated in a sense of best practice, not in an academic overanalysing of terminological diffusion.[10] Publications and/or guidelines on editing, revision and proof-reading often concern a mere modal framework, 'how revisers *ought* to go about their jobs or what jobs they *could* use' (Mossop 2007, online), and eventually best practices or workflows for revisers are often based on experience anyway.

Comparing the translation with the original text and ensuring that there are no errors left such as spelling mistakes, grammatical errors, omissions or ambiguities, is a well-established practice by the Translation Bureau of the Public Works and Government Services Canada. In their style guide long lists of possible errors in both writing and editing are produced. However, much of this list is aimed at text-production and not necessarily at translation projects in a computerized setting. The error categorization of the Canadian Translation Bureau proves that translation technology increased the speed of how editing (of errors) and translation merged: translation memory tools started to elaborate on their proprietary quality assurance functionalities (such as verification in SDL Trados Studio 2011). Companies have been working towards this trend too, as can be seen with Yamagata Europe's QA Distiller.

Whether in QA Distiller, in Studio or in any other TMS, detection of possible errors has become very much an automated feature of translation projects too. This greatly enhances the consistency of translator's output as well as his/her ability to be submitting a formally flawless target file, but it also provides a learning curve for translators to become more experienced in translation quality assurance and as such set themselves apart from those who do not.

In order to distinguish between the various forms of editing and the various identities editing can assume, a practical overview is reproduced below, whereby the various forms of editing are in fact allocated a position in the workflow.

Makoushina and Kockaert (2008) place editing of the translated files along with proof-reading and deem it a non-formal form of quality assurance. With this approach, editing 'after' the translation (either HT, human translation, or MT, machine translation), 'post-editing', and editing of source files, 'pre-editing', are differentiated clearly as stages in the translation workflow. As mentioned earlier, pre-editing and controlled language are not the scope of this article, as they are dealt with elsewhere, but (post-)editing still needs to be set apart from proof-reading.

Source files		Translated files		Final files	
Formal QA	Non-formal QA	Formal QA	Non-formal QA	Formal QA	Non-formal QA
Manual check QA software	Editing Controlled language Authoring memory, etc.	Manual check QA software	Proofreading Editing	Manual check QA software	In-country review

Figure 30.2 Editing stages in an overall quality assurance approach

Source: Makoushina and Kockaert (2008: 3)

Editing, revision and proof-reading

In 2006, the European Committee for Standardisation (CEN) published the EN15038 standard,[11] developed for Translation Service Providers. The standard aimed to cover the entire translation process, including quality assurance. The standard offered TSPs and their clients a breakdown of the entire translation provision in accurate definitions and standard description. Most importantly, the European standard required both a translator and a reviewer for each translation and differentiated between the two. Under EN15038 only translators with the appropriate background and competences can translate documents and it is the task of that translator to check the translation themselves.[12] A reviewer then is a subsequent person in the translation workflow who examines 'a translation for its suitability for the agreed purpose, and respect for the conventions of the domain to which it belongs' and who recommends corrective

measures, if necessary (European Quality Standard EN-15038:2006). A review can be distinguished from a revision in that in the case of the latter, a translation is examined with both source and target texts compared. According to the European standard, proof-reading is limited to checking of proofs.[13]

These concepts and their allocated positions in the translation workflow are often mimicked by the translation tools themselves. In the Editor window of SDL Trados Studio 2011, the status of each translated segment can be altered, including being translated and reviewed. This is similar to what XTM Cloud offers. Across Systems takes this even a step further and includes buttons for the various steps in the translation process and aligns them with the EN15038 standard.[14]

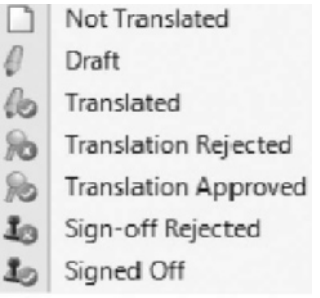


Figure 30.3 Segment status in SDL

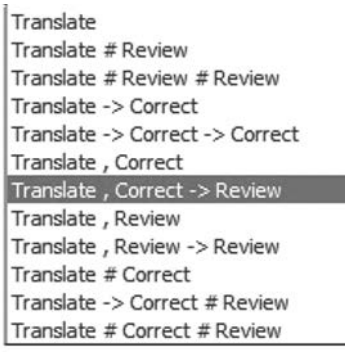


Figure 30.4 Various translation workflows possible in XTM Cloud

Source: Trados Studio (2011)

In EN15038, editing in any form (copy-editing, pre-editing, post-editing) is included in appendix only, as an added value service, but just how editing differs from review, revision and proof-reading is not very clear.[15] It can, however, be easily deduced from the descriptions what editing is and what it is not:

Table 30.2 How EN15038 could possibly set editing apart from review, revision and proof-reading

Elements editing shares with EN15038 stipulations of review, revision and proof-reading	*Elements editing does not share with EN15038 stipulations of review, revision and proof-reading*
Altering a translation for its suitability for the agreed purpose	Checking of proofs (even though it can be argued editing shares elements of checking of proofs on screen)
Matching the translation to the conventions of the domain to which it belongs	Recommendation of corrective measures (even though it can be argued editing proactively ensures these measures)
A level of comparing source and target text is involved	

Still, as already indicated by the various forms of editing, the above stipulations do not entail a set of practical guidelines on how editing is used in translation projects, be it in the strictest sense by means of a translation memory system or in a broader interpretation of translation technology. This then not only includes machine translation, but also social media (crowd-sourced translations or community translations), sometimes both are combined even (as if often the case with projects posted on platforms such as Transifex). But most importantly, editing alongside translation and/or language technology takes the shape of post-editing machine translation.

Post-editing and machine translation

Post-editing machine translation concerns the practical answer to the longstanding quest for the Holy Grail: machine translated material that is substantially good enough for communication and/or dissemination.[16] A valid example of how practical post-editing MT can be is Jeff Allen's *Creole MT*,[17] a publicly available MT system for the purpose of relief during the 2010 Haiti earthquake and its aftermath.

Understanding the choices translators make while working with translation technology such as translation memories can be of significant relevance on how to approach the influence of translation provided by machine translation. Even when translation scholars have considered the 'black box' of machine translation in the past, it was in opposition to Holmes's 'little black box of the translator's mind' (Holmes 1972: 72). However, especially when post-editing machine translation (PEMT) is concerned, the two in fact are more in juxtaposition and will be converging still more in the future. Above all, post-editing should be seen as a process of improving through modification (rather than revision) a machine-generated translation, often eyeing a minimum of effort on behalf of the post-editor.[18] The quicker the turn-around needs of a translation, the more likely the PEMT effort will be a fast one, also known as 'light post-editing'. More thorough modifications, with less urgency, aim to produce a better quality and is often known as 'full post-editing'. The latter category is the more common one, not least because it aims to obtain a quality level that is the same as if the entire text had been translated from scratch by the human translator.

The quality of a translation is a hotly debated issue, let alone the quality of a translation in which MT played a part, and subsequent post-editing. O'Brien (2010) rightly argues that the quality expectations differ depending on where a particular person is involved. Developers are very interested in automated quality metrics such as BLEU (Bilingual Evaluation Understudy), TER (Translation Edit Rate) or WER (Word Error Rate).[19] They are also very keen on getting usage feedback from the translator, improving the system they have developed with valuable input.[20] Buyers allocate PEMT projects to translators or TSPs because they hope for a faster turnaround. The overall translation cost might be similar to HT; if the PEMT approach

Source Text	Raw MT
Un vaste réseau qui piratait les codes de déverrouillage des téléphones portables a été démantelé, ont annoncé, dimanche 26 septembre, les enquêteurs.	A vast network hacked unlock codes for mobile phones has been dismantled, announced Sunday, Sept. 26, investigators.
Example of Light Post-Edit	A vast network which hacked unlock codes for mobile phones has been dismantled, it was announced Sunday, Sept. 26, by investigators.
Example of Full Post-Edit	A vast network which hacked security codes for mobile phones has been dismantled, according to an announcement by investigators on Sunday, Sept. 26.

Figure 30.5 Light and full post-editing of raw MT output

Source: O'Brien (2010: 5)

saves time, then that is a major benefit for the buyer already. The translators or TSPs hope that by increasing their productivity, they can also increase their client portfolio and/or market share. Two categories that are often overlooked are the project managers[21] and the account executives or sales. These people do not necessarily carry the need to be included in the list just now, but they are very crucial in the communication chain with the client and its subsequent users and as such cannot afford to be creating false expectations. In the end, much of the success of post-edited machine translations depends on how the users have perceived the quality of what was disseminated or communicated.

In the entire debate of considering raw MT output as fuzzy matches so as to gauge the probable workload for post-editors properly, Guerberof (2009) analysed findings of a small-scale research project that are very interesting. Translators were asked to post-edit TM segments of 80–90 per cent fuzzy matching on the one hand and SMT output on the other hand, as well as translate anew. In an analysis of all the errors produced in each of the three categories, new segments accounted for roughly one error in five. Intriguingly, a similar number of words to be post-edited triggered not many more errors. In fact, the errors in the final translation produced with the aid of a translation memory accounted for half of all the errors, i.e. editing fuzzy matches in a TMS triggers doubled the amount of errors compared to post-editing raw MT output.[22] Similarly, using the TM even slowed down productivity by 2.5 per cent, whereas MT increased this by 24.5 per cent, a combined difference of 27 per cent or nearly a third.

The re-usable nature of raw MT output has been confirmed by Fontes (2013), chair of the European Commission's MTUG (machine translation user group). In a survey across the Directorate-General for Translation experienced translators were asked to rate MT output quality. Of the 643 ratings of language pair combinations, 200 ratings confirmed that they had used MT for more than 75 per cent of their translation jobs. Asked to rate the output of the respective engines on a 0–4 scale, 726 ratings were delivered. One hundred and eighty-five people rated the MT quality as four or three, in which most segments were considered re-usable. Asked for the reasons why MT should be used, three of the five answers[23] (MT is a typing aid, MT is a source of inspiration for alternative translations available in the translation memory, MT is a quick draft) imply subsequent use of post-editing.

Post-editing guidelines

TAUS, the Translation Automation Society is one of the most authoritative sources on post-editing machine translation. Crucial to raising awareness among users of PEMT about the various issues involved, they have highlighted recommendations and post-editing guidelines.

On the recommendation of tuning your engine appropriately TAUS (2010) distinguishes between rule-based or statistical engines, whereby a high-level dictionary and linguistic coding is crucial for rule-based machine translation (RBMT) and clean, high-quality, domain-specific data are key to data-driven systems. The second TAUS recommendation is to ensure that the source text is written well, preferably written with later MT in mind even. As mentioned earlier: there is no post-editing machine translation without including pre-editing the source material.

One of the most obvious recommendations by TAUS 2010 is to train post-editors in advance. However, there is a major difference between training people to act as post-editors for a specific job with project-specific data and guidelines on the one hand and linguists on the other hand who receive more basic training because they work across projects and therefore need to adhere more to a common denominator. Moreover, including post-editing into the curriculum of higher education has proven a difficult feature.[24]

Providing generic guidelines for achieving quality that is in line with the project stipulations and the agreed expectations is not easy, as TAUS 2010 proves. Most guidelines, a dozen in total, remain very tentative and do not immediately constitute a checklist. However, in line with the quality assurance capacities of translation memories mentioned earlier, several guidelines can in fact be dealt with in the automated environment of a TMS:

Table 30.3 TAUS post-editing guidelines versus quality assurance in SDL

Selected guidelines for post-editing (TAUS 2010)	*Quality assurance in SDL Trados Studio 2011*
'Ensure that no information has been accidentally added or omitted.'	QA Checker 3.0: Segment verification Check for forgotten and empty translations Check for segments where source and target are identical Check for segments which are x% shorter / longer Segments to exclude
'Basic rules about spelling, punctuation and hyphenation apply.'	QA Checker 3.0: Inconsistencies (repeated words in target, unedited fuzzy matches) Punctuation Numbers, times, dates, measurements
'Ensure that key terminology is correctly translated and that untranslated terms belong to the client's list of "Do Not Translate" terms.'	QA Checker 3.0: Word List and Regular Expression Terminology Verifier (with a term base open)
'Ensure that formatting is correct.'	Any TMS strives towards maintaining exactly the same formatting between source and target. Most TMSs also include warning messages in case where there are differences.

Source: Trados Studio (2011)

In a combined approach of the above, the text segment represented below, which could have been reproduced in many other TMSs too, requires actions on both levels: in the TMS of Wordfast Anywhere (WFA) formatting has not been reproduced appropriately by Google Translate. A post-editor would need to restore the tags. However, this would have been picked up on already by the verification feature of WFA. The post-editor would have to restore some cultural elements to the source text and this example indeed triggers the copying across of the source segment.

So far, no proprietary environment for post-editing alone has been mentioned and even though they are around (such as the Post-Editing Tool (PET) by Wilker Aziz and Lucia Specia), it should be clear that post-editing can happen very well in the environment of a TMS. It should be noted that post-editing is also required in platforms for crowd-sourced translations such as Transifex, live subtitling with speech recognition or subtitling editors such as dotsub and YouTube Subtitler.

With post-editing material that has been provided by a translation memory, machine translation or even speech recognition, pricing methods are a tricky business. Three common options apply. Other than having a linguist available in-house (for public broadcasting and live captioning for instance), either a nominal fee is paid based on the time spent or a word rate is agreed, differentiating between re-use from the TM (see earlier categories of matches) and machine translation (which differs based on the training data and the input). Eventually PEMT is paid along the lines of fuzzy matching.

Figure 30.6 Tag differences returned by Google Translate in Wordfast Anywhere

Source: Text by *Le Monde*

Conclusion

While on the Eurostar into London, the author wanted to joke with friends who also use *Road Bike*, a cycling app. After travelling at about 285km/h on average for 5 minutes, the live tracking was stopped and as the 20660 kcal were about to be sent via Gmail, the app, which had been installed in Dutch along with the operating language of its Android 4.1 system, neatly indicated 'U gaat wel erg snel. Wellicht heeft u de verkeerde sport gekozen' [*You are going very fast. Perhaps you have chosen the wrong sport.* MT by Google Translate]. It would be very difficult to find out whether this segment had been localized into Dutch by a translator (who might have used machine translation for draft output and treat it as fuzzy matches), by machine translation *tout court* or by a community of users that master Dutch. Such a community can use a platform such as Transifex, which in its turn can have community members who base their work on machine translation. Although this anecdotal instance does not prove much, it will be recognized by millions of users, 99.9 per cent of whom are not translators or linguists. The world of translation technology, language technology, mobile technology and social media (the people networks, the cloud and the crowd, and subsequently the feed of social data too) are converging.

With the rapid uptake of machine translation at a low entry level, but also on mobile phones and on tablets, the perception of translation from the global user's perspective is changing

dramatically. The main problem in overcoming that threshold fear by translators to be incorporating machine translation in their workflow, and therefore post-editing, is that translators deem the process of translation sacred, whereas eventually the target text is only a product with a purpose that is relevant to a world outside their own. If the wider translation profession does not see the opportunity to still be maintaining a much cherished art and profession, too many users will discard the human translator and resort to MT output that has been post-edited by either a native speaker or someone who knows the subject really well. The latter can very well be someone who is trusted within the (online) user community.

Editing in translation technology applications is an elementary step in the well-sought increase in productivity. Any target text that is the product of a translation process should be considered complete only after careful revision and editing. Reviewing segment after segment whereby that process has been produced by a computer application can indeed be more cumbersome than editing a human translation. However, if translation as a process and the means to an end product, whether by a human, a machine or hybrid, needs post-edition and this is not mastered by the human translators themselves, then who will fight the corner of the added value of humans here?

Arguably most clashes between quality expectations and deliverables can be overcome beforehand. By examining raw MT output quality an appropriate price needs to be negotiated and an agreement needs to be reached about the final quality of the information to be post-edited. Even though these two recommendations are included in those by TAUS 2010, they in fact constitute common practice in projects that involve HT only or HT+TM. However, it is undeniable that the ongoing new paradigm of pervasive use of MT can indeed act as a technological turn that triggers an awareness HT has not been able to do for decades. Including MT output in translation projects offers an opportunity to start negotiating this awareness anew. It would be lethal to miss out on that.

Notes

1 Any historical review of translation memory systems will point at Trados MultiTerm and IBM Translation Manager emerging in 1992, Atril's Déja Vu in 1993 and Trados Workbench in 1994. The concept, however, emerged much earlier, with Peter Arthern already in 1979 stipulating that the use of unrestricted machine translation at the European Commission might very well be too early still, but that there was 'scope for post-edited machine translation of a restricted range of texts' (Hutchins 1998: 293).

2 Crucial in the perception of language and translation technology is Google Translate, which more than a year after it became a paid for service, had more than 200 million people using it monthly. By April 2012, the daily total number of words equalled that of 1 million books (Kerr 2012).

3 For an appreciation of the history of post-editing machine translation see Ignacio Garcia 2012.

4 For instance, re-use from a large TM on the basis of aligned source and equivalent target texts or MT output from an SMT that has been trained on the same or similar corpus of equivalent texts are perhaps distinctively different in technology but closely related in use.

5 Although terminology management and the inclusion of term recognition in TM systems is not discussed here, it should be made clear that terminology is not only key to the HT, but also to MT. A combined hybrid TM/MT + terminology management allows for an increased quality assurance and if maintained successfully also an increased consistency and thus quality. Also dictionary compilation is a skill crucial in the development of translation engines.

6 From here onwards, this contribution does not allocate much space to defining several of its key concepts, let alone analysing the differences between respective definition variants. The applied field of translation technology itself, a world of increased productivity, does not warrant such ponderings.

7 Biau Gil, however, attests that translators' perfomance is improved by an environment whereby the non-verbal elements of a text or its native format are visible in an interface that is similar to WYSIWYG, *What You See Is What You Get*: 'texts translated using WYSIWYG translation interfaces

include fewer errors than those translated using non-WYSIWYG interfaces' and that 'when translators use WYSIWYG translation interfaces they work faster than when they use non-WYSIWYG interfaces' (2007: 7).

8 One such purpose is to maintain the standard or open format in which the translation memory is contained. XLIFF (XML Localisation Interchange File Format) allows users of translation technology to pass on data between various tools during the translation or localization process. XLIFF Editors can be found among more familiar providers ot translation technology tools such as MultiTrans as well as through lesser known freeware, such as Transolution. Other file formats that drive the translation editing environment are for instance Poedit, which allows translators and users to edit cross-platform gettext catalogs (PO files). SRT Translator provides a translation memory in which Google Translate produces draft translations of subtitles.

9 The Copenhagen Business School has been particularly active in researching the cognitive processes while translating using a TM and the effects of segmentation on the productivity and quality of the translator. Dragsted (2004), Dragsted (2008), Jakobsen (2009), Christensen and Schjoldager (2010) and Christensen and Schjoldager (2011) are but a selected few. Other people who have contributed to this field are Bowker (2005), Guerberof (2009), O'Brien (2008), O'Brien (2011) and Pym (2011).

10 In analogy to doctors being the worst patients, translators have a similar ailment: perennial analysis of concepts, their definitions and denotations, and a subsequent ongoing debate about the slight differences.

11 EN 15038 was published in May 2006 and has been gaining acceptance ever since. It was accepted by 28 nations (all EU member states, except Bulgaria and Croatia, but it was accepted by non-EU Iceland, Norway and Switzerland) after its inception and acted as a benchmark in the European Union.

12 This check by the translator is also called self-editing.

13 The Language Resource Centre of the Aalborg University refers to proof-reading as follows: the process where 'we focus exclusively on orthography, typing errors, grammar and punctuation'. Vocabulary and spelling are proof-read so as to make them consistent. For English-language texts 'either British or American spelling is used, and not a mixture of the two varieties of English'. In the case of an ambiguous translation a comment is inserted explaining the problem, but the text itself will never be re-phrased. (LRC 2009, online)

14 Across Systems uses a slightly different terminology: the corrector and reviewer ensure checking, revision, reviewing and verification.

15 According to Mossop (2007), in editing a translation project, corrections and improvements are made whereby the purpose and the given readership of the text are prioritized. Revising is a very similar task, but this is then applied to draft translations. Trying to rename all the PEMT, post-editing machine translation, as PRMT (post-revision?), seems not immediately feasible. In light of Mossop (2007), it could be argued that post-editors revise the MT segments first and edit the text in its entirety next. In practice, this would hardly happen and texts are translated and subsequently edited on a segment-by-segment basis. These corrections to a translation in order to increase its quality are also known as Quality Assurance (QA), whereas any correction stage to detect flaws in a translation after it has been submitted is often referred to as Quality Control (QC). For an appreciation of QA and QC, see Makoushina and Kockaert (2008), Rasmussen and Schjoldager (2011) and European Union (2012).

16 John Hutchins differentiates between MT for the purpose of communication (light post-editing required only) and dissemination (full post-editing required) (Hutchins 2013).

17 For an appreciation of the language technology effort for distress relief in Haiti, see Munro (2010).

18 The description of post-editing is a combination of two definitions: post-editing is 'the process of improving a machine-generated translation with a minimum of manual labour' (TAUS 2010) and 'a process of modification rather than revision' (Loffler-Laurian 1985 in O'Brien 2010).

19 For an appreciation of machine translation evaluation metrics see Snover and Dorr (2006). Users can compare users Google Translate or Bing Translator through iBLEU.

20 This is where pre-editing re-emerges: by comparing the raw MT output with the source text, errors can be found and arguably a system behind types of errors too. Other than leaving things as they are, developers have two options: boost the engine by training it on new data or allowing document authors to pre-edit their source material so as to have an increased raw MT output quality.

21 For an appreciation of machine translation and project management, see Guerberof (2010).

22 When Guerberof categorized the errors according to five types (mistranslation, accuracy, terminology, language and consistency), post-editing raw MT output produced very similar numbers

of errors for language and consistency as the new segments did. With double the errors for mistranslation and accuracy, it should then come as no surprise that re-using and editing fuzzy matches from the TM in fact landed more than half the errors for the three approaches together, whereas MT only did for a quarter.

23 Other responses referred to an increase in productivity and a gain in time for more thorough research.

24 For an appreciation of teaching post-editing, see Allen (2001), Kenny and Way (2001), O'Brien (2002), Belam (2003) and Kliffer (2008).

References

Across Systems (2011) 'The EN 15038 Standard Workflow'. Available at: www.across.net/documentation/onlinehelp/across_en/acrossHaupt648.htm.

Allen, Jeff (2001) 'Post-editing: An Integrated Part of a Translation Software Program', *Language International* April, 26–29.

Aziz, Wilker and Lucia Specia (2012) 'PET: A Tool for Post-editing and Assessing Machine Translation', in *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, 28–30 May 2012, Trento, Italy, 99. Available at: www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf.

Belam, Judith (2003) 'Buying up to Falling Down: A Deductive Approach to Teaching Post-editing', in *Proceedings of MT Summit IX*, 23–27 September 2003, New Orleans, LA. Available at: http://www.dlsi.ua.es/t4/proceedings.html.

Biau Gil, José Ramon (2007) 'What You See Is What You Get? A Pilot Experiment on Access to Visual Information in Translation Interfaces', Paper presented at the 12th Annual Internationalisation and Localisation Conference, 26–28 September 2007, Dublin, Ireland. Available at: www.localisation.ie/resources/conferences/2007/presentations/Biau_LRC_XII/LRC_XII_slides.pps.

Bowker, Lynn (2005) 'Productivity vs. Quality: A Pilot Study on the Impact of Translation Memory Systems', *Localisation Focus* 4(1): 13–20.

Canada Translation Bureau (1997) 'The Canadian Style: A Guide to Writing and Editing'. Available at: http://bt-tb.tpsgc-pwgsc.gc.ca/btb.php?lang=eng&cont=791.

Chan, Sin-wai (2004) *A Dictionary of Translation Technology*, Hong Kong: The Chinese University Press.

Christensen, Tina Paulsen, and Anne Schjoldage (2010) 'Translation-Memory (TM) Research: What Do We Know and How Do We Know It?' *Journal of Language and Communication Studies* 44: 89–101.

Christensen, Tina Paulsen and Anne Schjoldage (2011) 'The Impact of Translation-Memory (TM) Technology on Cognitive Processes: Student-translators' Retrospective Comments in an Online Questionnaire', in Bernadette Sharp, Michael Zock, Michael Carl, and Arnt Lykke Jakobsen (eds) *Proceedings of the 8th International NLPCS Workshop: Special Theme: Human–Machine Interaction in Translation*, 20–21 August 2011, Copenhagen, Denmark, 119–130.

Crabbe, Stephen (2010) 'Controlled Languages for Technical Writing and Translation', in Ian Kemble (ed.) *The Changing Face of Translation: Proceedings of the 9th Annual Portsmouth Translation Conference 2009*, 7 November 2009, Portsmouth, the United Kingdom, Portsmouth: University of Portsmouth, 48–62. Available at: http://tinyurl.com/b8truqo.

Data Translations (2013) Services. Available at: www.datatra.be/en-services-revision-proof-reading.

Dotsub, http://dotsub.com.

Dragsted, Barbara (2004) 'Segmentation in Translation and Translation Memory Systems: An Empirical Investigation of Cognitive Segmentation and Effects of Integrating a TM System into the Translation Process', in *Cognition Distributed: How cognitive technology extends our minds*, Copenhagen: Copenhagan Business School.

Dragsted, Barbara (2008) 'Computer-aided Translation as a Distributed Cognitive Task', in Itiel Dror and Stevan Harnad (eds) *Cognition Distributed: How Cognitive Technology Extends Our Minds*, Amsterdam and Philadelphia: John Benjamins, 237–256.

European Quality Standard EN-15038: 2006. Available at: http://qualitystandard.bs.en-15038.com/ (last accessed 1 February 2013).

European Union (2012) 'Quantifying Quality Costs and the Cost of Poor Quality in Translation: Quality Efforts and the Consequences of Poor Quality in the European Commission's Directorate-General for Translation', Luxembourg: Publications Office of the European Union.

Fontes, Hilàrio Leal Fontes (2013) 'Evaluating Machine Translation: Preliminary Findings from the First DGT-wide Translators' Survey', in *Language and Translation: Machine Translation*, European Commission. Available at: http://goo.gl/66H3a.

Garcia, Ignacio (2012) 'A Brief History of Postediting and of Research on Postediting', *Revista Anglo Saxonica* 3(3): 291–310.

Google Translator Toolkit. Available at: translate.google.com/toolkit.

Google's Website Translator. Available at: http://translate.google.com/manager/website/?hl=en.

Guerberof, Ana (2009) 'Productivity and Quality in the Post-editing of Outputs from Translation Memories and Machine Translation', *Localisation Focus* 7(1): 11–21.

Guerberof, Ana (2010) 'Project Management and Machine Translation', *Multilingual* April/May. Available at: http://isg.urv.es/library/papers/2010_guerberof.pdf.

Holmes, James S. (1972) 'The Name and Nature of Translation Studies', Third International Congress of Applied Linguistics, Copenhagen. Available at: www.universita-mediazione.com/wp-content/uploads/2012/02/Materiale_Prof_Donadio_31_01_2012.pdf.

Hutchins, W. John (1998) 'The Origins of the Translator's Workstation', *Machine Translation* 13(4): 287–307.

Hutchins, W. John (2013) 'History and Methods of MT', Visiting Seminar, Imperial College London's Translation Unit, 27 February 2013.

ITR (2013) 'Document Translation'. Available at: www.itr.co.uk/documentation-translation.

Jakobsen, Arnt Lykke (2009) 'Instances of Peak Performance in Translation', *Lebende Sprachen* 50(3): 111–116.

Kenny, Dorothy and Andy Way (2001) 'Teaching Machine Translation and Translation Technology: A Contrastive Study', in *Proceedings of the Machine Translation Summit VII, Teaching MT Workshop*. Available at: http://doras.dcu.ie/15830/1/Teaching_Machine_Translation_%26_Translation_Technology.pdf.

Kerr, Dara (2012) 'Google Translate Boasts 64 Languages and 200M Users', *CNET Internet and Media News*, 26 April 2012. Available at: http://news.cnet.com/8301-1023_3-57422613-93/google-translate-boasts-64-languages-and-200m-users.

Kliffer, Michael (2008) 'Post-editing Machine Translation as an FSL Exercise', *Porta Linguarum* 9: 53–67.

Lagoudaki, Elina (2006) 'Translation Memory Systems: Enlightening Users' Perspective. Key Finding of the TM Survey 2006 Carried out during July and August 2006', unpublished doctoral dissertation, Imperial College London.

Lagoudaki, Elina (2009) 'Translation Editing Environments', in *Proceedings of the 12th MT Summit*, 26–30 August 2009, Ottawa, Ontario, Canada. Available at: http://goo.gl/sQHxS.

LionBridge (2013) Language Services. Available at: https://en-gb.lionbridge.com/translation-localization/language-quality.htm.

Madnani, Nitin (2013) iBLEU. Available at: https://code.google.com/p/ibleu.

Makoushina, Julia and Hendrik Kockaert (2008) 'Zen and the Art of Quality Assurance: Quality Assurance Automation in Translation: Needs, Reality and Expectations', in *Proceedings of the 13th International Conference on Translating and the Computer*, 27–28 November 2008, London ASLIB. Available at: http://goo.gl/l08EJ.

Mendez, José (1986) 'Machine Translation in Bureau Service', *Terminologie et Traduction* 1: 48–53.

Mossop, Brian (2006) 'Has Computerization Changed Translation?' *Meta* 51(4): 787–793. Available at: www.erudit.org/revue/meta/2006/v51/n4/index.html.

Mossop, Brian (2007) 'Empirical Studies of Revision: What We Know and Need to Know', *Journal of Specialised Translation* 8. Available at: www.jostrans.org/issue08/art_mossop.pdf.

Mossop, Brian (2007) *Revising and Editing for Translators*, 2nd edition, Manchester: St. Jerome Publishing.

Multitrans XLIFF Editor. Available at: http://multicorpora.com/products-services/options-and-add-ons/xliff-editor.

Munro, Robert (2010) 'Crowdsourced Translation for Emergency Response in Haiti: The Global Collaboration of Local Knowledge', in *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*, 31 October–4 November 2010, Denver, Colorado, the United States of America. Available at: http://amta2010.amtaweb.org/AMTA/papers/7-01-01-Munro.pdf.

O'Brien, Sharon (2002) 'Teaching Post-editing: A Proposal for Course Content', in *Proceedings of 6th EAMT Conference Workshop: Teaching Machine Translation*, 14–15 November 2002, Manchester, UK, 99–106. Available at: http://mt-archive.info/EAMT-2002-OBrien.pdf.

O'Brien, Sharon (2008) 'Processing Fuzzy Matches in Translation Memory Tools: An Eye-tracking Analysis', in Susanne Göpferich, Arnt Lykke Jakobsen, and Inger M. Mees (eds) *Looking at Eyes: Eye*

Tracking Studies of Reading and Translation Processing, Copenhagen: Copenhagen Business School, 79–102.

O'Brien, Sharon (2010) 'Introduction to Post-Editing: Who, What, How and Where to Next?'. *Conference paper Association of Machine Translation 2010*, 31 October – 4 November 2010, Denver, Colorado, the United States of America. Available at http://amta2010.amtaweb.org/AMTA/papers/6-01-ObrienPostEdit.pdf.

O'Brien, Sharon (ed.) (2011) *Cognitive Explorations of Translation*, London: Bloomsbury.

Och, Franz (2012) 'Breaking down the Language Barrier—Six Years in', Online blog post, Google official blog. Available at: http://googleblog.blogspot.be/2012/04/breaking-down-language-barriersix-years.html.

PoEdit. Available at: http://sourceforge.net/projects/poedit.

Post-Editing Tool. Available at http://pers-www.wlv.ac.uk/~in1676/pet/.

Pym, Anthony (2011) 'What Technology Does to Translating', *The International Journal for Translation and Interpreting Research* 3(1): 1–9. Available at: www.trans-int.org/index.php/transint/article/viewFile/121/81.

QA Distiller. Available at: www.qa-distiller.com.

Rasmussen, Kirsten Wölch and Anne Schjoldager (2011) 'A Survey of Revision Policies in Danish Translation Companies', *Journal of Specialised Translation* 15: 87–120. Available at: http://www.jostrans.org/issue15/art_rasmussen.pdf.

Snover, Matthew and Bonnie J. Dorr (2006) 'A Study of Translation Edit Rate with Targeted Human Annotation', in *Proceedings of the 7th Biennial AMTA Conference*, 8–12 August 2006, Cambridge, MA, USA, 223–231. Available at: http://goo.gl/33Et1.

SDL Trados Studio (2011) Available at: www.sdl.com/products/sdl-trados-studio.

Somers, Harold L. (2003) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins.

Somers, Nick (2001) 'Revision – Food for Thought', *Translation Journal Online* 5:1. Available at: https://sites.google.com/site/penidea/revision%E2%80%94foodforthought.

SRT Subtitler. Available at: http://sourceforge.net/projects/srt-tran/?source=directory.

TAUS (2010) 'MT Post-editing Guidelines'. Available at: www.translationautomation.com/images/stories/guidelines/taus-cngl-machine-translation-postediting-guidelines.pdf.

TAUS. Available at: www.translationautomation.com.

Transifex. Available at: www.transifex.com.

Transolution XLIFF Editor. Available at: http://sourceforge.net/projects/eviltrans.

Transperfect (2013) 'Human Translation Services'. Available at: www.transperfect.com/services/translation.html.

van de Poel, Kris, W.A.M. Carstens, and John Linnegar (2012) *Text Editing: A Handbook for Students and Practitioners*, New York: Academic and Scientific Publishing.

Verbeke, Charles A. (1973) 'Caterpillar Fundamental English', *Training and Development Journal* 27(2): 36–40.

Yamagata Europe (2013) *Quality Assurance*. Available at: www.yamagata-europe.com/en-gb/page/520/iso-certification.

31

INFORMATION RETRIEVAL AND TEXT MINING

Kit Chunyu

CITY UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Nie Jian-Yun

UNIVERSITY OF MONTREAL, CANADA

Information retrieval

Introduction

Anyone who has ever looked for any information via a web search has in fact experienced the most popular and powerful means of information retrieval ever available in human history. The World Wide Web has become the largest source of information on earth. Without a powerful web search engine for access to such a massive volume of data, one would find no way out of the problem of the information explosion in this information era. The information retrieval first envisioned by Bush (1945) as 'an enlarged intimate supplement' to a user's memory that 'may be consulted with exceeding speed and flexibility' has become part of our daily life that is characterized by extensive use of the World Wide Web (Berners-Lee *et al*. 1992).

Although web users search through the Web for required content in various kinds of media, including text, graphics, audio and video, canonical information retrieval deals only with texts. Other types of content apart from text are usually retrieved through their associated (or surrounding) texts, such as title, author, caption and/or other kinds of description. The term *information retrieval* (IR) was first coined by Mooers in his 1948 MIT master's thesis and subsequently introduced into the literature of documentation (Mooers 1950; Swanson 1988; Garfield 1997). However, IR as is generally recognized today deals with full text search instead of reference retrieval relying merely on certain specific types of information such as author, title and some keywords about a document (Sparck Jones 1981). The field started much later in the late 1950s, marked by the International Conference on Scientific Information held in Washington in 1958. IR can be defined in slightly different ways using similar terms. For example,

> Information retrieval is often regarded as being synonymous with document retrieval and nowadays, with text retrieval, implying that the task of an IR system is to retrieve documents or texts with information content that is relevant to a user's information need.
>
> (Sparck Jones and Willett 1997)

> Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).
>
> (Manning *et al.* 2008)

Information need is what a user intends to look for and a *query* is an (approximate) expression of information need in the form of free text to be input into an IR system to begin a search.

The development of IR so far can be divided into three stages. The first, roughly from the late 1950s to the mid-1970s, is a period of hatching and testing key ideas and basic techniques, including various IR models. The second, from the mid-1970s to the emergence of commercial search engines on the Web in the early 1990s, develops and advances operational systems that can cope with a massive volume of texts growing alongside computer capacity in terms of both storage and computing power, and puts them in large-scale evaluation in a competitive manner (Harman 1993; Voorhees and Harman 2005). The third, from the mid-1990s onward, is featured by the development of a web search for practical use on the Web, especially those search engines relying more on term-weighting schemes based on the cross-linkage of web pages, e.g., HITS (Kleinberg 1998) and PageRank (Brin and Page 1998), the most famous one.

General architecture of an IR system

The key components in the general architecture of an IR system are illustrated in Figure 31.1. The goal of IR is to find in a document collection what a user intends to find according to the information need expressed in an input query. A non-trivial task preceding that is to acquire and maintain a document collection of a certain size that demands automated means of

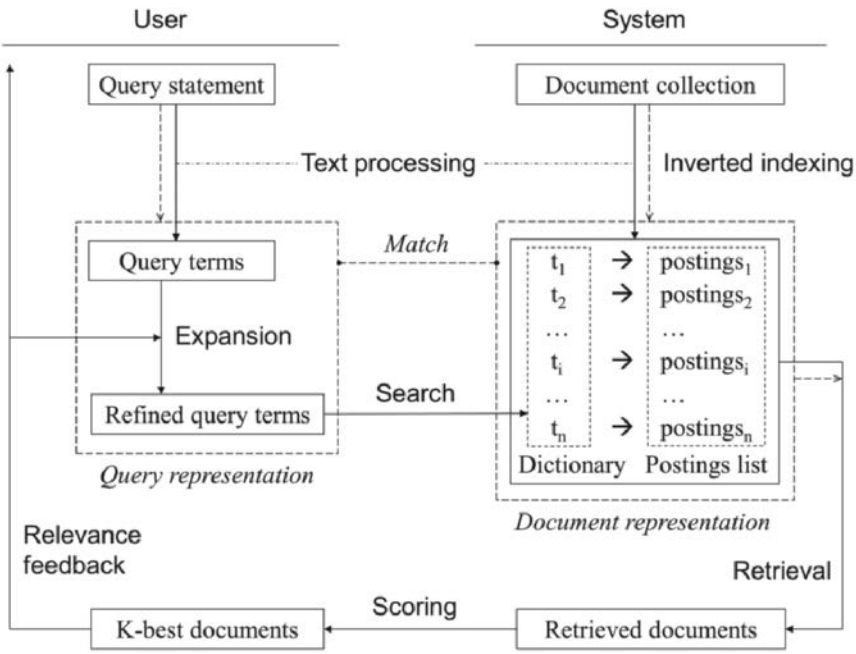


Figure 31.1 Key components of an information retrieval system

crawling, e.g., the collection of all webpages from the Web. It certainly demands adequate computing power and sufficient storage space besides advanced techniques (e.g., text processing and web crawling) for support. To focus on the core issues of IR, however, one may assume – as most researchers did in the past – the availability of such a document collection as the starting point of IR. There have been a good number of standard test collections of very large size for IR evaluation since the small but pioneering Cranfield collection, among which the most influential ones include TREC,[1] NTCIR,[2] and CLEF.[3]

Conceptually, an IR system achieves its goal by matching a user query statement against each document in the document collection, as depicted by dash lines in Figure 31.1. In order to facilitate such matching, necessary text processing has to be first applied to turn the texts in question into a comparable representation. The simplest and most popular form of such representation is called the *bag of words* model, which represents a text as a set of words, known as *index terms*. Even so, however, matching a query, as a bag of words, against each document, as another bag of words, one after another in a sequential manner throughout a large document set, is of too low efficiency to make a practically usable IR system. A technique generally used from a very early stage of IR (Firth 1958a, b; Nolan 1958; Leibowitz *et al*. 1958; as cited in Moore 1961) is to construct an *inverted index*, also known as *inverted file* or *inverted list*. An inverted index stores a list of documents (postings) for each index term, as we will see in more detail in the following subsection. With such an inverted index, the matching for a query is realized by combining the postings of different query terms, as depicted by solid lines in Figure 31.1. As a user's query statement is usually a very short description of the information needed (to be convinced of this, one only has to think about the two to three word queries generally used in web searches), *query expansion* or *query reformulation/rewriting* can be performed in order to enrich the query so as to retrieve more relevant documents that do not contain the same words as the query. Another way to refine the query is *relevance feedback*, which is aimed at formulating a better expression of information need by incorporating into the current query a few related terms extracted from the documents that are judged relevant by the user, or retrieved in the top of the first round of retrieval.

Indexing and inverted index

As mentioned before, search in IR has to be efficient given a very large number of documents. Inverted index, hailed as 'the first major concept in information retrieval' by Manning *et al*. (2008), was specifically devised to address this issue.

Given a document set, an inverted index can be built in a way as simple as compiling a list of postings for each index term recording all documents that contain it. The result of this compilation gives two parts of an inverted index: a *dictionary* consisting of all index terms, and a set of *postings lists*, each of which is associated to a particular term, as illustrated in Figure 31.1. In practice, a postings list is used to hold much more information than a list of bare document IDs, including, for instance, the number of occurrences of the term in each document that contains it, known as term frequency *tf*, their positions in the document (used in modern IR to calculate the proximity of query terms in a document), the total number of its occurrences in the whole document collection, known as collection frequency *f*, the number of documents that contains the term, known as its document frequency *df*, and so on. A postings list for position index may take a form as this:

$$term \rightarrow \{f, df\text{: } [\mathrm{ID}_1,\ tf_1\text{: } (\mathrm{pos}_1^1,\ \mathrm{pos}_2^1, \cdots)];\ [\mathrm{ID}_2,\ tf_2\text{: } (\mathrm{pos}_1^2,\ \mathrm{pos}_2^2, \cdots)]; \cdots\}$$

Accordingly, to search for documents containing two given terms is to intersect their postings lists for a set of common document IDs, and if needed, to determine whether they form a phrasal index term by examining their adjacency according to their positions.

Given the large scale of data in the dictionary and the postings list, a really efficient way of index construction and compression is needed for practical IR. Witten *et al.* (1999) provide a comprehensive coverage of these topics. Some more advanced treatments can be found elsewhere in the literature, e.g., the single-pass in-memory indexing (SPIMI) in Heinz and Zobel (2003) and a number of efficient storage allocation schemes, especially, the arrival rate scheme, in Luk and Lam (2007). Zobel and Moffat (2006) is recommended by Manning *et al.* (2008) 'as an in-depth and up-to-date tutorial on inverted indexes, including index compression'.

An important question in indexing is to determine which terms are to be kept as index terms. Not all the words in a language are deemed meaningful, i.e. bearing semantic meaning. Typical examples are function words that exist in any natural language. For example, we have 'a', 'an', 'the', 'to' etc. in English. A simple way to exclude these words from the index is to put them into a *stop list*.

Another important issue in natural language is that words in different forms may have the same or a similar meaning. For example, 'computer' and 'computing' are related to similar concepts. Therefore, one needs to normalize the word forms found in documents and queries. Lemmatization and stemming are two specific forms of normalization to reduce morphological variants. The former maps inflectional variants of a word to its *base form* or *lemma*, e.g., {'go', 'went', 'gone', 'goes', 'going'} to 'go'. The latter reduces words to their stems by removing their affixes, particularly, their suffixes, conflating a family of derivationally related words into an equivalence class – their stem, e.g., {'stems', 'stemming', 'stemmer', 'stemmers', 'stemmed'} to 'stem'. Computer programs for these kinds of processing are known as *lemmatizers* and *stemmers*, respectively. In principle, accurate morphological analysis is needed in order to support accurate identification of the lemma or the stem of a given word. In practice, however, stemming can be done by conditional transformation rules that successively transform or remove suffixes, as in the Porter stemmer[4] (Porter 1980), which is the most popular stemmer for English. This method has been extended to a number of other languages.

Term weighting and IR models

An IR model defines the representation of query and document, e.g., as a bag of words or even a graph with words as vertices, and the way the relevance of a document to a query is quantified or determined. The *Boolean model* represents a text as a conjunction of terms and a query as a Boolean expression of terms, e.g., a query q = *inverted* AND *file* AND (NOT *inverse*), and accordingly Boolean retrieval is to return documents that satisfy a given query, involving no term weighting in principle, and no document ranking. In practical uses of IR, document ranking is crucial and a good IR system should rank documents according to their relevance to the query.

An intuitive way of scoring the relevance of a document to a query is that the more term matches between them, the more relevant the document should be. However, not all the terms are equally meaningful, or representative of important and specific content. A widely held intuition for term weighting is to assume that frequent terms in a document (or a query) are important, and terms that do not appear in many other documents are specific. The *tf-idf* weighting schema combines both factors as follows:

$$tf\text{-}idf_{t,D} = tf_{t,D} \times idf_t$$

$$idf_t = \log \frac{N}{df_t}$$

where $tf_{t,D}$ – term frequency – is the frequency of term t in document D;[5] N is the total number of documents in the collection in question, df_t the document frequency of t (i.e., the number of documents that contain t), and idf_t the inverse document frequency (Sparck Jones 1972). A simple document scoring function can be defined as follows:

$$\text{score}(Q,D) = \sum_{t \in Q} \text{tf}_{t,Q} \times tf\text{-}idf_{t,D}$$

Many other IR models have been developed, in which documents are scored in different ways according to different principles.

The *vector space model* (Salton *et al.* 1975) treats queries and documents each as a vector in a high dimension space, in which each term is weighted using *tf-idf*. The relevance of a document to a query is estimated according to the similarity of their vectors: the greater their similarity, the greater the relevance. The *cosine similarity* may be adopted as a measure for this purpose:

$$\text{sim}(Q,D) = \frac{\vec{V}(Q) \cdot \vec{V}(D)}{\left|\vec{V}(Q)\right| \left|\vec{V}(D)\right|} = \vec{v}(Q) \cdot \vec{v}(D)$$

where the *dot* (or *inner*) *product* of two vectors, say, of k dimensions, is defined as $\vec{x} \cdot \vec{y} = \sum_{i=1}^{k} x_i y_i$, the *length* of a vector as $\left|\vec{x}\right| = \sqrt{\sum_{i=1}^{k} x_i^2}$, and a unit vector as $\vec{v}(x) = \vec{V}(x) / \left|\vec{V}(x)\right|$.

A *probabilistic model* aims at estimating the probability of relevance of a document to a query. The simplest probabilistic model is the *binary independence model* (BIM) (Robertson and Sparck Jones 1976; van Rijsbergen 1979), which assumes that terms are mutually independent of each other. Then, a document is ranked with the log *odds* of the event that the document is relevant (R) to a query vs. its being irrelevant ($\overline{R}$).

$$\log O(R|\, Q,D) \doteq \log \frac{P(R \mid Q,D)}{P(\overline{R} \mid Q,D)}$$

By the Bayes rule, we have

$$\log O(R|\, Q,D) = \frac{P(D|\, Q,R)P(R \mid Q)}{P(D|\, Q,\overline{R})P(\overline{R} \mid Q)} \propto \log \frac{P(D \mid Q,R)}{P(D \mid Q,\overline{R})},$$

for $\frac{P(R \mid Q)}{P(\overline{R} \mid Q)}$ is a constant. Assuming independence between terms, a document can be represented as a set of independent events – the presences and absences of terms. Let p_t and u_t represent respectively $P(t$ is present in $D \mid Q,R)$ and $P($t is present in $D \mid Q,\overline{R})$, and assume that only terms appearing in the query have an impact on the document's relevance, we have

$$\log O(R \mid Q,D) \propto \log \frac{\prod_{t \in Q \cap D} p_t \cdot \prod_{t \in Q \setminus D} (1-p_t)}{\prod_{t \in Q \cap D} u_t \cdot \prod_{t \in Q \setminus D} (1-u_t)}$$

Table 31.1 Contingency table of term occurrences

Number of documents	*Relevant*	*Irrelevant*	*Total*
Containing t	r	$df_t - \mathrm{r}$	df_t
Not containing t	$\mathrm{R} - \mathrm{r}$	$N - df_t - (\mathrm{R} - \mathrm{r})$	$N - df_t$
Total	R	$N - \mathrm{R}$	N

$$\log O(R \mid Q, D) \propto \log \frac{\prod_{t \in Q \cap D} p_t \cdot \prod_{t \in Q \setminus D} (1 - p_t)}{\prod_{t \in Q \cap D} u_t \cdot \prod_{t \in Q \setminus D} (1 - u_t)}$$

Adding a document-independent constant $\log \left(\prod_{t \in Q} \frac{1 - u_t}{1 - p_t}\right)$, we turn it into a neater form as follows.

$$\log O(R \mid Q, D) \propto \log \frac{\prod_{t \in Q \cap D} p_t \cdot \prod_{t \in Q \cap D} (1 - u_t)}{\prod_{t \in Q \cap D} u_t \cdot \prod_{t \in Q \cap D} (1 - p_t)} = \sum_{t \in Q \cap D} \log \frac{p_t (1 - u_t)}{u_t (1 - p_t)}$$

The logarithm in the sum is the term weight w_t for term t in the model. Given the contingency table of document counts in Table 31.1, we have $p_t = \mathrm{r}/\mathrm{R}$ and $u_t = (df_t - \mathrm{r})/(N - \mathrm{R})$. Accordingly, we have the following term weight, where 0.5 is added for smoothing (Robertson and Sparck Jones 1976).

$$w_t = \log \frac{\mathrm{r}\left(N - df_t - (\mathrm{R} - \mathrm{r})\right)}{\left(df_t - \mathrm{r}\right)\left(\mathrm{R} - \mathrm{r}\right)} \cong \log \frac{(\mathrm{r} + 0.5)\left(N - df_t - \mathrm{R} + \mathrm{r} + 0.5\right)}{\left(df_t - \mathrm{r} + 0.5\right)\left(\mathrm{R} - \mathrm{r} + 0.5\right)}$$

Some more sophisticated probabilistic models than BIM can be found in the literature, e.g., van Rijsbergen (1979) and Fuhr (1992). In practice, the contingency table is rarely available. To cope with this problem, Croft and Harper (1979) assume that p_t is the same for all query terms and hence $p_t/(1 - p_t)$ is a constant, and that almost all documents in a large collection are irrelevant to a query, giving an estimation of u_t by df_t/N (for t appears in df_t irrelevant documents among N). Accordingly, we have the following scoring function, with 0.5 for smoothing, which appears to be a variant of *idf* weighting.

$$\log O(R \mid Q, D) \propto \sum_{t \in Q \cap D} \log \frac{N - df_t}{df_t} \cong \sum_{t \in Q \cap D} \log \frac{N - df_t + 0.5}{df_t + 0.5}$$

As summarized in Robertson and Sparck Jones (1994), '[t]he idea of term weighting is selectivity: what makes a term a good one is whether it can pick any of the few relevant documents from the many non-relevant ones', and there are three kinds of data source available for weighting: (1) collection frequency: N, df_t and their combination into $idf_t = \log N / df_t$; (2) term frequency: $tf_{t,D}$; and (3) document length: $|D|$ and its ratio to average document length $|D|/L_{ave}$. They can be combined into a combined weight

$$w_{t,D} = \frac{idf_t \times (k_1 + 1) tf_{t,D}}{k_1 \left((1 - b) + b \frac{|d|}{L_{ave}} \right) + tf_{t,D}}$$

with the tuning parameters k_1 and b to calibrate the scaling of term frequency and document length, respectively. The idf_t certainly can be substituted with another one of its variants, e.g., the one above with 0.5 for smoothing. Among many options for term weighting, two widely used scoring schemes, namely, the *Okapi weighting* (Robertson *et al.* 1999) and the *pivoted normalization weighting* (Singhal *et al.* 1996, 1999), opt for the following term weights, and then combine three factors with the aid of the constant tuning parameters k_1 (between 1.0 and 2.0), b (usually 0.75), k_3 (between 0 and 1000) and s (usually 0.2) for scaling purpose.

Okapi:
$$\sum_{t\in Q\cap D} \log\frac{N-df_t+0.5}{df_t+0.5}\times\frac{(k_1+1)tf_{t,D}}{k_1\left((1-b)+b\dfrac{|D|}{L_{ave}}\right)+tf_{t,D}}\times\frac{(k_3+1)tf_{t,Q}}{k_3+tf_{t,Q}}$$

Pivoted normalization:
$$\sum_{t\in Q\cap D} \log\frac{N+1}{\mathrm{df}_{t,D}}\times\frac{1+\ln\left(1+\ln\, tf_{t,D}\right)}{(1-s)+s\dfrac{|D|}{L_{ave}}}\times tf_{t,Q}$$

In contrast to heuristic term weighting, *language modeling* provides a more principled approach to IR (Zhai 2008). Exploring the effective use of various language models in IR has been an active area of research since Ponte and Croft (1998). Instead of estimating document relevance, this approach aims at ranking documents according to the likelihood of a document D being what is looked for given a query Q, i.e.,

$$P(D|Q)=\frac{P(Q|D)P(D)}{P(Q)}\propto P(Q|D)P(D)$$

where $P(Q)$ is a constant. The document prior $P(D)$ can be used to favor some special text features or simply assumed, for the sake of simplicity, to be a uniform distribution. Thus, $P(Q|D)$ becomes the choice of scoring for ranking. If we opt for a unigram model, this probability can be decomposed, by the independence assumption, into that of each query term, which is then estimated by $P(t|\theta_D)$ with a language model θ_D derived from D. A popular choice for parameter estimation is the maximum likelihood estimator (MLE), which starts from using the relative frequency of a term in given data. In this way, we have

$$P(Q|D)=\prod_{t\in Q}P(t\mid D)\cong\prod_{t\in Q}P_{\mathrm{ML}}(t|\theta_D)=\prod_{t\in Q}\frac{tf_{t,D}}{|D|}$$

where $|D|$ denotes document length in number of words. This is called the *query likelihood model*, which ranks documents according to the probability that a query is generated by the model of each document. Accordingly, we can take its logarithm as a scoring function. Since a document is usually not large enough to train reliable parameters, an issue known as the *data sparseness* problem, smoothing is inevitably one of the most critical issues in language modeling for IR. Appropriate smoothing handles not only the zero probability of a query term unseen in a document but also the problem of overestimated probability for low-frequency terms, especially those occurring only once mostly by chance. A typical method for this is *linear interpolation*, also referred to as Jelinek-Mercer smoothing, to mix a document model θ_D with the collection model θ_C that is trained on the whole document collection:

$$P(t \mid D) = \lambda P_{\mathrm{ML}}(t \mid \theta_D) + (1-\lambda) P_{\mathrm{ML}}(t \mid \theta_C)$$

where $0 < \lambda < 1$. The setting of λ is critical and has to be carefully tuned through training. Then, we have the following estimation for how likely a document is what a user looks for with a particular query:

$$P(D \mid Q) \propto P(D) \prod_{t \in Q} [\lambda P_{\mathrm{ML}}(t \mid \theta_D) + (1-\lambda) P_{\mathrm{ML}}(t \mid \theta_C)]$$

A more general probabilistic similarity model for retrieval is formulated using the Kullback-Leibler (KL) divergence between the respective *query* and *document likelihood models* as follows:

$$-D(\theta_Q \mid\mid \theta_D) = -\sum_{t \in V} P(t \mid \theta_Q) \log \frac{P(t \mid \theta_Q)}{P(t \mid \theta_D)} \propto \sum_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_D)$$

where V is the vocabulary involved. The simplification made in the last step is due to the fact that $\Sigma_{t \in V} P(t \mid \theta_Q) \log P(t \mid \theta_Q)$ is a document-independent constant and can be ignored for document ranking. This model comparison approach is reported to outperform the approaches that use only a document or query model (Lafferty and Zhai 2001). Another language model is the *translation model* introduced into IR by Berger and Lafferty (1999) to deal with the problem of expression deviation (e.g., the use of synonyms) between queries and documents. It facilitates retrieval of documents containing alternativee terms with similar meanings to query terms by incorporating into the IR model a translation model, namely, a conditional probability $T(\cdot \mid \cdot)$ between terms:

$$P(Q \mid \theta_D) = \sum_{t \in Q} \prod_{w \in V} P(w \mid \theta_D)\ T(t \mid w)$$

This model is widely used in IR to incorporate relationships between terms, including in cross-language IR in which terms t and w are in two different languages.

In the traditional IR, each document is considered in isolation and its score to a query is only determined by its content words. The prior of a document (i.e. $P(D)$) is assumed to be uniform. This is counterintuitive, especially in the context of web searches. Some documents may be more popular, of higher quality or authority, than others, and therefore are preferred by users. Hyperlinks between web documents provide a way to estimate a document's popularity or authority: Each link to a document can be considered a vote in favor of it; the more votes a document receives, the more it is weighted. This idea is cast in the following PageRank algorithm (Brin and Page 1998): Assuming that a document d_i has links from a set $IN(d_i)$ of documents, its PageRank score $PR(d_i)$ is determined by

$$PR(d_i) = \frac{1-d}{N} + d \sum_{d_j \in IN(d_i)} \frac{PR(d_j)}{L(d_j)}$$

where d is a dumping factor, which is usually set at 0.85, N the total number of documents, and $L(d_j)$ the number of outbound links from d_j. This formula is used to update the PR scores (initially set to $1/N$) of documents (or pages) iteratively until reaching a stationary point. The resulting score $PR(d_i)$ can be used to estimate the prior of a document $P(d_i)$.

In a similar way, HITS (Kleinberg 1998) computes two scores for a page: authority and hub. The former estimates the value of the content of a page in terms of how many other pages (or hubs) link to it, and the latter the value of its links to other pages. An authoritative page means a page with many other pages referring to it, while a hub a directory page with many links to other authoritative pages. Starting with an initial value 1 for each page, the scores of authority and of hub are calculated as follows in a mutual recursion:

$$Auth(d_i) = \alpha \sum_{d_j \in IN(d_i)} Hub(d_j)$$

$$Hub(d_i) = \beta \sum_{d_j \in OUT(d_i)} Auth(d_j)$$

where $IN(d_i)$ and $OUT(d_i)$ denote the inlink and outlink pages of d_i, respectively, and α and β are two normalization factors. That is, for a page, its authority/hub score is determined by the sum of the hub/authority scores of its inlink/outlink pages.

In addition to content words and hyperlinks, a number of additional factors can also be utilized in a web search. For example, clickthrough data (the click behavior of users for a given query) is proven to be a useful resource that encodes some relevance relationship between a query and a (clicked) document. It is difficult to incorporate all these factors into a formal IR model. An alternative way is to consider them as defining features for use in combination to predict the relevance score of a document for a query. This has led to the new direction of *learning to rank* (L2R) (Liu 2009; Li 2011). Its original idea comes from Fuhr (1989, 1992), who tried to generalize the earlier probabilistic IR models by using a learning method to learn a probabilistic ranking model. The last decade has witnessed a significant progress in both research and applications of L2R. In general, L2R makes use of known relevance information to train a learning model to optimize the ranking in terms of a loss function, in a way to minimize the expected loss. Various machine learning methods (e.g. SVM) have been adapted for IR problems by transforming a desired ranking order to a list of binary preferences between documents or between document lists. The availability of relevance data from web search, especially, search log data from search engines, has made this supervised approach not only practically feasible but particularly appealing. A particularly strong advantage of this approach is that it treats a document as a bag of features (vs. a bag of terms) in its discriminative (vs. generative) learning. In theory, any useful feature conceivable can be integrated into a learning model for optimization on a large volume of training data with true answers, so as to yield a generalized IR model that subsumes classic IR weighting schemes as its features, especially those already proven to be particularly informative, e.g., *tf*, *idf*, normalized document length, PageRank and HITS scores. Interested readers may find a detailed description of L2R in Liu (2009) and Li (2011).

Evaluation

Comprehensive evaluation of an IR system is more complicated than one thinks at first glance, especially when subjectivity has to be involved in the judgment of document relevance to a user need. Nevertheless, the core of IR evaluation is to measure the retrieval effectiveness of an IR system, which is mostly conducted following the Cranfield paradigm using a *test set* consisting of (1) a set of queries as expressions of information needs and (2) a collection of documents in company with (3) relevance judgments specifying relevant documents to each

query, namely, the *gold standard* or true answers. Besides, a *development test set* may be provided for tuning system parameters towards expected performance. There have been a good number of standard test collections, including the most influential ones mentioned above (in the subsection 'General architecture of IR system').

Table 31.2 Contingency table of retrieved documents

Number of documents	*Relevant*	*Irrelevant*	*Total*
Retrieved	true positive: *tp*	false positive: *fp*	*tp* + *fp*
Not retrieved	false negative: *fn*	true negative: *tn*	*fn* + *tn*
Total	*tp* + *fn*	*fp* + *tn*	

The two most fundamental measures in IR evaluation are precision and recall. *Precision* and *recall* are the proportions of retrieved relevant documents to, respectively, all retrieved documents and all relevant documents. Given the contingency in Table 2, we have

$$P = \frac{tp}{tp + fp},\ R = \frac{tp}{tp + fn}$$

Sometimes, we also use *F-measure*, which is the weighted harmonic mean of precision and recall:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad \text{, or} \quad F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{with} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

where $\alpha \in [0,1]$ and accordingly $\beta^2 \in [0,\infty]$. With $\beta^2 = 1$ (equivalently $\alpha = 0.5$), we have the default balanced F-measure (also called F1-measure) as follows:

$$F_{\beta=1} = \frac{2PR}{P + R}$$

As an IR system returns a ranked list, precision and recall change according to the number (k) of documents one picks from the list. A common practice is to draw a precision-recall curve by considering more and more documents. As k gets larger, we get more relevant documents included in the top-k and a lower precision, resulting in a curve descending along recall: the larger the recall, the lower the precision. To provide a single measure on the quality of IR systems, average precision at 11 points of recall is often used, i.e. the average of the precisions at 0.0, 0.1, …, 1.0 recall (Teufel 2007). Another widely used measure, the *mean average precision* (MAP), is defined as follows:

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|D_q|} \sum_{d \in D_q} P(R_d)$$

where Q is a query set, D_q the set of documents relevant to $q \in Q$, and R_d the set of top-k ranked retrieved documents up to d.

In the scenario of using multiple levels (or labels) of relevance judgment, e.g., {perfect, excellent, good, fair, and bad} each with a score (say, 4−0), as used in most recent works on learning to rank, a popular performance measure is the *normalized discounted cumulative gain* (NDCG) introduced by Järvelin and Kekäläinen (2000, 2002). The NDCG at position k (i.e., over the top-k retrieved documents) for a set of queries Q is defined as

$$NDCG(Q,k) = \frac{1}{|Q|}\sum_{q\in Q} Z_{k,q} \sum_{d\in \text{top}-k} \frac{2^{score(q,d)}-1}{\log_2\left(1+\text{pos}(d)\right)}$$

where *pos* (·) is the position of a document in the top-k list in question, $Z_{k,q}$ is a normalization factor calculated to ensure that the *NDCG* at k for a perfect ranking is 1, and the numerator and denominator of the inner fraction are the *gain* and the *position discount* function, respectively.

Query expansion and relevance feedback

An initial query from a user is often not a good enough expression of information need. One of the main reasons for this is that the same thought can be paraphrased in different ways using different words. The words used in a user query are not always the only and the best search terms used in relevant documents. The goal of query expansion is to extend an original query by incorporating other related terms that could be used in relevant documents. Towards this goal, two key issues need to be dealt with: how to select related terms for addition to a query, and how to combine these terms with those already in the query.

The first obvious way to find related terms to expand a q uery is to use a manually constructed thesaurus. For many language analysis tasks, one needs relations between words or terms. Among the available resources of this kind that are readily useable for IR, a typical example is WordNet[6] (Miller *et al.* 1990) which provides various semantic relations between words and compound terms (e.g. synonymy, hyponymy, hypernymy, etc.). For example, 'data processing system' is a synonymous term with 'computer', which has 'PC' (a-kind-of 'computer') as a hyponym. Using such a resource, one can append synonyms or other types of related term to initial query terms to form an expanded query, e.g., expand the query 'computer' with 'data processing system' and 'PC'. Voorhees (1993, 1994) first attempted to use WordNet in IR experiments on TREC collections, but the expected advantages did not concretize, in that when related terms were added to expand a query, retrieval effectiveness was degraded rather than increased. Careful analyses revealed that coverage and ambiguity were the two main problems that limit the usefulness of this kind of resource in IR. Like other lexical resources, WordNet has only a partial coverage of concepts and terms used in documents and queries, missing many others, and it does not deal with term ambiguity either. For example, 'computer' has two meanings in WordNet: as a machine or as a human expert. When a query containing 'computer' is expanded with all its meanings in WordNet, a certain amount of noise (i.e., unrelated terms) is unavoidably brought in. Some later studies (Mandala *et al.* 1999; Cao *et al.* 2005) obtained positive results by means of selecting or weighting related WordNet terms with the aid of corpus statistics.

Another approach to query expansion, which is widely used in IR, exploits term co-occurrences in the document collection in question (Qiu and Frei 1993), based on the assumption that two terms that co-occur often are likely to be related. It helps improve retrieval effectiveness significantly. Since it is also a topic in text mining for IR, we will present its details later in the section on text mining.

The above approaches, one relying on general lexical knowledge and the other on query-independent co-occurrence statistics, are two typical *global expansion* methods. However, it is often observed that a strong co-occurring term in the collection in question is not always appropriate for use to expand a given query. Consider, for instance, a query on 'Java hotel' and a collection mainly consisting of computer science documents. Most co-occurring terms with 'Java' are Java language related terms, which are inappropriate for use to expand this query. A way to remedy this is to perform local analysis, extracting expansion terms only from top-ranked retrieved documents (Xu and Croft 2000). More specifically, the first round of retrieval identifies a small set of documents, from which a set of related terms are extracted and used as expansion terms. Compared to global methods, this method benefits from a filtering of the documents retrieved with an initial query. In general, the top documents so retrieved are more likely to be related to the query than others in the same collection, and thus the expansion terms extracted from them are more related to the topic of the query.

In contrast to global expansion that may use any other resources, a *local expansion* method expands a query using only documents retrieved for this particular query. Typically, it selects a set of strongly related terms only from top-ranked retrieved documents, using various criteria such as *tf*, *idf*, etc., or co-occurrences with query terms. Since user judgment of relevance is rarely available, the best one can do is to assume the relevance, the so-called pseudo relevance, of a few top-ranked retrieved documents and incorporate their terms into a query. Query expansion carried out this way is called *pseudo* (or *blind*) *relevance feedback*. Experiments confirm that local expansion of this kind outperforms global expansion. However, it has a drawback of performing two rounds of retrieval, which may not be practical in a real situation, e.g., web search.

In general, query expansion may exploit all available relations between terms (e.g., thesaurus, ontology) and all possible connections between queries and documents (e.g., user relevance judgments, clickthrough data) or between documents (e.g., hyperlinks). A special case is *relevance feedback*, in which a user is asked to judge the relevance (or irrelevance) of some returned documents for a query, such that the query can be extended to a new one by incorporating some terms extracted from relevant documents while excluding those from irrelevant documents. As mentioned above, true relevance feedback is usually unavailable, and the best we can resort to is pseudo-relevance feedback, which assumes the relevance of top-ranked documents. User clickthrough is another form of implicit relevance feedback from users: when choosing to click on a document, a user often (although not always) considers it to be potentially relevant. A simple way to exploit clickthrough data is to assume that terms in clicked documents are related to those in the query in question. This idea has motivated a number of studies on mining term relationships from clickthrough data (Wen *et al.* 2001; Cui *et al.* 2002; Baeza-Yates and Tiberi 2007; Gao *et al.* 2010), which will be presented in more detail in the section on text mining. Similarly, anchor texts pointing to a document also reflect some relations of them with the terms in the document.

The next key issue in query expansion is how to combine selected expansion terms with original ones in a query. Most approaches are based on, or derived from, the Rocchio formula (1965/1971) popularized by the SMART system (Salton 1971), which was developed for relevance feedback to a vector space model. Given an initial query vector $\vec{q}_0$, a set $\mathcal{R}$ of relevant documents and a set $\overline{\mathcal{R}}$ of irrelevant documents (as judged by users), the new query vector to be produced by the relevance feedback with these documents is defined as:

$$\vec{q}_1 = \alpha \vec{q}_0 + \frac{\beta}{|\mathcal{R}|} \sum_{\vec{d} \in \mathcal{R}} \vec{d} - \frac{\gamma}{|\overline{\mathcal{R}}|} \sum_{\vec{d} \in \overline{\mathcal{R}}} \vec{d}$$

where α, β, and γ are the weights to balance the three vectors in reflecting the true use need. The meaning of this formulation is straightforward: it moves the query closer towards the centroid of relevant documents and away from that of the irrelevant documents, in hopes that there are more relevant documents around the former centroid than any other place and hence more of them can be retrieved by the new query. To reflect the observation that relevant documents are more useful than irrelevant ones, most IR systems have $\alpha = 1$ and $\beta > \gamma$ (e.g., $\beta = 0.75$ and $\gamma = 0.15$).

The same formula has been used for pseudo-relevance feedback, with top-k retrieved documents as $\mathcal{R}$, and no irrelevant document (i.e. $\gamma = 0$). Unlike true relevance feedback that usually leads to improved retrieval effectiveness, pseudo-relevance feedback often produces less consistent results: when top retrieved documents are truly related to an initial query, it often yields better effectiveness; otherwise, a query drift phenomenon (i.e., the resulting new query departures from the original intent of the initial query) is often observed. An example to illustrate this problem is the query 'Java public transportation', which may retrieve many documents about 'Java programming language'—using the terms in these documents for query expansion inevitably makes the resulting query drift away from the originally intended information need.

In the framework of probabilistic IR, when true relevance feedback is available, it is used to build (or update) the contingency table so as to obtain more precise probability estimations. When pseudo-relevance feedback is used, it may be exploited as follows: Let $\mathcal{R}$ be the set of top-ranked documents that are retrieved from the document collection C, of size N, and $\mathcal{R}_t \subset \mathcal{R}$ be the subset of documents containing term t, and assume that the remaining $N - |\mathcal{R}|$ documents in the collection are irrelevant, the p_t and u_t defined above can be estimated as

$$p_t = \frac{|\mathcal{R}_t|}{|\mathcal{R}|}, \quad u_t = \frac{\mathrm{df}_t - |\mathcal{R}_t|}{N - |\mathcal{R}|}$$

Accordingly, the term weight for t is

$$w_t = \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \log \frac{|\mathcal{R}_t|}{|\mathcal{R}| - |\mathcal{R}_t|} \cdot \frac{N - |\mathcal{R}| - (df_t - |\mathcal{R}_t|)}{df_t - |\mathcal{R}_t|}$$

Since $N \gg |\mathcal{R}|$ and $df_t \gg |\mathcal{R}_t|$, it can be approximated as follows, with 0.5 for smoothing (as we did before):

$$w_t \cong \log \frac{|\mathcal{R}_t| + 0.5}{|\mathcal{R}| - |\mathcal{R}_t| + 0.5} + \log \frac{N - df_t + 0.5}{df_t + 0.5}$$

where the second log may be further approximated by $idf_t = \frac{N}{df_t}$, assuming $N \gg df_t$. Then, comparing it with log $O(R|Q,D)$ formulated earlier, we can see that relevance feedback adds the first log to t's weight.

Query expansion and pseudo-relevance feedback have also been widely used in language models for IR. Recall that in the formulation using KL-divergence, a language model is built, respectively, for a query and for a document. Query expansion aims at building a new language model for a user query. It is typically implemented as an interpolation between an original query model θ_{Q_0} and a feedback (or expansion) model $\theta_{\mathcal{R}}$ that accommodates new terms from top-ranked documents or a thesaurus:

$$P(t \mid \theta_{Q_1}) = (1-\alpha)\mathrm{P}(t \mid \theta_{Q_0}) + \alpha P(t \mid \theta_{\mathcal{R}})$$

where $\alpha \in [0,1]$ is a parameter to control the contribution of the two models.

There are various ways to formulate an expansion model $\theta_{\mathcal{R}}$. If a set of term relationships have been determined (e.g., extracted from a document collection or obtained from a thesaurus), and let the relationship between two terms t_i and t be expressed as a probability function $P(t \mid t_i)$, then the model $\theta_{\mathcal{R}}$ can be defined as:

$$P(t \mid \theta_{\mathcal{R}}) = \sum_{t_i \in V} P(t \mid t_i) P(t_i \mid \theta_{Q_0})$$

In theory, this formulation could also be applicable to pseudo-relevance feedback, using the set of top-ranked documents to construct the model $\theta_{\mathcal{R}}$, say, using MLE. However, it does not work well, because top-ranked documents contain many terms that are not necessarily related to the query in use. In addition, they also contain many terms that are generally frequent in a language and hence ineffective in the discrimination of relevant documents from irrelevant ones. A better way to construct $\theta_{\mathcal{R}}$ is to isolate a part of term distribution mass in feedback documents that is different from a general language model, i.e., specific to the query. This idea is implemented in the mixture model (Zhai and Lafferty 2001), which assumes that the feedback documents $\mathcal{R}$ are generated by a mixture of two models: a query-specific feedback (or topic) model $\theta_{\mathcal{R}}$ and a general (or background) language model, which is usually approximated by the collection model θ_C. The log likelihood of $\mathcal{R}$ under this model is

$$\log P(\mathcal{R} \mid \theta'_{\mathcal{R}}) = \sum_{t \in V} f_{t,\mathcal{R}} \log[(1-\lambda) P(t \mid \theta_{\mathcal{R}}) + \lambda P(t \mid \theta_C)]$$

where f_t, $\mathcal{R}$ is the frequency of t in $\mathcal{R}$, and $\lambda \in [0,1]$ the weight for the background model. With λ set to a constant, the topic model $\theta_{\mathcal{R}}$ can be trained in a way to use the Expectation-Maximization (EM) algorithm to maximize the above log likelihood (see Zhai and Lafferty (2001) for detailed EM updates).

Alternatively, their divergence minimization method pursues a $\theta_{\mathcal{R}}$ as close to the language model of each document in $\mathcal{R}$ and as far away from the background model θ_C as possible, using C as an approximation of the set of nonrelevant documents:

$$\theta_{\mathcal{R}} = \arg\min_{\theta} \frac{1}{|\mathcal{R}|} \sum_{d \in \mathcal{R}} D(\theta \,||\, \theta_d) - \lambda D(\theta \,||\, \theta_C)$$

where $\lambda \in [0,1]$ is a weighting parameter. Once determined, this $\theta_{\mathcal{R}}$ can be interpolated with θ_{Q_0} to produce a new query model θ_{Q_1} for ranking documents.

However, it is showed in Cao *et al.* (2008) that terms extracted by the above methods are not always beneficial when added to a query. A case study on three TREC collections reveals that only about 17 percent of terms so extracted are truly useful, while about 30 percent of them are harmful, i.e. lowering retrieval effectiveness. It is thus necessary to perform a further selection or reweighting of expansion terms among all candidates. A classification method, based on a set of features, can be used to determine if a candidate term is a good one, in order to further improve retrieval effectiveness.

Another well-known method to incorporate feedback documents is the *relevance model* (Lavrenko and Croft 2001). Its basic idea is to consider top-ranked retrieved documents to be i.i.d. (independent and identically distributed) samples of relevance. Using these samples, a relevance model $\theta_{\mathcal{R}}$ is defined as follows:

$$P\left(t|\,\theta_{\mathcal{R}}\right) = P\left(t|\,Q,\mathcal{R}\right) = \sum_{d\in\mathcal{R}} P(t\mid\theta_d)P(\theta_d\mid Q) \propto \sum_{d\in\mathcal{R}} P(t\mid\theta_d)P(Q\mid\theta_d)P\left(\theta_d\right)$$

where $P(Q|\,\theta_d)$ is indeed the original ranking score of a document. One may further assume a uniform $P(\theta_d)$ for simplification.

There are a number of other approaches to query expansion derived from the above ones. Interested readers can refer to Carpineto and Romano (2012) for a comprehensive survey of them. A few representative ones by means of text mining will be discussed later in the text mining section. In practice, query expansion techniques can be used in alternative forms, including query suggestion, query rewriting, and query reformulation, all of which are intended to suggest a better query formulation. Besides the resources mentioned above for query expansion, the search history of a user can also be analyzed for use to determine which query formulation is preferable. Nonetheless, technical details on these related tasks are not permitted in this chapter due to limited space.

Cross-language information retrieval

A special form of information retrieval is *cross-language information retrieval* (CLIR), which aims at retrieving documents in a language different from that of a query. CLIR research started in the early 1970s (e.g., Salton (1970)) and has been an active area of research since the late 1990s when TREC introduced a cross-language track in 1997 involving English, French and German. NTCIR started CLIR experiments between English and Asian languages (mainly Chinese, Japanese and Korean) in 1999 and CLEF for European languages in 2000. On top of traditional monolingual IR, CLIR has a language barrier, an extra difficulty, to overcome by some means of *translation*, in order to bring documents and queries into a comparable representation as if they were in the same language. One may opt to perform either query translation or document translation. Experiments (Franz *et al.* 1999; McCarley 1999) show that either achieves a comparable level of effectiveness. Nevertheless, query translation is more commonly adopted, for its flexibility in adapting to new languages of interest and the efficiency of translating a smaller amount of texts.

Translation can be performed in several ways using different resources and methods (Oard and Dorr 1996; Nie 2010). The simplest way is to use a bilingual dictionary to turn each query word into its translation words as stored in the dictionary, or into a selection of the translation words based on some coherence criteria (Grefenstette 1999; Gao *et al.* 2001; Liu *et al.* 2005; Adriani and van Rijsbergen 2000). It is certainly very convenient to use a machine translation (MT) system, if available, to translate a query simply as a text into a target language (Franz *et al.* 1999; McCarley 1999; Savoy and Dolamic 2009), so that it can be used as a query in monolingual IR. CLIR is typically cast as a problem of MT + monolingual IR, although several recent studies have started to investigate IR-specific translation, with a focus on examining the utility of MT results (Türe *et al.* 2012; Ma *et al.* 2012; Türe and Lin 2013). Another option is to use parallel and/or comparable texts more directly. From a large corpus of parallel texts, translation relations can be extracted automatically with a statistical translation model for use in CLIR (Nie *et al.* 1998; Nie *et al.* 1999; Kraaij *et al.* 2003). The simplest way

to do so is to train an IBM model 1 (Brown *et al.* 1993), which assumes that word alignment between two parallel sentences is independent of word order and position. This assumption is certainly invalid for translation in general, but it corresponds well to the traditional word-bag model of IR, whose retrieval results are also independent of word order and position. This assumption has been questioned in both IR (Metzler and Croft 2005; Bendersky *et al.* 2010; Shi and Nie 2010) and CLIR (Türe *et al.* 2012; Türe and Lin 2013; Ma *et al.* 2012). It turns out that a more sophisticated phrase-based translation model can produce better query translations and hence lead to better cross-language retrieval results.

In addition to parallel texts, comparable texts (i.e., bilingual or multilingual texts about the same topics) that are available in an even larger amount can also be utilized, e.g., Wikipedia (or news) articles in different languages about the same concepts (or events). A number of studies have attempted to exploit comparable texts to facilitate query translation (Sheridan and Ballerini 1996; Franz *et al.* 1999; Braschler and Schäuble 2000; Moulinier and Molina-Salgado 2003). In general, it is unrealistic to apply the same word alignment process for parallel texts to comparable texts. A more flexible cross-language similarity function is instead needed. However, it is also unrealistic to expect it to work as well as a translation model, as shown in CLIR experiments. Nevertheless, comparable texts can be used not only as a last means for rough translation of user queries, especially in the scenario of no parallel text available, but also as complementary resources to available parallel texts for additional gain in CLIR (e.g., Sadat *et al.* 2003). The use of Wikipedia is a special case of exploiting comparable texts to facilitate CLIR (Nguyen *et al.* 2008): in addition to text contents on similar topics in different languages, its organization structure and concept descriptions can also be utilized to further enhance the mining of translation relations.

Query translation for CLIR is not merely a translation task. It is intended to produce a query expansion effect by means of including multiple and related translation words (Oard and Dorr 1996; Nie 2010). This has been proven useful for both general IR (Carpineto and Romano 2012) and CLIR. Simple use of MT to translate user queries is often not a sufficient solution. A number of recent studies (Türe *et al* 2012; Türe *et al.* 2013; Ma *et al.* 2012) have shown that opening the MT 'blackbox' to allow the use of multiple translation candidates and their appropriate weighting in CLIR is indeed more advantageous than using a single best translation. Last but not least, however, whatever approach to CLIR is opted for, there is an acute need to infer translation relations between words and/or phrases, towards which a very first step is to mine parallel/comparable texts from the Web. A number of attempts to do such mining will be presented in a later section.

Text mining

Text mining (TM) is also known as *knowledge discovery in texts* (KDT; Feldman and Dagan 1995) or *text data mining* (TDM; Hearst 1999), referring to the process and/or the study of the (semi) automated discovery of novel, previously unknown information in unstructured texts. Both TM and IR are aimed at facilitating our access to information, but they differ in a number of ways. What IR returns to a user is some known and overt information that can be directly read off from the documents it retrieves in relevance to the user's query, and the relevance is estimated by computing the similarity of a document and a query or the likelihood that a document is looked for with a query. Unlike IR, TM is not for locating any wanted information in a large collection of texts in response to a query. Instead, the goal of TM is to infer new knowledge, mostly as covert information about facts, patterns, trends or relationships of text entities, which is hidden in, and hence inaccessible via the comprehension and literary interpretation of, texts. Despite no query being involved, TM serves a certain information need, in the sense that the

novel knowledge it uncovers needs to be of good quality for use to serve a particular purpose or application, e.g., term correlation information to facilitate query expansion in IR, and term translation options and respective probabilities to enable statistical machine translation (SMT) and CLIR. If we refer to the content that can be obtained from a text by reading as overt information and to the rest that cannot be so obtained as covert, a clear-cut boundary between IR and TM is that IR accesses the former and TM the latter. It is thus conceivable that any information access to texts beyond the reach of fully fledged IR may have to be facilitated by TM. Serving the general goal of TM to make the covert visible, the development of visualization tools has been an indispensable and popular sub-area of TM since the very beginning.

TM is considered a variation or extension of *data mining* (DM), which is also known as *knowledge discovery in database* (KDD) and whose goal is to find implicit, previously unknown, and non-trivial information, of potential use for certain purpose or interest, from large structured databases. Instead of working on databases, TM works on unstructured texts. The view of TM as a natural extension of DM can be found in the early work by Feldman and Dagan (1995), that once a certain structure or relation can be imposed onto text entities of interest, e.g., a conceptual hierarchy, traditional DM methodologies can be applied. A typical DM process can be conceptually divided into three stages: (1) pre-processing, (2) mining and (3) result validation. The first stage is to prepare a set of target data for a mining algorithm to work on, mostly focusing on data selection, necessary transformation and noise filtering. For the purpose of validation, available data is usually divided into a training and a test set, so that the patterns of interest mined by a mining algorithm from the training set are evaluated on the test set. A common practice of evaluation is to apply the mined patterns to a target application that the mining is aimed at bettering, and then measure the performance gain of using the mined patterns. As in language modeling, a common problem in TM is over-fitting, that the patterns found in training data have a rare or too low a chance to present elsewhere, such as in test data. Drawing on statistical inference and machine learning, a mining approach may fall into one of the following categories:

1 regression, to model data with the least amount of error;
2 anomaly/deviation detection, to identify unusual records or trends in data, e.g., deviations from normal credit card usage, revealing possible frauds;
3 association/dependency modeling, to detect relationships between variables, e.g., customers' purchasing habits, such as items customarily bought together;
4 clustering, to find groups (or categories) and/or structures in data in terms of a certain similarity;
5 classification, to assign known (or predefined) categories or structures to new data;
6 summarization, to infer a more compact representation for data, usually by means of finding regularities in data or estimating the importance of data; and
7 sequencing or sequence pattern mining, to infer significant co-occurring patterns of data items in a data stream.

This names a few among many others. Several representative ones are presented below for a bird's eye view of the whole field.

Categorization, clustering and information extraction

Considering TM as a natural extension of DM, we have text categorization, text clustering and information extraction as typical TM tasks. Interestingly, however, not all scholars agree with this view. For example, Hearst (1999, 2003) holds a purist position against this view while

defining what TM is, for the reason that these tasks do not lead to any genuine discovery of new, heretofore unknown information, in the sense that anything in a text already known to its author is not new! She points out that 'mining' as a metaphor to imply 'extracting precious nuggets of ore from otherwise worthless rock' mismatches the real essence of TM. The best known example of real TM is Swanson's (1987, 1991) work on deriving novel hypotheses on causes of rare diseases from biomedical literature, e.g., the hypothesis of magnesium deficiency as a possible cause of migraine headache. Besides the aforementioned TM tasks, what she also puts under the label of mining-as-ore-extraction include automatic generation of term associations from existing resources (such as WordNet) for IR query expansion (Voorhees 1994) and automatic acquisition of syntactic relations from corpora (Manning 1993).

The gap between Hearst's definition and the work by many researchers in the field suggests that two issues concerning the word 'new' are worth examining: (1) the continuum of newness (or the degree of novelty), e.g., wholly vs. partly new, and (2) new to whom, e.g., to everyone vs. to a particular user (or agent). To avoid too narrow a scope of research, it is necessary to relax the definition of TM to this weaker one: the (semi)automated acquisition of information from texts to enrich or add to an existing pool of information (or knowledge), which at the beginning can be empty or the whole of human knowledge. In this way, Hearst's purist definition becomes a special case, and the two meanings of the term 'text mining', namely, mining texts (or text nuggets) from some resources (e.g., large corpora) and mining hidden (i.e., not directly readable) information from texts, are covered, corresponding to two different types of TM that resort to different methodologies.

The former, that mines text nuggets for critical information or special knowledge, is more fundamental and popular, and relies more on basic text processing operations for recognition of surface string patterns. A simple and typical example of this is to extract strings of particular patterns from texts (such as e-mail addresses, phone numbers, URLs, and so forth from webpages) that are new to an interested user. Information extraction (IE) to find targeted types of information to fill in predefined slots (e.g., an event frame: who did what to whom, where, when, how, why, and so forth) and mining personal data to compose or complement a personal profile are other two examples involving natural language processing of various degrees of complexity. In particular, *named entity recognition* (NER) to identify names of various types (e.g., person, organization, place, and so forth) and their variants (e.g., full names, nicknames, abbreviations, etc.) plays a critical role in IE, underlain by basic natural language processing techniques, e.g., part-of-speech (POS) tagging. Associating a recognized name with its true referent, e.g., 'Ford' with a company or a person, and differentiating concepts under the same word or term may be tackled by means of categorization or clustering, depending on the availability of candidate entities, which usually resorts to advanced statistical inference.

The latter type of TM to dig out concealed information in texts is more challenging and attracts a more serious research effort. It needs to go beyond the basic language processing that supports the former, and rely more on logical reasoning and/or statistical inference. Swanson's aforementioned work demonstrates the effectiveness of logical reasoning. For statistical inference, the starting point is to derive (co-)occurrence frequencies of text units (e.g., words) from a large corpus (i.e., a collection of texts), for use in statistical measurement, test, and/or distribution modeling. In the case of using a machine learning model, the features in use need to be extracted from text units with regard to their context, to produce training data for model training. No matter what methodologies are employed, the two types of TM manifest their common and different characteristics clearly in specific tasks in almost all popular areas of TM in recent years.

Summarization

In language technology, many undertakings that are now viewed as typical branches of TM in fact originated independently of DM and developed into standalone disciplines before the emergence of, or in parallel with, TM. For example, originating from Luhn (1958), *text summarization*, also known as *automatic summarization*, has developed into a popular research area for exploration of various approaches (mainly in two categories: extraction vs. abstraction) to producing various types (e.g., indicative, informative, vs. critical) of summaries with different orientations (i.e., query-based vs. query-independent) in one of two major dimensions (i.e., single- vs. multi-document). Three stages are identified in a full process of summarization (Sparck Jones 1999; Hovy and Lin 1999; Hovy 2005):

1 *topic identification*, to identify the key content of the text(s) to be summarized by identifying the most important text units (including words, phrases, sentences, paragraphs, etc.) in terms of some predefined criterion of importance and returning the *n* best ones in respect to a requested summary length;
2 *interpretation*, to fuse and then represent the identified topics in an abstract representation or formulation (e.g., event template, lexical chain, concept network/relation, etc.), using prior domain knowledge and involving other words or concepts than those in the input text(s); and
3 *summary generation*, to turn the unreadable abstract representation to a coherent summary output in human-readable text form using language generation techniques.

The involvement of interpretation distinguishes an abstraction from an extraction approach. The latter simply extracts the most important portions (e.g., key phrases, sentences, etc.) of text and combines them into a summary through a 'smoothing' process to eliminate such dysfluencies as redundant repetitions of nouns, name entities and even clauses. The popularly used criteria for topic identification include frequency, position (e.g., such locations as headings, titles, first paragraphs/sentences, etc.), cue phrases, query and title, lexical connectedness, discourse structure, and combined scores of various models. All these can be used as features, together with other textual ones (e.g., uppercases, sentence length, proper names, dates, quotes, pronouns, etc.), in a classifier or a machine learner for estimating the probabilities of text portions for inclusion into a summary (Kupiec *et al.* 1995; Lin 1999). In particular, Lin (1999) shows that the most useful summary length is 15–35 per cent of that of an original text. In another dimension, multi-document summarization has to deal with more challenges beyond single documents, including cross-document overlaps and inconsistencies, in terms of both timeline and thematic content.

Summarization evaluation is complicated and remains one of the greatest challenges in this area. The common practice is to compare machine generated summaries against ideal ones prepared by human (e.g., evaluators). ROUGE (Lin and Hovy 2003; Lin 2004) is the most popular metrics for this purpose, which is defined to quantify the quality of a candidate summary in terms of its n-gram overlaps with a reference summary. Besides, two widely used measures, namely, *compression ratio* (CR) and *retention ratio* (RR), are defined respectively as the proportions of the length and information of a summary to its original text(s). In general, we assume that the smaller the CR and the larger the RR, the better the summary.

Sequence mining

Good examples of sequencing include DNA sequence analysis, stock trend predication, and language pattern recognition; and a good example of the latter is unsupervised lexical learning to model how language learners discover words from language data from scratch without a priori knowledge and teaching (Olivier 1968; Brent and Cartwright 1996; Brent 1999; Kit 2000, 2005; Venkataraman 2001). The basic idea is to segment a sentence into chunks, namely word candidates, that yield the greatest probability of the whole sentence, computed as the product of some conditional probability of each chunk. Theoretically, most existing works follow the *minimum description length* (MDL) principle (Solomonoff 1964; Rissanen 1978, 1989; Wallace and Boulton 1968; Wallace and Freeman 1987).[7] Technically, the learning becomes an issue of mining string patterns, mostly by means of formulating an optimization algorithm (e.g., the EM algorithm) to infer a probabilistic model (i.e., a set of candidate chunks and respective probabilities) on a given set of child-directed speech data (i.e., a set of utterances transcribed into speech transcription or plain text), such that the optimal chunks into which an utterance is segmented by the model coincide with what we call words. Alternatively, *description length gain* (DLG) is formulated as an empirical goodness measure for word candidates in terms of their compression effect (Kit and Wilks 1999) and later applied to simulate the manner of lexical learning that preverbal infants take to acquire words, by means of pursuing an optimal sum of compression effect over candidate chunks (Kit 2000, 2005). This approach is particularly successful in simulating language learning infants' two basic strategies to acquire new words that are widely recognized in psycholinguistics: a bottom-up strategy combines speech elements (or characters) into a word candidate, and a top-down strategy first recognizes clumps of frequently co-occurring words as word-like items and then divides them into smaller candidate chunks recursively when upcoming evidence favors this division, e.g., leading to further description length gain.

Biomedical text mining

Biomedical text mining is one of the most active areas of TM, having formed the biggest community of researchers and developed the largest volume of specialized resources:

1 MEDLINE/PubMed[8] database of journal citations and abstracts, and the Medical Subject Headings (MeSH),[9] maintained by the U.S. National Library of Medicine (NLM);
2 a good number of datasets derived from this primary one over various periods or special areas, e.g., the OHSUMED test collection,[10] the TREC Genomics Track[11] data, the GENIA[12] corpus of annotated MEDLINE abstracts, the BioCreAtive[13] collections, and the PennBioIE[14] corpus; and
3 many knowledge resources, e.g., the Metathesaurus[15] and the Semantic Network[16] of Unified Medical Language System (UMLS)[17] of NLM that unify over 100 controlled vocabularies such as dictionaries, terminologies and ontologies, the Pharmacogenomics Knowledge Base (PharmGKB),[18] the Neuroscience Information Framework,[19] and the Gene Ontology.[20]

This field has undergone rapid development in response to the exponentially growing volume of biomedical literature and data, including clinical records. Its general purpose is to facilitate information access, beyond ordinary IR, to the massive volume of specialized texts, e.g., retrieving explicitly expressed relations, facts or events of interest, and further exploit such texts

for discovery of unrecognized hidden facts, via the generation of hypotheses for further investigation. Its main tasks include NER, extraction of relations and events, summarization, question answering, and literature-based discovery. Grouped together with the extraction of relations and events under the banner of IE, NER is in fact the very initial step for almost all biomedical text processing, conceptually corresponding to the tokenization phase of general-purpose NLP but practically requiring to go beyond, whilst also based on, morphological processing to tackle a more complicated problem of identifying names (and terms) of various types in the biomedical domain, including gene and protein names, disease names and treatments, drug names and dosages, and so forth, most of which are compounds composed of several words. The main challenges in NER come not only from the growth of new names alongside the rapid growth of scientific discoveries but also from our slack use of existing names, giving rise to many problems such as synonyms (several names referring to the same entity) and polysemous acronyms and abbreviations (one abbreviated name referring to more than one entity or concept). Thus, NER is not merely to identify the boundaries of a name entity in the text, but to further carry out *entity normalization* to map a recognized entity to its canonical, preferred name (i.e., its unique concept identifier). The main approaches to NER can be grouped into the following categories (Krauthammer and Nenadic 2004; Leser and Hakenberg 2005; Simpson and Demner-Fushman 2012):

1 dictionary-based approach, which demands a comprehensive list of names and also has to resort to approximate string matching to deal with various kinds of variants;
2 rule-based approach, which uses a set of man-made rules or string patterns to describe the structures of names and their contexts;
3 statistical approach, especially machine learning, which exploits a classifier (e.g., support vector machine) or a sequencing model (e.g., hidden Markov model, maximum entropy model, or conditional random fields) to predict the position (e.g., beginning (B), inside (I) and outside (O)) of a word in a name entity, trained on annotated data with various kinds of lexical information (e.g., orthographical characteristics, affix, POS) as features; and
4 hybrid approach, which integrates (1) or (2) with (3) above, or combines several machine learning models.

Approaches of these categories are also applied to extracting pair-wise relations between two entities, including interactions between protein and protein, genes and proteins, genes and diseases, proteins and point mutations, and so forth, and relations of diseases and tests/treatments. The starting point of relation extraction is that a high co-occurrence frequency of two entities indicates a higher chance of a relation (or association) between them, e.g., the association between diseases and drugs, and then other means are applied to determine the type and direction of the relation. Besides rule-based approaches, which use rules or patterns manually prepared by experts or automatically derived from annotated corpora, machine learning approaches are commonly used to identify and classify these relations of interest, especially those between diseases and treatments. Nevertheless, further advances certainly have to count on advanced NLP techniques such as syntactic and semantic parsing, especially dependency parsing and semantic role labeling, to enable the utilization of specific syntactic patterns and/or semantic roles of words in predicate-argument structures.

Event extraction is to identify event structures, each of which consists of a verb (or nominalized verb, e.g., 'expression'), termed a *trigger*,[21] that specifies an event of some type (e.g., binding, positive regulation),[22] and one or more than one name (or another event) as event *argument* of some role (e.g., cause, theme). For example, in a nested event like '*X* gene

expression is activated by *Y*', we have 'activated' as a trigger, and *Y* and another event 'expression' (with '*X* gene' as theme) as its cause and theme, respectively. In general, an event extraction procedure goes through three stages, namely, trigger detection to identify trigger words, argument detection and role assignment to determine if a name entity or trigger is an argument and what role it plays, and event construction to form an event structure for a trigger using available arguments. This works in a similar way as semantic parsing via semantic role labeling: identifying predicates, their argument candidates and respective roles, in such a way as to form well-instantiated predicate-argument structures. Usually, a machine learning model is trained for each stage. However, to cope with the problem of cascading errors, i.e., an error in an earlier stage resulting in many in a later stage, the joint prediction of triggers and arguments is necessary, following the common practice of semantic parsing in this direction.

Unlike question answering that is basically to extend existing IR techniques to retrieving highly specialized information, via extracting relevant snippets of text as candidate answers and then returning the top-ranked ones, *literature-based discovery* is a genuine TM task beyond extraction of relations and events to uncover hidden, previously unknown or unrecognized relationships between entities (or concepts) in scientific literature. It was pioneered by a series of Swanson's prototypical examples of hypothesis generation, based on his observation of the 'complementary structures in disjoint literatures' (Swanson 1991). This series of manually generated hypotheses include the hidden connections between fish oil and Raynaud's syndrome (Swanson 1986), migraine and magnesium (Swanson 1988), somatomedin C and arginine (Swanson 1990), and also the potential use of viruses as biomedical weapons (Swanson *et al.* 2001). A prototypical pattern to generalize these discoveries is: given a known characteristic *B* (e.g., stress, spreading cortical depression, high platelet aggregability, and so forth) of a disease *C* (e.g., migraine) presented in a body of literature and the effect(s) of a substance *A* (e.g., magnesium) on *B* in another 'complementary but disjoint' body of literature, we can infer a hidden $A-C$ relationship, i.e., *A* may be a potential medication for *C*. What bridges such a hidden connection is a shared co-occurrent *B* of two terms *A* and *C* in two disjoint literatures. Two modes of discovery for this kind of second-order association (vs. the first-order relation between two co-occurring entities) are further distinguished by Weeber *et al.* (2001), namely, closed vs. open discovery: the former to find *B* term(s) to bridge a hypothesized $A-C$ connection and the latter to find $B-C$ relations (in another domain) given some $A-B$ relations already known. Existing approaches to automatic literature-based discovery can be categorized into three categories:

1 co-occurrence based, which relies on the statistics (e.g., frequency, log likelihood, or some information theoretic score) of second-order (or shared) co-occurrences of two biomedical entities (e.g., gene symbols, concepts);
2 semantic based, which further builds on (1) above by applying semantic information (e.g., UMLS semantic types) to filter out uninteresting or spurious candidate relations; and
3 graph based, which constructs a graph representation for various kinds of association among biomedical entities (e.g., gene-disease) to allow uncovering indirect associations along paths in the graph.

Literature-based discovery systems are evaluated in terms of how well they can replicate known discoveries (e.g., Swanson's ones or those in recent publications), as measured by precision and recall. Comprehensive reviews of existing research in this field can be found in Cohen and Hersh (2005), Zweigenbaum *et al.* (2007) and Simpson and Demner-Fushman (2013).

Opinion mining

Opinion mining, also known as *sentiment analysis* and many other names in literature, has been another very active research area of TM since around the turn of this century. It aims at analyzing and collecting, from unstructured opinionated texts such as those from social media websites, people's subjective evaluations, views or opinions, in the form of expressing certain sentiment, attitude, emotion, mood or other type of affect,[23] towards entities, issues or topics of interest (e.g., products, services, events, individuals, organizations, etc.) and/or their aspects or attributes. Effective access to mainstream social opinions or public sentiment has a profound impact on decision making in many domains, especially politics, the economy, business and management. A good number of recent studies using Twitter data have demonstrated this from various perspectives. For example, a 'relatively simple' analysis of Twitter sentiment by O'Connor *et al.* (2010) replicates highly similar results of traditional polls (e.g., consumer confidence and presidential job approval polls), illustrating a strong correlation of the two, and Bollen *et al.* (2011) show that the GPOMS Calm time series lagged by 3 days exhibit an amazing congruence with the DJIA closing values.[24]

The goal of opinion mining is to detect opinions in real texts (e.g., product reviews). An opinion is a claim (or statement) in the form of verbal expression held (1) by someone, called *opinion holder* or *source*, (2) about something (e.g., a movie, product, service, event or individual, formally referred to as an *entity* or *object*), called *topic* or *opinion target*, (3) at some time (4) associated with certain semantic orientation or (sentiment) *polarity*, which may be positive, negative or neutral, or of some strength or intensity (e.g., represented by 1 to 5 stars or some other numeric rating score) (see Kim and Hovy 2004; Wiebe *et al.* 2005; Liu 2010, 2012; among many others). It is described in Kim and Hovy (2004) as a quadruple [Topic, Holder, Claim, Sentiment]. However, a common situation is that an opinion is specifically targeted at some particular (sub)part(s) or attribute(s)/feature(s) of an entity. With *aspect* as a general term to refer to the (sub)parts and attributes in the hierarchical decomposition of an entity, an opinion is defined in Liu (2012) as a quintuple (e, a, s, h, t), where s is the sentiment on the aspect a of the target entity e,[25] h the holder of the opinion, and t the time that the opinion is expressed. The sentiment may be expressed by a categorical and/or a numerical value for its polarity and/or intensity, respectively. Accordingly, the general task of opinion mining can be divided into subtasks to identify these components of the quintuple in an opinionated document. Since the early works by Hatzivassiloglou and McKeown (1997), Turney (2002), and Pang *et al.* (2002) on detecting the polarity of adjectives, product reviews and movie reviews, respectively, there has been so large a volume of literature in the field on opinion holders' attitudes towards topics of interest that sentiment analysis is equated with the detection of attitudes, mostly represented in terms of polarities and/or intensity scores (instead of a set of types as in Scherer's typology).[26]

Most existing works on sentiment analysis can be categorized into two types, namely, text-oriented vs. target-oriented. The former focuses on identifying the sentiment polarities, or predicting the rating scores, of opinionated/subjective text units of various sizes, ranging from documents to sentences, phrases and words. Since it is formulated as a classification problem, various supervised and unsupervised machine learning approaches have been applied, including naïve Bayes, MaxEnt, SVM, CRFs, etc. Conceptually, one may conceive two levels of classification, one to detect whether a text carries any opinion, i.e., opinionated (vs. factual) or subjective (vs. objective), and the other to determine a text's sentiment polarity or score (e.g., 1 to 5 stars), although the latter subsumes the former in principle. Extracting features from input text for machine learning of such classification is a critical task after sentiment-aware

tokenization and POS tagging to deal with the irregularities of web texts (e.g., Twitter) and retain useful tags, expressions and symbols (e.g., emoticons).[27] Many special forms of verbal expression, such as negation and subjunctive mood, known as *sentiment shifters*, that critically deflect the sentiment of lexical items, require special treatment, e.g., marking negated words (in between a negation word and the next punctuation) with 'NOT_' (Das and Chen 2001). Besides known sentiment words and phrases, the most important features, other effective features include n-grams (and their frequencies), POS tags (especially adjectives, adverbs (e.g., *hardly* and *barely*) and negation words (e.g., *never*, *neither* and *nobody*)), syntactic dependency, and rules of opinion (e.g., negation of positive means negative and vice versa, mentioning a desirable fact implies positive; see Liu (2010) for more). Interestingly, however, it is shown in Pang *et al.* (2002) that using unigrams outperforms bigrams and using their presence outperforms their frequencies in classifying movie reviews with naïve Bayes and SVM models.

Certainly, this does not necessarily imply the denial of the significance of (co-) occurrence statistics. The Turney (2002) algorithm is a successful demonstration of this in inferring the *semantic orientation* (SO) of extracted phrases (by a set of predefined POS tag patterns and constraints) in terms of their *pointwise mutual information* (PMI), with one positive and one negative reference word (namely, 'excellent' and 'poor') as

$$\mathrm{SO}(\mathrm{phrase}) \equiv \mathrm{PMI}(\mathrm{phrase},\ \text{"excellent"}) - \mathrm{PMI}(\mathrm{phrase},\ \text{"poor"}),$$

where the PMI of two terms to measure their association is

$$\mathrm{PMI}(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}.$$

Then, a review is classified into positive ('recommended') or negative ('not recommended') in terms of the average SO over all its phrases. This early work not only illustrated an unsupervised approach to sentiment document classification, but also inspired many subsequent works on mining sentiment lexicons, by utilizing statistics of co-occurrence (in various co-occurring patterns) with a seed set of polarity-labeled words. Before this, Hatzivassiloglou and McKeown (1997) followed the intuition, that adjectives in *and*/*but* conjunctions have a similar/different semantic orientation, to mine adjective polarity by means of supervised learning: first, expanding a seed set of adjectives with predetermined polarities to conjoined adjectives, predicting their polarity (dis)similarity using a log-linear regression model, and then clustering the resulting graph (with hypothesized polarity links between adjectives) in two groups, namely, positive vs. negative. Starting from a seed set of words with known polarities, a basic corpus-based strategy is to follow available hints (e.g., syntactic patterns) to enlarge this set with their synonyms and antonyms iteratively until no more are available (Hu and Liu 2004). In contrast, a dictionary-based approach compiles sentiment words from existing dictionaries (e.g., WordNet) that provide tractable relations of synonyms and antonyms.

Unlike text-oriented sentiment analysis that assumes a default (or unconcerned) target for a text (document) under analysis, aspect-oriented sentiment analysis is conducted at a much finer level of granularity to pinpoint the exact aspect (or feature) of a target entity that an opinion is about, and its sentiment polarity or strength. Specifically, it involves two main subtasks, namely, identification of opinion target, mostly by aspect extraction, and classification (or quantification) of aspect sentiment. The rules of thumb to follow include the assumptions that every piece of opinion has a target and that an opinionated text usually focuses on one opinion

target and hence mentions its aspects in a more prominent way than other texts. A simple but effective strategy to extract explicitly mentioned aspects is to take frequent nouns and noun phrases, by means of POS tagging and frequency thresholding, and the nearest (infrequent) ones to a sentiment word (Hu and Liu 2004). This strategy can be further enhanced by other statistical means or constraints, e.g., PMI between a candidate and known hints (such as meronymy discriminators for a target entity) (Popescu and Etzioni 2005), occurrence in a subjective sentence or co-occurring with sentiment words (Blair-Goldensohn *et al.* 2008), and dependency relations (Zhuang *et al.* 2006). The fact that sentiment words and aspects tend to co-occur with each other is commonly exploited by researchers, e.g., Ghani *et al.* (2006) and Qiu *et al.* (2011). Starting from a seed set of sentiment words, Qiu *et al.* extended the bootstrapping method to *double propagation* of both sentiment words and aspects along their dependency relations, combining sentiment lexicon acquisition and aspect extraction into one. Aspect extraction can also be tackled as an information extraction problem with supervised learning, using sequential labeling models such as HMM and CRFs. Also in this direction of research, Li *et al.* (2010) extended the linear-chain CRFs to a few variants, namely, skip-chain, tree and skip-tree CRFs, so as to exploit rich structure features to facilitate extraction of both aspects and opinions. Other kinds of syntactic or semantic information can be exploited as well, e.g., semantic role (Kim and Hovy 2006) and coreference (Stoyanov and Cardie 2008). Topic modeling (Hofmann 1999; Blei *et al.* 2003), which outputs a probability distribution over words (of certain semantic coherence) as a topic, is a principled statistical method of conceptual and mathematical elegance that many researchers have followed to attempt unsupervised extraction of aspects and sentiment words, but its weaknesses (e.g., hard to tell apart aspects and sentiment words, insensitive to locally frequent but globally infrequent words) needs to be overcome in order to have more practical use in sentiment analysis (Liu 2012). Besides *explicit* aspects, *implicit* aspects are perhaps more challenging to detect, because they have no overt form, but are only implied by certain expressions, especially adjective phrases (e.g., 'heavy' indicates weight and 'expensive' price). Their detection becomes a task of mapping sentiment words to explicit (or known) aspects. Co-occurrence association is usually the primary criterion for this mapping using various strategies, e.g., clustering (Su *et al.* 2008; Hai *et al.* 2011).

Extraction of opinion holder and time is also an NER problem. Usually, the author and publishing (or posting) time of a text (such as a review or blog) are, respectively, the default holder and time of an opinion extracted from the text, unless they are explicitly stated. The latter case is a typical NER issue to be tackled by strategies of various complexity, including (1) heuristics, e.g., only consider person and organizations as candidates (Kim and Hovy 2004), (2) sequential labeling, e.g., using CRFs with surrounding words' syntactic and semantic attributes as key features (Choi *et al.* 2006), and (3) other machine learning models, e.g., using MaxEnt to further rank heuristically selected candidates (Kim and Hovy 2006) or using SVM to classify them (Johansson and Moschitti 2010). SRL is also an effective means to facilitate this task (Bethard *et al.* 2004; Kim and Hovy 2006), especially in joint recognition of opinion holders and expressions.

Since, in different domains, not only are opinions expressed with different words, but the same words may express different sentiments, a sentiment model trained with opinion data (labeled or unlabeled) in one domain, called the *source* domain, is usually not directly applicable or adaptable to another, called the *target* domain. The purpose of *domain adaptation* is to enable this with minimum resources. Towards this goal, Gamon and Aue (2005) illustrated the outperformance of semi-supervised learning with EM training over a number of other strategies using SVM. Typically, this learning approach uses a small amount of labeled data combined

with a large amount of unlabeled data, both from a target domain, for training. Most subsequent research by others focused on selecting domain independent features (words) to enable the adaptation. Yang *et al.* (2006) selected highly-ranked features in labeled data from two domains as common features for across-domain transfer learning, so as to facilitate sentence-level opinion detection in a target domain that lacks labeled training data. Blitzer *et al.* (2006, 2007) applied structural correspondence learning (SCL) to cross-domain sentiment analysis, first choosing a set of *pivot* features (in terms of their frequency and mutual information with a source label) and then establishing a feature correspondence between two domains by computing the correlation of pivot and non-pivot features.

Cross-language sentiment analysis looks to be an issue of adaptation across two languages as if they were two domains, but is argued to be qualitatively different from the usual cross-domain adaptation caused by domain mismatch (Duh *et al.* 2011). It aims to utilize both monolingual and bilingual tools and resources to accomplish sentiment analysis in another language. Given that most available sentiment analysis tools and resources are developed for English, one strategy to achieve this aim is to convert English resources into a foreign language to analyze foreign texts, and the other is to have foreign texts automatically translated into English for analysis. Mihalcea *et al.* (2007) found that projecting sentiment annotation from English into Romanian by virtue of a parallel corpus is more reliable than translating a sentiment lexicon with the aid of bilingual dictionaries. To leverage English resources for Chinese sentiment analysis, Wan (2008) opted to work on multiple MT outputs, and then combined their sentiment analysis results with ensemble methods. Wan (2009) further demonstrated a co-training approach that makes use of labeled English texts, unlabeled Chinese texts, and their MT output counterparts in the other language, to train two SVM classifiers and combine them into one for Chinese sentiment classification. Besides, other approaches can be applied to tackle this problem too, e.g., transfer learning using the SCL method (Wei and Pal 2010). In a multilingual setting, unsupervised methods such as topic modeling can be used to create multilingual topics (Boyd-Graber and Resnik 2010) or mulitlingual aspect clusters (i.e., semantic categories of product-features) (Guo *et al.* 2010).

Besides the key issues briefly introduced above, there are many others in the field that cannot be accommodated in this short subsection, such as discourse analysis for sentiment analysis, comparative opinion mining, summarization and presentation/visualization of mined opinions, opinion spam detection, and estimation of opinion/review quality/sincerity, to name but a few. Interested readers may refer to the books/surveys by Shanahan *et al.* (2006), Pang and Lee (2008) and Liu (2012) for a more detailed discussion. Pang and Lee (2008) also provide plenty of information about publicly available resources and evaluation campaigns, including:

1 annotated datasets, such as Cornell Movie-Review Datasets[28] (Pang *et al.* 2002; Pang and Lee 2004, 2005), Customer Review Datasets[29] (Hu and Liu 2004; Ding *et al.* 2008), and MPQA Corpus[30] (Wiebe *et al.* 2005);
2 past evaluations, such as TREC Blog Track[31] and NTCIR-6~-8[32] (on multilingual opinion analysis tasks (MOAT) in Japanese, English and Chinese);
3 lexical resources, such as General Inquirer[33] (Stone 1966), OpinionFinder's Subjectivity Lexicon[34] (Wilson *et al.* 2005) and SentiWordNet[35] (Esuli and Sebastiani 2006; Baccianella *et al.* 2010); and
4 a few pointers to online tutorials and bibliographies.

Besides, there are also language resources and open evaluations for other languages than English, e.g., NTU Sentiment Dictionary[36] (Ku *et al.* 2006) and COAE[37] (Zhao *et al.* 2008; Xu *et al.* 2009) for Chinese.

Text mining for IR—mining term relations

TM has long been applied to IR, focused on deriving term relationships (or association) to facilitate query expansion. Assuming that a relationship between terms t_1 and t_2 can be found and expressed as $P(t_2 \mid t_1)$, measuring the extent to which t_1 implies t_2, a new query representation can then be built upon an old one by adding a set of expansion terms which are related to its terms. In language modeling, this means the construction of a new query model θ_{Q_1} by interpolating an existing query model θ_{Q_0} with another one, constructed with selected expansion terms $\theta_{\mathcal{R}}$. In a similar manner of expanding a user query to a new model using feedback documents, we can further build an expansion query model using term relations mined from available texts.

The relation most widely used in IR is the term co-occurrence (Crouch 1990; Qiu and Frei 1993; Jing and Croft 1994). Its underlying assumption is that the more frequently two terms co-occur, the more closely they are related. In general, such term relationships can be quantified by the following conditional probability (or a similarity measure following the same idea):

$$P\left(t_i \mid t_j\right) = \frac{cooc\left(t_i, t_j\right)}{\sum_{t \in V} cooc\left(t, t_j\right)}$$

where *cooc* means the frequency of co-occurrence within a certain context. Various types of context can be used to derive co-occurrence statistics, e.g., within the same text, paragraph, sentence, or a text window of some fixed size (such as 10 words). Too large a context (e.g. the same text) could bring in much noise—unrelated terms are extracted, while too narrow a context may miss useful relations. In IR experiments, it turns out that a relatively small context (such as sentence or a text window of 10 words) works well. In addition to the above conditional probability, one can also use other measures to quantify the relationship between terms such as mutual information, log-likelihood ratio, χ-square statistics, and so on.

To extract many other useful relations such as synonyms that cannot be extracted this way (because many synonyms are rarely used in the same context), we need to resort to second-order co-occurrences: two terms are deemed to be related if they occur in similar contexts, i.e., co-occur with similar words. This idea was practiced in Lin (1998) in a way to define word context in terms of certain syntactic relations (e.g., verb–object): two words are considered related if they are linked to similar words within a similar syntactic context. In this way, words in the same category (e.g., clothes, shoes, hat, etc., that appear as objects of the verb 'wear') can be extracted. In Bai *et al.* (2005), this idea was incorporated into *information flow* (IF): the context of a word is defined by its surrounding words, and two words are considered similar if their context vectors are similar. Word relations extracted this way by means of information flow were successfully applied to query expansion, achieving a large performance improvement. More specifically, the following document scoring function is used:

$$\text{score}\left(\text{Q},\text{D}\right) = \lambda \sum_{t \in Q} P_{\text{ML}}\left(t \mid \theta_Q\right) \log P(t \mid \theta_D) + \left(1-\lambda\right) \sum_{t \in V} P_{\text{IF}}\left(t \mid \theta_Q\right) \log P\left(t \mid \theta_D\right)$$

where $P_{IF}(t|\theta_Q) = \Sigma_{q \subseteq Q} P_{IF}(t|q)P(q|\theta_Q)$, with q to be a subset of query terms and $P_{IF}(\cdot|\cdot)$ to be defined in terms of the degree of information flow.

To tackle the noise problem with co-occurrence relation (i.e., frequently co-occurring terms are not necessarily truly related) and word ambiguity (e.g., the relation between 'java' and 'language' does not apply to a query on 'java tourism'), phrases are used instead of words in relation extraction. Multi-word phrases are in general less ambiguous, and thus a term (word or phrase) related to a phrase in a query has a higher chance of being truly related to the query. Among the several studies following this idea, Bai *et al.* (2007) extended the *context-independent* co-occurrence relation between single words to the *context-dependent* co-occurrence relation of a word to a set of words. This idea can be traced back to Schütze and Pedersen (1997). Using more than one word as a context to determine related words imposes a stronger contextual constraint, resulting in words in a less ambiguous relation (e.g., the relation between 'java, program' and 'language' is more certain than the one between 'java' and 'language', in that a query containing both 'java' and 'program' is more likely to be related to 'language'). In Bai *et al.* (2007), the number of context words was restricted to two, so as to minimize the complexity of the extraction process. Accordingly, the above conditional probability can be extended to estimate the strength of a word–words relation, as follows:

$$P\left(t_i | t_j, t_k\right) = \frac{cooc\left(t_i, t_j, t_k\right)}{\sum_{t \in V} cooc\left(t, t_j, t_k\right)}$$

Experiments on several TREC collections show that this word−words relationship brings in more performance gain than the word−word one. According to the survey of Carpineto and Romano (2012), this is one of the best performing methods for query expansion.

Among the long list of query expansion methods surveyed in Carpineto and Romano (2012), all four best performing ones (including the above two) on TREC collections utilize some forms of context-dependent relation. (1) Liu *et al.* (2004) rank documents first by *phrase* similarity and then *word* similarity, using (a) machine-learned distances of window size for identifying phrases of different types, (b) WordNet for query expansion, (c) a combination of local and global correlations for pseudo relevance feedback, and (d) a variant of *Okapi* weighting (with document length replaced by L2-norm, i.e., the vector length, of a document) for similarity scoring. (2) Bai *et al.* (2005) integrate a set of information flow relationships of terms into language modeling with KL-divergence for document scoring as mentioned above. (3) Bai *et al.* (2007) further generalize the above query model by the following interpolation:

$$P\left(t|\theta_Q\right) = \sum_{i \in X} \alpha_i P\left(\theta_Q^i\right) \text{ , with } \sum_{i \in X} \alpha_i = 1$$

where α_i are mixture weights. It integrates a set of component models $X = \{0, F, Dom, K\}$, including the original query model θ_Q^0, the feedback model θ_Q^F on retrieved documents, a domain model θ_Q^{Dom} on predefined domain documents, and a knowledge model θ_Q^K on the *context dependent* relations of terms in the form $P(t_i | t_j, t_k)$ as above. Each model has a scoring function defined as:

$$\text{score}\left(\text{Q,D}\right) = \sum_{i \in X} \alpha_i \text{ score}_i\left(Q, D\right) = \sum_{i \in X} \alpha_i \sum_{t \in V} P(t | \theta_Q^i) \log P\left(t|\theta_D\right)$$

With other language models trained on respective data, say, using the EM algorithm, the knowledge model is defined as:

$$P\left(t|\theta_Q^K\right) = \sum_{(t_j t_k)\in Q} P\left(t_i|t_j,t_k\right)P(t_j t_k|\theta_Q) = \sum_{(t_j t_k)\in Q} P\left(t_i|t_j,t_k\right)P(t_j|\theta_Q)P(t_k|\theta_Q)$$

where $(t_j, t_k) \in Q$ and $P(t_i|t_j, t_k)$ is defined as above. (4) Metzler and Croft (2005, 2007) generalize the relevance model into a general discriminative model, using the Markov random field (MRF) model to integrate a large variety of scoring schemes as its arbitrary features (such as *tf* and *idf* of word and phrase, document length, and PageRank, among many others) for scoring a document with respect to a query, and utilizing term dependencies by combining the scoring of unigrams, bigrams and collocations (i.e., co-occurring words within a small fixed-size window) into expansion term likelihood computation on feedback documents. These four methods all demonstrate the power of effective use of term dependencies, regardless of the difference of their scoring schemes.

Besides a document collection, modern search engines also benefit from rich interactions with users, which are recorded in *query logs*. Data items in query logs usually include, among others, user ID (or IP address), submitted query, time, search results (usually URLs plus some key information) presented to, and a few of them clicked by, a user. User clickthrough information preserved in the last item is of particular importance, for there is reason to assume that a user must have a certain preference for clicked documents over unclicked ones, meaning that the former ones are more likely to be relevant to the query in question than the latter ones (Joachims 2002). The more such clickthrough by users, the stronger the signal of the relevance.

Clickthrough information can be exploited to facilitate IR in several ways. (1) It can be used to identify similar queries to a given query following the 'co-click' assumption, that queries sharing the same or similar clicked documents are similar (Wen *et al.* 2001; Baeza-Yates and Tiberi 2007; Craswell and Szummer 2007). Implicit relations of this kind between queries have been widely used in commercial search engines to suggest alternative queries to users. (2) A large number of user clicks on a document may imply a certain semantic relatedness between the query terms in use and the terms in the clicked document. However, it is risky to assume that such a query term is related to any term in the document, for it would result in false relationships between unrelated query and document terms. A more common (and safer) practice is to assume that trustable relations only exist between terms in a query and in a document title, and then estimate their relation probability (Cui *et al.* 2002), or train a statistical translation model for this, taking a query and a clicked document title as a pair of parallel texts for the training (Gao *et al.* 2010). (3) Other useful features can also be extracted from clickthrough data to help document ranking, in a way to favor the popularly clicked documents (or similar ones) in the ranked list output for a query (Joachims 2002).

Several sets of query log data have been made available for IR experiments: MSN query logs,[38] Sogou query logs in Chinese,[39] Yandex query logs,[40] MSRBing image retrieval challenge,[41] and the AOL query logs, which was the first publicly released dataset of this kind but later retracted. However, larger amounts of query logs and clickthrough information beyond these datasets are only available within search engine companies. To simulate this kind of large-scale clickthrough, Dang and Croft (2010) used anchor texts and links in webpages, assuming the former as queries and the latter as clicked documents. Their experiments showed that even such simulated data can bring some nice improvements to IR effectiveness, but certainly not on a par with true clickthrough data.

Text mining for CLIR—mining cross-language term relations

CLIR imposes a strong demand for translation relations between terms (words or phrases) across languages. In the sense that such cross-language relations encode translation knowledge, i.e., how likely a word (or phrase) in one language is to be the translation for a word (or phrase) in another language, inferring them from parallel texts can be considered a kind of bilingual text mining. A principled way to do this is to resort to an SMT model in the series, called IBM models, defined in Brown *et al.* (1993), which provides a mathematical foundation for all SMT methods. It is assumed that an IBM model generates a translation (i.e., a target sentence) from a source sentence by (1) first determining its length, (2) then determining the position alignment between the two sentences, and (3) finally filling appropriate translation words into the slots in the target sentence.

The most popular translation model used in CLIR is IBM model 1, or IBM 1 for short, thanks to its underlying assumption that a word is translated in isolation from its context, i.e., independently of its position and other words in a sentence. This is an assumption that does not hold for human translation but which corresponds so well to the bag-of-words assumption for IR. The IBM model 1 is formulated as follows for a source sentence S and a target sentence T:

$$P(S \mid T) = \frac{\in}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} P(s_j \mid t_i)$$

where l and m are, respectively, the length of T and S, $\in$ is a constant meaning the probability to produce a sentence of m words from T, and $P(s_j \mid t_i)$ is the lexical translation probability between two words t_i and s_j. Note that in SMT we choose the translation $\hat{T}$ that maximizes $P(T \mid S)$, i.e.,

$$\hat{T} = \arg\max_T P(T \mid S) \propto \arg\max_T P(S \mid T) P(T)$$

where $P(T)$ is a language model to estimate the likelihood of T in target language, and $P(S \mid T)$ the translation model from T to S; and that a lexical translation model $P(s_j \mid t_i)$ is usually trained on a large set of parallel sentences using the EM algorithm, in a way to maximize the translation likelihood of the parallel sentences.

While fully fledged SMT demands a more sophisticated translation model than IBM 1, in order to take into account word order and word position, CLIR has relied on IBM 1 successfully so far, thanks to the fact that the state-of-the-art IR is largely rooted in the word-bag model, sharing similar assumptions as IBM 1. Several experiments have shown that such a simple translation model trained on large scale parallel data can be competitive to high-performance MT systems in supporting CLIR (e.g., Kraaij *et al.* 2003).

As query translation does not need as strict a translation as MT, translation relationships to be utilized in CLIR can be relaxed to *cross-language co-occurrence*. The idea is that the more often a pair of source and target words co-occurs in parallel sentences, the stronger the translation relationship they have. Accordingly, this relationship can be estimated as:

$$P(s \mid t) \propto \frac{cooc(s,t)}{\sum_s cooc(s,t)}$$

However, this estimation has an innate weakness: when *s* is a frequent term (e.g., a function word), P(s|t) will be too strong, lowering CLIR effectiveness. In contrast, the alignment of *t* with a frequent *s* in IBM 1 is gradually weakened along EM iterations. Even so, many other measures effectively used in monolingual text mining can be extended to the bilingual case without iterative training, e.g., mutual information, log-likelihood ratio, etc.

Moreover, this less strict co-occurrence measure can be straightforwardly applied to comparable texts, on which an IBM model can hardly be trained. As mentioned before, less strict translation (or cross-language) relations of this kind trained on comparable texts can improve CLIR to some extent, as shown in the experiments of Sheridan and Ballerini (1996) and Braschler and Schäuble (2000), although their effectiveness is lower in comparison with an IBM model trained on parallel texts. However, when parallel texts are not available, they can be used as a second-choice substitute. Even when parallel texts are available, translation relations trained on comparable texts can also be used as a beneficial complement (Sadat *et al.* 2003).

Mining bilingual texts from the Web

The training of statistical MT models critically depends on the availability of a large amount of bilingual parallel texts (or bitexts). For resource-rich languages such as European languages, many manually compiled large-scale parallel corpora are available, including the Canadian Hansard, the earliest large parallel corpus in English and French, popularly used in MT, and the European Parliament documents in several European languages. In contrast, however, bitexts are inadequately available for many other languages such as Arabic, Chinese, Indian languages, and so on. This was, and to a great extent still is, the case. A possible way out from this situation seems to be allowed by the flourishing of the Web, where more and more websites provide information in several languages, mostly through bilingual or parallel webpages. The Web has been a huge repository of various kinds of texts, including parallel texts (Grefenstette 1999; Kilgarriff and Grefenstette 2003; Resnik and Smith 2003). What we need to do is to identify and then extract available parallel texts automatically via web mining.

A good number of attempts have been made in this direction to illustrate the feasibility and practicality of automatically acquiring parallel corpora from bilingual (or multilingual) websites, resulting in respective web miners for parallel texts, including STRAND (Resnik 1998, 1999; Resnik and Smith 2003), BITS (Ma and Liberman 1999), PTMiner (Nie *et al.* 1999; Chen and Nie 2000), PTI (Chen *et al.* 2004), WPDE (Zhang *et al.* 2006), the DOM tree alignment model (Shi *et al.* 2006), PupSniffer (Kit and Ng 2007; Zhang *et al.* 2013), PagePairGetter (Ye *et al.* 2008), and Bitextor (Esplà-Gomis and Forcada 2010).

The basic strategy they follow is to utilize the characteristic organization patterns of parallel webpages, including inter-page links, similarity of intra-page structures, file and URL naming conventions, and other features that reveal any such pattern. For example, many parallel pages are either linked to from a common entry page or to each other mutually. An entry page usually contains many close links with anchor texts (such as 'English version' and 'Version française') to indicate the language of a linked-to page, providing strong hints about parallel pages. This structure is exploited in STRAND (Resnik 1998, 1999). Another common structure is that parallel pages contain mutual links to each other, pointing to their counterparts in the other language, with an anchor text to indicate language (e.g., 'English version'). This structure is used in PTMiner (Nie *et al.* 1999; Chen and Nie 2000).

Also, two parallel pages are often found to have similar names or URLs, e.g., 'www.xyz.org/intro_en.html' vs. its French counterpart 'www.xyz.org/intro_fr.html'. Their only

difference is the segments indicating their languages, i.e. their URL pairing pattern 'en:fr'. This kind of widespread characteristic of parallel pages was widely used in the previous attempts to determine possible parallel pages in a bilingual website, relying on predefined pairing patterns such as {'e:c', 'en:ch', 'eng:chi', …} for English–Chinese. A problem with such an *ad hoc* method is that hand-crafted heuristics can never exhaust all possibilities, leaving many true parallel pages untouched.

Automatic discovery of URL pairing patterns of this kind was attempted in Kit and Ng (2007) and further extended in Zhang *et al.* (2013). Among the fundamental tasks involved in web mining for parallel texts, namely, (1) identifying bilingual (or multilingual) websites and retrieving their webpages, (2) matching retrieved webpages into parallel pairs, and (3) extracting parallel texts from matched pairs for alignment at a finer level of granularity. Since there have been matured techniques of web crawling and text alignment to deal with (1) and (3) respectively, (2) is the most vital one at the core of the whole mining process. Compared to similarity analysis of HTML structure and/or webpage content, it is preferable to match candidate webpages by pairing up their URLs using automatically inferred URL pairing patterns (or keys). The basic idea to achieve this is as follows: given two sets of URLs for webpages in a bilingual website, each in a language, a candidate key is generated from each pair of candidate URLs by removing their common prefix and suffix, and then its linking power, defined as the number of URL pairs that it can match, is used as the objective function to search for the best set of keys that can find the largest number of webpage pairs within the website. A best-first strategy to let candidate keys compete, in a way that a more powerful key matches URLs first, results in correct discovery of 43.7 percent true keys (at precision 67.6 percent) that matches 98.1 percent true webpage pairs (at precision 94.8 percent), with the aid of an empirical threshold to filter out weak keys (Kit and Ng 2007). Later, this approach is extended to work on a large set of bilingual websites, digging out more webpage pairs by extending the notion of linking power to global credibility to rescue many weak (but true) keys, uncovering bilingual webpages from the deep web by generating crawler-unreachable counterparts of unmatched URLs using found keys, and also incorporating PageRank based analysis of bilingual website relationship into this framework for discovery of more bilingual websites beyond an initial seed set for mining more bitexts. This automatic approach is simple and effective, but it is not designed to deal with machine-generated webpages, e.g., from a (text) database, whose URLs are randomly generated without any pairing patterns.

Notes

1 At http://trec.nist.gov/data.html.
2 NII Test Collections for IR system, at http://research.nii.ac.jp/ntcir/data/data-en.html.
3 The CLEF Initiative – Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum, at http://www.clef-initiative.eu.
4 http://tartarus.org/martin/PorterStemmer.
5 However, it is questionable whether a term ten times more frequent than another in a document would really make a ten times more significant contribution to the relevance of the document. Different scaling strategies can be applied to adjust the above term weighting, e.g., using the logarithm of term frequency $1 + \log tf$ (or 0, if $tf = 0$), the probabilistic *idf*: $\max\{0, \log[(N - df_t)/df_t]\}$. A systematic presentation of a good number of principal weighting schemes can be found in Salton and Buckley (1988), Singhal *et al.* (1996), and Moffat and Zobel (1998).
6 http://wordnet.princeton.edu.
7 See Grünwald (2007) for an extensive introduction to the MDL principle.
8 http://www.ncbi.nlm.nih.gov/pubmed, with a resources guide at http://www.nlm.nih.gov/bsd/pmresources.html.
9 http://www.nlm.nih.gov/mesh/meshhome.html.

10 http://trec.nist.gov/data/t9_filtering.html.
11 http://ir.ohsu.edu/genomics.
12 http://www.nactem.ac.uk/aNT/genia.html.
13 http://biocreative.sourceforge.net.
14 http://bioie.ldc.upenn.edu.
15 http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html.
16 http://www.nlm.nih.gov/research/umls/knowledge_sources/index.html#semantic.
17 http://www.nlm.nih.gov/research/umls.
18 http://www.pharmgkb.org.
19 http://www.neuinfo.org.
20 http://www.geneontology.org.
21 Obviously it corresponds to *predicate* or *semantic head* in semantic parsing.
22 Nine event types are listed in the BioNLP Shared Task 2011 GENIA Event Extraction (GENIA) site, at https://sites.google.com/site/bionlpst/home/genia-event-extraction-genia.
23 See Scherer (1984) for a typology of affective states, including emotion, mood, interpersonal stances, attitudes, and personality traits.
24 GPOMS: Google-Profile of Mood States, which measures mood in terms of six dimensions, namely, Calm, Alert, Sure, Vital, Kind and Happy. DJIA: the Dow Jones Industrial Average.
25 A conventional aspect *a=general* is reserved for an opinion targeted on an entity as a whole.
26 'Attitudes: relatively enduring, affectively colored beliefs, preferences, and dispositions towards objects or persons (*liking*, *loving*, *hating*, *valuing*, *desiring*)' (Scherer 1984).
27 More technical details (and source codes) are available from Christopher Potts' Sentiment Symposium Tutorial at http://sentiment.christopherpotts.net/ and from CMU's Twitter NLP and Part-of-Speech Tagging site at http://www.ark.cs.cmu.edu/TweetNLP/.
28 http://www.cs.cornell.edu/people/pabo/movie-review-data/.
29 http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.
30 http://www.cs.pitt.edu/mpqa/.
31 http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG/.
32 http://research.nii.ac.jp/ntcir/index-en.html.
33 http://www.wjh.harvard.edu/~inquirer/.
34 http://www.cs.pitt.edu/mpqa/.
35 http://sentiwordnet.isti.cnr.it/.
36 http://nlg18.csie.ntu.edu.tw:8080/opinion/.
37 Chinese Opinion Analysis Evaluation 2008–2012 at http://ir-china.org.cn/Information.html and COAE 2013 at http://ccir2013.sxu.edu.cn/COAE.aspx.
38 http://research.microsoft.com/en-us/um/people/nickcr/wscd09/.
39 http://www.sogou.com/labs/dl/q.html.
40 http://switchdetect.yandex.ru/en/datasets.
41 http://web-ngram.research.microsoft.com/GrandChallenge/.

References

Adriani, Mirna and Keith van Rijsbergen (2000) 'Phrase Identification in Cross-Language Information Retrieval', in *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) (RIAO 2000)*, 520–528.

Aggarwal, Charu C. and ChengXiang Zhai (eds) (2012) *Mining Text Data*. New York: Springer.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani (2010) 'SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining', in *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, 2200–2204.

Baeza-Yates, Ricardo and Alessandro Tiberi (2007) 'Extracting Semantic Relations from Query Logs', in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2007)*, 76–85.

Bai, Jing, Dawei Song, Peter Bruza, Jian-Yun Nie, and Guihong Cao (2005) 'Query Expansion Using Term Relationships in Language Models for Information Retrieval', in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005)*, 688–695.

Bai, Jing, Jian-Yun Nie, Hugues Bouchard, and Guihong Cao (2007) 'Using Query Contexts in Information Retrieval', in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, 15–22.

Bendersky, Michael, Donald Metzler, and W. Bruce Croft (2010) 'Learning Concept Importance Using a Weighted Dependence Model', in *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM2010)*, 31–40.

Bethard, Steven, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky (2004) 'Automatic Extraction of Opinion Propositions and their Holders', in *Proceedings of 2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 22–24.

Berger, Adam and John Lafferty (1999) 'Information Retrieval as Statistical Translation', in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1999)*, 222–229.

Berners-Lee, Tim, Robert Cailliau, Jean-François Groff, and Bernd Pollermann (1992) 'World-Wide Web: The Information Universe', *Electronic Networking: Research, Applications and Policy* 1(2): 74–82.

Blair-Goldensohn, Sasha, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, and Jeff Reynar (2008) 'Building a Sentiment Summarizer for Local Service Reviews', in *Proceedings of WWW Workshop on NLP in the Information Explosion Era (NLPIX 2008)*.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* 3: 993–1022.

Blitzer, John, Ryan McDonald, and Fernando Pereira (2006) 'Domain Adaptation with Structural Correspondence Learning', in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, 120–128.

Blitzer, John, Mark Dredze, and Fernando Pereira (2007) 'Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification', in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (*ACL 2007*), 440–447.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011) 'Twitter Mood Predicts the Stock Market', *Journal of Computational Science* 2(1): 1–8.

Boyd-Graber, Jordan and Philip Resnik (2010) 'Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation', in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, 45–55.

Braschler, Martin and Peter Schäuble (2000) 'Using Corpus-based Approaches in a System for Multilingual Information Retrieval', *Information Retrieval* 3(3): 273–284.

Brent, Michael R. (1999) 'An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery', *Machine Learning* 34(1-3): 71–105.

Brent, Michael R. and Timothy A. Cartwright (1996) 'Distributional Regularity and Phonological Constraints Are Useful for Segmentation', *Cognition* 61(1-2): 93–125.

Brin, Sergey and Larry Page (1998) 'The Anatomy of a Large-scale Hyper-textual Web Search Engine', *Computer Networks and ISDN Systems* 30(1-7): 107–117.

Brown, Peter E., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational Linguistics* 19(2): 263–311.

Bush, Vannevar (1945) 'As We May Think', *The Atlantic Monthly* 176: 101–108.

Cao, Guihong, Jian-Yun Nie, and Jing Bai (2005) 'Integrating Word Relationships into Language Models', in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2005)*, 298–305.

Cao, Guihong, Jian-Yun Nie, Jianfeng Gao, and Stephan Robertson (2008) 'Selecting Good Expansion Terms for Pseudo-relevance Feedback', in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2008)*, 243–250.

Carpineto, Claudio and Giovanni Romano (2012) 'A Survey of Automatic Query Expansion in Information Retrieval', *ACM Computing Surveys* 44(1): Article 1, 50 pages.

Chen, Jisong, Rowena Chau, and Chung-Hsing Yeh (2004) 'Discovering Parallel Text from the World Wide Web', in *Proceedings of the 2nd Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation*, 157–161.

Chen, Jiang and Jian-Yun Nie (2000) 'Automatic Construction of Parallel English-Chinese Corpus for Cross-language Information Retrieval', in *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-NAACL 2000)*, 21–28.

Chen, Jiang and Jian-Yun Nie (2000) 'Parallel Web Text Mining for Cross-language IR', in *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) (RIAO 2000)*, 62–77.

Choi, Yejin, Eric Breck, and Claire Cardie (2006) 'Joint Extraction of Entities and Relations for Opinion Recognition', in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, 431–439.

Cohen, Aaron M. and William R. Hersh (2005) 'A Survey of Current Work in Biomedical Text Mining', *Briefings in Bioinformatics* 6(1): 57–71.

Craswell, Nick and Martin Szummer (2007) 'Random Walks on the Click Graph', in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, 239–246.

Croft, W. Bruce and D.J. Harper (1979) 'Using Probabilistic Models on Document Retrieval without Relevance Information', *Journal of Documentation* 35(4): 285–295.

Crouch, Carolyn J. (1990) 'An Approach to the Automatic Construction of Global Thesauri', *Information Processing and Management* 26(5): 629–640.

Cui, Hang, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma (2002) 'Probabilistic Query Expansion Using Query Logs', in *Proceedings of the 11th International Conference on World Wide Web (WWW 2002)*, 325–332.

Dang, Van and Bruce W. Croft (2010) 'Query Reformulation Using Anchor Text', in *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010)*, 41–50.

Das, Sanjiv and Mike Chen (2001) 'Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards', in *Proceedings of the 8th Asia Pacific Finance Association Annual Conference (APFA)*, 37–56.

Ding, Xiaowen, Bing Liu, and Philip S. Yu (2008) 'A Holistic Lexicon-based Approach to Opinion Mining', in *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM 2008)*, 231–240.

Duh, Kevin, Akinori Fujino, and Nagata, Masaaki (2011) 'Is Machine Translation Ripe for Cross-lingual Sentiment Classification?' in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HTL 2011)*, 429–433.

Esuli, Andrea and Fabrizio Sebastiani (2006) 'SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining', in *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, 417–422.

Esplà-Gomis, Miquel and Mikel L Forcada (2010) 'Combining Content-based and URL-based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor', *The Prague Bulletin of Mathematical Linguistics* 93:77–86.

Feldman, Ronen and Ido Dagan (1995) 'Knowledge Discovery in Textual Databases (KDT)', in *Proceedings of the First International Conference on Knowledge Discovery (KDD-95)*, 112–117.

Firth, Frank E. (1958a) 'An Experiment in Mechanical Searching of Research Literature with RAMAC', in *Proceedings of the May 6-8, 1958, Western Joint Computer Conference: Contrasts in Computers (IRE-ACM-AIEE '58 (Western))*, 168–175.

Firth, Frank E. (1958b) *An Experiment in Literature Searching with the IBM 305 RAMAC*, San Jose, California: IBM.

Franz, Martin, J. Scott McCarley, and Salim Roukos (1999) 'Ad hoc and Multilingual Information Retrieval at IBM', in *Proceedings of the 7th Text REtrieval Conference (TREC-7)*, 157–168.

Fuhr, Norbert (1989) 'Optimum Polynomial Retrieval Functions Based on the Probability Ranking Principle', *ACM Transactions on Information Systems* 7(3): 183–204.

Fuhr, Norbert (1992) 'Probabilistic Models in Information Retrieval', *The Computer Journal* 35(3): 243–255.

Gamon, Michael and Anthony Aue (2005) 'Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms', in *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, 57–64.

Gao, Jianfeng, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, Changning Huang (2001) 'Improving Query Translation for Cross-language Information Retrieval Using Statistical Models', in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2001)*, 96–104.

Gao, Jianfeng, Xiaodong He, and Jian-Yun Nie (2010) 'Clickthrough-based Translation Models for Web Search: From Word Models to Phrase Models', in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, 1139–1148.

Garfield, Eugene (1997) 'A Tribute to Calvin N. Mooers, A Pioneer of information Retrieval', *Scientist* 11(6): 9–11.

Ghani, Rayid, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano (2006) 'Text Mining for Product Attribute Extraction', *ACM SIGKDD Explorations Newsletter* 8(1): 41–48.

Grefenstette, Gregory (1999) 'The World Wide Web as a Resource for Example-based Machine Translation Tasks', in *Translating and the computer 21: Proceedings of the 21st International Conference on Translating and the Computer*, London: Aslib.

Grünwald, Peter D. (2007) *The Minimum Description Length Principle*. Cambridge, MA: MIT Press.

Guo, Honglei, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, and Zhong Su (2010) 'OpinionIt: A Text Mining System for Cross-lingual Opinion Analysis', in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, 1199–1208.

Hai, Zhen, Kuiyu Chang, and Jung-jae Kim (2011) 'Implicit Feature Identification via Co-occurrence Association Rule Mining', in *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*, LNCS 6608, Springer, 393–404.

Harman, K. Donna (ed.) (1993) *The First Text REtrieval Conference* (TREC-1), NIST Special Publication 500–207, Gaithersburg, MD: National Institute of Standards and Technology.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown (1997) 'Predicting the Semantic Orientation of Adjectives', in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*, 174–181.

Heinz, Steffen and Justin Zobel (2003) 'Efficient Single-pass Index Construction for Text Databases', *Journal of the American Society for Information Science and Technology* 54(8): 713–729.

Hearst, Marti A. (1999) 'Untangling Text Data Mining', in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (ACL 1999), 3–10.

Hearst, Marti (2003) 'What is Text Mining?', unpublished essay, UC Berkeley, available at http://people.ischool.berkeley.edu/~hearst/text-mining.html.

Hofmann, Thomas (1999) 'Probabilistic Latent Semantic Indexing', in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1999)*, 50–57.

Hovy, Eduard (2005) 'Text Summarization', in R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 583–598.

Hovy, Eduard and Chin-Yew Lin (1999) 'Automated Text Summarization in SUMMARIST', in Inderjeet Mani and Mark T. Maybury (eds), *Advances in Automatic Text Summarisation*, Cambridge, MA: MIT Press, 81–97.

Hu, Minqing and Bing Liu (2004) 'Mining and Summarizing Customer Reviews', in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, 168–177.

Järvelin, Kalervo and Jaana Kekäläinen (2000) 'IR Evaluation Methods for Retrieving Highly Relevant Documents', in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, 41–48.

Järvelin, Kalervo and Jaana Kekäläinen (2002) 'Cumulated Gain-based Evaluation of IR Techniques', *ACM Transactions on Information Systems* 20(4): 422–446.

Jing, Yufeng and W. Bruce Croft (1994) 'An Association Thesaurus for Information Retrieval', in *Proceedings of the 4th International Conference on Computer Assisted Information Retrieval (Recherche d'Informations Assistee par Ordinateur) (RIAO1994)*, 146–160.

Joachims, Thorsten (2002) 'Optimizing Search Engines Using Clickthrough Data', in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, 133–142.

Johansson, Richard and Alessandro Moschitti (2010) 'Reranking Models in Fine-grained Opinion Analysis', in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 519–527.

Kilgarriff, Adam and Gregory Grefenstette (2003) 'Introduction to the Special Issue on the Web as Corpus', *Computational Linguistics* 29(3): 333–347.

Kim, Soo-Min and Eduard Hovy (2004) 'Determining the Sentiment of Opinions', in *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, 1367–1373.

Kim, Soo-Min and Eduard Hovy (2006) 'Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text', in *Proceedings of COLING-ACL 2006 Workshop on Sentiment and Subjectivity in Text*, 1–8.

Kim, Jin-Dong, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii (2003) 'GENIA Corpus–A Semantically Annotated Corpus for Bio-textmining', *Bioinformatics* 19 (suppl. 1): i180–i182.

Kit, Chunyu (2000) *Unsupervised Lexical Learning as Inductive Inference*, PhD thesis, University of Sheffield.

Kit, Chunyu (2005) 'Unsupervised Lexical Learning as Inductive Inference via Compression', in James W. Minett and William S-Y. Wang (eds) *Language Acquisition, Change and Emergence: Essays in Evolutionary Linguistics*, Hong Kong: City University of Hong Kong Press, 251–296.

Kit, Chunyu and Jessica Y. H. Ng (2007) 'An Intelligent Web Agent to Mine Bilingual Parallel Pages via Automatic Discovery of URL Pairing Patterns', in *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops: Workshop on Agents and Data Mining Interaction (ADMI 2007)*, 526–529.

Kit, Chunyu and Yorick Wilks (1999) 'Unsupervised Learning of Word Boundary with Description Length Gain', in Miles Osborne and Erik T. K. Sang (eds) *CoNLL99: Computational Natural Language Learning*, 1–6.

Kleinberg, Jon M. (1998) 'Authoritative Sources in a Hyperlinked Environment', in *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 668–677. An extended version was published in *Journal of the Association for Computing Machinery* 46(5): 604–632. Also as IBM Research Report RJ 10076, May 1997.

Kraaij, Wessel, Jian-Yun Nie, and Michel Simard (2003) 'Embedding Web-based Statistical Translation Models in Cross-language Information Retrieval', *Computational Linguistics* 29(3): 381–420.

Krauthammer, Michael and Goran Nenadic (2004) 'Term Identification in the Biomedical Literature', *Journal of Biomedical Informatics* 37(6): 512–526.

Ku, Lun-Wei, Yu-Ting Liang, and Hsin-Hsi Chen (2006) 'Tagging Heterogeneous Evaluation Corpora for Opinionated Tasks', in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 667–670.

Kupiec, Julian, Jan Pedersen, and Francine Chen (1995) 'A Trainable Document Summarizer', in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1995)*, 68–73.

Lafferty, John and Chengxiang Zhai (2001) 'Document Language Models, Query Models, and Risk Minimization for Information Retrieval', in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2001)*, 111–119.

Lavrenko, Victor and W. Bruce Croft (2001) 'Relevance-based Language Models', in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2001)*, 120–127.

Leibowitz, Jacob, Julius Frome, and Don D. Andrews (1958) *Variable Scope Patent Searching by an Inverted File Technique*, Patent Office Research and Development Reports No. 14, U. S. Department of Commerce, Washington, DC, 17 November 1958.

Leser, Ulf and Jörg Hakenberg (2005) 'What Makes a Gene Name? Named Entity Recognition in the Biomedical Literature', *Briefings in Bioinformatics* 6(4): 357–369.

Li, Hang (2011) *Learning to Rank for Information Retrieval and Natural Language Processing*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 4(1):1–113.

Li, Fangtao, Chao Han, Minlie Huang, Xiaoyan Zhu,Ying-Ju Xia, Shu Zhang, and Hao Yu (2010) 'Structure-aware Review Mining and Summarization', in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 653–661.

Lin, Chin-Yew (1999) 'Training a Selection Function for Extraction', in *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM 1999)*, 55–62.

Lin, Chin-Yew (2004) 'ROUGE: A Package for Automatic Evaluation of Summaries', in *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out*, 74–81.

Lin, Chin-Yew and Eduard H. Hovy (2003) 'Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics', in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, 71–78.

Lin, Dekang (1998) 'Automatic Retrieval and Clustering of Similar Words', in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, 768–774.

Liu, Bing (2010) 'Sentiment Analysis and Subjectivity', in Nitin Indurkhya and Fred J. Damerau (ed.) *Handbook of Natural Language Processing*, 2nd edition, Boca Raton, FL: Chapman & Hall/CRC, 627–666.

Liu, Bing (2012) *Sentiment Analysis and Opinion Mining*. San Rafao, CA: Morgan & Claypool Publishers.

Liu, Shuang, Fang Liu, Clement Yu, and Weiyi Meng (2004) 'An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases', in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, 266–272.

Liu, Tie-Yan (2009) 'Learning to Rank for Information Retrieval', *Foundations and Trends in Information Retrieval* 3(3): 225–331.

Liu, Tie-Yan, Xu Jun, Tao Qin, Wenying Xiong, and Hang Li (2007) 'LETOR: Benchmark Dataset for Research on Learning to Rank for Information Retrieval', in *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007),* 3–10.

Liu, Yi, Rong Jin, and Joyce Y. Chai (2005) 'A Maximum Coherence Model for Dictionary-based Cross-language Information Retrieval', in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 2005)*, 536–543.

Luhn, Hans Peter (1958) 'The Automatic Creation of Literature Abstracts', *IBM Journal of Research and Development* 2(2): 159–165.

Luk, Robert W. P. and Wai Lam (2007) 'Efficient In-memory Extensible Inverted File', *Information Systems* 32(5): 733–754.

Ma, Xiaoyi and Mark Liberman (1999) 'BITS: A Method for Bilingual Text Search over the Web', in *Proceedings of Machine Translation Summit VII*, 538–542.

Ma, Yanjun, Jian-Yun Nie, Hua Wu, and Haifeng Wang (2012) 'Opening Machine Translation Black Box for Cross-language Information Retrieval', in *Information Retrieval Technology: Proceedings of the 8th Asia Information Retrieval Societies Conference (AIRS 2012)*, Berlin/Heidelberg: Springer, 467–476.

Mandala, Rila, Takenobu Tokunaga, and Hozumi Tanaka (1999) 'Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion', in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 1999)*, 191–197.

Manning, Christopher D. (1993) 'Automatic Acquisition of A Large Sub Categorization Dictionary From Corpora', in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL 1993)*, 235–242.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008) *Introduction to Information Retrieval*, Cambridge University Press.

McCarley, J. Scott (1999) 'Should We Translate the Documents or the Queries in Cross-language Information Retrieval?' in *Proceedings of the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, 208–214.

Metzler, Donald and W. Bruce Croft (2005) 'A Markov Random Field Model for Term Dependencies', in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, 472–479.

Metzler, Donald and W. Bruce Croft (2007) 'Latent Concept Expansion Using Markov Random Fields', in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, 311–318.

Mihalcea, Rada, Carmen Banea, and Janyce Wiebe (2007) 'Learning Multilingual Subjective Language via Cross-lingual Projections', in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, 976–983.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller (1990) 'WordNet: An Online Lexical Database', *International Journal of Lexicography* 3(4): 235–244.

Moffat, Alistair and Justin Zobel (1998) 'Exploring the Similarity Space', *SIGIR Forum* 32(1): 18–34.

Moore, Robert T. (1961) 'A Screening Method for Large Information Retrieval Systems', in *Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference (IRE-AIEE-ACM'61 (Western)),* 259–274.

Mooers, Calvin E. (1950) 'Coding, Information Retrieval, and the Rapid Selector', *American Documentation* 1(4): 225–229.

Moulinier, Isabelle and Hugo Molina-Salgado (2003) 'Thomson Legal and Regulatory Experiments for CLEF 2002', in C.A. Peters (ed.), *Advances in Cross-Language Information Retrieval: 3rd Workshop of the Cross-Language Evaluation Forum (CLEF 2002)*, Berlin/Heidelberg: Springer, 155–163.

Nguyen, Dong, Arnold Overwijk, Claudia Hauff, Dolf R. B. Trieschnigg, Djoerd Hiemstra, and Franciska De Jong (2008) 'WikiTranslate: Query Translation for Cross-lingual Information Retrieval Using only Wikipedia', in *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum (CLEF'08)*, Berlin/Heidelberg: Springer, 58–65.

Nie, Jian-Yun (2010) *Cross-Language Information Retrieval*, Synthesis Lectures on Human Language Technologies, San Rafael, CA: Morgan & Claypool Publishers, 3(1): 1–125.

Nie, Jian-Yun, Pierre Isabelle, and George Foster (1998) 'Using a Probabilistic Translation Model for Cross-language Information Retrieval'', in *Proceedings of the 6th Workshop on Very Large Corpora*, 18–27.

Nie, Jian-Yun, Michel Simard, Pierre Isabelle, and Richard Durand (1999) 'Crosslanguage Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web', in *Proceedings*

of the 22nd Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1999), 74–81.

Nolan, J.J. (1958) *Principles of Information Storage and Retrieval Using a Large Scale Random Access Memory*. San Jose, CA: IBM.

Oard, Douglas and Bonnie J. Dorr (1996) *A Survey of Multilingual Text Retrieva*l, Research Report UMIACS-TR-96-19 CS-TR-3615, 31 pages, University of Maryland.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith (2010) 'From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series', in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2010),* 122–129.

Olivier, Donald Cort (1968) *Stochastic Grammars and Language Acquisition Mechanisms*, PhD thesis, Harvard University.

Pang, Bo and Lillian Lee (2004) 'A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts', in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, 271–278.

Pang, Bo and Lillian Lee (2005) 'Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales', in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 115–124.

Pang, Bo and Lillian Lee (2008) 'Opinion Mining and Sentiment Analysis', *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002) 'Thumbs up? Sentiment Classification Using Machine Learning Techniques', in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2002), 79–86.

Ponte, Jay M. and Croft, W. Bruce (1998) 'A Language Modeling Approach to Information Retrieval', in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1998)*, 275–281.

Popescu, Ana-Maria and Oren Etzioni (2005) 'Extracting Product Features and Opinions from Reviews', in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005)*, 339–346.

Porter, Martin F. (1980) 'An Algorithm for Suffix Stripping', *Program* 14(3): 130–137.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen (2011) 'Opinion Word Expansion and Target Extraction through Double Propagation', *Computational Linguistics* 37(1): 9–27.

Qiu, Yonggang and Hans-Peter Frei (1993) 'Concept based Query Expansion', in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1993)*, 160–169.

Resnik, Philip (1998) 'Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text', in D. Farwell, L. Gerber, and E. Hovy (eds) *Machine Translation and the Information Soup: 3rd Conference of the Association for Machine Translation in the Americas (AMTA 1998)*, Springer, 72–82.

Resnik, Philip (1999) 'Mining the Web for Bilingual Text', in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, 527–534.

Resnik, Philip and Noah A Smith (2003) 'The Web as a Parallel Corpus', *Computational Linguistics* 29(3): 349–380.

Rissanen, Jorma (1978) 'Modeling by Shortest Data Description', *Automatica* 14(5): 465–471.

Rissanen, Jorma (1989) *Stochastic Complexity in Statistical Inquiry Theory*, River Edge, NJ: World Scientific Publishing.

Robertson, Stephen E. and Karen Sparck Jones (1976) 'Relevance Weighting of Search Terms', *Journal of the American Society for Information Science* 27(3): 129–146.

Robertson, Stephen E. and Karen Sparck Jones (1994) *Simple, Proven Approaches to Text Retrieval*, Technical Report UCAM-CL-TR-356, Computer Laboratory, University of Cambridge.

Robertson, Stephen E., Stephen Walker, and Micheline Hancock-Beaulieu (1999) 'Okapi at TREC–7: Automatic Ad Hoc, Filtering, VLC and Filtering Tracks', in *Proceedings of the 7th Text REtrieval Conference (TREC-7)*, 253–264.

Rocchio, J.J. (1965/1971) 'Relevance Feedback in Information Retrieval', reprinted in Gerald Salton (ed.) (1971) *The SMART Retrieval System – Experiments in Automatic Document Processing*, Englewood Cliffs, NJ: Prentice Hall, 313–323.

Sadat, Fatiha, Masatoshi Yoshikawa, and Shunsuke Uemura (2003) 'Learning Bilingual Translations from Comparable Corpora to Cross-language Information Retrieval: Hybrid Statistics-based and Linguistics-based Approach', in *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages (IRAL)*, 57–64.

Salton, Gerard (1970) 'Automatic Processing of Foreign Language Documents', *Journal of the American Society for Information Science*, 21(3): 187–194.

Salton, Gerard (ed.) (1971) *The SMART Retrieval System – Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice Hall.

Salton, Gerard and Christopher Buckley (1988) 'Term-weighting Approaches in Automatic Text Retrieval', *Information Processing and Management* 24(5): 513–523.

Salton, Gerald, Anita Wong, and Chung-Shu Yang (1975) 'A Vector Space Model for Automatic Indexing', *Communications of the ACM* 18(11): 613–620.

Savoy, Jacques and Ljiljana Dolamic (2009) 'How Effective is Google's Translation Service in Search?', *Communications of the ACM* 52(10): 139–143.

Scherer, Klaus R. (1984) 'Emotion as a Multicomponent Process: A Model and Some Cross-cultural data', *Review of Personality and Social Psychology* 5: 37–63.

Schütze, Hinrich and Jan O. Pedersen (1997) 'A Co-occurrence-based Thesaurus and Two Applications to Information Retrieval', *Information Processing and Management* 33(3): 307–318.

Shanahan, James G., Yan Qu, and Janyce Wiebe (2006*) Computing Attitude and Affect in Text: Theory and Applications*, the Information Retrieval Series, Vol. 20, New York: Springer.

Sheridan, Páraic and Jean Paul Ballerini (1996) 'Experiments in Multilingual Information Retrieval Using the SPIDER System', in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1996)*, 58–65.

Shi, Lixin and Jian-Yun Nie (2010) 'Using Various Term Dependencies according to their Utilities', in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, 1493–1496.

Shi, Lei, Cheng Niu, Ming Zhou, and Jianfeng Gao (2006) 'A DOM Tree Alignment Model for Mining Parallel Data from the Web', in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, 489–496.

Simpson, Matthew S. and Dina Demner-Fushman (2012) 'Biomedical Text Mining: A Survey of Recent Progress', in Charu C. Aggarwal and Chengxiang Zhai (eds) *Mining Text Data*, New York: Springer, 465–517.

Singhal, Amit, Chris Buckley, and Mandar Mitra (1996) 'Pivoted Document Length Normalization', in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1996)*, 21–29.

Singhal, Amit, Gerard Salton, and Chris Buckley (1996) 'Length Normalization in Degraded Text Collections', in *Proceedings of the Annual Symposium on Document Analysis and Information Retrieval (SDAIR 1996)*, 149–162.

Singhal, Amit, John Choi, Donald Hindle, David Lewis, and Fernando Pereira (1999) 'AT&T at TREC-7', in *Proceedings of the 7th Text REtrieval Conference (TREC-7)*, 239–252.

Solomonoff, Ray J. (1964) 'A Formal Theory of Inductive Inference', *Information Control*, 7:1–22 (part 1), 224–256 (part 2).

Sparck Jones, Karen (1972) 'A Statistical Interpretation of Term Specificity and its Application in Retrieval', *Journal of Documentation* 28(1): 11–21; reprinted in *Journal of Documentation* 60(5): 493–502.

Sparck Jones, Karen (ed.) (1981) *Information Retrieval Experiment*. London: Butterworths.

Sparck Jones, Karen (1999) 'Automatic Summarising: Factors and Directions', in Inderjeet Mani and Mark T. Maybury (eds) *Advances in Automatic Text Summarisation*, Cambridge, MA: MIT Press, 1–12.

Sparck Jones, Karen and Peter Willet (1997) *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann.

Stone, Philip J. (1966) *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: MIT Press.

Stoyanov, Veselin and Claire Cardie (2008) 'Topic Identification for Fine-grained Opinion Analysis', in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, 817–824.

Su, Qi, Xinying Xu, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen, and Zhong Su (2008) 'Hidden Sentiment Association in Chinese Web Opinion Mining', in *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, 959–968.

Swanson, Don R. (1986) 'Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge', *Perspectives in Biology and Medicine* 30(1): 7–18.

Swanson, Don R. (1987) 'Two Medical Literatures That Are Logically but Not Bibliographically Connected', *Journal of the American Society for Information Science* 38(4): 228–233.

Swanson, Don R. (1988) 'Migraine and Magnesium: Eleven Neglected Connections', *Perspectives in Biology and Medicine* 31(4): 526–557.

Swanson, Don R. (1988) 'Historical Note: Information Retrieval and the Future of an Illusion', *Journal of the American Society for Information Science* 39(2): 92–98.

Swanson, Don R. (1990) 'Somatomedin C and Arginine: Implicit Connections between Mutually Isolated Literatures', *Perspectives in Biology and Medicine* 33(2): 157–186.

Swanson, Don R. (1991) 'Complementary Structures in Disjoint Science Literatures', in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1991)*, 280–289.

Swanson, Don R., Neil R. Smalheiser, and A. Bookstein (2001) 'Information Discovery from Complementary Literatures: Categorizing Viruses as Potential weapons', *Journal of the American Society for Information Science and Technology*, 52(10): 797–812.

Teufel, Simone (2007) 'An Overview of Evaluation Methods in TREC Ad-hoc Information Retrieval and TREC Question Answering', in Laila Dybkjaer, Holmer Hemsen, and Wolfgang Minker (eds) *Evaluation of Text and Speech Systems*, Dordrecht: Springer, 163–186.

Türe, Ferhan, Jimmy J. Lin and Douglas W. Oard (2012) 'Combining Statistical Translation Techniques for Cross-language Information Retrieval', in *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, 2685–2702.

Türe, Ferhan and Jimmy J. Lin (2013) 'Flat vs. Hierarchical Phrase-based Translation Models for Cross-language Information Retrieval', in *Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR-2013)*, 813–816.

Turney, Peter (2002) 'Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, 417–424.

van Rijsbergen, Cornelis Joost (1979) *Information Retrieval*, 2nd edition. London: Butterworths.

Venkataraman, Anand (2001) 'A Statistical Model for Word Discovery in Transcribed Speech', *Computational Linguistics* 27(3): 351–372.

Voorhees, Ellen M. (1993) 'Using WordNet to Disambiguate Word Senses for Text Retrieval', in *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (SIGIR 1993)*, 171–180.

Voorhees, Ellen M. (1994) 'Query Expansion Using Lexical-semantic Relations', in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994)*, 61–69.

Voorhees, Ellen M. and Donna K. Harman (eds) (2005) *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge, MA: MIT Press.

Wallace, C. S. and D. M. Boulton (1968) 'An Information Measure for Classification', *Computer Journal* 11(2): 185–194.

Wallace, C. S. and P. R. Freeman (1987) 'Estimation and Inference by Compact Coding', *Journal of the Royal Statistical Society*, Series B, 49(3): 240–265.

Wan, Xiaojun (2008) 'Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis', in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, 553–561.

Wan, Xiaojun (2009) 'Co-Training for Cross-lingual Sentiment Classification', in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, 235–243.

Weeber, Marc, Henny Klein, Lolkje T. W. de Jong-van den Berg, and Rein Vos (2001) 'Using Concepts in Literature-based Discovery: Simulating Swanson's Raynaud–Fish Oil and Migraine–Magnesium Discoveries', *Journal of the American Society for Information Science and Technology* 52(7): 548–557.

Wei, Bin and Christopher Pal (2010) 'Cross Lingual Adaptation: An Experiment on Sentiment Classifications', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, 258–262.

Wen, Ji-Rong, Jian-Yun Nie, and Hong-Jiang Zhang (2001) 'Clustering User Queries of a Search Engine', in *Proceedings of the 10th International Conference on World Wide Web (WWW 2001)*, 162–168.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie (2005) 'Annotating Expressions of Opinions and Emotions in Language', *Language Resources and Evaluation* 39(2-3): 165–210.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann (2005) 'Recognizing Contextual Polarity in Phrase-level Sentiment Analysis', in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005)*, 347–354.

Witten, Ian H., Alistair Moffat, and Timothy C. Bell (1999) *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd edition, San Francisco: Morgan Kaufmann Publishers.

Xu, Hongbo, Tianfang Yao, Xuanjing Huang, Huifeng Tang, Feng Guan, and Jin Zhang (2009) 'Overview of Chinese Opinion Analysis Evaluation 2009', in *Proceedings of the 2nd Chinese Opinion Analysis Evaluation (COAE 2009)*. (In Chinese.)

Xu, Jinxi and W. Bruce Croft (2000) 'Improving the Effectiveness of Information Retrieval with Local Context Analysis', *ACM Transactions on Information Systems* 18(1): 79–112.

Yang, Hui, Luo Si, and Jamie Callan (2006) 'Knowledge Transfer and Opinion Detection in the TREC2006 Blog Track', in *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*.

Ye, Sha-ni, Ya-juan Lv, Yun Huang, and Qun Liu (2008) 'Automatic Parallel Sentence Extraction from Web', *Journal of Chinese Information Processing* 22(5): 67–73. (In Chinese.)

Zhai, ChengXiang (2008) *Statistical Language Models for Information Retrieval*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 1(1): 1–141.

Zhai, Chengxiang and John Lafferty (2001) 'Model-based Feedback in the Language Modeling Approach to Information Retrieval', in *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM 2001)*, 403–410.

Zhang, Chengzhi, Xuchen Yao, and Chunyu Kit (2013) 'Finding more Bilingual Webpages with High Credibility via Link Analysis', in *Proceedings of the 6th Workshop on Building and Using Comparable Corpora (BUCC 2013)*, 138–143.

Zhang, Ying, Ke Wu, Jianfeng Gao, and Phil Vines (2006) 'Automatic Acquisition of Chinese-English Parallel Corpus from the Web', in *Advances in Information Retrieval, Proceedings of 28th European Conference on IR Research (ECIR 2006)*, Springer, 420–431.

Zhao, Jun, Hongbo Xu, Xuanjing Huang, Songbo Tan, Kang Liu, and Qi Zhang (2008) 'Overview of Chinese Opinion Analysis Evaluation 2008', in *Proceedings of the 1st Chinese Opinion Analysis Evaluation (COAE 2008)*, 1–20. (In Chinese.)

Zhuang, Li, Feng Jing, and Xiaoyan Zhu (2006) 'Movie Review Mining and Summarization', in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM 2006)*, 43–50.

Zobel, Justin and Alistair Moffat (2006) 'Inverted Files for Text Search Engines', *ACM Computing Surveys* 38(2): Article 6.

Zweigenbaum, Pierre, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen (2007) 'Frontiers of Biomedical Text Mining: Current Progress', *Briefings in Bioinformatics* 8(5): 358–375.

32

LANGUAGE CODES AND LANGUAGE TAGS

Sue Ellen Wright

KENT STATE UNIVERSITY, THE UNITED STATES

Introduction

Most people are unaware that every time they log on the web or use a computer program, watch a TV program, or use a cell phone, they invoke some sort of language code. Sometimes the codes become visible even to the inexperienced user if a colleague who mostly works in another language sends a word-processing file that starts generating spelling errors because it is checking English text against a dictionary for a different language. Or we find we cannot follow a link that happens to use Russian or Chinese characters in a Western European computing environment, an issue that may disappear with the further implementation of the World Wide Web's (W3C) *Internationalization Tagset* (W3C, 2013).

Chances are that unless the user knows how language codes, or the related system of language tags, work, the problems described above may end in frustration. In any event, somewhere, embedded in the code that governs the Web or generates any screen display, a simple string of characters provides the information fuel that keeps the engine of the Web running: for instance, en-US, zh-CN, deu-CHE, esp-MEX – American English, Mainland Chinese, Swiss German, or Mexican Spanish. In the problematic cases cited above, adjusting either the code, the spell-checker setting, or the character set used in a given environment may be required to set things right in the computing environment.

This chapter will examine what a language is as well as other related concepts, such as language families and groups, language variants, and dialects. The first part of the chapter provides an outline-like handy reference to the many stakeholders and standards used to encode or otherwise characterize languages and language variants, while the second part provides a more detailed review of the history and future directions affecting the creation, maintenance, and application of language identifiers.

Language codes – what is a language?

Printed text relies on humans to know or guess the language of the text. Every translator experiences how difficult this can be for non-linguists, for instance when uninformed clients send a Catalan text to a Spanish translator. Computers, however, need to be "told" explicitly what to expect, although many utilities exist today for parsing text to determine its probable language. Protocols governing content representation, especially in Web environments,

"declare" language, locale (country or region), script, and other related information using a *language code* or a *language tag*. The language codes are presented in the International Organization for Standardization ISO 639 six-part family of standards. Not until Part 4 appeared did it seem necessary to the responsible committee to actually define what a language is for purposes of the codes: a "systematic use of sounds, characters, symbols or signs to express or communicate meaning or a message between humans." (ISO 639-4, 3.6) The standard clarifies, however, that this definition may not apply outside the context of the standard. The notion of a language is profiled with respect to that of dialects, a "language variant specific to a geographical region or group of language users" (3.8). While these concepts are so widely accepted that it took the committee in question many years to get round to putting the definitions in writing, the principles for assigning a particular manner of speaking (writing or thinking) to one or the other designation is not that transparent.

There may be general agreement that, for instance, English, German, Japanese, and Russian are languages, but as soon as the view moves away from the most common idioms, the certainty of what a language is can stand on shaky ground. Speakers of English and Chinese who have not studied each other's languages can be quite clear on the distinction: their respective speech conventions are mutually unintelligible. This notion of non-intelligibility is often used, even by linguists and in the code standards, to distinguish languages from dialects, but the assertion itself is also tenuous.

Speakers of Norwegian and Swedish, for instance, are capable of chatting together, each speaking his or her own language, but understanding most of what is said and delighting in their shared heritages. In contrast, Croatian and Serbian are also mutually intelligible, but they are traditionally represented by different scripts (Latin and Cyrillic, respectively), and for political and historical reasons, linguists and ordinary speakers alike are working hard to profile what was once designated as a single language in ways that underscore differences. This/these language(s) is/are (depending on one's perspective) being consciously split apart, invoking linguistic particularism to express political discord.

Should this trend sound regrettable or even bizarre, it is not unprecedented. After the American Revolution at the beginning of the nineteenth century, American lexicographers and linguists consciously introduced the spelling "reforms" that today distinguish written American English from its British cousin. Nevertheless, with some degree of change in lexis (word and term usage) and pronunciation over the course of two and a half centuries, American, Canadian, British, and even Australian, Indian, and South African English(es) remain more or less mutually comprehensible varieties of the same language. Although experts do indeed define dialects of English (Spanish, Russian), these huge speech communities are classified as recognized language variants of the same mother tongues, the same macrolanguages.

By the same token, although German asserts itself as a single individual language with many dialects, a speaker of a so-called "low German" dialect in Hamburg finds the everyday speech of a Bavarian mountaineer to be quite incomprehensible. The "high" German learned carefully throughout the German school system and spoken by educated people across Germany, Austria and even, with some stress and strain, part of Switzerland (as well as part of Belgium and in pockets of a German-speaking diaspora scattered about the world), is actually an artificial construct painstakingly fashioned at the end of the eighteenth and the beginning of the nineteenth century in an effort both to quell French influence and to unify a splintered nation with no single dominant political center. (Göttert, 2010) The situation with Italian is similar. Chinese is even more dramatic. Although it claims a highly portable, unified writing system, actual spoken variants differ even more dramatically than spoken dialects of German or Italian.

So if intelligibility or the lack thereof is not the determining criterion for defining a language, what is? An often quoted Yiddish definition claims *"a shprakh iz a diyalekt mit an armey un a flot"* (a language is a dialect with an army and a navy), despite the fact that the Yiddish language itself has never had either.[1] All joking aside, the distinction between the two concepts is essentially colored by historical and geo-political precedents. When all is said and done, a language variety is a *language* (as opposed perhaps to a *dialect*) if its speakers – or perhaps even some external researchers – have decided this is so, with the result that some mutually intelligible languages, like the Scandinavian languages already cited, or for instance, Portuguese and some dialects of Spanish, are classified as individual languages, while some unintelligible dialects are nonetheless considered to be dialects of larger macrolanguages. National boundaries may be cited as dividing lines in some cases, but idioms such as Basque, Catalan, and Kurdish, for instance, assert themselves as languages that straddle national and provincial borders and throughout their checkered histories have experienced both ethnic and linguistic suppression from various sides. Other languages like Yiddish, Ladino, and the languages of the Roma distinguish themselves as borderless migrant languages ever in motion and often subject to discrimination and persecution.

In some cases, external influences related to the global history of colonialism and various waves of cultural and political hegemony affect the choice of language identification. Current Iranian preference for "Persian" as the name of their language, in rejection of "Farsi," reflects a negative reaction to both Arabic (from which the latter stems) and western misunderstandings. Some may argue that these distinctions are political and not warranted as "scientific" linguistic factors, but personal, historical, and political forces play a significant role in defining both language and in some cases national identity. Nevertheless, even the most vehement nationalist positions on language as incontrovertibly associated with specific nation states are dangerously untenable in the long term. Even French, with its valorization of a single variant (the language of the Île de France) must at the pragmatic level tolerate the existence of regional dialects. Spain under Franco and Libya under Gadhafi promoted the ascendance of a single national language within their borders, only to see repressed regional languages and dialects break free at the first opportunity. Hence the juxtaposition of language with country should not be interpreted as an immutably aggregated entity, but rather as a theme with determiner – French *as spoken in Canada* – and nothing more. Nor dare we view the tendency of languages, like blood, to seep across boundaries to penetrate deep into the tissue of neighboring lands as an extension of the sovereignty of the mother country, an error that can have serious consequences if taken too literally, as evidenced by the history of Germany in the twentieth century.

How then, if it is so difficult to determine what a language is, or what a dialect is, when, and by whom have languages been classified and coded? What is the range of application for these codes? What determines their form and content?

Stakeholders and standards

Despite the ignorance of language identifiers on the part of the general public described in the introduction, there are several communities of practice (CoPs) that have a strong vested interest in naming and citing languages, as well as assigning designators to them, commonly called *language codes* and *language tags*. These CoPs are dedicated to the maintenance and stability of the codes as well as to responding to changing insights and attitudes, both scholarly and political. The following section groups these CoPs together with their standards and resources in order to provide a roadmap designed to traverse a potentially forbidding landscape.

Terminologists: ISO TC 37, *Terminology and other language and content resources*, and Infoterm together maintain the Registration Authority for

- *ISO 639-1:2002, Codes for the representation of names of languages — Part 1: Alpha-2 code.*[2]
  Alpha-2 codes are represented in lowercase and are for the most part formed mnemonically based on the name of a language as it is expressed in that language, e.g., ***en, es, de*** for English, Spanish (español) and German (Deutsch). The current Alpha-2 code comprises 136 language identifiers.

Librarians and information scientists: ISO TC 46, *Information and documentation*, and the US Library of Congress (LoC), maintain the Registration Authority for

- ISO 639-2:1998, *Codes for the representation of names of languages — Part 2: Alpha-3 code.*

and for

- ISO 639-5:2008, Codes for the representation of names of languages — Part 5: Alpha-3 code for language families and groups.
  Alpha-3 codes are formed mnemonically where possible, with the note that there are some English-based forms that exist in parallel with native-tongue forms, e.g., *eng, fra/fre, spa/esp, ger/deu*. The collection comprises 464 language identifiers.
  LoC publishes information on these codes (including 639-1 two-letter codes) in its ISO 639-2 Registration Authority page, http://www.loc.gov/standards/iso639-2/, and in the *MARC Code List for Languages*, http://www.loc.gov/marc/languages/.

OCLC (originally called the Ohio College Library Center), which calls itself "the nonprofit, membership, computer library service and research organization," specifies procedures for entering MARC language codes in various fields of the OCLC search record in its *OCLC Bibliographic Formats and Standards 047, Language Code (R)*, available at https://www.oclc.org/bibformats/en/0xx/041.html.

Translators and field linguists: SIL International maintains the Registration Authority for

- ISO 639-3:2007, *Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages*).
  The 639-3 codes follow the same pattern as 639-2, with exception that there are no ambiguous English-based designators. It provides language identifiers for 7,105 languages.
- SIL International maintains an easy-to access and interpret table documenting ISO 639, Parts 1-3 and 5 at *ISO 639 Code Tables*, http://www-01.sil.org/iso639-3/codes.asp?order=639_2&letter=a, and provides detailed information on each individual code point at http://www.ethnologue.com/world.

"Voluntary transnational researchers" working under the auspices, sequentially, of Linguasphere, Geolang, and the Observatoire linguistique, comprise the Registration Authority for

- ISO 639-6:2009, *Codes for the representation of names of languages – Part 6: Alpha-4 code for comprehensive coverage of language variants*
  The standard describes procedures for documenting the Alphha-4 codes for the purpose of representing dialects and varieties, but unfortunately, it offers no examples or models for the actual codes. Information on these codes is supposed to be available at the Geolang link, http://www.geolang.com/iso639-6/, but this URL appears to have been dormant

for some time. PDF representations of the original descriptors can be viewed and downloaded at the Observatoire linguistique site, http://www.linguasphere.info/lcontao/fichier-pdf.html, but examples of the actual four-letter codes are currently inaccessible. It should also be noted that despite the claim of multiple researchers, the website clarifies that much of the effort reflected in the collection is the work of a single person, David Dalby. The standard is due for ISO review in 2014.

The Linguists List maintains a list of extinct languages which are linked to its MultiTree resource, where the language or dialect is represented by a composite code, whose makeup depends on the declared components for that particular dialect or language. For instance, Middle Franconian Middle High German is classified as a dialect with the assigned code of *gmh-mfr* for *German Middle High – Middle FRanconian*. See:

- *Multitree*, "a searchable database of hypotheses on language relationships," http://multitree.org/

Linguists List also maintains:

- *Linguist List Codes for Ancient and Extinct Languages*, http://multitree.org/codes/extinct.html
- *Linguist List Codes for Constructed Languages,* http://linguistlist.org/forms/langs/GetListOfConstructedLgs.cfm
  Constructed languages reflect various efforts to artificially create languages, either for research purposes or as part of a fantasy universe (e.g., Tolkien's Elven languages). This list does not include famous constructed languages such as Esperanto and Klingon, which actually have their own official identifiers in ISO 639-3.

The Open Language Archives Community (OLAC), an international partnership of institutions and individuals creating a worldwide virtual library of language resources (closely affiliated with the Linguists List), defines the *OLAC Language Extension* as a Current Best Practice for tagging syntax used in encoding both the language of a document and the language discussed in a document using ISO 639-3 codes together with Linguist List extensions, as specified in: http://www.language-archives.org/REC/language.html. This tagging scheme is used in the OLAC Language Archives and elsewhere to identify language resources for purposes of information search and retrieval.

A sample OLAC notation for the language of document (in this case one written in German) is `<dc:language xsi:type="olac:language" olac:code="deu"/>`

or for the language that is the *topic* of a document (about the Suafa dialect of the Lao language of the Solomon Islands):

```
<dc:subject xsi:type="olac:language" olac:code="llu">Suafa dialect</dc:subject>
```

This tag name is potentially confusing to linguists accustomed to making the distinction between *subject language* (the language of a document) and *object language* (the language under discussion).

The ***Research Data Alliance Working Group on Standardisation of Data Categories and Codes*** is launching an effort to expand codes for the identification of languages and language varieties,

together with categories for describing the content of resources. Their work is based on the contention that current language codes require expansion at both the macro-language, but also at the dialect and variety levels (Musgrave, 2014). See https://rd-alliance.org/working-groups/standardisation-data-categories-and-codes-wg.html. The working group is currently exploring their needs with regard to the data points they will include in their resource; hence it is premature to cite a format for their notation.

The United Nations and the International Organization for Standardization (ISO): together under the auspices of the ISO 3166 Maintenance Agency in Geneva, Switzerland, administer

- ISO 3166-1:2006, *Codes for the representation of names of countries and their subdivisions — Part 1: Country codes; Part 2: Country subdivision code, and Part 3: Code for formerly used names of countries*.

  Country codes are expressed with capital letters to distinguish them from language codes and are available as Alpha-2, Alpha-3, and three-digit numerical forms, e.g. ***BE, BEL, 056, Belgium***. As shown above in the Introduction, language codes can be combined with country codes in an effort to represent national and regional variants.

Computer scientists, internationalization specialists, Internet and World Wide Web designers, under the auspices of the following entities maintain a variety of standards and normative recommendations:

- The ***World Wide Web Consortium (W3C)*** specifies that web pages and online resources shall utilize the *lang* (HTML) *xml:lang* (XML) attributes to identify the language of Web content. W3C does not specify the form of these elements; it only requires them, as stated in:
- W3C, *The global structure of an HTML document*, Chapter 8.1, "Specifying the language of content: the *lang* attribute," http://www.w3.org/TR/html401/struct/dirlang.html#adef-lang
- W3C, *Extensible Markup Language (XML)* 1.0 (Fifth Edition); W3C Recommendation 26 November 2008. http://www.w3.org/TR/REC-xml/

The Internet Engineering Task Force (IETF, BCP 47) specifies the syntax for creating language tags used as the values for *lang* and *xml:lang* in:

- *Tags for Identifying Languages*, https://tools.ietf.org/html/bcp47 (current edition as of 2013-12: IETF RFC 5646, Addison Phillips and Mark Davis, editors). Language tags can be as simple as a single two letter code or may indicate additional information, such as language, variety, script, region, e.g., ***zh-cmn-Hans-CN*** *(Chinese, Mandarin, Simplified script, as used in (mainland) China).*

The Internet Assigned Numbers Authority (IANA) assigns and maintains a

- *Registry of Language Subtags* designed to meet the private needs of individuals for whom the standard language tags do not reflect a particular desired level of specificity, http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry

The ***Unicode Consortium*** maintains the

- *Unicode Standard,* which mirrors the parallel ISO standard: *ISO 10646, Information technology – Universal Coded Character Set (UCS).* These standards provide "a character coding system designed to support the worldwide interchange, processing, and display of the written texts of the diverse languages and technical disciplines of the modern world,"

including support for classical and historical texts of many written languages. Unicode encoding replaces the earlier cumbersome system, which was based on the old *American Standard Code for Information Interchange (ASCII)* standard (ANSI_X3.4-1968 / 1986) and its multipart ISO companion ISO standards, ISO/IEC 8859 series, *Information technology – 8-bit single-byte coded graphic character sets*. Unicode maintains an extensive website providing access to codes, as well as a wealth of other information, at: http://www.unicode.org/versions/Unicode6.3.0/

Unicode also maintains:

- ISO 15924, *Codes for the representation of names of scripts* and
- The *Common Locale Data Registry (CLDR)*, which augments the language tags with locale-related information needed in computing environments, particularly on the web. These *locale IDs* can express a variety of information, such as currencies, time related information, region-specific spelling and capitalization rules, transliteration rules, keyboard layouts, and more, expressed using Unicode's *UTS #35: Unicode Locale Data Markup Language (LDML)*. Primary resources are available at: http://cldr.unicode.org/ and http://www.unicode.org/reports/tr35/, respectively.

A typical langauge tag might look like the example shown in Figure 32.1 below.

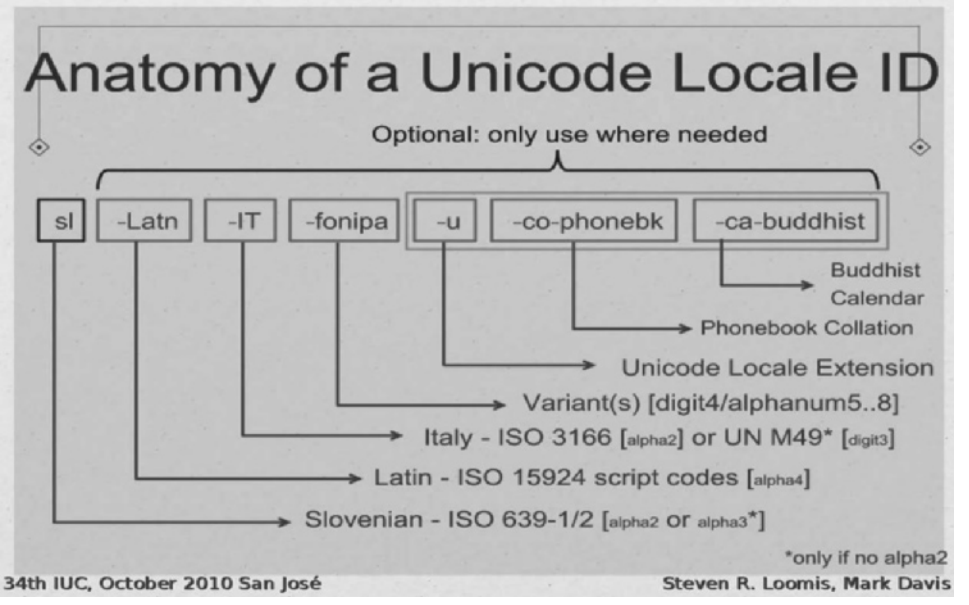


Figure 32.1 Unicode Locale ID taken from the CLDR

Note: where the code implies: Slovenian | represented in Latin script | as spoken in Italy | classified by the variant code fonipa | assigned the Unicode Locale Extension –u | collated according to phonebook rules | [and whimsically] subject to the Buddhist calendar conventions (Image © S.R. Loomis and M. Davis, 2010) https://docs.google.com/presentation/d/1rJaqrzxlywkiQDKS6JAenzdts3sPYVI3giMpcUOWkHs/present#slide=id.i89

Proprietary variants

Google Web Interface and Search Language Codes can be used to specify search in a specific language (as opposed to using the advanced search features provided by the standard search interface). A typical instantiation might read: http://www.google.com/search?hl=en . Detailed instructions are available at https://sites.google.com/site/tomihasa/google-language-codes.

Microsoft posts its [MS-LCID]: *Windows Language Code Identifier (LCID) Reference* on the MSDN Microsoft Developer Website. "The LCID structure is used to identify specific languages for the purpose of customizing software for particular languages and cultures." The identifiers consist of a primary language identifier and a sub-language identifier and are considered to be para-linguistic in that they do not relate to specific linguistic features differentiating language varieties (Constable and Simon, 2000). They are used to specify formatting for dates, times, and numbers, as well as sorting based on language elements. The code itself consists of a six digit string associated with either a simple or possibly complex IETF language tag. For instance, *0x0803 for ca-ES-valencia,* for the Valencian variety of Catalan. The list is posted at: http://msdn.microsoft.com/en-us/library/cc233965.aspx.

This encoding system powers the familiar dropdown lists used in many multilingual computing applications, such as the selection of specific keyboard layouts, or the specification of regional language variants used to classify Machine Translation or Translation Memory resources (see Figures 32.2 and 32.3).

IBM also specifies a short list of National Language Codes, using what appear to be country code conventions (all caps): ***I ITA Italia***; the list can be accessed at

- http://publib.boulder.ibm.com/infocenter/cicsts/v3r1/index.jsp?topic=%2Fcom.ibm.cics.ts31.doc%2Fdfhp4%2Ftopics%2Fdfhp4_nlscodes.htm

Figure 32.2 Sub-languages drop-down menu, MultiTerm™ 2011

Figure 32.3 Language keyboard selection menu, Microsoft Word™

Historical development

Historically speaking, the language codes as we know them today evolved out a need to save space and time within the framework of both terminology documentation and library cataloguing procedures in an era when both terminographers and lexicographers on the one hand and library cataloguers on the other recorded documentary information about terms or library holdings (books and other objects) on relatively small paper or cardboard *fiche* or catalogue cards. Terminologists working at Infoterm and in similar environments developed a system of two-letter codes (ISO 639-1) as identifiers for a relatively limited set of "familiar" languages. Even with the best of intentions, this collection is nonetheless limited to the number of languages it can accommodate because of the simple mathematical principle afforded the possible arithmetic permutations of the number "2" coupled with the 26 letters of the Latin alphabet. The alpha-2 codes were the first to be standardized, and they set the base length for language identifier fields in many legacy computing environments.

Librarians, both at the US Library of Congress and across Europe, were confronted with a broader collection of languages than required by the early terminologists. Viewed from the mathematical perspective, their three-letter-code solution immediately provides for a broader range of values and became ISO 639-2 in 1998. The introduction of this new collection caused some consternation, particularly in the computing community. First of all, the introduction of three-letter codes was problematic for systems built on two-character data fields. Needless to say, cries for expansion of the two-letter codes as an alternative were met with the logical admonition to "do the math." A further complication involves the presence of several instances of ambiguity: the original LoC designators were developed for use in English in the US, with a resulting set of mnemonic codes that reference the English names for languages, a practice that is mirrored in the European library community with parallel codes specified mnemonically according to the native-language name for those languages. There are 21 of these items, and

they are referred to as the *Bibliographic* (B) and the *Terminological* (T) codes, with the latter preferenced for non-bibliographical purposes. For instance, B codes *fre, ger,* and *spa* correspond to T counterparts *fra, deu*, and *esp* for French (français), German (Deutsch), and Spanish (español), respectively.

Although 639-2 is not necessarily limited by numerical permutations, it imposes its own constraints in that it has as its purpose the creation of codes for languages for which "a significant body of literature" exists. The original requirements for requesting a new code be added to the collection specified the holding of at least 50 documents in the language by one to five "agencies" (ISO 639-2). This stipulation rules out the inclusion of spoken languages without any tradition of written literature, and could actually pose the danger of becoming a kind of catch-22 clause in an environment where it becomes increasingly difficult to publish anything without assigning a language code to it, a factor of particular concern in Web environments.

ISO 639-2 also includes a number of so-called "collective language codes" designed for assignment to documents in language groups "where a relatively small number of documents exist or are expected to be written, recorded or created." ISO 639-5 represents a refinement of sorts of this approach by providing a segregated list of Alpha-3 codes for language families and groups, which provides broad classifiers for languages that did not have their own three-letter codes in 639-2, or for which it is sometimes expedient to cite related sublanguages as aggregates. One of the criticisms that has been voiced in this particular regard is that for pragmatic cataloging purposes, the original Alpha 3 set lumped some languages together that were not linguistically or culturally related. While this approach might work for library collections, it can cause sincere distress when the code is used to classify, for instance, the mother tongue of school children when a family finds itself classified with a designator that would more properly refer to a despised ethnic rival language.

Another concept introduced in 639-2 is the notion of the "macrolanguage," with Arabic cited as a primary example. In this case, the macro-language designation reflects the overriding notion of Arabic, both the classical language of the Koran and so-called Modern Standard Arabic (MSA, *al-fusha*), being the common language across the Arabic speaking region, in much the same way that standard German is a common language in an essentially diglossic area where everyone speaks both the "high" form of the literary and school language, as well as his or her own local dialect. An essential difference here, however, is that High German is indeed mastered, written and spoken by the vast majority of individuals who consider themselves to be German speakers. MSA in contrast is a written language that is not generally spoken and that is not mastered outside the ranks of the highly educated, while the spoken vernaculars are many and varied across the region. The upshot of these concerns is that collective and macrolanguages must be carefully scrutinized on an individual basis when using the codes for pragmatic applications.

Whereas the first two parts of ISO 639 dealt with major world languages used in technical publications requiring terminological documentation and that have significant bodies of catalogable literature, ISO 639-3 continues the three-letter code tradition with the intent to "support a large number of the languages that are known to have ever existed" (Scope statement). Administered by SIL International, the collection treats languages, both living, dead, and endangered. In contrast to the earlier standards, Part 3 is backed up by a more substantial database called *Ethnologue*, which includes data fields documenting the population of language speakers, geographical location, language maps showing geographic distribution, status, classification, dialects, use, resources, writing system, and other comments. With the exception of the language family designators, the three sets are configured such that Part 1 is a subset of 2, and 2 a subset of 3.

As illustrated above, ISO 639 Parts 1–3 combine with other related standards to form locale-specific two-part codes that reflect geographical variants. From the outset, Part 1 already provided for regional encoding by combining the Alpha-2 language codes with Alpha-2 country codes taken from ISO 3166-1, yielding the example: *elevator (en-US), lift (en-GB)*. In like manner, Part 2 specifies the combinatory pattern: spool of thread (eng-US), bobbin of cotton (eng-GB). Obviously, this principle informs the generation of more complex language codes and CLDR notations.

The country codes are specified in ISO 3166, again a multipart standard, which defines two- and three-letter codes, as well as numeric codes. The codes are based on country names cited in either the UN Terminology database (UNTerm) or by the UN Statistics Division. The standard itself is maintained by an ISO-supported Maintenance Agency. Country codes are coordinated with Part 1 code assignments, so there may in some cases be discontinuities with 2 and 3. Of further interest is ISO 15924, the standard for script codes, which is maintained by the Unicode Consortium. Together these standards comprise a base for forming the regional locale identifiers according to the series of intertwined standards, common practices, and open-source data repositories cited in the outline in Section 2.

IETF RFC 1766 of 1995, also named *Tags for the Identification of Languages* (Alvastrand, http://www.ietf.org/rfc/rfc1766.txt), specified the use of a language tag consisting either of a two-letter language code or the five-character locale string (ex.: en-GB) suggested in 639-1. Already in 1766 the standard allowed for the expansion of the language code to include not only the country code, but also dialect or variant information, codes for languages not listed in 639-1, and script variations. At that point in time, anyone wishing to use a language code for any language not present in the then valid 639-1 could apply to the Internet Assigned Numbers Authority (IANA) to receive an Alpha-3 code that could be used for this purpose. When 639-3 (2007) was adopted, all extra IANA assigned three-letter codes rolled over to 639-3 codes. This initial IETF language tag standard specified the values used for the SGML *lang* attribute (Standard Generalized Markup Language, the parent standard that has spawned both HTML and XML).

In February 1998 the World Wide Web Consortium (W3C) published the first edition of the Extensible Markup Language (XML) standard, which specified that the *xml:lang* attribute shall be used to identify language information in XML documents, citing IETF 1766. The present Fifth Edition was updated in February 2013 to comply with the currently valid version of the IETF language tag standard, now referenced as BCP 47, which at the time of writing actually references IETF RFC 5646, which has replaced all previous versions of the document. BCP 47 is a stable designation for the committee responsible for maintaining the IETF's language tag standard, but each successive revision receives a new number, which can cause user confusion, although each document is clearly referenced to its predecessor and successor documents. As shown above, according to this version of the standard, the language tags can become quite complicated (and expressive) and can cover a wide range of applications and scenarios. It also provides rules for negotiating the Alpha-2/Alpha-3 anomaly in order to protect legacy data while at the same time facilitating the use of the newer three-letter codes.

The Unicode Common Locale Data Repository (CLDR) is based on Unicode's UTS # 35 Locale Data Markup Language (LDML), which is designed for the interchange of locale data. The CLDR provides a broad range of information designed to provide software companies with machine-parsable data associated with different languages and regions, such as formatting and parsing for dates, times, numbers, and currency values; scripts used with specific languages; time designations; pluralization rules, sorting and searching rules; writing direction and transliteration rules; country information, calendar preference, postal and telephone codes, and

much more. Selected sets of relevant information can be downloaded from the Unicode website in LDML format free of charge.

Linguistic considerations

As discussed above, the assignment of language codes to modes of spoken and written language have in the past been based on political and historical precedents, and to a certain extent on the basis of the wishes of individuals requesting a particular code. The codes that exist reflect the needs of specific groups of stakeholders working in a wide range of environments at different times and under different conditions. Constable and Simon openly explain that many different factors contribute to the definition of codes and identifiers (Constable and Simon, 2000; Constable, 2001a & b). Musgrave notes a distinction between "insider views of the relevant distinctions and outsider views" (Musgrave, 2014). This consideration applies particularly to languages of limited distribution that at least originally or even to the present day have not been the object of scientific or scholarly study by native speakers of those languages. The originators of field linguistics, that is, Jakob and Wilhelm Grimm (they of *Fairy Tales* fame) set out to document their own German language by transcribing field samples from a wide variety of speakers using many different dialects. They were at that time probably the most highly versed scholars capable of understanding these utterances and classifying them according to logical principles as they perceived them within the framework of their understanding of language development, etymology, and the history of the Indo-European languages (which they invented along the way). This kind of linguistic self-introspection holds true over time for the large developed languages, all of which have become the source of serious study by their own native speakers, by people capable of both understanding and classifying what they see and hear.

Beyond the boundaries of these highly developed systems, however, the smaller languages, isolated in jungles and mountains, proliferating across rugged landscapes like Papua New Guinea, or sandwiched into small enclaves in countries like Nigeria, both of which can lay claim to large numbers of languages compared to the real estate involved (in contrast to the vast territories claimed by hegemonic languages like English, Arabic, Chinese, Spanish, and Russian, for instance) became the subject of study by outsiders looking in. These people were intrepid explorers, missionaries, and field linguists. They painstakingly mastered new idioms without the benefit of foreign language teachers or dictionaries and grammars – they wrote their own as they went along, sometimes inventing new scripts and sometimes stretching existing ones to fit patterns for which they were ill suited. Often they might not have a good overview of how neighboring (or for that matter, far-flung) languages were related, or of how many variants actually might exist within a language family. They might in some cases assign a name to a language based on names they had learned from neighboring groups, ignorant of what the speakers of a language call it themselves.

In addition, political and historical forces have militated to hide languages in plain sight or to work to eradicate them. For instance, not only has the "Berber" slur been historically assigned to a class of languages in North Africa, under Gadhafi's regime, until recent times, the very existence of Tamazight languages and dialects was denied for political reasons. Indeed, there is a pattern across history for dictatorial rulers to deny the recognition of subjugated languages – witness Franco's Spain. Colonial languages such as English in the United States and Australia attempted to snuff out of the survival of undesirable native languages by removing children from their homes in order to prevent the native tongues from being passed on to the next generation. Even assertive Francophone Canadians do not always honor the language

rights of First Peoples. These kinds of historical events and practices are a part of the legacy data that underlies at least some of the language codes. Despite the fact that the current administrators of the codes are themselves conscientious, expert linguists, viewed from new perspectives, sometimes the old codes may not meet the needs of new research. Musgrave, for instance, defines seven different aboriginal Australian languages where the 639-3 only recognizes two, and they are not defined the same way or with the same names. The old question brought up early in this chapter concerning what distinguishes a dialect from a language also plays a role – if political and historical aspects are stripped away, leaving purely linguistic considerations, the categorization of language and dialect may look very different. This is the argumentation that underlies the incipient approach of the Research Data Alliance Working Group.

Regardless of whether future research provides very new and different insights, the current system, particularly the main three-part language codes together with the OLAC tags, represent a massive legacy investment that is bound to persist by sheer weight of its presence throughout the Web and existing data systems. The codes and other related standards provide a significant component to the mortar that holds together the building blocks of our information systems; even where they are linguistically inaccurate, they play a pragmatic role.

ISO TC 37 and ISO TC 46, which jointly administer the ISO 639 family of standards, have long contemplated some sort of merger of the standards, and yet the way going forward is not entirely clear. No viable successor entity (UNICODE? UNESCO?) has stepped forward with a sound business plan and an equitable strategy for balancing the needs and interests of the many divergent stakeholders and providing a secure framework for maintaining the codes and providing ready access to language-related information. Certainly, steps can be taken to create a more transparent process for adjudicating changes and additions to current systems. The evolution of a new linguistically oriented methodology may feed new codes into the existing system, or it may develop a parallel system for language variants that can be used in certain specialized environments. One thing is certain: the computing environment as we know it relies on a certain level of stability across the broad range of the various codes. They comprise a component of the foundation layer of the Internet model. Going forward any changes and additions need to maintain a clear balance between stakeholder needs and scientific research.

Definitions of key terms

language systematic use of sounds, characters, symbols or signs to express or communicate meaning or a message between humans

language variant, language variety identified language that differs from other languages
Note: Language varieties or variants can be classified as individual languages or dialects, as well as major regional clusters, such as American or Indian English.

dialect language variant specific to a geographical region or a group of language users
language variety (ISO 639-4)

language code combination of characters used to represent a language or languages

language identifier language symbol, which in the language codes of Parts 1, 2, 3, and 5 of ISO 639 is composed of two or three letters

language tag indicator of the language of text or other items in HTML and XML documents
Note: The *lang* attribute specifies language tags in HTML and the *xml:lang* attribute specifies them for XML. (See W3C Internationalization, *Language Tags in HTML and XML*, http://www.w3.org/International/articles/language-tags/)

locale cultural and linguistic setting applicable to the interpretation of a character string

macrolanguage language that for some purpose may be subdivided into two or more individual languages (ISO 639-4)

Notes

1 The Wikipedia entry http://en.wikipedia.org/wiki/A_language_is_a_dialect_with_an_army_and_navy points to the many varied versions of this statement and its uncertain provenance.

2 ISO standards are available as downloadable PDF files for purchase from the ISO Store, http://www.iso.org/iso/home/store.htm, or from national standards bodies. As noted below, however, specific language code, country code, and language tag information is available online at the official web addresses cited below. Users should be cautioned to seek out official websites because the many secondary postings found online may be outdated or otherwise incorrect or incomplete. Some published standards (e.g., ISO 639 Part 3 and Part 6) do not print the entire collection and only set down rules and guidelines.

References

Constable, Peter, and Simons, Gary (2000) "Language identification and IT: Addressing problems of linguistic diversity on a global scale." SIL International, http://www-01.sil.org/silewp/2000/001/SILEWP2000-001.pdf .

Constable, Peter (2002a) "An analysis of ISO 639: preparing the way for advancements in language identification standards". SIL International, http://www.sil.org/resources/publications/entry/7847.

——(2002b) "Toward a model for language identification: defining an ontology of language-related categories SIL Electronic Working Papers." SIL International http://www-01.sil.org/silewp/abstract.asp?ref=2002-003.

Göttert, Karl Heinz (2010) *Deutsch: Biografie einer Sprache*. Berlin: Ullstein Verlag.

Musgrave, Simon (2014) "Improving Access to Recorded Language Data." *D-Lib Magazine*, Vol. 20, No. 1/2. http://doi.org/10.1045/january2014-musgrave.

W3C *Internationalization Tag Set (ITS) Version 2.0;* W3C Recommendation 29 October 2013, http://www.w3.org/TR/its20/.

33
LOCALIZATION

Keiran J. Dunne

KENT STATE UNIVERSITY, THE UNITED STATES

Localization is an umbrella term that refers to the processes whereby digital content and products developed in one locale are adapted for sale and use in one or more other locales. Although the term 'localization' has been in use since the early 1980s, confusion persists as to what exactly it means. To understand localization, it is necessary to consider when, why and how it arose, the ways it has changed over time, and its relationship to translation and internationalization. Thus, this chapter will examine localization and its evolution from the 1980s to present.

The practice of translation remained relatively unchanged from the dawn of writing until the commoditization of the PC and the advent of mass market software ushered in the digital revolution in the 1980s. As increasing numbers of computers appeared in homes and business offices, 'typical users were no longer professional computer programmers, software engineers or hardware engineers' (Uren, Howard and Perinotti 1993: ix). U.S.-based software companies quickly realized that by developing products such as spreadsheet programs and word processors that average people could use for work or leisure, they could sell to a vastly larger potential market. Targeting average users instead of computer professionals was not without challenges, however.

> While experienced professionals had become adept in detecting bugs and working around them, the new users expected, indeed demanded, that the software they bought operate exactly as described in the manuals. Benign acceptance of anomalies in the operation of software could no longer be tolerated.
>
> (Uren, Howard and Perinotti 1993: x)

Initial efforts by software publishers to develop this embryonic mass market thus focused on improving software reliability and user-friendliness.

U.S.-based software companies soon broadened the scope of their marketing efforts beyond the domestic market to target international users. Expansion into international markets required that software publishers offer products in languages other than English. 'For a software product to have wide market acceptance in a non-English-speaking environment, it was essential to convert the software so that users saw a product in their own language and firmly based in their own culture' (Uren, Howard and Perinotti 1993: x). Software publishers thought that adapting their products for international markets was merely a matter of 'translating' software. As a result, initial attempts to adapt software for international users were characterized as 'translation on the computer for the computer' (van der Meer 1995). However, it soon became clear to practitioners that this work was 'related to, but different from and more involved than, translation' (Lieu 1997). Indeed, the scope of the undertaking was not confined to translation

of text in the user interface but rather encompassed all target market requirements for culturally dependent representation of data, including but not limited to the following:[1]

- character sets, scripts and glyphs for the representation of various writing systems
- encodings to enable the storage, retrieval and manipulation of multilingual data
- text comparison, searching and sorting (collation)
- line and word breaking
- calendars (e.g., Buddhist, Coptic, Gregorian, Hebrew lunar, Hijri, Japanese Emperor Year, Julian, Year of the Republic of China and Tangun Era calendars)
- date formats (MM/DD/YYYY, DD/MM/YYYY, YYYY-MM-DD, etc.; for example, 5/11/2014 would be read as May 11, 2014 in the United States but as November 5, 2014 in Italy)
- time formats (12-hour vs. 24-hour clock; use of AM and PM)
- number formats, digit groupings and decimal separators (period vs. comma)
- paper sizes (A3, A4, legal, letter)
- units of measurement (metric vs. imperial)

In software engineering, these local market requirements are referred to using the hypernym 'locale'. Locales are expressed as language-country pairs. Thus, 'French-Canada is one locale, while French-France is another' (Cadieux and Esselink 2002). The need to account not only for translation but also for 'locale' explains why and how the process of adapting software for international markets came to be known as 'localization' in the early 1980s. The scope and scale of this new activity expanded so rapidly that within less than a decade localization was perceived as an industry unto itself, as reflected by the creation of the Localization Industry Standards Association in 1990 (Lommel 2007: 7).

The costs of adapting software products for other locales seemed like a small price to pay given the sizable international markets and potential revenues to which localized products could enable access. Software publishers approached localization in different ways: some performed the work using in-house teams and some outsourced the work to specialized service providers, whereas others assigned responsibility for localization to in-country subsidiaries or distributors (Esselink 2003b: 4). Despite the ostensible differences between these approaches, they all shared one fundamental characteristic: in each case localization was performed apart from, and subsequent to, the development of the original, domestic-market products. 'This separation of development and localization proved troublesome in many respects,' observes Esselink (2003b: 4).

First, the software provided to localization teams often could not be localized because it lacked certain fundamental capabilities, such as the ability to display target-language scripts and writing systems. In such cases, the localization teams had to send the software back to the development teams for implementation of the necessary capabilities, such as support for the display of Asian languages or of right-to-left scripts for languages such as Arabic and Hebrew. Second, translatable text was typically embedded in the software source code. Identifying and locating translatable text was very difficult for localization teams that had not participated in the development of the software (see Figure 33.1). Finally, and perhaps most critically, localization required that changes be made directly to the source code of the software. To understand why and how working directly in source code caused problems, it is important to note that software source code is the raw material from which the running copy of a program is created. In other words, source code must be compiled, or built, into a machine-readable (binary) executable file, which in turn must be tested (and debugged, if any bugs are found) before the software can be released for use (see Figure 33.2).

```
IDD_PEN_WIDTHS DIALOG  0, 0, 203, 65
STYLE DS_SETFONT | DS_MODALFRAME | WS_POPUP | WS_VISIBLE | WS_CAPTION |
    WS_SYSMENU
CAPTION "Pen Widths"
FONT 8, "MS Sans Serif"
BEGIN
    DEFPUSHBUTTON   "OK",IDOK,148,7,50,14
    PUSHBUTTON      "Cancel",IDCANCEL,148,24,50,14
    PUSHBUTTON      "Default",IDC_DEFAULT_PEN_WIDTHS,148,41,50,14
    LTEXT           "Thin Pen Width:",IDC_STATIC,10,12,70,10
    LTEXT           "Thick Pen Width:",IDC_STATIC,10,33,70,10
    EDITTEXT        IDC_THIN_PEN_WIDTH,86,12,40,13,ES_AUTOHSCROLL
    EDITTEXT        IDC_THICK_PEN_WIDTH,86,33,40,13,ES_AUTOHSCROLL
END
```

Figure 33.1 Source-code representation of the dialog box shown in Figure 33.5(a)

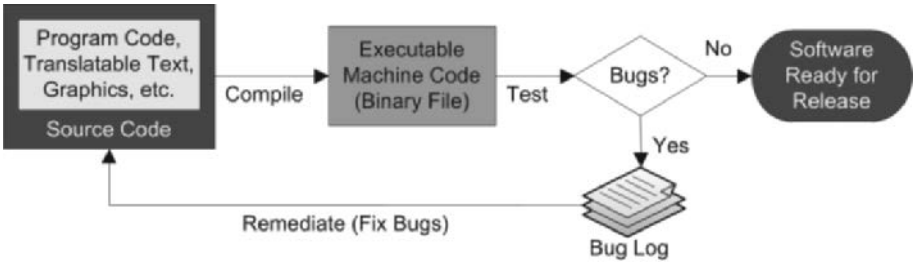


Figure 33.2 Because source code must be compiled and tested anew whenever it is modified, localizing directly in source code is labor-intensive

Identifying translatable text can be difficult for non-programmers. In the example shown in Figure 33.1, items in quotation marks are translatable except for the name of the default font (MS Sans Serif). Each group of four digits separated by commas represents layout coordinates.

Working directly in source code had profound ramifications for localization. Indeed, the adaptation of software products for other locales did not merely entail a few changes to compiled, tested, and debugged versions of programs that had already been released to the domestic market. Instead, localization of a given program required that a separate set of source code be maintained and that a different executable be compiled, tested and debugged for each target locale. Consequently, creating N localized versions of a program required that the publisher maintain N + 1 sets of source code: one for each target locale plus one for the domestic market. In addition, each set of source code had to be localized, compiled, tested, debugged, updated and managed separately (Luong, Lok, Taylor and Driscoll 1995: 3). For instance, a U.S.-based publisher that wanted to market a product in three international locales, such as German-Germany, French-France, and Japanese-Japan, was required to manage four different versions of source code in parallel, one set for the domestic English-United States locale plus one set for each of the three international locales (see Figure 33.3). The process of compiling and testing localized software products as distinct from source-locale versions is called localization engineering (Esselink 2002).

Creating, maintaining and supporting multiple localized versions in parallel proved to be time-consuming and expensive. Testing and debugging software was inherently labor-intensive and costly even without adding localization as a variable. Seminal work on software engineering economics by Boehm had demonstrated that 'uncorrected errors become exponentially more costly with each phase in which they are unresolved' (1981: 8). The exponential cost increase of error correction was exacerbated by the *post hoc* approach to localization in which the adaptation of software for other locales – and thus the discovery of localization bugs – did not begin until the development of the domestic-market versions had been completed. This cost

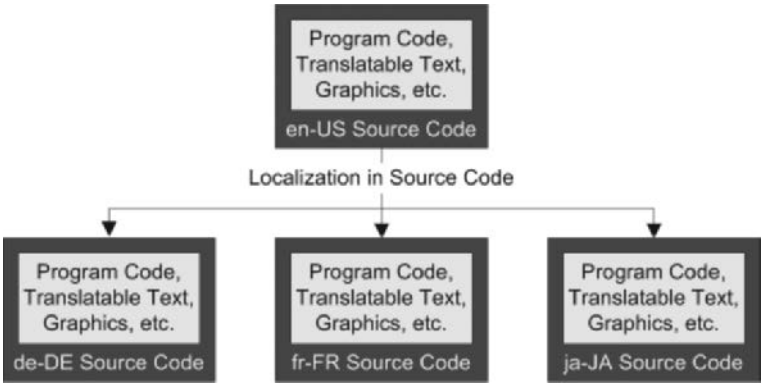


Figure 33.3 When localization is performed in the source code, it is necessary to maintain a separate set of code for each target locale plus one set for the domestic market

multiplier problem was compounded by the management of a distinct set of source codes for each target locale, since a bug discovered in one set of code might need to be fixed in all other sets. Indeed, localization engineering has traditionally involved 'quite a bit of bug fixing,' as Esselink observes (2002: 4). Not surprisingly, complexity soon established itself as a hallmark of localization (Esselink 2000b). Ultimately, most of the problems posed by early localization efforts stemmed from a single root cause: the failure to effectively plan for the reuse of software source code across multiple locales.[2]

Most software and hardware companies that made forays into international markets quickly concluded that localization and translation were not an integral part of their business. As Esselink (2000a: 5) observes, '[t]he increasing size and complexity of localization projects soon forced companies to an outsourcing model. Most software publishers simply did not have the time, knowledge or resources to manage multilingual translation or localization projects.' As a result, most companies decided that it would be more efficient to outsource the adaptation of software products for international markets to external language service providers as project work. In addition to the adaptation of the software application, a typical software localization project might also involve the translation and/or adaptation of various other components such as sample files, demos and tutorials, Help systems, printed and online user documentation, as well as marketing collateral (see Figure 33.4). The fact that these components were authored in a variety of digital formats, some of which needed to be compiled and tested prior to release, meant that localization involved a number of new forms of work in addition to traditional translation, including software and online Help engineering and testing, conversion of documentation to different formats, translation memory creation and management, as well as project management (Esselink 2000b; Esselink 2003a: 69; Dunne and Dunne 2011). Localization thus required that translators possess strong instrumental and technical skills in addition to traditional translation and domain expertise. 'Throughout the 1990s, the localization industry tried to turn translators into semi-engineers', recalls Esselink (2003b: 7).

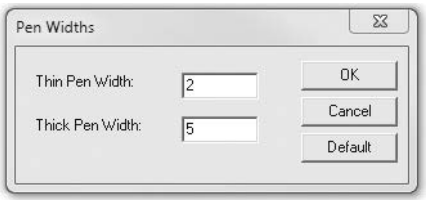
Outsourcing shifted the challenges of managing complex multilingual localization projects to external service providers but did not address the fundamental problem of the duplication of effort required to manage multiple sets of source code in parallel. Faced with the challenge of controlling the complexity and cost of the localization process, software publishers in the late 1980s and early 1990s began to realize that 'certain steps could be performed in advance to make localization easier: separating translatable text strings from the executable code, for

Figure 33.4 The scope of a traditional software localization project may encompass a number of components in addition to the software application itself

Source: Adapted from Esselink (2000a: 10)

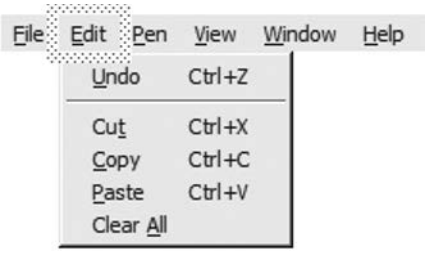
example. This was referred to as *internationalization* or *localization-enablement*' (Cadieux and Esselink 2002). Internationalization is an engineering process that precedes localization and entails the separation of '[a]ll the culturally and linguistically sensitive software components … from the core of the application' (Hall 1999: 298). In practice, the scope of internationalization is typically confined to the linguistic and culturally dependent contents of the user interface that may require adaptation, which are collectively designated using the hypernym 'resources.' When a piece of software is properly internationalized, '[t]here is no programming code in the [resources] nor is there any [translatable] text in the program code' (Uren, Howard and Perinotti 1993: 60). Resources in a typical desktop software application may include the following:

- *Accelerators*: keyboard shortcuts that enable direct execution of commands. Accelerators are typically associated with a Function key or with a combination of the Ctrl key plus a specific keyboard letter. For example, pressing the F1 Function key in a typical Windows application launches the Help, while pressing Ctrl+C executes the Copy command.
- *Dialog boxes*: secondary windows that allow the user to perform a command and/or that ask the user to supply additional information (see Figure 33.5(a)). Common examples include the 'Save As' and 'Print' dialog boxes. Dialog box resources also contain the coordinates that govern the layout and display of the user interface (see Figure 33.1).
- *Icons*: images that symbolize and provide clickable shortcuts to programs, files and devices (see Figure 33.5(b)).
- *Menus*: lists of options or commands that display at the top of the main program window (see Figure 33.5(c)). Secondary menus, called 'context' or 'popup' menus, display when the user clicks the right-hand mouse button.
- *String tables*: 'string' is short for 'string of characters,' and designates any text that is stored and manipulated as a group. Strings include button captions, dialog box titles, error messages, menu items, tool tips, and so on. Menu and dialog box strings can often be visually represented in a WYSIWYG (what you see is what you get) editor during localization, whereas string tables typically cannot (see Figure 33.5(d)).
- *Toolbars*: raster graphics that contain toolbar button images, typically in bitmap (*.bmp) or Portable Network Graphics (*.png) format (see Figure 33.5(e)).



(a)

(b)



(c)

Value	Caption
57632	Erase the selection\nErase
57633	Clears the drawing
57634	Copy the selection and put it on the Clipboard\nCopy
57635	Cut the selection and put it on the Clipboard\nCut
57636	Find the specified text\nFind
57637	Insert Clipboard contents\nPaste
57640	Repeat the last action\nRepeat

(d)

(e)

Figure 33.5 Typical resources in a software application include (a) one or more dialog boxes; (b) a program icon and a document icon (left and right images, respectively); (c) one or more menus; (d) a string table; and (e) one or more toolbars.[3] See also Figures 33.8 and 33.9

The creation of a standardized way to represent culturally dependent user interface material and store it independently from the functional program code greatly facilitated the localization process. No longer was it necessary to modify the source code or to compile, test, and debug each target version of a program separately (Luong *et al.* 1995: 3). Instead, the development team could simply extract the resources from the binary executable file, provide them to a localization team that would translate the text and perform all other necessary modifications and then return the target resources to the developers, who would integrate them into copies of the binary executable file to create the necessary target version or versions (see Figure 33.6). This extraction-adaptation-integration process is one of the defining characteristics of localization.

By enabling the use of a single set of source code to support multiple target locales, internationalization diminished the effort and cost associated with localization and increased the speed and accuracy with which it can be accomplished (Schmitz 2007: 51). It soon occurred to software developers that they could not only embed resources in a program's executable file and bind them directly to the application code, but also externalize the resources, store them in a dedicated file called a satellite assembly and dynamically link this external resource file to the application. To localize an application developed using this internationalization strategy, one simply creates a localized version of the resource file(s) for each target locale. For example, a publisher that created a program for the German-Germany locale and wanted to market localized versions in the United States and the People's Republic of China would create a localized version of the resource assembly for each locale (see Figure 33.7).

The advent of software internationalization coincided with, and was facilitated by, the broad shift from procedural and structured programming to object-oriented programming in the 1980s and the 1990s. Procedural and structured programming languages, which predominated

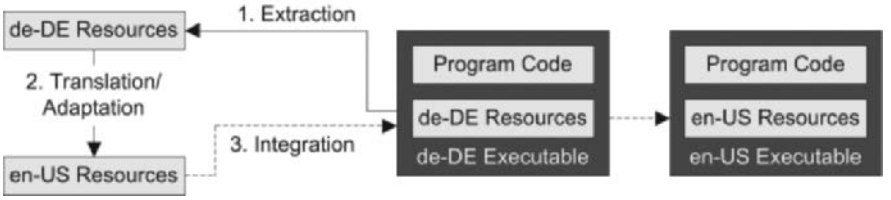


Figure 33.6 Internationalization enables the logical separation of the culturally dependent contents of the user interface from the functional core of the program, transforming localization into a simpler process of resource replacement

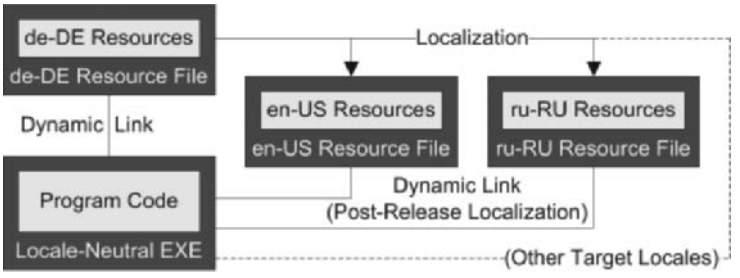


Figure 33.7 Externalizing resources, storing them in dedicated files and linking them dynamically to a locale-neutral program core is the logical culmination of software internationalization strategies

during the 1960s and 1970s, were adequate for small, relatively simple standalone applications. However, as applications expanded in terms of size, complexity, and the degree of interaction with other systems, procedural and structured languages began to show their limitations (Clark 2013: 2–5). The larger a program became the harder it was to maintain, and it was difficult to modify one aspect of existing functionality without negatively impacting the system as a whole. Programmers needed a comprehensive understanding of how a program worked and could not focus their efforts on discrete functions. In addition, the absence of standardized notational ways to represent and encode functions hindered the portability and re-usability of a software source code, with the result that programs were typically built from scratch.

Object-oriented programming effectively resolved these problems. In object-oriented programming, data and functions that use those data are grouped and encapsulated in logical structures named objects. One object's encapsulated data and functions can be used, or invoked, by other functions or programs. Communication between objects in a program is carried out through messages. An object's interface is defined by the messages that it can send and receive. In object-oriented programming, sending a message to an object is also called setting a property of that object. Objects are defined by classes, which determine their code, data and the messages they can send and receive (i.e., their properties). Individual objects inherit all of the properties and functions of the class to which they belong. Inheritance enables the creation of 'child' objects and subclasses that inherit all of the properties and functions of the original class of the 'parent' object. Inheritance facilitates software maintenance, updates and debugging because a change made to one instance of an object is applied to all instances of objects in that class. Objects can also be reused in and across programs, which simplifies the development of new programs.

Object-oriented programs are not written, but rather drawn in integrated WYSIWYG ('what you see is what you get') development environments using a variety of objects including menus, dialog boxes, forms, and user controls such as command buttons, check boxes, and text labels, among others. From the standpoint of object-oriented programming, internationalization standardizes the representation, definition and storage of the inventory of user interface objects as classes of resources. It follows that localization is properly understood as the modification of the properties of objects. For example, translating a command button caption entails the modification of the Caption property of the Button object (see Figure 33.8).

Internationalization not only eliminated the need to maintain a separate set of source code for each supported locale, but also clarified the respective roles of programmers, engineers and translators.

> [Internationalization] allows programmers and engineers to focus on code and translators to focus on translation. It means the software with all its complex logic does not have to be touched just because you want to add another language; all you have to do is translate some files.
>
> (Uren, Howard and Perinotti 1993: 63)

Indeed, most locale-dependent aspects of data storage, retrieval, manipulation and presentation can be managed today through internationalization capabilities built into host operating systems and/or by using development frameworks and runtime environments that offer robust support for internationalization, such as Java (Deitsch and Czarneki 2001) or Microsoft's .NET Framework (Smith-Ferrier 2007). If a program has been properly internationalized and all translatable text has been externalized from the source code, the localizer works only on resource files and cannot access or modify the program's functional code. Consequently, the

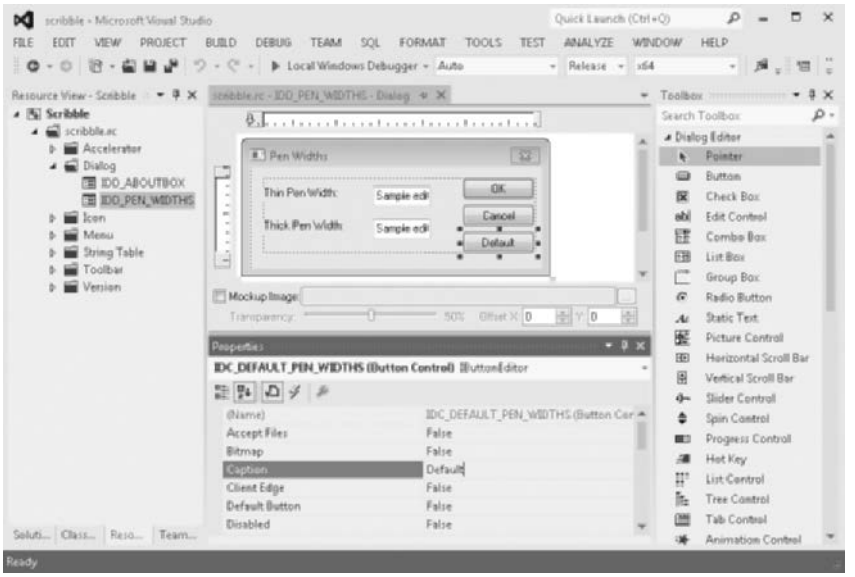


Figure 33.8 An object-oriented program interface is drawn using classes of user control objects (right-hand pane); composite interface objects are defined and stored as resources (left-hand pane). Localization of object-oriented software is properly understood as the modification of the properties of objects, such as the command button caption 'Default' (upper and lower middle panes). This image depicts the creation of the sample application Scribble (see Figures 33.1 and 33.5) in Microsoft® Visual Studio® 2012

Source: Used with permission from Microsoft

nuts-and-bolts work of software localization now primarily involves the translation of strings in menus, dialog boxes and string tables. Dialog boxes and user controls such as buttons and labels may also require resizing to account for translation-related string expansion or shrinkage, depending on the source and target languages involved. Visual localization tools facilitate these tasks by enabling localizers to view resources in context and by enabling them to use translation memory and terminology databases as they work (see Figure 33.9).

Internationalization and the use of satellite resource files allow software publishers to proactively address potential locale-related problems during the software development process, well in advance of localization. This state of affairs begs the question of how – and perhaps even if – localization differs from translation today. The fact that the translation of strings comprises the bulk of the work in current practice suggests that the term 'localization' has come full circle and once again essentially means 'translation on the computer, for the computer.' The blurring of the boundaries between translation and localization can also be seen as evidence of a convergence of these processes as authoring and publishing undergo an evolution similar to that of software localization (Esselink 2003b).

Authoring and publishing were generally separate processes and professions until the 1980s and 1990s, when the advent of digital authoring tools such as word processors turned authors into desktop publishers who were able not only to create digital content, but also to control the manner of its presentation (Rockley, Kostur and Manning 2003: 165). However, the desktop- and document-based approach to authoring and publishing hindered content reuse in much the same way as early localization efforts that required modifications to source code.

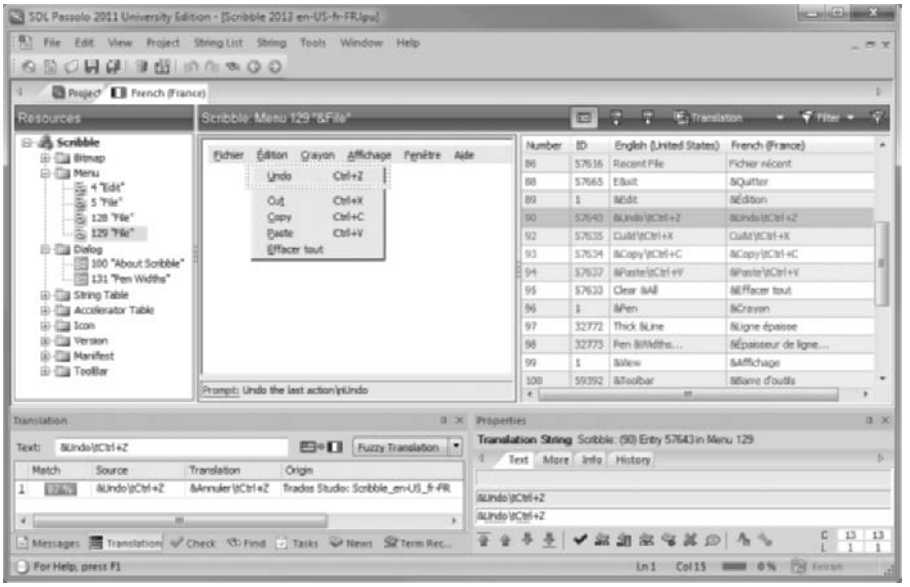


Figure 33.9 Localization of a sample application named Scribble using a visual localization tool. The left-hand pane displays the resource tree, the middle pane displays the selected resource in WYSIWYG mode, and the right-hand pane displays the corresponding source and target strings in tabular format

Repurposing content stored in documents created using word processors and other traditional desktop publishing software is generally a labor-intensive, manual process. A common rule of thumb is that technical communicators who use word processors and document-based authoring tools spend as much as half of their time formatting documents (e.g., Bartlett 1998). 'To reuse the content, authors must apply formatting that is appropriate for each output. Stripping and reapplying formatting is tricky and usually not 100% effective. Format conversions always require correction by hand or complicated scripting' (Rockley *et al.* 2003: 165). Further complicating content reuse efforts was the need to manage multiple versions of document source files in parallel. Practitioners soon discovered that file management and version control become exponentially more complex as the number of parallel versions of source files and the number of target languages increase. These problems were exacerbated by the Web and its widespread adoption as an enterprise communication platform. Desktop-based authoring and publishing could not keep up with the speed of change on the Web, nor could it meet demand for content in an increasingly wide range of formats for use on an expanding array of devices, from PCs and laptops to PDAs, tablets and smart phones.

The challenges of document-based content reuse were very similar to those associated with early software localization efforts. Thus it is perhaps unsurprising that the strategies adopted to facilitate content reuse are very similar to strategies developed to facilitate software localization. Just as internalization simplified localization by logically separating the culturally and linguistically dependent aspects of the user interface from the functional core of a program, content reuse strategies are predicated on the separation of content from presentation. This approach is called *single sourcing*: 'Single sourcing implies that there is a single source for content; content is written once, stored in a single source location, and reused many times' (Rockley *et al.* 2003: 15).

The implementation of single sourcing typically involves XML-based authoring strategies (Savourel 2001: 7; Rockley *et al.* 2003: 159-171). XML (eXtensible Markup Language) is a meta-markup language that provides a universal, application-independent mechanism for representing text in a structured format. XML was created in response to the challenges of content reuse associated with large-scale digital publishing. As stated in a December 1997 World Wide Web Consortium Press release announcing the publication of version 1.0 of XML as a proposed recommendation, 'XML is primarily intended to meet the requirements of large-scale Web content providers for industry-specific markup, vendor-neutral data exchange, media-independent publishing, one-on-one marketing, workflow management in collaborative authoring environments, and the processing of Web documents by intelligent clients' (W3C 1997). Whereas HTML is a presentational markup language that specifies how content should be displayed, XML is a semantic markup language that specifies what content means. Because XML provides a structured, semantic representation of content, and does not focus on the presentational aspects of document, authoring of content in XML is often referred to as *structured authoring*. Formatting of XML in documents is specified by style directives stored in separate files and applied dynamically in response to user demand. In this way, the same content can be processed and output in any format for which the organization has a defined set of style rules, such as webpages (HTML), PDF documents, Word documents, as well as Eclipse Help, HTML Help, Java Help and WebHelp, to cite but a few examples. Today, the single sourcing of technical and procedural documentation is often implemented using the XML-based DITA (Darwin Information Typing Architecture) standard (Bellamy, Cary and Schlotfeldt 2012).

The implementation of single sourcing also often involves the use of content management systems (Rockley *et al.* 2003: 178–191), which are centralized, server-based repositories 'designed to manage "information chunks" (generically known as "content"), usually no longer than a couple of paragraphs' (Biau Gil and Pym 2006: 11). Information chunks, such as DITA topics, are dynamically assembled into documents in response to user requests, typically via a web interface. As is the case with XML-based authoring, content stored in a content management system (CMS) can generally be output in various formats.

Single sourcing, structured authoring, and 'chunking' can be thought of as applications of the concepts of object orientation and internationalization in the fields of authoring and publishing. Once written, a given information object can be reused systematically; once translated, the target language versions of that same information object can also be reused systematically. Because content is separated from form, it can be processed using as many different style directives as needed to publish it in the desired output formats (e.g., print, help, web, and mobile devices) without having to modify the content. Just as object orientation and internationalization facilitated the modularization and reuse of software, single sourcing, structured authoring and chunking facilitate the modularization and reuse of content.

Translation of XML content and of information chunks is not 'localization' as the process has been traditionally understood, because it does not entail modification of the properties of objects in a software user interface. Nevertheless, 'content translation projects are now often considered as localization projects simply because of the complex environments in which the content is authored, managed, stored and published,' as Esselink has pointed out (2003b: 7). Complexity was once a defining characteristic of software localization projects, but now characterizes large-scale translation projects as well.

At a more fundamental level, the complexity of software localization and content translation is due largely to the fact that translators and localizers do not work on *linear* text but rather on decontexualized text strings or chunks. Working on text without context not only complicates

the translation decision-making process, but arguably calls into question the very possibility of understanding the text as a whole and the pragmatic act of communication of which it is an ostensible artifact. 'In understanding text, a reader must not only be able to integrate information within sentences but also make connections across sentences to form a coherent discourse representation,' as Rayner and Sereno observe (1994: 73). However, it is not always possible for translators to make connections across sentences while working on software strings. 'Due to their non-linear structure and lack of narrative thread, software programmes cannot be "read" in the same way as [traditional documents]' (Dunne 2009: 197). This also holds true for XML content and information chunks. In single sourcing projects, the 'document' does not exist until it is created dynamically in response to a user request (typically from an end-user). On a surface level, the translation of strings and information chunks may seem technologically simpler than traditional localization because translators do not have to compile or test target files. However, the translation of strings and information chunks is cognitively more complex because reading and comprehending text without context and 'texts without ends' (Biau Gil and Pym 2006: 11) requires translators to construct a situation model of a text that does not yet exist. In other words, the industry is no longer trying to turn translators into semi-engineers, but translators still need to understand the architecture of the components from which software localization project deliverables are created, such as software resource files, and Help topics, tables of contents and indexes. As Esselink observes, 'it looks likely that while translators will be able and expected to increasingly focus on their linguistic tasks … the bar of technical complexity will be raised considerably as well' (2003b: 7).

Notes

1 For more information on target-market requirements, see Giammarresi 2011, especially 39–40.
2 For a case study that illustrates some of the problems that can occur in the absence of an organized approach to internationalization, see Margulies 2000.
3 These resources are derived from a sample application called Scribble developed by the author using Visual Studio 2010 C++ sample files. MSDN Archive, Visual C++ MFC Samples for Visual Studio 2010, http://archive.msdn.microsoft.com/vcsamplesmfc (accessed Sep. 8, 2012).

References

Bartlett, P.G. (1998) 'The Benefits of Structured XML Authoring for Content Management', in Graphic Communications Association (U.S.), Organization for the Advancement of Structured Information Systems (OASIS) (ed.) *XML 98 Conference Proceedings*, 15–18 November 1998, Chicago, IL/New York: Graphic Communications Association. Available at: http://www.infoloom.com/gcaconfs/WEB/chicago98/bartlett.HTM.

Bellamy, Laura, Michelle Carey, and Jenifer Schlotfeldt (2012) *DITA Best Practices: A Roadmap for Writing, Editing, and Architecting in DITA*, Upper Saddle River, NJ: IBM Press.

Biau Gil, Jose Ramon, and Anthony Pym (2006) 'Technology and Translation (A Pedagogical Overview)', in Anthony Pym, Alexander Perekrestenko, and Bram Starink (eds) *Translation Technology and Its Teaching*, Tarragona, Spain: Intercultural Studies Group, Universitat Rovira I Virgili, 5–19.

Boehm, Barry W. (1981) *Software Engineering Economics*, Englewood Cliffs, NJ: Prentice-Hall.

Cadieux, Pierre and Bert Esselink (2002) 'GILT: Globalization, Internationalization, Localization, Translation,' *Globalization Insider* 11. Available at: http://bit.ly/1uE03QT.

Clark, Dan (2013) Beginning C# Object-Oriented Programming, N.p.: Apress.

Deitsch, Andy and David Czarnecki (2001) *Java Internationalization*, Sebastopol, CA: O'Reilly.

Dunne, Keiran J. (2006) 'Putting the Cart behind the Horse: Rethinking Localization Quality Management', in Keiran J. Dunne (ed.) *Perspectives on Localization*, Amsterdam and Philadelphia: John Benjamins, 1–11.

Dunne, Keiran J. (2009) 'Assessing Software Localization: For a Valid Approach', in Claudia V. Angelelli and Holly E. Jacobson (eds) *Testing and Assessment in Translation and Interpreting Studies*, Amsterdam and Philadelphia: John Benjamins, 185–222.

Dunne, Keiran J. and Elena S. Dunne (2011) *Translation and Localization Project Management: The Art of the Possible*, Amsterdam and Philadelphia: John Benjamins.

Esselink, Bert (2000a) *A Practical Guide to Localization*, revised edition, Amsterdam and Philadelphia: John Benjamins.

Esselink, Bert (2000b) 'Translation versus Localization', *Tranfree* 10 (15 January 2000). Available at: http://www.translatortips.net/tranfreearchive/tf10-localization-one.html.

Esselink, Bert (2002) 'Localization Engineering: The Dream Job?' *Revista Tradumàtica* 1 (October): 2–5. Available at: http://www.fti.uab.es/tradumatica/revista/articles/besselink/besselink.PDF.

Esselink, Bert (2003a) 'Localisation and Translation', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 67–86.

Esselink, Bert (2003b) 'The Evolution of Localization', *The Guide to Localization*, Supplement to *MultiLingual Computing and Technology* 14(5): 4–7. Available at: http://www.multilingual.com/downloads/screenSupp57.pdf.

Giammarresi, Salvatore (2011) 'Strategic Views on Localization Project Management: The Importance of Global Product Management and Portfolio Management', in Keiran J. Dunne and Elena S. Dunne (eds) *Translation and Localization Project Management: The Art of the Possible*, Amsterdam and Philadelphia: John Benjamins, 17–49.

Hall, Patrick A.V. (1999) 'Software Internationalization Architectures', in Gregory E. Kersten, Zbigniew Mikolajuk, and Anthony Gar-On Yeh (eds) *Decision Support Systems for Sustainable Development in Developing Countries: A Resource Book of Methods and Applications*, Boston, MA: Kluwer Academic Publishers, 291–304.

Lieu, Tina (1997) 'Software Localization: The Art of Turning Japanese', *Computing Japan* 4 (12). Available at: http://www.japaninc.com/cpj/magazine/issues/1997/dec97/local.html.

Lommel, Arle (2007) *The Globalization Industry Primer*, Romainmôtier, Switzerland: Localization Industry Standards Association.

Luong, Tuoc V., James S.H. Lok, David J. Taylor, and Kevin Driscoll (1995) *Internationalization: Developing Software for Global Markets*, New York: John Wiley & Sons.

Margulies, Benson I. (2000) 'Your Passport to Proper Internationalization', *Dr Dobb's* (May 1). Available at: http://drdobbs.com/your-passport-to-proper-internationaliza184414603.

Rayner, Keith and Sara C. Sereno (1994) 'Eye Movements in Reading: Psycholinguistic Studies', in Morton A. Gernsbacher (ed.) *Handbook of Psycholinguistics*, San Diego, CA: Academic Press, 57–81.

Rockley, Ann, Pamela Kostur, and Steve Manning (2003) *Managing Enterprise Content: A Unified Content Strategy*, Indianapolis, IN: New Riders.

Savourel, Yves (2001) *XML Internationalization and Localization*, Indianapolis, IN: Sams Publishing.

Schmitz, Klaus-Dirk (2007) 'Indeterminacy of Terms and Icons in Software Localization', in Bassey Edem Antia (ed.) *Indeterminacy in Terminology and LSP*, Amsterdam and Philadelphia: John Benjamins, 49–58.

Smith-Ferrier, Guy (2007) *.NET Internationalization: The Developer's Guide to Building Global Windows and Web Applications*, Upper Saddle River, NJ: Addison-Wesley.

Uren, Emmanuel, Robert Howard, and Tiziana Perinotti (1993) *Software Internationalization and Localization: An Introduction*, New York: Van Nostrand Reinhold.

van der Meer, Jaap (1995) 'The Fate of the Localization Industry and a Call to Action', *The LISA [Localization Industry Standards Association] Forum Newsletter* 4(4): 14–17.

W3C (World Wide Web Consortium) (1997) 'W3C Issues XML1.0 as a Proposed Recommendation', *World Wide Web Consortium Press Release* (8 December 1997). Available at: http://www.w3.org/Press/XML-PR.

34

NATURAL LANGUAGE PROCESSING

Olivia Kwong Oi Yee

CITY UNIVERSITY OF HONG KONG, HONG KONG, CHINA

Introduction

Natural language processing (NLP) concerns the handling and understanding of human languages by computers with two major goals. One is to enable human–computer interaction with human languages as the medium. The other is to build language application systems which require considerable human language abilities and linguistic knowledge. The focus of NLP is thus often on practical tools, and the design and implementation of computational systems allowing mostly textual natural language input and output. It is therefore closely related to but somehow distinguished from computational linguistics and speech recognition and synthesis, while they are often considered together within the larger umbrella of speech and language processing (e.g. Jurafsky and Martin 2009).

Among the many applications, machine translation (MT) is apparently the most typical and well-known example. In fact, MT has a critical role in NLP research from day one, as projects in the 1950s have set out with the ambition to build systems that were capable of automatically translating texts from one language into another. The efforts have unfortunately turned sour, as remarked by Charniak and McDermott (1985) as the 'sad story of machine translation', when the famous ALPAC report[1] issued in 1966 pronounced the failure of MT research thus far and recommended further funding should support more basic research in computational linguistics and down-to-earth studies toward the improvement of translation. Despite the unfavourable outcome, the lesson learned is important in at least two regards. On the one hand, translations produced by systems which relied entirely on a bilingual dictionary and simple syntactic methods could barely meet the most basic of professional requirements, even when generously considered. On the other hand, processing language by computers is much more complicated than once imagined and the knowledge required is diverse and enormous. Without satisfactorily addressing the smaller and intermediate sub-problems, it is hard to make any substantial achievement on the more sophisticated and demanding tasks like translation.

It turns out that research on MT came under the spotlight again after about two decades, during which research on a variety of other NLP problems has borne some important progress and insight. Especially with the development of the first statistical machine translation (SMT) system (Brown *et al*. 1990: 79–85), it marked the significance of statistical approaches in NLP, and since then MT as well as many other NLP areas have embarked on a fast track of

development. Three related factors have played a critical role in this course of evolution in pushing language technology forward: (1) the availability of large electronic corpora, (2) the rise of statistical and machine learning approaches, and (3) the fast-growing web technology.

Importance of corpus data

One of the foremost and notorious problems in NLP is often known as the knowledge acquisition bottleneck. Language understanding typically needs diverse and large amount of knowledge, linguistic or otherwise. This thus gives rise to three questions: What knowledge is necessary? How should knowledge be represented for computational purposes? Where can we adequately obtain such knowledge? Knowledge crafting and representation in early systems has been in the artificial intelligence (AI) fashion. To balance between the details required and the time and labour incurred, it is often limited in scale and domain. For instance, the classic SHRDLU system was designed to communicate with the user and perform accordingly, but only within the pre-defined 'blocks world' (Winograd 1973: 152–186). Knowledge represented in the form of plans and scripts is often situated in specific domains and scenarios (e.g. Schank and Abelson 1977). Realizing the limitation, in the 1980s researchers went for automatic or semi-automatic means for extracting knowledge, particularly lexico-semantic knowledge, from existing language resources such as machine-readable dictionaries (e.g. Boguraev and Briscoe 1989). With the availability of large electronic text corpora, such as the Brown Corpus and the British National Corpus in the 1960s and 1990s respectively, and structurally annotated corpora like the Penn Treebank later, and the even larger gigaword corpora today, lexical information can be more conveniently gathered on a large scale, capitalizing on the occurrence patterns of words exhibited in corpora. This has given rise to an area of research on automatic lexical acquisition, aiming to acquire a variety of lexical information including domain-specific monolingual and bilingual lexicons, significant collocations, subcategorization information, semantic similarities, selectional preferences, and others (e.g. Church and Hanks 1990: 22–29; Resnik 1993). In addition to knowledge acquisition, large corpora are often directly used for training statistical NLP systems, as a source from which probabilities for particular linguistic phenomena are estimated with respect to the statistical language models underlying the systems.

As far as translation is concerned, the necessary linguistic knowledge often comes in the form of bilingual lexicons, manually constructed or automatically acquired, and SMT systems are essentially trained on bilingual or multi-lingual corpora. Parallel corpora, referring to the same textual content existing simultaneously in two languages, are particularly valuable. The Canadian Hansard which consists of records of the proceedings of the Canadian parliament in both English and French is a typical example (e.g. Brown *et al.* 1990: 79–85). The Hong Kong Hansard consisting of records of the proceedings of the Legislative Council of Hong Kong, as well as the bilingual laws and the bilingual court judgments in Hong Kong are exemplary English–Chinese parallel corpora (e.g. Wu 1994: 80-87; Kit *et al.* 2005: 71–78; Kwong *et al.* 2004: 81–99). While parallel corpora were relatively scarce, comparable corpora which consist of textual data from two languages on similar contents or topics but are nevertheless not a direct translation of each other could be the next resort (e.g. Fung 1998: 1–17). More recently, sources of parallel corpora are no longer restricted to government or official documents, but span much wider origins. For instance, the NTCIR[2] workshops have a track on patent MT, providing English–Chinese and English–Japanese patent documents for training and testing MT systems (Goto *et al.* 2011: 559-578). Strassel *et al.* (2011) have compiled large English–Chinese and English–Arabic corpora with text and speech data for the development of MT

systems, and the data cover newswire and broadcast sources. The internet has also become an important source for mining parallel texts, especially for less common language pairs (e.g. Resnik and Smith 2003: 349–380).

Dominance of statistical and machine learning algorithms

Another notorious problem for NLP is the handling of ambiguity which exists at different levels of linguistic analysis. For tasks like translation which definitely require a thorough understanding of a text, the ambiguity problem is particularly relevant. Lexical ambiguities such as part-of-speech ambiguity and word sense ambiguity, as well as higher level syntactic and semantic ambiguities, which exist for most NLP tasks in general, will also need to be handled in MT. For cases where there are lexical gaps in one language, such as when the target language has lexical distinctions that are absent in the source language, disambiguation becomes even more critical to arrive at a correct lexical choice in the target language during translation. NLP tools and applications must be able to robustly process different input texts and resolve various kinds of ambiguities therein.

As demonstrated in the second edition of the *Handbook of Natural Language Processing*, approaches in NLP are often categorized into the more classical symbolic approaches and the more contemporary empirical and statistical approaches (Indurkhya and Damerau 2010). Symbolic approaches mostly follow the AI tradition, with manually crafted procedural knowledge. They were later developed into more general knowledge-based or rule-based approaches, for which knowledge may be in the form of rules, or more declarative as in semantic lexicons like WordNet (Miller 1995: 39-41), ontologies like SUMO (Pease *et al.* 2002), etc., which may be handcrafted or (semi-)automatically acquired from existing linguistic resources. Empirical and statistical approaches, also known as stochastic approaches, are data-driven. They often require large annotated corpora as training data for estimating the probabilities of various linguistic units or phenomena with respect to particular statistical language models. According to Charniak (1993), a statistical approach in its purest form is to note statistical regularities in a corpus. Statistical methods thus evaluate different possible outcomes probabilistically and output the one with highest value as the result. Statistical language models, however, may assume different levels of complexity with different dependency assumptions. An important group of algorithms is based on machine learning, which attempts to learn patterns for classification from a set of linguistic features extracted from texts. Learning can be supervised, with annotated data, or unsupervised, with no or just minimal annotated data to start with.

As large text corpora are becoming more accessible, machine learning has become the dominant approach in many NLP tasks. In many cases, hybrid methods are adopted, with machine learning algorithms at the core, supplemented by post-processing the results with rules. Statistical methods have an advantage with its scalability. Its general coverage regardless of the frequency or rarity of individual linguistic phenomena overcomes the severe limitation of rule-based systems, as the efforts involved in crafting the rules often confine the resulting systems to toy systems. Statistical methods remove this hurdle, although they do not necessarily model human cognitive processes. Allen (1995) gives detailed descriptions of symbolic approaches to NLP, while a comprehensive account of statistical NLP can be found in Manning and Schütze (1999).

Web technology as a catalyst

The development of web technology has to a certain extent catalysed the development of statistical NLP. On the one hand, the use of XML as a standard protocol for data markup and document encoding has made it easier to share data over the internet and improved interoperability of language resources. On the other hand, the global popularity of the internet has made the sharing of resources much more convenient and allowed crowdsourcing for data preparation or even annotation (e.g. Chklovski and Mihalcea 2002: 116–123). Web 2.0 has led to a surge of user-generated content over the World Wide Web, and web-crawling techniques have enabled quick collection of mega-size textual data. With such facilities just at our fingertips (e.g. Baroni and Bernardini 2004: 1313–1316), mining the web for large corpora has thus formed a trend, beating traditional corpus compilation in terms of time, quantity and variety, although materials gathered from the web have to be considerably cleaned and used with caution (Kilgarriff and Grefenstette 2003: 334–347). In addition, cloud computing has enabled storage of virtually unlimited data and running of applications without confinement to specific physical locations, and the evolution of web and mobile applications has brought about more different modes of deployment of language technology, accessible to a great many users.

This section has thus given a bird's eye view of natural language processing. In the next section, we look into selected NLP tasks and applications bearing on translation technology in the broadest sense.

NLP tasks and translation technology

Common text pre-processing tasks

In most natural language processing applications, one of the first steps is to tokenize the input text, that is, to locate word boundaries in the input text and break it into basic token units. This tokenization process is relatively less demanding for Indo-European languages like English and French, as words in the texts are already delimited by spaces. Despite this, the process is not always straightforward as certain punctuations and symbols could have multiple functions depending on the context of their individual occurrences (Mikheev 2003: 201–218). For example, a period amidst digits should be considered part of the decimal number instead of a token delimiter. In other cases multi-word expressions might be more properly considered a single unit. For instance, New York may more intuitively be treated as one unit instead of two. The definition and boundary of tokens are further blurred with most web texts where emoticons and informal spellings are abundant.

Tokenization is much more important for processing many Asian languages where word boundaries are implicit. In the case of Chinese, the notion of word has always been under debate and there is no standard definition for what constitutes a Chinese 'word'. Various criteria are usually considered, including syntax (e.g. bound/free morphemes and word structures), semantics (e.g. fixed phrases, idioms, and emergent meanings), frequency, length, etc. In practice, disyllabic words apparently dominate. The Maximum Matching Algorithm is a simple knowledge-based method for Chinese word segmentation, which requires only a dictionary containing all legitimate words in the language to start with. During the process, the input text is compared against the dictionary to find the longest matching word and the word boundary is thus marked. Sproat *et al.* (1996: 377–404) used statistical methods together with finite state techniques for the purpose. Recent approaches include hybrid methods and character-based segmentation with machine learning (e.g. Xue and Converse 2002: 63–69).

The SIGHAN[3] International Chinese Segmentation Bakeoff, organized since 2003, has provided a common platform for evaluating system performance, measured in terms of In-Vocabulary and Out-Of-Vocabulary segmentation scores (Sproat and Emerson 2003: 133–143). A comprehensive review of the development of Chinese word segmentation research can be found in Huang and Zhao (2007: 8–19).

Equally important is part-of-speech tagging, which assigns the most appropriate lexical category label to individual words in a text. Traditional rule-based tagging relies on a dictionary containing all possible part-of-speech tags for individual words and a large set of manually devised disambiguation rules for eliminating incompatible tags. Transformation-based tagging, which is also known as Brill tagging, is data-driven in the sense that the input text is first tagged with the most frequent tag for each word, followed by applying a large set of transformation rules in a particular order to modify the tags under particular contexts. These transformation rules were learned from a large set of training data (Brill 1995: 543–565). Hidden Markov Model taggers are by far the most common stochastic approach (e.g. Cutting *et al.* 1992: 133–140; Chang and Chen 1993: 40–47).

Sentence structures are often worked out as a next step for subsequent processing. The process of analysing a sentence in terms of its syntactic structure is known as parsing, which can be done in a top-down or bottom-up manner. Based on the grammar formalisms adopted, common frameworks include phrase-structure parsing and dependency parsing. The former renders a sentence into a tree of phrasal constituent structures whereas the latter delineates the dependency relations among constituents in a sentence. Probabilistic parsing trains a parser with the structural annotations in a Treebank for the probabilities of particular constituent structures and provides a means to resolve structural ambiguities (e.g. Charniak 1997: 598–603; Kübler *et al.* 2009). SIGPARSE organizes biennial conferences on parsing technologies and CoNLL[4] has held shared tasks on multi-lingual parsing and dependency parsing from 2006 to 2009.

Word sense disambiguation

Word sense disambiguation (WSD) refers to the process of identifying word meanings in a discourse for words which have multiple form-meaning possibilities. The importance of disambiguating word senses has already been realized by the time MT emerged as ambitious projects in the 1950s. On the one hand, a thorough understanding of the source text is needed to resolve the word sense ambiguities therein, and vice versa. On the other hand, word sense ambiguities often surface as translation differences. For example, 'duty' in English should be translated to 'devoir' or 'droit' in French depending on whether the word is used in its 'obligation' or 'tax' sense respectively (e.g. Gale *et al.* 1992: 233–237). Automatic WSD largely depends on the knowledge sources capturing the various semantic (and possibly other) relations among words available to a system, and subsequently its ability to uncover and deploy these relations among words in a text.

Current mainstream practice often treats WSD as a classification task. Systems thus attempt to assign the most appropriate sense among those given in a particular sense inventory, typically some dictionary or lexical resource, to individual words in a text. As for many NLP tasks in general, WSD methods are conventionally classified into AI-based, knowledge-based, and corpus-based methods, corresponding closely with the predominant kind of lexical resources during various historical periods. The survey by Ide and Veronis (1998: 1–40) documents the development of WSD before the end of the last millennium. Knowledge-based approaches, supervised approaches and unsupervised approaches are discussed in details in Mihalcea (2006),

Màrquez *et al.* (2006) and Pedersen (2006) respectively. Navigli (2009: 1–69) gives detailed technical descriptions of various algorithms.

The performance evaluation of WSD systems has been more or less standardized in the last decade with the SENSEVAL and SEMEVAL exercises (e.g. Edmonds and Cotton 2001: 1–6). It turns out that state-of-the-art systems are mostly based on combinations of multiple classifiers, and voting schemes combining several learning algorithms outperform individual classifiers (Mihalcea *et al.* 2004: 25–28). Notwithstanding the many encouraging results, there is still room for research on WSD, not only as a task in itself, but also more importantly regarding its contribution to real language processing applications like machine translation and information retrieval. Despite the all-time conviction that some sort of WSD is needed for nearly every NLP application, with a few exceptions like Chan *et al.* (2007: 33–40) and Specia *et al.* (2006: 33–40), most have denied the contribution of imperfect WSD components to real NLP applications (e.g. Krovetz and Croft 1992: 115–141; Sanderson 1994: 142–151). Errors in WSD often adversely affect the application, and many have attributed this to the inappropriateness of the sense inventories and sense granularity used in basic WSD experiments for real NLP applications (e.g. Ide and Wilks 2006; McCarthy 2006: 17–24; Resnik 2006). Kwong (2012) suggested going beyond external factors like resources and algorithms and considering some intrinsic properties of words such as part-of-speech and concreteness for lexically sensitive disambiguation.

Automatic transliteration

Transliteration takes a name in a source language and renders it in a target language, in a phonemically similar way. Proper names including personal names, place names, and organization names, make up a considerable part of naturally occurring texts, and even the most comprehensive bilingual lexicon cannot capture all possible proper names. The accurate rendition of personal names thus means a lot to machine translation accuracy and intelligibility, and cross-lingual information retrieval, especially between dissimilar languages such as English and Chinese, English and Japanese, and English and Hindi. There are basically two categories of work on machine transliteration: acquiring transliteration lexicons from parallel corpora and other resources (e.g. Lee *et al.* 2006: 67–90; Jin *et al.* 2008: 9–15; Kuo *et* al. 2008: 16–23) and generating transliteration for personal names and other proper names.

Traditional systems for transliteration generation often consider phonemes as the basic unit of transliteration (e.g. Knight and Graehl 1998: 599–612; Virga and Khudanpur 2003: 57–64). Li *et al.* (2004: 159–166) suggested a grapheme-based Joint Source-Channel Model within the Direct Orthographic Mapping framework, skipping the middle phonemic representation in conventional phoneme-based methods, and modelling the segmentation and alignment preferences by means of contextual n-grams of the transliteration units. Their method was shown to outperform phoneme-based methods. In fact, transliteration of foreign names into Chinese is often based on the surface orthographic forms, as exemplified in the transliteration of Beckham, where the supposedly silent *h* in "*h*am" is taken as pronounced. Models based on characters (e.g. Shishtla *et al.* 2009: 40–43), syllables (e.g. Wutiwiwatchai and Thangthai 2010: 66–70), as well as hybrid units (e.g. Oh and Choi 2005: 451–461), are also seen. In addition to phonetic features, others like temporal, semantic, and tonal features have also been found useful in transliteration (e.g. Tao *et al.* 2006: 250–257; Li *et al.* 2007: 120–127; Kwong 2009: 21–24).

The shared task on transliteration generation organized by the Named Entities Workshop (NEWS) series suggested that an appropriate transliteration should meet three criteria, namely

phonemic equivalence between the source name and the target name, conformity of target name to the phonology of the target language, and user intuition considering cultural and orthographic conventions in the target language (Zhang *et al.* 2011: 1–13). The report of the shared task in NEWS 2010 (Li *et al.* 2010: 1–11) highlighted two particularly popular approaches for transliteration generation among the participating systems. One is phrase-based statistical machine transliteration (e.g. Song *et al.* 2010: 62–65; Finch and Sumita 2010: 48–52) and the other is Conditional Random Fields, which treats the task as one of sequence labelling (e.g. Das *et al.* 2010: 71–75). Besides these popular methods, for instance, Huang *et al.* (2011: 534–539) used a non-parametric Bayesian learning approach, and Kwong (2011: 11–19) proposed a simple syllable-based method for direct transliteration by chunks.

The task of transliteration can be quite mechanical on the one hand, but can also be highly variable on the other. In the case of English-to-Chinese transliteration, for instance, homophones are abundant in Chinese and the choice and combination of characters for the Chinese rendition is relatively free. Apparently transliteration follows a more standard practice in mainland China but exhibits more variations in the Hong Kong context. Besides linguistic and phonetic properties, many other social and cognitive factors such as dialect, gender, domain, meaning, and perception, also play a role in the naming process. Evaluating systems based on a given set of 'correct' transliterations may therefore not be entirely satisfactory, as there might be options outside this set which are also acceptable. Nevertheless, effective systems developed under such a paradigm should be helpful to organizations like news agencies and mass media, which are likely to encounter many new foreign names every day.

Text alignment

Automatic text alignment refers to taking two parallel input texts, or bitexts, and outputting their segmentation at different granularities such as sentences or words with the corresponding segments between the two texts identified. It is tightly linked to machine translation and translation lexicons.

The relation between translation lexicons and parallel text alignment is especially close as the extraction of translation lexicons from parallel corpora depends, to a certain extent, on parallel text alignment at the word level. In addition, together they provide foundational resources for machine translation research. The relation between parallel text alignment and machine translation, especially SMT, is even more intimate as the alignment model often forms part of an SMT model. Bilingual word alignment and thus extraction of translation lexicons were usually carried out statistically or via lexical criteria. The former relies on large corpora to be effective, and the latter depends on existing bilingual dictionaries which often only cover general terms.

Bilingual sentence alignment on Indo-European language pairs has conventionally been based statistically on sentence length (e.g. Gale and Church 1991: 177–184), or lexically on cognates (e.g. Simard *et al.* 1992: 67–81) and correspondence of word position (e.g. Kay and Röscheisen 1993: 121–142; Piperidis *et al.* 1997: 57–62). While the length criterion was found to work surprisingly well between English and Chinese, Wu (1994: 80–87) supplemented it with lexical criteria by identifying fixed words or phrases with consistent translations first. Word alignment is often done statistically, leveraging the translation association or token co-occurrences between the source language and the target language (e.g. Wu and Xia 1995: 285–313; Melamed 1997: 490–497). In practice, sentence alignment is not always distinctly separated from word alignment. In fact, apart from Gale and Church's length-based method, most others also simultaneously tackle word alignment to some extent. More technical details

and comparisons of different alignment models can be found in Och and Ney (2003: 19–51) and Wu (2010).

Translation lexicon extraction

Parallel corpora, aligned or otherwise, are important resources for extracting translation lexicons. One may start by acquiring monolingual collocations before extracting their translation equivalents based on certain statistical association criteria (e.g. Wu and Xia 1995: 285–313; Smadja *et al.* 1996: 1–38; Gaussier 1998: 444–450). When the corpus is too small for statistical methods and contains many general words, existing bilingual dictionaries may prove useful for word alignment and lexicon extraction (e.g. Ker and Chang 1997: 63–96). However, the coverage is often limited with existing lexical resources, even when several are used in combination (e.g. Huang and Choi 2000: 392–399). Using a third, pivot language as a bridge in word alignment could be an alternative (e.g. Borin 2000: 97–103; Mann and Yarowsky 2001: 151–158). However, it was applied to alignment between Slavic languages or Indo-European languages, and it seems difficult to imagine an effective bridge between very different languages like English and Chinese. Hybrid methods may produce better results. For instance, Piperidis *et al.* (1997: 57–62) first aligned sentences statistically, and then used a variety of information including part-of-speech categories, noun phrase grammars, and co-occurrence frequency to identify translation equivalents at word or multi-word level.

Instead of parallel corpora, Fung (1998: 1–17) tried to extract bilingual lexicons from comparable corpora, which is supposedly more difficult. She compared the context vector of a given English word with the context vectors of all Chinese words for the most similar candidate. During the process, a bilingual dictionary was used to map the context words in the two languages. About 30 per cent accuracy was achieved if the top-one candidate was considered, reflecting the inferiority of non-parallel corpora for bilingual lexicon extraction. Recent studies along this line have focused on improving the quality of the comparable corpora, using better association measures for comparing the context vectors, and addressing the polysemy problem with the bilingual dictionary used in the process, amongst others (e.g. Laroche and Langlais 2010: 617–625; Li and Gaussier 2010: 644–652; Bouamor *et al.* 2013: 759–764).

While most translation lexicon extraction methods do not particularly address domain specificity, Resnik and Melamed (1997: 340–347) suggested that a domain-specific translation lexicon could be obtained by filtering out general terms from the results. They compared the extracted lexicon entries against a machine readable dictionary and discarded the terms found in both. Kwong *et al.* (2004: 81–99) capitalized on the high consistency exhibited in bilingual Hong Kong court judgments to align and extract bilingual legal terminology based on context profiles, overcoming the problem of small corpus size and domain specificity.

Machine translation

Machine translation has always been considered one of the most important and typical applications of natural language processing. This type of sophisticated language processing task is often described as 'AI-complete' (e.g. Rich and Knight 1991), which means all difficult problems identified in artificial intelligence are relevant and the task can only be achieved when these problems are resolved. For a task like translation, which is more properly considered an art than a science, it requires deep understanding of the source language and the input text, and a sophisticated command of the target language. One may also need to possess a poetic

sense and be creative for literary translation. In addition to basic linguistic (lexical, syntactic, semantic, pragmatic and discourse) knowledge, common sense and encyclopaedic knowledge is particularly difficult to quantify and adequately represent. To produce high-quality translation without human intervention, which is the original and ambitious goal of MT research, was soon found to be unrealistic. The goals thus have to be toned down, such as aiming only at rough translation to be post-edited by humans and limiting the content to small sublanguage domains.

Traditional wisdom of MT depicts several levels of intermediary representation between the source text and the target text: Direct Translation, Transfer Approach, and Interlingua Approach. There are thus three phases in MT, namely analysis, transfer, and generation. Direct translation does little analysis but simply word-to-word translation from a source language to a target language. Lexical transfer considers the contrastive lexical differences between the source and target languages. Syntactic transfer considers the contrastive syntactic differences between the source and target languages. The source sentence is analysed according to the source language syntactic representation, which is converted to the corresponding target language syntactic representation, and the target sentence is generated accordingly. The interlingua approach uses a somewhat language-independent semantic representation to bridge the source sentence analysis and target sentence generation.

The first statistical machine translation system by Brown *et al.* (1990: 79–85) offered a new perspective to view the MT problem. Based on an aligned parallel corpus, automatic translation is achieved by finding the best translation, that is, the translation giving the highest probability $P(T|S)$ where T is the target sentence and S is the source sentence. Fitting it into the noisy-channel model, this is equivalent to finding the translation which gives the highest value for $P(S|T)P(T)$. These two terms analogously quantify two important criteria in translation: faithfulness and fluency respectively. The faithfulness model and the fluency model can be of different complexity or sophistication. Subsequent research on SMT has developed the original word-based model into phrase-based models (e.g. Marcu and Wong 2002: 133–139; Koehn *et al.* 2003: 48–54). Despite the apparent superiority and dominance of the statistical approach, traces of the transfer models are found in contemporary SMT research, such as syntax-based translation model (e.g. Liu *et al.* 2006: 609–616) as well as SMT model incorporating predicate-argument structures (e.g. Zhai *et al.* 2012: 3019–3036). As Costa-jussà *et al.* (2013: 1–6) remarked, there is a clear trend toward hybrid MT, where more linguistic knowledge is incorporated into statistical models, and data-driven methods are combined with rule-based systems.

MT is too significant a topic to be sufficiently discussed in a short section. For instance, example-based machine translation has also been an important empirical method enabled by large corpora (e.g. Sato and Nagao 1990: 247–252). An account of early MT can be found in Hutchins and Somers (1992), and more recent development of SMT is discussed in Koehn (2010). Maturing corpus processing techniques as well as statistical and hybrid approaches to various subtasks have again led to large projects and investments on machine translation and related language technology (e.g. Olive *et al.* 2011). We have nevertheless not covered here the area of computer-aided translation, which refers to various translation tools supporting human translators, such as online dictionaries, terminology banks, and translation memory, often integrated into the translator workstation (Hutchins 1998: 287–307).

Concluding remarks

We have thus taken a quick glimpse of natural language processing in general and a snapshot of its tasks and applications particularly relevant to translation. Although it is unrealistic to expect high quality fully automated translation by machine, the progress and development in NLP and MT is still remarkable in the last few decades. Translation technology developed so far is helpful, not only as tools to assist human translators for professional translations for various purposes, but also for anyone who is happy with some rough but immediate translation over the internet.

Translation is conventionally assessed in terms of faithfulness, fluency (or intelligibility) and elegance. Current statistical machine translation systems intend to model faithfulness and fluency. There is still plenty of room for improvement, but we may still wonder when MT systems can read between the lines. After all, to comply with professional standards, the translation for the simple sentence 'Thank you for reading' is expected to be radically different from that for 'Thank you for nothing', despite their surface similarities. It may be impractical to think of elegance and the artistic nature of translation at this moment, but it may not be impossible for other areas in NLP such as sentiment analysis to become mature enough to one day be incorporated into translation systems.

Notes

1 The ALPAC report, titled 'Language and Machine: Computers in Translation and Linguistics', was issued in 1966 by the Automatic Language Processing Advisory Committee established by the U.S. government.

2 The NTCIR is a series of evaluation workshops organized by the National Institute of Informatics of Japan, aiming at enhancing research in information access technologies. NLP applications evaluated include information retrieval, question answering, sentiment analysis, machine translation, etc.

3 SIGHAN is the Special Interest Group on Chinese Language Processing of the Association for Computational Linguistics.

4 SIGPARSE is the Special Interest Group on Natural Language Parsing of the Association for Computational Linguistics. CoNLL is the series of Conference on Computational Natural Language Learning, organized by the Special Interest Group on Natural Language Learning (SIGNLL) of the Association for Computational Linguistics.

References

Allen, James (1995) *Natural Language Understanding (Second Edition)*, Redwood City, CA: The Benjamin/Cummings Publishing Company, Inc.

Baroni, Marco and Silvia Bernardini (2004) 'BootCaT: Bootstrapping Corpora and Terms from the Web', in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 26–28 May 2004, Lisbon, Portugal, 1313–1316.

Boguraev, Bran and Ted Briscoe (eds) (1989) *Computational Lexicography for Natural Language Processing*, London: Longman.

Borin, Lars (2000) 'You'll Take the High Road and I'll Take the Low Road: Using a Third Language to Improve Bilingual Word Alignment', in *Proceedings of the 18th International Conference on Computational Linguistics*, 31 July–4 August 2000, Saarbrucken, Germany, 97–103.

Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum (2013) 'Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora', in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 4–9 August 2013, Sofia, Bulgaria, 759–764.

Brill, Eric (1995) 'Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging', *Computational Linguistics* 21(4): 543–565.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (1990) 'A Statistical Approach to Machine Translation', *Computational Linguistics* 16(2): 79–85.

Chan, Yee Seng, Hwee Tou Ng, and David Chiang (2007) 'Word Sense Disambiguation Improves Statistical Machine Translation', in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, 25–27 June 2007, Prague, Czech Republic, 33–40.

Chang, Chao-Huang and Cheng-Der Chen (1993) 'HMM-based Part-of-speech Tagging for Chinese Corpora', in *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, 22 June 1993, Ohio State University, Columbus, OH, 40–47.

Charniak, Eugene and Drew McDermott (1985) *Introduction to Artificial Intelligence*, Reading, MA: Addison-Wesley.

Charniak, Eugene (1993) *Statistical Language Learning*, Cambridge, MA: The MIT Press.

Charniak, Eugene (1997) 'Statistical Parsing with a Context-free Grammar and Word Statistics', in *Proceedings of the 14th National Conference on Artificial Intelligence (AAAI-97)*, 27–31 July 1997, Menlo Park, CA: AAAI Press, 598–603.

Chklovski, Timothy and Rada Mihalcea (2002) 'Building a Sense Tagged Corpus with Open Mind Word Expert', in *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, 11 July 2002, University of Pennsylvania, Philadelphia, PA, 116–123.

Church, Kenneth Ward and Patrick Hanks (1990) 'Word Association Norms, Mutual Information, and Lexicography', *Computational Linguistics* 16(1): 22–29.

Costa-jussà, Marta R., Rafael E. Banchs, Reinhard Rapp, Patrik Lambert, Kurt Eberle, and Bogdan Babych (2013) 'Workshop on Hybrid Approaches to Translation: Overview and Developments', in *Proceedings of the 2nd Workshop on Hybrid Approaches to Translation*, 8 August 2013, Sofia, Bulgaria, 1–6.

Cutting, Doug, Julian Kupiec, Jan Pedersen, and Penelope Sibun (1992) 'A Practical Part-of-speech Tagger', in *Proceedings of the 3rd Conference on Applied Natural Language Processing*, 31 March – 3 April 1992, Trento, Italy, 133–140.

Das, Amitava, Tanik Saikh, Tapabrata Mondal, Asif Ekbal, and Sivaji Bandyopadhyay (2010) 'English to Indian languages machine transliteration system at NEWS 2010', in *Proceedings of the 2010 Named Entities Workshop*, 16 July 2010, Uppsala University, Uppsala, Sweden, 71–75.

Edmonds, Philip and Scott Cotton (2001) 'SENSEVAL-2: Overview', in *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, 5–6 July 2001, Toulouse, France, 1–6.

Finch, Andrew and Eiichiro Sumita (2010) 'Transliteration Using a Phrase-based Statistical Machine Translation System to Re-score the Output of a Joint Multigram Model', in *Proceedings of the 2010 Named Entities Workshop*, 16 July 2010, Uppsala University, Uppsala, Sweden, 48–52.

Fung, Pascale (1998) 'A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora', *Lecture Notes in Artificial Intelligence*, Volume 1529, Springer, 1–17.

Gale, William A. and Kenneth W. Church (1991) 'A Program for Aligning Sentences in Bilingual Corpora', in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, 18–21 June 1991, University of California, Berkeley, CA, 177–184.

Gale, William A., Kenneth W. Church, and David Yarowsky (1992) 'One Sense per Discourse', in *Proceedings of the Speech and Natural Language Workshop,* 23–26 February 1992, New York, 233–237.

Gaussier, Eric (1998) 'Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora', in *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, 10–14 August 1998, University of Montreal, Montreal, Quebec, Canada, 444–450.

Goto, Isao, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou (2011) 'Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop', in *Proceedings of the 9th NTCIR Workshop Meeting*, 6–9 December 2011, Tokyo, Japan, 559–578.

Huang, Chang-ning and Hai Zhao (2007) 'Chinese Word Segmentation: A Decade Review', *Journal of Chinese Information Processing* 21(3): 8–19.

Huang, Jin-Xia and Key-Sun Choi (2000) 'Chinese-Korean Word Alignment Based on Linguistic Comparison', in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, 1–8 October 2000, Hong Kong, 392–399.

Huang, Yun, Min Zhang, and Chew Lim Tan (2011) 'Nonparametric Bayesian Machine Transliteration with Synchronous Adaptor Grammars', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 19–24 June 2011, Portland, OR, 534–539.

Hutchins, John (1998) 'The Origins of the Translator's Workstation', *Machine Translation* 13(4): 287–307.

Hutchins, John and Harold L. Somers (1992) *An Introduction to Machine Translation*, London: Academic Press.

Ide, Nancy and Jean Veronis (1998) 'Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art', *Computational Linguistics* 24(1): 1–40.

Ide, Nancy and Yorick Wilks (2006) 'Making Sense About Sense', in Eneko Agirre and Philip Edmonds (eds), *Word Sense Disambiguation: Algorithms and Applications*, Dordrecht: Springer.

Indurkhya, Nitin and Fred J. Damerau (2010) *Handbook of Natural Language Processing* (Second Edition), Boca Raton, FL: Chapman & Hall.

Jin, Chengguo, Seung-Hoon Na, Dong-Il Kim, and Jong-Hyeok Lee (2008) 'Automatic Extraction of English-Chinese Transliteration Pairs Using Dynamic Window and Tokenizer', in *Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, 11–12 January 2008, Hyderabad, India, 9–15.

Jurafsky, Daniel and James H. Martin (2009) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Second Edition), Upper Saddle River, NJ: Prentice-Hall.

Kay, Martin and Martin Röscheisen (1993) 'Text-translation Alignment', *Computational Linguistics* 19(1): 121–142.

Ker, Sue Jin and Jason S. Chang (1997) 'Aligning More Words with High Precision for Small Bilingual Corpora', *Computational Linguistics and Chinese Language Processing* 2(2): 63–96.

Kilgarriff, Adam and Gregory Grefenstette (2003) 'Introduction to the Special Issue on the Web as Corpus', *Computational Linguistics* 29(3): 334–347.

Kit, Chunyu, Xiaoyue Liu, King Kui Sin, and Jonathan J. Webster (2005) 'Harvesting the Bitexts of the Laws of Hong Kong from the Web', in *Proceedings of the 5th Workshop on Asian Language Resources (ALR-05)*, 14 October 2005, Jeju, Korea, 71–78.

Knight, Kevin and Jonathan Graehl (1998) 'Machine Transliteration', *Computational Linguistics* 24(4): 599–612.

Koehn, Philipp (2010) *Statistical Machine Translation*, Cambridge: Cambridge University Press.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu (2003) 'Statistical Phrase-based Translation', in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 27 May – 1 June 2003, Edmonton, Canada, 48–54.

Krovetz, Robert and W. Bruce Croft (1992) 'Lexical Ambiguity and Information Retrieval', *ACM Transactions on Information Systems* 10(2): 115–141.

Kübler, Sandra, Ryan McDonald, and Joakim Nivre (2009) *Dependency Parsing*, San Rafael, CA: Morgan and Claypool.

Kuo, Jin-Shea, Haizhou Li, and Chih-Lung Lin (2008) 'Mining Transliterations from Web Query Results: An Incremental Approach', in *Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, 11–12 January 2008, Hyderabad, India, 16–23.

Kwong, Oi Yee (2009) 'Homophones and Tonal Patterns in English-Chinese Transliteration', in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2–7 August 2009, Singapore, 21–24.

Kwong, Oi Yee (2011) 'English-Chinese Name Transliteration with Bi-directional Syllable-based Maximum Matching', in *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, 16–18 December 2011, Singapore, 11–19.

Kwong, Oi Yee (2012) *New Perspectives on Computational and Cognitive Strategies for Word Sense Disambiguation*, Springer Briefs in Speech Technology, New York: Springer.

Kwong, Oi Yee, Benjamin K. Tsou, and Tom B.Y. Lai (2004) 'Alignment and Extraction of Bilingual Legal Terminology from Context Profiles', *Terminology* 10(1): 81–99.

Laroche, Audrey and Philippe Langlais (2010) 'Revisiting Context-based Projection Methods for Term-translation Spotting in Comparable Corpora', in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 23–27 August 2010, Beijing, China, 617–625.

Lee, Chun-Jen, Jason S. Chang, and Jyh-Shing Roger Jang (2006) 'Extraction of Transliteration Pairs from Parallel Corpora Using a Statistical Transliteration Model', *Information Sciences* 176: 67–90.

Li, Bo and Eric Gaussier (2010) 'Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora', in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 23–27 August 2010, Beijing, China, 644–652.

Li, Haizhou, Min Zhang, and Jian Su (2004) 'A Joint Source-channel Model for Machine Transliteration', in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, 21–26 July 2004, Barcelona, Spain, 159–166.

Li, Haizhou, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong (2007) 'Semantic Transliteration of Personal Names', in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 23–30 June 2007, Prague, Czech Republic, 120–127.

Li, Haizhou, A Kumaran, Min Zhang, and Vladimir Pervouchine (2010) 'Report of NEWS 2010 Transliteration Generation Shared Task', in *Proceedings of the 2010 Named Entities Workshop*, 16 July 2010, Uppsala University, Uppsala, Sweden, 1–11.

Liu, Yang, Qun Liu, and Shouxun Lin (2006) 'Tree-to-string Alignment Template for Statistical Machine Translation', in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 17–21 July 2006, Sydney, Australia, 609–616.

Mann, Gideon S. and David Yarowsky (2001) 'Multipath Translation Lexicon Induction via Bridge Languages', in *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, 2–7 June 2001, Pittsburgh, PA, 151–158.

Manning, Christopher D. and Hinrich Schütze (1999) *Foundations of Statistical Natural Language Processing*, Cambridge, MA: The MIT Press.

Marcu, Daniel and William Wong (2002) 'A Phrase-based, Joint Probability Model for Statistical Machine Translation', in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 6–8 July 2002, Philadelphia, PA, 133–139.

Màrquez, Lluís, Gerard Escudero, David Martínez, and German Rigau (2006) 'Supervised Corpus-based Methods for WSD', in Eneko Agirre and Philip Edmonds (eds), *Word Sense Disambiguation: Algorithms and Applications*, Dordrecht: Springer.

McCarthy, Diana (2006) 'Relating WordNet Senses for Word Sense Disambiguation', in *Proceedings of the EACL-2006 Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, 4 April 2006, Trento, Italy, 17–24.

Melamed, I. Dan (1997) 'A Word-to-word Model of Translational Equivalence', in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL'97)*, 7–12 July 1997, Madrid, Spain, 490–497.

Mihalcea, Rada (2006) 'Knowledge-based Methods for WSD', in Eneko Agirre and Philip Edmonds (eds), *Word Sense Disambiguation: Algorithms and Applications*, Dordrecht: Springer.

Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff (2004) 'The SENSEVAL-3 English Lexical Sample Task', in *Proceedings of SENSEVAL-3, the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 25–26 July 2004, Barcelona, Spain, 25–28.

Mikheev, Andrei (2003) 'Text Segementation', in Ruslan Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*, London: Oxford University Press, 201–218.

Miller, George A (1995) 'WordNet: A Lexical Database for English', *Communications of the ACM* 38(11): 39–41.

Navigli, Roberto (2009) 'Word Sense Disambiguation: A Survey', *ACM Computing Surveys* 41(2): 1–69.

Och, Franz Josef and Hermann Ney (2003) 'A Systematic Comparison of Various Statistical Alignment Models', *Computational Linguistics* 29(1): 19–51.

Oh, Jong-Hoon and Key-Sun Choi (2005) 'An Ensemble of Grapheme and Phoneme for Machine Transliteration', in Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong (eds), *Natural Language Processing – IJCNLP 2005*, Springer, LNAI Vol. 3651, 451–461.

Olive, Joseph, Caitlin Christianson, and John McCary (eds) (2011) *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, New York: Springer.

Pease, Adam, Ian Niles, and John Li (2002) 'The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications', in *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, 29 July 2002, Edmonton, Canada.

Pedersen, Ted (2006) 'Unsupervised Corpus-based Methods for WSD', in Eneko Agirre and Philip Edmonds (eds), *Word Sense Disambiguation: Algorithms and Applications*, Dordrecht: Springer.

Piperidis, Stelios, S. Boutsis, and Iason Demiros (1997) 'Automatic Translation Lexicon Generation from Multilingual Texts', in *Proceedings of the 2nd Workshop on Multilinguality in Software Industry: The AI Contribution (MULSAIC'97)*, 25 August 1997, Nagoya, Japan, 57–62.

Resnik, Philip (1993) 'Selection and Information: A Class-based Approach to Lexical Relationships.', Doctoral Dissertation, Department of Computer and Information Science, University of Pennsylvania.

Resnik, Philip (2006) 'WSD in NLP Applications', in Eneko Agirre and Philip Edmonds (eds), *Word Sense Disambiguation: Algorithms and Application*, Dordrecht: Springer.

Resnik, Philip and I. Dan Melamed (1997) 'Semi-automatic Acquisition of Domain-specific Translation Lexicons', in *Proceedings of the 5th Conference on Applied Natural Language Processing*, 31 March – 3 April 1997, Washington, DC, 340–347.

Resnik, Philip and Noah A. Smith (2003) 'The Web as a Parallel Corpus', *Computational Linguistics* 29(3): 349–380.

Rich, Elaine and Kevin Knight (1991) *Artificial Intelligence*, New York: McGraw-Hill Book Company.

Sanderson, Mark (1994) 'Word Sense Disambiguation and Information Retrieval', in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 3–6 July 1994, Dublin, Ireland, 142–151.

Sato, Satoshi and Makoto Nagao (1990) 'Toward Memory-based Translation', in *Proceedings of the 13th International Conference on Computational Linguistics*, 20–25 August 1990, Helsinki, Finland, 247–252.

Schank, Roger C. and Robert P. Abelson (1977) *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*, Hillsdale, NJ: Lawrence Erlbaum.

Shishtla, Praneeth, Surya Ganesh V, Sethuramalingam Subramaniam, and Vasudeva Varma (2009) 'A Language-independent Transliteration Schema Using Character Aligned Models', in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, 7 August 2009, Singapore, 40–43.

Simard, Michel, George F. Foster, and Pierre Isabelle (1992) 'Using Cognates to Align Sentences in Bilingual Corpora', in *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT (TMI-92)*, 25–27 June 1992, Montreal, Canada, 67–81.

Smadja, Frank, Vasileios Hatzivassiloglou, and Kathleen McKeown (1996) 'Translating Collocations for Bilingual Lexicons: A Statistical Approach', *Computational Linguistics* 22(1): 1–38.

Song, Yan, Chunyu Kit, and Hai Zhao (2010) 'Reranking with Multiple Features for Better Transliteration', in *Proceedings of the 2010 Named Entities Workshop*, 16 July 2010, Uppsala University, Uppsala, Sweden, 62–65.

Specia, Lucia, Maria das Graças Volpe Nunes, Mark Stevenson, and Gabriela Castelo Branco Ribeiro (2006) 'Multilingual versus Monolingual WSD', in *Proceedings of the EACL-2006 Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together,* 4 April 2006, Trento, Italy, 33–40.

Sproat, Richard and Thomas Emerson (2003) 'The First International Chinese Word Segmentation Bakeoff', in *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, 11–12 July 2003, Sapporo, Japan, 133–143.

Sproat, Richard, William Gales, Chilin Shih, and Nancy Chang (1996) 'A Stochastic Finite-state Word-Segmentation Algorithm for Chinese', *Computational Linguistics* 22(3): 377–404.

Strassel, Stephanie, Caitlin Christianson, John McCary, William Staderman, and Joseph Olive (2011) 'Data Acquisition and Linguistic Resources', in Joseph Olive, Caitlin Christianson, and John McCary (eds), *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, New York: Springer.

Tao, Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat, and ChengXiang Zhai (2006) 'Unsupervised Named Entity Transliteration Using Temporal and Phonetic Correlation', in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 22–23 July 2006, Sydney, Australia, 250–257.

Virga, Paola and Sanjeev Khudanpur (2003) 'Transliteration of Proper Names in Cross-lingual Information Retrieval', in *Proceedings of the ACL2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, 12 July 2003, Sapporo, Japan, 57–64.

Winograd, Terry (1973) 'A Procedural Model of Language Understanding', in Roger C. Schank and Kenneth Mark Colby (eds), *Computer Models of Thought and Language*, San Francisco, CA: Freeman, 152–186.

Wu, Dekai (1994) 'Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria', in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, 27–30 June 1994, Las Cruces, NM, 80–87.

Wu, Dekai (2010) 'Alignment', in Nitin Indurkhya and Fred J. Damerau (eds), *Handbook of Natural Language Processing* (Second Edition), Boca Raton, FL: Chapman & Hall.

Wu, Dekai and Xuanyin Xia (1995) 'Large-scale Automatic Extraction of an English-Chinese Translation Lexicon', *Machine Translation* 9(3–4): 285–313.

Wutiwiwatchai, Chai and Ausdang Thangthai (2010) 'Syllable-based Thai-English Machine Transliteration', in *Proceedings of the 2010 Named Entities Workshop*, 16 July 2010, Uppsala University, Uppsala, Sweden, 66–70.

Xue, Nianwen and Susan P. Converse (2002) 'Combining Classifiers for Chinese Word Segmentation', in *Proceedings of the 1st SIGHAN Workshop on Chinese Language Processing*, 1 September 2002, Taipei, 63–69.

Zhai, Feifei, Jiajun Zhang, Yu Zhou, and Chengqing Zong (2012) 'Machine Translation by Modeling Predicate Argument Structure Transformation', in *Proceedings of the 24th International Conference on Computational Linguistics*, 8–15 December 2012, Mumbai, India, 3019–3036.

Zhang, Min, Haizhou Li, A Kumaran, and Ming Liu (2011) 'Report of NEWS 2011 Machine Transliteration Shared Task', in *Proceedings of the 3rd Named Entities Workshop (NEWS 2011)*, 12 November 2011, Chiang Mai, Thailand, 1–13.

35
ONLINE TRANSLATION

Federico Gaspari

UNIVERSITY OF BOLOGNA AT FORLÌ, ITALY

Overview

This chapter concerns key aspects related to online translation, and focuses on the relationship between translators and the web. The 'Introduction' section offers an overview of the first internet-based communication channels used by the early online communities of language and translation professionals, and charts their subsequent evolution. The section on 'The ecosystem of online translation' presents a range of web-based resources for translators, including online (meta-)dictionaries, glossaries, terminology databases and shared translation memories, highlighting their key features. The following part covers online translation tools and internet-based translation environments, such as browser-based applications that support translation projects from start to finish, allowing the deployment of communal or proprietary translation memories and glossaries, as well as the integration of online machine translation (MT) for subsequent post-editing to boost productivity. The chapter continues with a description of the key features of the most popular online translation marketplaces, given their growing importance in creating business opportunities for translators.

Finally, the section on 'Online translation in the Web 2.0' is devoted to the latest developments of user-generated translation in the Web 2.0 scenario, and reviews high-profile online collaborative translation projects as well as crowdsourcing efforts. This leads to an assessment of the translation crowdsourcing model, in which volunteer (amateur) translators are involved in projects for the localization of popular social media platforms of which they are themselves users. Each section discusses representative examples of relevant websites, online tools, web resources, internet-based projects or services, with a bias towards those available in English, which are analyzed from the perspective of translators. While the inclusion or exclusion of items from the various categories does not imply any endorsement or criticism, or any implicit judgment of their quality, an attempt is made to identify common trends and interesting specific features that have played a role in the development of online translation, discussing their pros and cons.

Introduction: the origins of online translation

The gradual impact of the internet on translation

As has been the case for virtually all other professions and businesses, the internet has had a profound impact on translation, dramatically accelerating the process that started in the 1980s,

when personal computers became widely available to translators, initially as sophisticated typewriters. This was the prelude to rapid developments in the following three decades that were bound to affect the daily work of translators and, more broadly, increase the impact of technology on the ways in which translations are produced, circulated and finally used by clients and target readers (Cronin 2013). The internet affected translation only indirectly until the late 1990s, due to two main reasons: first of all, until that time the internet was not yet widely available across the globe, with the exception of relatively few users working mostly for government agencies, academic institutions, multinational companies and large organizations in a small number of countries around the world; second, the internet still had very little to offer to translators, compared to the impressive array of online resources, tools and opportunities that can be found online today.

For most of the 1990s, then, one could talk of online translation only insofar as translation agencies and professionals progressively started to use email in order to receive source texts from their clients, and deliver back their translations (O'Hagan 1996), with some pioneers setting up (often multilingual) websites to advertise their services. This was clearly a very limited use of the internet, which was largely confined to taking advantage of its speed and convenience for transferring files electronically and for promotional purposes. The associated benefits gradually led to the decline in the use of fax machines and floppy disks, which had been the primary means to exchange original and translated documents during the final stages of the pre-internet era. The early transition to the use of email by translators (and their clients) motivated by basic file transfer needs encouraged more and more professionals to get online towards the late 1990s. This, in turn, laid the foundation for the first online communities of translators, who started to use email and the then rather rudimentary internet-based communication facilities to discuss work-related issues.

Early translation-related newsgroups and online forums

Usenet-based newsgroups were early online communities organized in hierarchical categories around topics of interest to their members, and designed to share textual messages via the nascent internet infrastructure. Users could read and post messages for their preferred categories, thus becoming part of virtual distributed communities: this made it possible for translators to exchange information among themselves and to consult more knowledgeable colleagues, for instance to seek advice on terminological issues, payment practices, etc. Members could generally sign up to these newsgroups free of charge, and it was not uncommon for individuals to be members of multiple (in some cases partially overlapping) communities, depending on their interests. This was a significant turning point for translators, as for the first time it was no longer necessary to attend conferences, workshops or meetings of translators' associations to be instantly connected to fellow professionals. Indeed, physical distance became irrelevant, and such online asynchronous discussion systems allowed the interaction of translators and interpreters over the internet across time zones, regardless of where they were based – crucially, due to their multilingual skills, international contacts of this nature were more feasible and natural for translators and interpreters than for any other professionals.

One of the first newsgroups of specific interest to translators, sci.lang.translation, was set up in late 1994 as an unmoderated newsgroup, where related postings were presented as threads. Its aims included facilitating discussion among members on professional issues such as the activities of translators' organizations, accreditation procedures, dictionaries and other useful resources, terminology problems, training opportunities, etc. Subsequently, with the development of the internet and associated technologies, several Usenet-type newsgroups

evolved into discussion forums, which had additional user-friendly features designed to make the exchange of postings smoother. Some of them subsequently migrated to more easily accessible platforms – sci.lang.translation, for example, became part of Google Groups,[1] and it still serves as a forum for discussions on a vast number of translation-related subjects. Although the volume of traffic has increasingly moved to other channels (see the next section), there are still several popular online forums about translation, languages and linguistics, such as those hosted by WordReference.com (more information in the section 'Online dictionaries and meta-dictionaries') and others that are part of websites serving as online translation marketplaces (see the section 'Online translation marketplaces').

From translators' mailing lists to web social media

Around the mid-1990s, the first mailing lists for (and run by) translators started to appear. They were different from newsgroups and discussion forums in some respects: mailing lists were usually moderated, and instead of logging on to a website to view the latest threads, subscribers received the posts of fellow members directly into their email inbox, and they could also reply and send new messages to other subscribers from their email client. Mailing lists could have memberships of variable sizes, from a few dozens to several thousands, and as a consequence their email traffic varied dramatically. These email-based communication channels were no longer public, strictly speaking, in that one had to join the mailing lists to start receiving and posting messages. Mailing lists, especially those with many members and heavy traffic, adopted policies to manage the flow of information efficiently; these included conventions concerning the subject lines of emails posted to the group, allowing for quick filtering of relevant messages and management of threads, and the possibility of receiving overview summaries (so-called 'digests') of messages posted during a day or a week, to avoid the constant trickle of emails on diverse topics.

One of the earliest mailing lists for translators, which still retains a sizeable international membership, is Lantra-L.[2] This is a generalist mailing list, i.e. it does not impose any restriction on membership and broadly welcomes discussions on any topic related to translation and interpreting, concerning any language and language combination, although messages are predominantly in English, which in itself guarantees a large readership. Other mailing lists have chosen different approaches, adopting more restrictive criteria in terms of their membership and remit, e.g. by focusing exclusively on a single target language or language pair, concentrating on specific professional areas (say, literary as opposed to technical or specialized translation), or encouraging membership from translators based in certain countries. Of course, mailing lists can also be open only to the members of an existing translators' community or organization, for instance a regional or national professional association, and as such serve as restricted communication channels among members, e.g. to circulate newsletters or information on upcoming events.

Since the mid-1990s newsgroups, online discussion forums and mailing lists have had remarkable appeal for translators, in particular because they suddenly offered an unprecedented antidote to the oft-lamented isolation and loneliness of translation professionals. These online communication channels created a sense of community, giving individuals access to the collective wealth of expertise of translators spread around the world, at virtually no cost once a user had a computer with internet access (Plassard 2007: 643–657; Williams 2013: 92). Wakabayashi (2002) and Alcina-Caudet (2003: 634–641) stress the importance of helping translation students and budding professionals to familiarize themselves with these email-based online communities, due to the large repository of expertise to which they give access. In a similar vein, several translators also maintain blogs to share their views on professional matters,

and McDonough Dolmaya (2011a: 77–104) analyzes a convenience sample of 50 such translation blogs to discover how they are used to discuss translation-related problems, generate business contacts and socialize with colleagues around the world.

With the passing of time, however, this scenario is changing: just as the increased popularity of personal email accounts and the availability of user-friendly clients boosted the success of translators' mailing lists around the turn of the twenty-first century, they currently seem to be losing ground due to the growing role of the web social media. For instance, the popular online social networking platform Facebook[3] supports the fast and easy creation of (public, restricted or private) groups and shared pages to discuss any topic and to circulate information among members of the online community, encouraging cross-platform interaction. LinkedIn[4] serves a similar purpose, but with a much stronger focus on professional networking, hence its specific relevance to practising translators and players in the translation and localization industry. Finally, the microblogging site Twitter[5] allows its users to follow real-time updates on issues of interest to them, reading and exchanging brief online posts that can be organized and prioritized according to personal preferences (the section 'Crowdsourced online translation projects' reviews crowdsourced online translation projects promoted by Facebook, LinkedIn and Twitter). The visibility of translation and the presence of translators on these and other web social media can be expected to grow in the foreseeable future, while the activity of traditional online discussion forums and mailing lists seems to be gradually, and probably irreversibly, dwindling.

The ecosystem of online translation

Online resources for translators

With the start of the twenty-first century, online translation became a reality in parallel with the gradual growth of the internet into a widely used multi-faceted environment to create, share and circulate digital contents in multiple languages for increasingly diverse audiences. In spite of all its multimedia contents (videos, graphics, sound, etc.), the web is still a heavily textual and language-intensive medium – one need only think of how search engines are used. In addition, the multilingual and multicultural nature of its global user population arguably makes the internet a natural environment for translation: it is therefore not surprising that a multitude of valuable online resources are available for translators, some of which are created by members of the translation community. Their actual use by, and relevance for, professionals varies depending on their working languages and specialisms, but it is fair to say that hardly any translator today can afford to neglect the importance of online resources. Indeed, they are increasingly considered part and parcel of the translators' standard working environment, as traditional libraries and documentation centres are unlikely to match the quantity and diversity of relevant information that is available on the internet – as Cronin (2013: 8) puts it, 'the working practices of translators have been changed beyond recognition in terms of the access to many different kinds of knowledge that are afforded by the infrastructure of the internet'. The double challenge for translators consists in finding quickly high-quality websites and valuable online sources when they need them, and in using them properly.

Online dictionaries and meta-dictionaries

At the most basic level, translators can take advantage of online resources that are made available to the general public, such as monolingual and multilingual dictionaries, some of which are of very high quality and may be available free of charge, including the online versions of leading

traditional printed dictionaries for an impressive variety of languages and language pairs. Apart from these general-purpose online resources, there are others that are more specifically geared towards the needs of translators, including technical and specialized dictionaries. In addition, the web hosts a number of meta-dictionaries, which aggregate several online reference works (sometimes also including encyclopedias) in many languages drawing on a huge amount of background information, and which support multiple simultaneous word searches.[6]

Lexicool.com,[7] for example, is an online directory listing thousands of free internet-based bilingual and multilingual dictionaries and glossaries, which also offers other translation-oriented language resources covering several technical and specialized domains. The directory, which is run by an international team of linguists and is based in France, has steadily grown since it was put online in 1999. Users can search the available dictionaries and glossaries by selecting a combination of criteria including the language(s) of interest to them, the specific subject of their search and any keywords to identify the domain more precisely. Another popular collection of resources among translators is hosted by WordReference.com,[8] which was also launched in 1999 and provides free online bilingual dictionaries as well as a variety of other contents of interest to translators. This website offers two main sets of online resources: first of all, bilingual dictionaries primarily of English in combination with other languages, even though WordReference.com also aims to serve other languages, including Spanish, French, Italian, Portuguese, etc.; second, the site also provides a wide range of language-specific and translation-related forums on a number of topics, especially concerning the meaning and translation of words, phrases and idioms in several languages (Cronin 2013: 74–75). The archives of these forums are publicly available for perusal, but some of them require registration before a user can post a question or reply to an existing message.

WordReference.com combines the traditional discussion forum format (see also the section on 'Online Translation Marketplaces' for other websites offering similar virtual meeting spaces) with the constantly expanding availability of online dictionaries, thus bringing together lexicographic information and expert advice on usage and translation issues. One potential problem in this respect is that users who volunteer answers may not be great experts of the languages or fields in question, and it might be difficult for inexperienced professionals to identify forum members who can give valuable advice. On the one hand, the translators posting questions and requests for help via online discussion forums, for example concerning terminology, must provide sufficient background information and context with their queries, so that others are effectively able to provide relevant and helpful answers. On the other hand, though, the ultimate responsibility for using wisely the advice received from such mostly anonymous online communities rests with the translators themselves, as they are accountable to clients for the quality of their work. This underscores the need to scrutinize the trustworthiness and reliability of information obtained from internet-based sources, including via professional discussion forums and online translators' communities, due to the woeful lack of robust screening procedures linked to the professional credentials of online forum members.

Online terminology databases and glossaries

There are also several online resources for translators which are vetted and made available by reputable sources following rigorous quality checks, thus providing more reliable and authoritative materials to translators conducting research for a specific assignment. This is the case, for example, of IATE,[9] an online multilingual terminology database that was developed from the European Union's inter-institutional terminology assets, which were built up over many years by translators, terminologists and linguists working in various EU institutions and

agencies (Ball 2003). IATE started to be used internally by the EU in 2004, and the database was made publicly available online in mid-2007, following significant efforts in terminology collection, validation and standardization. Its records are constantly extended with new entries concerning EU-related terminology in all the official EU languages from a wide variety of areas, including law, agriculture, economics, computing, etc.

IATE can be consulted online by anybody via a user-friendly interface available in all the official EU languages, which offers a number of intuitive features, including the possibility of automatically loading the user's search preferences for source and target languages; a star-rating system showing how reliable an equivalent term in the target language is considered; term definitions, cross-references to the sources consulted for term validation and examples of real uses in context. In addition, users can submit queries for terms selecting specific search domains from dozens of options (ranging from 'air and space transport' to 'wood industry'). IATE is a high-profile publicly available resource containing excellent multilingual terminological data that can be used by any translator, not only those working in or for the EU institutions.

There are countless other web-based terminological databases and online glossaries covering an extremely wide range of languages and subjects, which can be found on the internet using search engines or based on the recommendations and advice of colleagues. Of course, nothing prevents translators from mining the web directly for linguistic and terminological information, without consulting pre-compiled lexicographic online resources. A number of tools have been created to support translators in their efficient online terminology research. For example, IntelliWebSearch[10] is a freeware application developed by a professional translator to speed up the search for terms on the internet when working on a translation (see also e.g. Durán Muñoz 2012: 77–92). Again, caution is required to differentiate trustworthy, valuable webpages from online sources of dubious or variable quality.

Online shared translation memories and parallel corpora

In addition to online (meta-)dictionaries, terminological databases and glossaries, the internet also offers other language resources whose use entails a higher level of technical expertise on the part of translators, but whose benefits can be very significant. This is the case for a number of multilingual translation memory repositories that have been made available online as part of different projects. Three examples provided by the EU are reviewed here for illustration purposes, starting with the DGT-Translation Memory,[11] a large database that has been updated every year since 2007 and contains the multilingual parallel versions of the Acquis Communautaire (i.e. the entire EU legislation, including all the treaties, regulations and directives) in the official EU languages (Steinberger *et al.* 2012: 454–459). All the multilingual versions of each text are aligned at the sentence level, which results in an impressive set of combinations: more than 250 language pairs, or over 500 language pair directions; this is particularly remarkable because some of the languages covered are not particularly well supported in terms of parallel corpora or language resources (as is the case for relatively rare language combinations such as Maltese–Finnish or Estonian–Slovene). The parallel texts can be downloaded in the widely adopted Translation Memory eXchange (TMX) format, thus enabling translators to import the aligned parallel data into their preferred translation memory software.

Similarly to the case of IATE, the DGT-Translation Memory is released online as part of the European Commission's effort to support multilingualism, language diversity and the reuse of information produced within the EU institutions. These high-quality multilingual parallel texts can be used by translators when working with translation memory software, to ensure

that identical segments in new source texts do not have to be retranslated from scratch several times (or to avoid them ending up being translated inconsistently into the same target language); in addition, similar past translations can be leveraged to speed up the current job, while ensuring consistency across translation projects. Collections of multilingual parallel data such as the DGT-Translation Memory can also be used for a variety of other purposes, such as training statistical MT systems, extracting monolingual or multilingual lexical and semantic resources, developing language models for data-driven systems with a linguistic or translation component, etc. Other similar, but smaller, parallel data sets publicly released online by the EU are the ECDC-TM,[12] which contains documents in 25 languages from the European Centre for Disease Prevention and Control, and the EAC-TM,[13] consisting of texts in 26 languages from the Directorate General for Education and Culture.

The OPUS project[14] is dedicated to assembling a growing collection of publicly available parallel corpora covering dozens of languages, including multiple domains such as law, administration (again using mostly EU texts), film subtitles, software and computing documentation, newspaper articles, biomedical papers, etc. (Tiedemann 2012: 2214–2218). Overall, the parallel corpora of the OPUS collection cover nearly 4,000 language pairs, for a total of over 40 billion words, distributed in approximately 3 billion parallel translation units (aligned sentences and fragments), and the collection is constantly growing. These multilingual parallel corpora provided by the OPUS project can be downloaded from the web in different formats (native XML, TMX and plain text format), so that translators can use them with their own translation memory software.

In addition, the OPUS project website offers an online interface to directly interrogate the parallel corpora, for example to generate (multilingual aligned) concordances from the selected corpora: users can consult their chosen languages in parallel, or query a single version of a given corpus to consider one language at a time. Although it is difficult to retrieve information on which version of the parallel texts was the original one, all the translations were done by human professionals, thus guaranteeing high-quality standards across languages. To accompany these online linguistic resources, the OPUS project website also makes available a range of other tools to further annotate and process the data, so that they can be used for multiple purposes, as noted above for the DGT-Translation Memory.

These resources created by the EU institutions and by the OPUS project have been made available online to the advantage of different users and for a variety of purposes, but it is easy to see how translators in particular can benefit from them, also because they are constantly growing with updates and extensions. In addition, these translation memories and parallel corpora come from reputable sources and underwent a number of quality checks. There exist several similar online resources, which however tend to be subject to a more restrictive regime in terms of use and distribution, partly because they were created within commercial entities, and are covered by copyright and other intellectual property constraints. Moreover, in some cases, the reliability of potentially useful resources for translators found online may be unclear in terms of provenance, copyright status and quality requirements.

Online translation platforms, environments and tools

Online platforms to share translation memories

MyMemory[15] claims to be the world's largest free online translation memory, but in fact it does not limit itself to offering a collection of linguistic data, because it also provides a web platform to upload, store and manage a repository of translation memories for different language

pairs and various domains. MyMemory is free to use for registered members, and its multilingual translation memories have been assembled by collecting parallel texts from different institutions, including the EU and the UN, as well as from other sources, particularly crawling multilingual websites covering several domains – it should be noted that this resource is not as well documented as the ones reviewed in the section 'Online shared translation memories and parallel corpora'.

After registering with MyMemory, users can download translation memories customized to match the texts that they intend to translate, and use them in their own computer-assisted translation environments. Interestingly, when its existing translation memory databases fail to come up with relevant matches for a document uploaded by a user to be translated, MyMemory provides raw output from statistical MT as a translation draft. Differently from the resources examined in the section 'Online shared translation memories and parallel corpora', MyMemory also invites users to directly contribute to the growth of its online repository by uploading additional translation memories or by editing existing materials via its platform, although it is not clear what guidelines or quality controls apply in such cases. In addition, MyMemory's online platform also provides the possibility of searching terms and translated segments among its resources in multiple language pairs and domains.

Online translation environments and tools

Google Translator Toolkit[16] (GTT) is a web application that supports translators, not necessarily only professionals, but also amateur bilinguals who wish to translate documents or webpages. GTT was released in 2009 and handles several document formats; it also includes a specific facility to import Wikipedia entries that a user wishes to translate into another language, taking care of layout and formatting issues and letting the user concentrate on the translation of the text. This browser-based translation environment, which is reported to support hundreds of source and target languages, enables users to set up translation projects and gives them the option to upload relevant glossaries and translation memories, if they so wish. Interestingly, and similarly to what was noted in the section 'Online platforms to share translation memories' for MyMemory, GTT can also be configured so that it offers a draft of the target text based on MT output provided by Google Translate,[17] which the user can then correct and improve (on the development of the early free online MT systems see Yang and Lange 1998: 275–285; Yang and Lange 2003: 191–210; Gaspari and Hutchins 2007: 199–206).

Effectively, then, GTT is a cloud-based customizable translation memory environment, with an integrated terminology lookup function and an optional statistical MT facility supported by Google Translate, of which the user can take advantage when no matches are found in the translation memories. One feature of interest to translators is that when the translation of a document is completed, the updated translation memory created during the project can be downloaded from the GTT online environment and imported by users in TMX format along with the finished target text, thus allowing for its reuse in subsequent projects and even with different offline translation memory tools. Since GTT was made available for free to users, there has been speculation that this might be a veiled attempt by Google to acquire high-quality glossaries and parallel texts from translators to feed its own proprietary MT system, also relying on users' corrections of the raw MT output to improve its performance (Garcia and Stevenson 2009: 16–19). It should be noted, however, that users can choose not to share the resources that they upload online into GTT when working on a translation project.

Other cloud-based translation environments have been developed, which represent an increasingly popular option to run distributed translation projects with centrally managed

resources and assets. A couple of notable examples in this area are MemSource Cloud,[18] a translation environment that supports translation memory and terminological resources as well as integrated MT, and the Translation Workspace,[19] an on-demand translation productivity solution providing access to translation memory and glossary assets offered through the GeoWorkz[20] portal by Lionbridge,[21] a large translation and localization provider.

Boitet *et al.* (2005) and Bey *et al.* (2008: 135–150) describe the design and development of an online translation environment called BEYTrans. This wiki-based platform is aimed at volunteer translators working on shared projects via the internet, especially those involved in well-organized mission-oriented translation communities, say to translate the technical documentation related to open-source software development projects. The BEYTrans environment enables users to access and manage a range of online language tools and resources, including multilingual dictionary and glossary lookup facilities and community-specific translation memories to match the requirements of particular translation projects. In addition, this wiki-based translation platform also has an integrated facility that supports Internet searches for translated text fragments in the target language corresponding to parts of the source text. Similarly to GTT, the output of free web-based MT services can also be provided: BEYTrans users are expected to check and improve it where necessary.

Utiyama *et al.* (2009) and Kageura *et al.* (2011: 47–72) present three projects related to a Japan-based open online translation aid and hosting service called Minna no Hon'yaku.[22] This was launched in 2009, initially to facilitate voluntary translation work for NGOs and with subsequent ramifications into the commercial domain. QRedit is the translation-aid text editor of this online translation platform, and was designed specifically to support the work of online distributed communities of translators (Abekawa and Kageura 2007: 3–5). The overall platform enables users to contribute to translation projects involving a set of language pairs with a strong Asian focus, offering them access to useful online language resources and translation-related tools. These include integrated dictionary lookup, a terminology management system, a bilingual concordancer and a bilingual term extraction facility. Depending on the type of project (as some of them have a more commercial, rather than philanthropic or humanitarian, nature), the platform may also be partly open to non-professional translators, who have to pass a language proficiency test before they can contribute to the translations, and then they receive payment for their work – the quality requirements are adjusted accordingly.

Online translation marketplaces

This section gives an overview of a small sample of particularly popular online marketplaces for translation and related services, which evolved as the distributed communities of translators (and their clients) grew over time and converged onto the web to conduct parts of their business. A long-standing and well-established online platform devoted to connecting professionals, agencies and clients specifically in the translation and localization industry is Aquarius,[23] which was launched in 1995 and aims to support the outsourcing of translation and localization projects. The site includes some features to match offer and demand, i.e. a list of jobs where potential contractors can give quotes for available projects, and a directory of professionals that can be searched by clients. In addition, the Aquarius website has an area with forums where subscribers can post professional questions and elicit feedback or advice from other registered members.

Another professional networking site for translators is ProZ.com,[24] which was founded in 1999 and now gathers a large international community of translators, interpreters, translation companies and clients. The ProZ.com site is a virtual marketplace where professionals can

outsource and accept assignments, give feedback on their experience with clients, exchange information on their work, request help on the translation of specific terminology, be referred to training opportunities, etc. ProZ.com has an interesting mechanism to incentivize and reward successful high-quality online interactions among its members. When someone receives help from a fellow ProZ.com member in reply to a request for assistance concerning the translation of a difficult term or phrase, the author of the most useful answer is rewarded with KudoZ points; these are favorable ratings that are used to establish the professional reputation of ProZ.com members, and subsequently to rank professionals in the directory of language service providers from which clients can pick contractors (Perrino 2009: 55–78).

TranslatorsCafé.com[25] was launched in 2002 and is another popular networking site hosting an online community and marketplace for translators. Not only does it provide access to a range of language- and translation-related news and resources such as online forums, but it also offers a platform to connect with other professionals and to contact registered translation agencies looking for translation and interpreting services. Similarly to the policy adopted by ProZ.com, TranslatorsCafé.com also encourages users to provide feedback and comments on the online activities of fellow members, in particular by rating the quality and usefulness of answers to the questions they post. McDonough (2007: 793–815) proposes a framework for categorizing and describing such online translation networks, and discusses TranslatorsCafé.com as a case study of a practice-oriented translation network; the analysis of the most typical and interesting online behavior of its members reveals the dynamics governing this and other similar virtual professional communities, also showing that less than 10 percent of registered users ever posted a message in the discussion forum, and that threads normally had many more views than replies (ibid.: 809).

There are several other portals, online translation marketplaces and networking sites specifically designed to match translators and their clients, e.g. GoTranslators,[26] Langmates.com[27] and TranslationDirectory.com.[28] They all have different strengths and specific features, depending on the areas in which they specialize and on the specific types of professionals, agencies and clients that they aim to serve. Nowadays online translation marketplaces are an important arena bringing together professionals who can secure projects and clients who can source some of their service providers. Given their vital role in an increasingly fragmented and competitive industry, the significance of online translation marketplaces in matching offer and demand is likely to grow in the future.

Pros and cons of online translation resources, tools and services

Due to the very nature of the internet, online resources such as (meta-)dictionaries, glossaries, terminological databases, translation memory repositories, etc. are subject to constant and potentially rapid change, which presents both advantages and disadvantages. On the positive side, materials published online can be constantly updated, extended and refreshed at a fraction of the costs and time needed for their printed counterparts, also lending themselves to quick searching and further electronic processing. This helps, for example, the timely inclusion in major online dictionaries of neologisms, accepted borrowings, etc., as well as the constant creation of new web-based technical glossaries, or the expansion of already existing terminological databases, translation memory and corpus repositories, covering an ever growing range of languages and specialized domains. As has been mentioned above, these online resources can then be accessed very easily via web interfaces, or conveniently downloaded for further offline use by professional translators.

On the other hand, however, the unstable nature of the internet means that valid information sources and online resources on which translators rely for research and documentation purposes may suddenly disappear without any warning, or the quality of their contents may deteriorate over time (e.g. by becoming obsolete, due to the lack of updates or thorough quality checks, which may not always be transparent to users). Another danger is that the multitude of resources, tools and services available online to translators may make it difficult to select the reliable high-quality ones, especially for relatively inexperienced professionals working with widely used languages; as a result, the task of screening websites and online resources on the basis of their quality may turn out to be a time-consuming activity.

A final word of caution concerns the confidentiality of data circulated or shared on the web. For example, MyMemory offers a mechanism to protect privacy, by removing people's and brand names from the translation memories that registered users upload onto its online platform to be shared. However, sensitive information may go beyond these details, for example in the case of documents concerning financial matters, legal disputes or medical conditions, or describing inventions, patents, industrial applications, etc. Translators and translation agencies have a duty to protect the privacy and interests of their clients, therefore they should carefully consider all the confidentiality issues and legal implications that may arise, before passing on texts and translation resources to third parties over the internet (Drugan and Babych 2010: 3–9).

Online translation in the Web 2.0

Online collaborative translation

The web 2.0 emphasizes the social and collaborative dimensions of the internet, with users becoming active producers of dynamic online contents distributed and shared across platforms, rather than purely passive consumers of static pages found in websites authored by others. Crucial to the achievement of this vision is the removal (or at least the reduction) of language barriers: users are not only empowered to author online contents in their preferred languages, but they also contribute to their translation (O'Hagan 2011: 11–23; Cronin 2013: 99ff.). Given its diverse and multilingual nature, Wikipedia[29] epitomizes the key role of users as both authors and translators in making its entries accessible in multiple languages to communities of internet users with different national, linguistic and cultural backgrounds. What is worth noting here is that online translation is performed not only by professionals or trained translators, but also more and more often by multilingual amateurs as a hobby or volunteer activity.

Désilets *et al.* (2006: 19–31) discuss the design and implementation of processes and tools to support the multilingual creation and maintenance of multilingual wiki contents; their discussion is not restricted to Wikipedia, but also applies in principle to any wiki site with collaboratively created and translated multilingual contents. Désilets (2007; 2011) outlines a number of challenges for translators and consequences for translation arising from the model of massive online collaboration leading to the distributed and user-participated creation and translation of web content, especially due to the availability of open and shared online translation-oriented resources. Désilets and van der Meer (2011: 27–45) describe current best practices to manage collaborative translation projects successfully. Gough (2011: 195–217) investigates the trends that have developed in the translation industry as part of the transition towards the web 2.0: her sample of 224 respondents reveals that professional translators have a vague awareness and limited understanding of such trends, and also that they make marginal use of open online tools, engaging little in collaborative translation processes. McDonough Dolmaya (2012: 167–191) presents a survey among Wikipedia volunteer translators to explore

the perception of collaborative translation efforts, also discussing the role of the organizations driving them, while Fernández Costales (2012: 115–142) illustrates the range of motivations leading volunteers to become involved in web-based collaborative translation projects.

Perrino (2009: 55–78) reviews a number of tools specifically designed to support online collaborative translation in the web 2.0 scenario, enabling what he calls 'user-generated translation', including Traduwiki,[30] Der Mundo[31] and Cucumis,[32] in addition to WordReference.com (covered in the section 'Online dictionaries and meta-dictionaries'), Proz.com and TranslatorsCafé.com (both discussed in the section 'Online translation marketplaces'). Perrino (2009: 68–70) has a rather negative assessment of the actual usefulness of integrating web-based MT services into these tools for online collaborative translation, even though this trend seems to be continuing in the latest online translation platforms and environments (cf. the sections 'Online platforms to share translation memories' and 'Online translation platforms, environments and tools'). His conclusion is that overall web-based tools and environments supporting distributed online collaborative translation efforts are below the standards that would be expected by those translating user-generated content in the web 2.0 era, thus pointing to the need for improvement in this area.

There have been attempts to efficiently translate and synchronize the multilingual versions of wiki sites by maximizing the impact of online MT and other state-of-the-art natural language processing techniques, e.g. by the European project CoSyne[33] (Gaspari *et al.* 2011: 13–22; Bronner *et al.* 2012: 1–4). Other endeavors have investigated ways in which online MT can support the creation of multilingual content for Wikipedia and other wiki sites by translating entries from English which are then checked and post-edited by users fluent in the target language, e.g. WikiBhasha,[34] a multilingual content creation tool for Wikipedia developed by Microsoft Research from the previous WikiBABEL project[35] (Kumaran *et al.* 2010); the multiple language versions of Wikipedia entries thus created can also provide parallel data to develop domain-focused statistical MT systems.

Crowdsourced online translation projects

DePalma and Kelly (2011: 379–407) address the project management issues faced by four commercial companies that pioneered voluntary community translation, including Facebook, whose crowdsourced translation effort was launched in late 2007. In early 2008 the social networking platform inaugurated its versions in Spanish and German with contents translated for free by volunteer users, and dozens of other languages were added with this crowdsourcing approach by the end of the same year. This early example of grassroots volunteer web localization applied to a major social media platform showed that enthusiastic users were willing to work for free to extend the multilingual accessibility of their online communities. Consistently with Facebook's community-focused ethos, in the case of competing translations for specific strings and textual contents, a voting mechanism was in place to measure the popularity of alternatives and assign points to the most prolific and successful volunteer translators (Ellis 2009: 236–244). Dramatic gains in speed were reported thanks to this crowdsourcing model, compared to the time that it would have taken to localize Facebook in new languages using traditional methods (Kelly *et al.* 2011: 87).

However, some control over translation quality was still retained centrally, as the whole process powered by distributed online volunteer translators was overseen and checked by professionals before Facebook released the new language versions. Using a monolingual comparable corpus methodology, Jiménez-Crespo (2013a: 23–49) shows that the interface textual units of the crowdsourced non-professionally translated version of Facebook in

Peninsular Spanish are more similar to those found in 25 native social networking sites originally produced in Spain, than to other social media platforms that had been professionally localized from English. This indicates the qualitative soundness of the approach adopted by Facebook: the specialist insider knowledge possessed by bilingual users of this social networking environment enabled them to produce localized versions matching the norms of the target language and meeting the expectations of the target user population, even though they were not trained or professional translators.

Crowdsourced user-driven translation projects have also been successfully pursued by other major social media platforms such as LinkedIn and Twitter (Garcia 2010; Fernández Costales 2011). However, professional translators' associations and other interest groups have heavily criticized these attempts at leveraging the expertise and time of enthusiastic bilingual internet users providing a critical mass of online amateur translators, decrying the callousness of profit-making enterprises relying on volunteers to effectively increase their revenues (O'Hagan 2013: 513). McDonough Dolmaya (2011b: 97–110) investigates the ethical issues surrounding voluntary crowdsourced translations for the web, also discussing how these projects affect the perception of translation, with a special focus on minority languages; she finds that the for-profit or not-for-profit status of the organizations behind these crowdsourced translation efforts is not the only consideration in ethical terms, and places emphasis on how these projects are managed and presented to the public (see also Baer 2010).

Conclusion: the future of (online) translation

The technological and social evolution of the internet seems to be an unstoppable process, for example with the development of the semantic web, and the role of online translation is crucial to its growth. More and more users from different countries and with a variety of linguistic and cultural backgrounds obtain online access and become active on web social media and in internet-based projects. While this extends the potential for ubiquitous collaborative translation and large-scale crowdsourcing efforts (European Commission 2012), it also raises thorny issues in terms of professionalism, quality, accountability, public perception, etc. (Sutherlin 2013: 397–409). With these trends set to continue for the foreseeable future, online translation will remain an exciting area for volunteer and professional translators, translation scholars and researchers. Jiménez-Crespo (2013b: 189) predicts 'the expansion of online collaborative translation for all types of translation', also forecasting the future growth of online crowdsourcing services and exchange marketplaces providing free volunteer translations as part of web-based projects. Whilst the current importance of online translation is there for all to see, the ways in which professional and amateur translators will be able to harness its potential and maximize the opportunities that it offers still remain to be discovered.

Notes

1 https://groups.google.com/forum/#!forum/sci.lang.translation.
2 http://segate.sunet.se/cgi-bin/wa?A0=LANTRA-L.
3 www.facebook.com.
4 www.linkedin.com.
5 www.twitter.com.
6 http://dictionary.reference.com, www.onelook.com and www.thefreedictionary.com.
7 www.lexicool.com.
8 www.wordreference.com.
9 http://iate.europa.eu.
10 www.intelliwebsearch.com.

11 http://ipsc.jrc.ec.europa.eu/index.php?id=197.
12 http://ipsc.jrc.ec.europa.eu/?id=782.
13 http://ipsc.jrc.ec.europa.eu/?id=784.
14 http://opus.lingfil.uu.se/index.php.
15 http://mymemory.translated.net.
16 translate.google.com/toolkit.
17 http://translate.google.com.
18 www.memsource.com.
19 www.lionbridge.com/solutions/translation-workspace.
20 www.geoworkz.com.
21 www.lionbridge.com.
22 http://en.trans-aid.jp.
23 www.aquarius.net.
24 www.proz.com.
25 www.translatorscafe.com.
26 www.gotranslators.com.
27 www.langmates.com.
28 www.translationdirectory.com.
29 www.wikipedia.org.
30 http://traduwiki.org.
31 www.dermundo.com.
32 www.cucumis.org.
33 http://cordis.europa.eu/fp7/ict/language-technologies/project-cosyne_en.html.
34 www.wikibhasha.org.
35 http://research.microsoft.com/en-us/projects/wikibabel.

References

Abekawa, Takeshi and Kyo Kageura (2007) 'QRedit: An Integrated Editor System to Support Online Volunteer Translators', in *Proceedings of the Digital Humanities 2007 Conference*, 4–8 June 2007, University of Illinois, Urbana-Champaign, IL, 3–5.

Alcina-Caudet, Amparo (2003) 'Encouraging the Use of E-mail and Mailing Lists among Translation students', *Meta* 48(4): 634–641.

Baer, Naomi (2010) 'Crowdsourcing: Outrage or Opportunity?' *Translorial: Journal of the Northern California Translators Association – Online Edition* February 2010. Available at: http://translorial.com/2010/02/01/crowdsourcing-outrage-or-opportunity.

Ball, Sylvia (2003) 'Joined-up Terminology – The IATE System Enters Production', in *Proceedings of the Translating and the Computer 25 Conference*, 20–21 November 2003, ASLIB, London, UK.

Bey, Youcef, Christian Boitet, and Kyo Kageura (2008) 'BEYTrans: A Wiki-based Environment for Helping Online Volunteer Translators', in Elia Yuste Rodrigo (ed.) *Topics in Language Resources for Translation and Localisation*, Amsterdam and Philadelphia: John Benjamins, 135–150.

Boitet, Christian, Youcef Bey, and Kyo Kageura (2005) 'Main Research Issues in Building Web Services for Mutualized, Non-commercial Translation', in *Proceedings of the 6th Symposium on Natural Language Processing, Human and Computer Processing of Language and Speech (SNLP)*, Chiang Rai, Thailand. Available at: http://panflute.p.u-tokyo.ac.jp/~bey/pdf/SNLP05-BoitetBeyKageura.v5.pdf.

Bronner, Amit, Matteo Negri, Yashar Mehdad, Angela Fahrni, and Christof Monz (2012) 'CoSyne: Synchronizing Multilingual Wiki Content', in *Proceedings of the 8th Annual International Symposium on Wikis and Open Collaboration – WikiSym 2012*, 27–29 August 2012, Linz, Austria, 1–4. Available at: www.wikisym.org/ws2012/Bronner.pdf.

Cronin, Michael (2013) *Translation in the Digital Age*, London: Routledge.

DePalma, Donald A. and Nataly Kelly (2011) 'Project Management for Crowdsourced Translation: How User-translated Content Projects Work in Real Life', in Keiran J. Dunne and Elena S. Dunne (eds) *Translation and Localization Project Management: The Art of the Possible*, Amsterdam and Philadelphia: John Benjamins, 379–407.

Désilets, Alain (2007) 'Translation Wikified: How Will Massive Online Collaboration Impact the World of Translation?' in *Proceedings of the Translating and the Computer 29 Conference*, 29–30 November 2007, ASLIB, London, UK.

Désilets, Alain (2011) *Wanted: Best Practices for Collaborative Translation*, Montreal, Ottawa: Institute for Information Technology, National Research Council Canada. Available at: www.taus.net/downloads/finish/56-public-reports/482-wanted-best-practices-for-collaborative-translation.

Désilets, Alain and Jaap van der Meer (2011) 'Co-creating a Repository of Best-practices for Collaborative Translation', in Minako O'Hagan (ed.) *Linguistica Antverpiensia: Special Issue on Translation as a Social Activity*, 10: 27–45.

Désilets, Alain, Lucas Gonzalez, Sebastien Paquet, and Marta Stojanovic (2006) 'Translation the Wiki Way', in *Proceeding of WikiSym '06 – 2006 International Symposium on Wikis*, 21–23 August 2006, Odense, Denmark, 19–31.

Drugan, Joanna and Bogdan Babych (2010) 'Shared Resources, Shared Values? Ethical Implications of Sharing Translation Resources', in Ventsislav Zhechev (ed.) *Proceedings of the 2nd Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry*, 4 November 2010, Denver, CO, USA, 3–9. Available at: www.mt-archive.info/JEC-2010-Drugan.pdf.

Durán Muñoz, Isabel (2012) 'Meeting Translators' Needs: Translation-oriented Terminological Management and Applications', *The Journal of Specialised Translation* 18: 77–92. Available at: www.jostrans.org/issue18/art_duran.pdf.

Ellis, David (2009) 'A Case Study in Community-driven Translation of a Fast-changing Website', in Nuray Aykin (ed.) *Internationalization, Design and Global Development: Proceedings of the 3rd International Conference IDSG 2009*, 19–24 July 2009, San Diego, CA/Berlin: Springer Verlag, 236–244.

European Union, European Commission, Directorate-General for Translation (2012) *Crowdsourcing Translation: Studies on Translation and Multilingualism*, Luxembourg: Publication Office of the European Union. Available at: http://bookshop.europa.eu/en/crowdsourcing-translation-pbHC3112733/.

Fernández Costales, Alberto (2011) '2.0: Facing the Challenges of the Global Era', in *Proceedings of Tralogy, Session 4 – Tools for translators / Les outils du traducteur*, 3–4 March 2011, Paris, France. Available at: http://lodel.irevues.inist.fr/tralogy/index.php?id=120.

Fernández Costales, Alberto (2012) 'Collaborative Translation Revisited: Exploring the Rationale and the Motivation for Volunteer Translation', *Forum – International Journal of Translation* 10(3): 115–142.

Garcia, Ignacio and Vivian Stevenson (2009) 'Google Translator Toolkit: Free Web-based Translation Memory for the Masses', *Multilingual* (September 2009) 16–19.

Garcia, Ignacio (2010) 'The Proper Place of Professionals (and Non-professionals and Machines) in Web Translation', *Revista Tradumàtica*, December 2010, Issue 08. Available at: www.fti.uab.cat/tradumatica/revista/num8/articles/02/02.pdf.

Gaspari, Federico and W. John Hutchins (2007) 'Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects', in *Proceedings of Machine Translation Summit XI*, 10–14 September 2007, Copenhagen Business School, Copenhagen, Denmark, 199–206. Available at: www.hutchinsweb.me.uk/MTS-2007.pdf.

Gaspari, Federico, Antonio Toral, and Sudip Kumar Naskar (2011) 'User-focused Task-oriented MT Evaluation for Wikis: A Case Study', in Ventsislav Zhechev (ed.) *Proceedings of the 3rd Joint EM+/CNGL Workshop Bringing MT to the User: Research Meets Translators*, 14 October 2011, European Commission, Luxembourg, 13–22. Available at: www.computing.dcu.ie/~atoral/publications/2011_jec_usereval_paper.pdf.

Gough, Joanna (2011) 'An Empirical Study of Professional Translators' Attitudes, Use and Awareness of Web 2.0 Technologies, and Implications for the Adoption of Emerging Technologies and Trends', in Minako O'Hagan (ed.) *Linguistica Antverpiensia: Special Issue on Translation as a Social Activity* 10: 195–217.

Jiménez-Crespo, Miguel A. (2013a) 'Crowdsourcing, Corpus Use, and the Search for Translation Naturalness: A Comparable Corpus Study of Facebook and Non-translated Social Networking Sites', *Translation and Interpreting Studies* 8(1): 23–49.

Jiménez-Crespo, Miguel A. (2013b) *Translation and Web Localization*, London and New York: Routledge.

Kageura, Kyo, Takeshi Abekawa, Masao Utiyama, Miori Sagara, and Eiichiro Sumita (2011) 'Has Translation Gone Online and Collaborative? An Experience from Minna No Hon'yaku', in Minako O'Hagan (ed.) *Linguistica Antverpiensia: Special Issue on Translation as a Social Activity* 10: 47–72.

Kelly, Nataly, Rebecca Ray, and Donald A. DePalma (2011) 'From Crawling to Sprinting: Community Translation Goes Mainstream', in Minako O'Hagan (ed.) *Linguistica Antverpiensia: Special Issue on Translation as a Social Activity* 10: 75–94.

Kumaran, A., Naren Datha, B. Ashok, K. Saravanan, Anil Ande, Ashwani Sharma, Srihar Vedantham, Vidya Natampally, Vikram Dendi, and Sandor Maurice (2010) 'WikiBABEL: A System for Multilingual Wikipedia Content', in *Proceedings of the Collaborative Translation Workshop: Technology, Crowdsourcing,*

and the Translator Perspective at the AMTA 2010 Conference, Association for Machine Translation in the Americas, 31 October 2010, Denver, Colorado, the United States of America. Available at: http://amta2010.amtaweb.org/AMTA/papers/7-01-04-KumaranEtal.pdf.

McDonough, Julie (2007) 'How Do Language Professionals Organize Themselves? An Overview of Translation Networks', *Meta* 52(4): 793–815.

McDonough Dolmaya, Julie (2011a) 'A Window into the Profession: What Translation Blogs Have to Offer Translation Studies', *The Translator: Studies in Intercultural Communication* 17(1): 77–104.

McDonough Dolmaya, Julie (2011b) 'The Ethics of Crowdsourcing', in Minako O'Hagan (ed.) *Linguistica Antverpiensia: Special Issue on Translation as a Social Activity* 10: 97–110.

McDonough Dolmaya, Julie (2012) 'Analyzing the Crowdsourcing Model and Its Impact on Public Perceptions of Translation', *The Translator: Studies in Intercultural Communication* 18(2): 167–191.

O'Hagan, Minako (1996) The Coming Industry of Teletranslation: Overcoming Communication Barriers through Telecommunication, Clevedon: Multilingual Matters.

O'Hagan, Minako (2011) 'Community Translation: Translation as a Social Activity and Its Possible Consequences in the Advent of Web 2.0 and Beyond', in Minako O'Hagan (ed.) *Linguistica Antverpiensia: Special Issue on Translation as a Social Activity* 10: 11–23.

O'Hagan, Minako (2013) 'The Impact of New Technologies on Translation Studies: A Technological Turn?' in Carmen Millán and Francesca Bartrina (eds) *The Routledge Handbook of Translation Studies*, London and New York: Routledge, 503–518.

Perrino, Saverio (2009) 'User-generated Translation: The Future of Translation in a Web 2.0 Environment', *The Journal of Specialised Translation* July 2009: 55–78. Available at: www.jostrans.org/issue12/art_perrino.php.

Plassard, Freddie (2007) 'La traduction face aux nouvelles pratiques en réseaux', *Meta* 52(4): 643–657.

Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter (2012) 'DGT-TM: A Freely Available Translation Memory in 22 Languages', in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 21–27 May 2012, Istanbul, Turkey, 454–459.

Sutherlin, Gwyneth (2013) 'A Voice in the Crowd: Broader Implications for Crowdsourcing Translation during Crisis', *Journal of Information Science* 39(3): 397–409.

Tiedemann, Jörg (2012) 'Parallel Data, Tools and Interfaces in OPUS', in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 21–27 May 2012, Istanbul, Turkey, 2214–2218.

Utiyama, Masao, Takeshi Abekawa, Eiichiro Sumita, and Kyo Kageura (2009) 'Minna no Hon'yaku: A Website for Hosting, Archiving, and Promoting Translations', in *Proceedings of the Translating and the Computer 31 Conference*, 19–20 November 2009, ASLIB, London, the United Kingdom.

Wakabayashi, Judy (2002) 'Induction into the Translation Profession through Internet Mailing Lists for Translators', in Eva Hung (ed.) *Teaching Translation and Interpreting 4: Building Bridges*, Amsterdam and Philadelphia: John Benjamins, 47–58.

Williams, Jenny (2013) *Theories of Translation*, London: Palgrave Macmillan.

Yang, Jin and Elke D. Lange (1998) 'Systran on AltaVista: A User Study on Real-time Machine Translation on the Internet', in David Farwell, Laurie Gerber, and Eduard Hovy (eds) *Proceedings of the 3rd AMTA Conference*, 21–28 October 1998, Langhorne, Philadelphia/New York: Springer, 275–285.

Yang, Jin and Elke D. Lange (2003) 'Going Live on the Internet', in Harold L. Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 191–210.

36
PART-OF-SPEECH TAGGING

Felipe Sánchez-Martínez

UNIVERSITAT D'ALACANT, SPAIN

Introduction

Part-of-speech (PoS) tagging is a well-known problem and a common step in many natural language processing applications such as machine translation, word sense disambiguation, and syntactic parsing. A PoS tagger is a program that attempts to assign the correct PoS tag or lexical category to all words of a given text, typically by relying on the assumption that a word can be assigned a single PoS tag by looking at the PoS tags of neighbouring words.

Usually PoS tags are assigned to words by looking them up in a lexicon, or by using a morphological analyser (Merialdo 1994). A large portion of the words found in a text (around 70 per cent in English) have only one possible PoS, but there are ambiguous words that have more than one possible PoS tag; for example, the English word *book* can be either a noun (*She bought a book for you*) or a verb (*We need to book a room*).

This chapter reviews the main approaches to PoS tagging and the methods they apply to assign PoS tags to unknown words. It also elaborates on the design of the tagset to be used, and reviews two different approaches that have been applied to automatically infer the tagset from corpora. Among the many different applications that PoS tagging has in natural language processing and computational linguistics, it elaborates on the use of PoS taggers in rule-based and statistical machine translation. The chapter ends by providing pointers to free/open-source implementations of PoS taggers.

Approaches to part-of-speech tagging

Different approaches have been applied in order to obtain robust general-purpose PoS taggers to be used in a wide variety of natural language processing and computational linguistic applications; most of these approaches are statistical, but there are also approaches based on the application of rules. This section overviews the main approaches to PoS tagging, some of which are also reviewed by Feldman and Hana (2010), who provide a comprehensive review of the main approaches to PoS tagging with emphasis on the tagging of highly inflected languages.

Hidden Markov Model. Among the different statistical approaches to PoS tagging, hidden Markov model (HMM; Rabiner 1989: 257–286; Baum and Petrie 1966: 1554–1563) is one of the most used. An HMM is a statistical model in which it is assumed that one can make predictions (e.g. assign a PoS tag to a word) based solely on the current (hidden) state, on the

previous states (e.g. PoS tags of previous words) and on the observable output (e.g. word) emitted from the current state. In addition to PoS tagging, HMMs are used for a wide variety of applications such as speech recognition, optical character recognition, and machine translation, just to name a few.

In an HMM, states are not directly visible; only observable outputs generated by the states are visible (see Figure 36.1). Each state has a probability distribution over the possible observable outputs; therefore, the sequence of observable outputs generated by an HMM gives some information about the underlying sequence of hidden states. The parameters of an HMM are the state transition probabilities, i.e. the probability of being in a hidden state at time t given the hidden states at previous times, and the emission probabilities, i.e. the probability of generating an observable output from a hidden state.

When an HMM is used to perform PoS tagging, each HMM state is made to correspond to a different PoS tag, and the observable outputs are made to correspond to word classes, which, in general, may be any suitable partition of the vocabulary; using word classes instead of the words themselves makes it easier to collect reliable statistics. Typically a word class is an *ambiguity class* (Cutting *et al*. 1992: 133–140), i.e. the set of all possible PoS tags that a word could receive. However, some frequent words may be chosen to have their own word classes (ibid.), i.e. a word class holding a single word, to better deal with their distributional peculiarities across texts. Having *lexicalized* states for very frequent words is also possible (Pla and Molina 2004: 167–189; Kim *et al*. 1999: 121–127), in that case the possible PoS tags for the selected words are made to correspond to HMM states that are different to those used for the rest of the words.

The PoS ambiguity is solved by assigning to each word the PoS tag represented by the corresponding state in the sequence of states that maximizes, given a set of HMM parameters previously estimated, the probability of the sequence of word classes observed; this can be efficiently done by means of a dynamic programming algorithm as described by Viterbi (1967: 260–269; Manning and Schütze 1999: 332). The model assumes that the PoS tag of each word depends solely on the PoS tag of the previous n words, and therefore this model is referred to as n-th order HMMs.

The parameters of an HMM can be estimated in a supervised way from hand-tagged corpora via the maximum-likelihood estimate (MLE) method (Gale and Church 1990: 283–287). A hand-tagged corpus is a text in which each PoS ambiguity has been solved by a human expert; such tagged corpora are expensive to obtain, and therefore they may not be available. The MLE method estimates the transition and emission probabilities from frequency counts (e.g. the number of times PoS tag s_i is seen before PoS tag s_j) collected from the hand-tagged corpus.

When hand-tagged corpora are not available, the HMM parameters can be estimated in an unsupervised way by using untagged corpora as input to the Baum-Welch expectation-maximization algorithm (Baum 1972: 1–8; Manning and Schütze, 1999: 333). An untagged

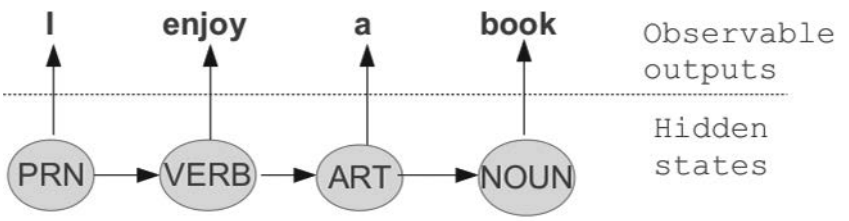


Figure 36.1 Example of state transitions (horizontal arrows) and output emissions (vertical arrows) in a Hidden Markov Model

corpus (Merialdo 1994: 155–171) is a text in which each word has been assigned the set of all possible PoS tags that it could receive independently of the context. Untagged corpora can be automatically obtained if a morphological analyser or a lexicon is available; in an untagged corpus ambiguous words receive more than one PoS tag. Baum-Welch is an iterative algorithm that needs an initial HMM which may be randomly chosen, estimated from untagged corpora (Kupiec 1992: 225–242) or obtained from a small amount of hand-tagged corpora as explained above. Once an initial model is chosen the method works by giving the highest probability to the state transitions and output emissions used the most. In this way a revised model that is more accurate is obtained after each iteration.

Maximum-entropy Model. Maximum-entropy models (Berger *et al*. 1996: 39–71) allow the integration of information coming from different sources in a principled way in the form of features to build classifiers to identify the category (e.g. PoS tag) to which a new instance (e.g. word in context) belongs. Features encode any information that might help the task at hand, and may be of the form '*suffix of word at position j is "ing" and its PoS tag is* VERB-GERUND'. They represent constraints imposed to the model, usually take binary values and get activated when seen in data, i.e. they equal 1 only when the condition they express is true.

Each feature has a weight associated to it that measures its contribution to the probability of an event, such as the event of a word being assigned a certain PoS tag in a given context. Training a maximum entropy model means finding the set of weights that makes the resulting probabilistic model have the maximum entropy of all possible models that satisfy the constraints, which can be shown to be equivalent to maximizing the likelihood of the training data. Having a model that maximizes entropy, i.e. uncertainty, means that no assumptions are made over the training data: the model maximizes the likelihood of the training corpus and at the same time it has the least possible bias towards it, thus making no assumptions about the missing information in the corpus. Maximum-entropy models have been used for word sense disambiguation, text categorization, machine translation, and language modelling, among others.

For maximum-entropy models to be used for PoS tagging one needs to specify a set of features and run a training algorithm, such as generalized iterative scaling (Darroch and Ratcliff 1972: 1470–1480), over a hand-tagged corpus. Ratnaparkhi (1996: 133–142) uses a set of feature templates from which the actual features to be used are generated by scanning all pairs of contexts and PoS tags in the training corpus. These features ask yes/no questions on the context of word at position *j* and constraint the tag in that position to a certain PoS tag, as in the example shown above.

Tagging is performed one sentence at a time. All the candidate PoS tags' sequences of the sentence being tagged are enumerated and the PoS tag sequence with the highest probability is chosen. When enumerating the candidate tag sequences a lexicon providing the allowed PoS tags for each word is used to avoid generating meaningless candidates. The search for the best tag sequence is done by means of a beam search algorithm that works from left to right and maintains, as it sees a new word, a set with the *n* best candidate sequences up to the position of that word. To estimate how promising a candidate PoS tag sequence is, the algorithm uses the conditional probability of a PoS tag given its context. The complexity of this tagging algorithm grows linearly with the product of the sentence length, the number of allowed PoS tags, the average number of features that are activated per event (PoS tag and context) and *N*, the size of the beam.

Support Vector Machines. Support vector machines (SVM, Cristianini and Shawe-Taylor 2000) are a supervised machine learning method for binary classification that learns an hyperplane in an *n*-dimensional space that has the largest possible distance to the nearest *training*

sample, and linearly separates positive samples from negative ones; a training sample is a point of *n* feature values like those used in maximum entropy models. The hyperplane is chosen to have the largest possible distance to the nearest training sample because this has proved to provide a good generalization of the classification bounds. When the training samples are not linearly separable, kernel functions (Shawe-Taylor and Cristianini 2004) that map each sample into a higher dimensional space may be used in the hope that they will be linearly separable in that higher dimensional space. SVM has been used for text classification, hand-writing recognition and word sense disambiguation, just to name a few applications.

PoS tagging is a multi-class classification problem, whereas SVM are useful for binary classification: a binarization of the problem is therefore needed for SVM to be applied for PoS tagging. Giménez and Màrquez (2003: 153–163) apply a simple binarization approach that consists of training a different SVM per PoS tag which discriminates between that tag and all the rest; then, when tagging, they select the PoS tag most confidently predicted among all the possible ones according to the predictions provided by all the binary classifiers. For training they only consider the set of allowed PoS tags provided by a lexicon for each word and consider a word occurrence (training sample) as a positive sample for the tag assigned to it in the training hand-tagged corpus, and as a negative sample for the rest of allowed PoS tags for that word.

Transformation-based Error-driven Learning. Most of the approaches to PoS tagging are statistical mainly because they achieve high performance (tagging error rate on the English WSJ corpus is below 4 per cent) without having to perform any deep analysis of the input text, and because they are easy to train from corpora; thus allowing PoS taggers for the disambiguation of new languages or types of text to be learned when enough corpora are available. However, the knowledge learned by statistical PoS taggers is indirectly coded in large statistical tables which makes it hard, or even impossible, to fix any recurrent tagging error detected after training because it is the result of the combination of different probabilities together.

A rule-based PoS tagger achieving competitive results is described by Brill (1992: 152–155, 1995a: 543–565, 1995b: 1–13) whose transformation-based error-driven learning (TBEDL) approach is capable of acquiring a set of human-readable rules both from a small amount of tagged corpus (Brill 1992: 152–155, 1995a: 543–565) and also from untagged corpora (Brill 1995a: 543–565, 1995b: 1–13); the method may even use untagged corpora to infer an initial set of rules, and then improve the tagging it provides by using a small hand-tagged corpus.

Brill's method works by first using an initial PoS tagger which makes naïve decisions such as selecting for each word the most-frequent PoS tag in a tagged corpus, or no decision at all when no tagged corpus is available, thus providing for each word the set of allowed PoS tags, and then learning patching rules in an iterative process that tries to reduce the tagging errors as much as possible. At each iteration a different rule candidate is evaluated and added to the set of rules learned so far; if the tagging error decreases the rule is eventually added to the final set of rules, otherwise discarded. The tagging error is easily computed over a hand-tagged corpus, or approximated from an untagged corpus by taking advantage of the distribution of non-ambiguous words over the untagged corpus. Each candidate rule is an instance of one of the rule templates provided to the TBEDL algorithm. These templates are of the form '*change tag A to tag B when the preceding (or following) word is tagged Z*', for the supervised learning, and of the form '*change tag (A or B or C) to tag D when the preceding (or following) word is tagged Z*' for the unsupervised learning.

One of the advantages of the TBEDL approach is that one can try as many rule templates as one can devise without affecting tagging performance, because the method only includes in the final set of patching rules those that have proved to improve tagging performance. Another

advantage is that the inferred rules are easy to post-edit, and therefore it is easier than with statistical methods to fix recurrent tagging errors. A possible drawback when learning the rules from untagged corpora is that the resulting PoS tagger may not be capable of resolving the PoS ambiguity in all cases, leaving after the application of the patching rules some words with more than one PoS tag. In those cases the method just picks one, either the first one or a random one. The final set of patching rules can be coded using finite-state transducers (Roche and Schabes 1995: 227–253) which makes the complexity of the resulting PoS tagger linear with the length of the input texts, i.e. faster than any of the statistical taggers described above.

Other approaches. There are other approaches to PoS tagging such as the one by Sánchez-Villamil *et al*. (2004: 454–463) which uses a fixed-width context window of word classes – like those used in HMM-based PoS taggers – around the word to tag. In this approach PoS tagging is performed by evaluating the probability of a given PoS given a fixed number of word classes to the left and to the right, in addition to the word class to which the word to disambiguate belongs. The parameters of the model can be estimated either from a hand-tagged corpus (supervised training), or from an untagged corpus (unsupervised training), and the tagger can directly be implemented as a finite-state machine.

Schmid (1994: 172–176) uses a multilayer perceptron neural network in which each unit in the output layer correspond to one PoS tag, and there is a unit in the input layer per PoS tag t and context word position to take into account; the activation of an input unit represents the probability of the corresponding context word being assigned the PoS tag t; no hidden layers are used. When tagging, the unit in the output layer with the highest activation indicates the PoS tag to assign. Training is supervised: output unit activations are set to zero except for the unit which corresponds to the correct tag, and a modified back-propagation algorithm is used.

Tagging unknown words

In the previous section we assumed that for every word the set of allowed PoS tags was known, but this is not actually the case when we face new texts in which unknown words are likely to appear. Different approaches can be applied for tagging unknown words; the most simple one consists in considering the set of all possible PoS tags as the set of allowed PoS tags for unknown words and let the tagger decide. Alternatively, one can restrict this set to the set of open categories, that is, the set of PoS tags (categories) which are likely to grow by addition of new words to the lexicon: nouns, verbs, adjectives, adverbs and proper nouns. Another option is to run a morphological guesser to reduce the set of allowed PoS tags (Mikheev 1996: 327–333) before running any specific tagging method.

More sophisticated approaches to tagging unknown words by using their lexical aspects get integrated into the specific method being used for tagging. The TnT HMM-based PoS tagger (Brants 2000) uses suffix analysis (Samuelsson 1993: 225–237) to set tag probabilities according to the ending of the words; an approach that seems to provide good results with highly inflected languages. The probability $p(t \mid s)$ of a tag t given a suffix s of n letters is estimated from all the words in the training hand-tagged corpus that share the same suffix; the value of n depends on the length of the word and is chosen to have the largest possible suffix from which evidences are found in the training corpus. These probabilities are then integrated through Bayesian inversion into the HMM as the emission probability $p(s \mid t)$ of an unknown word with suffix s being emitted from HMM state (PoS tag) t.

In the maximum-entropy approach, lexical aspects of the words to be tagged are easily integrated in the form of features to properly deal with unknown words. The feature templates used by Ratnaparkhi (1996: 133–142) differentiate between *rare* and regular words, and assume

that unknown words behave pretty much in the same way as the rare ones with respect to how their spelling helps to predict their PoS tags. Rare words are those that occur less than five times in the training corpus. The feature templates for the rare words take into account the prefix and suffix of the word up to a length of four letters, and whether the word contains numbers, hyphens or capitalized letters. When tagging a new text the features for the rare words are also used to help predict the PoS tag of the unknown words.

Nakagawa *et al.* (2001: 325–331) and Giménez and Màrquez (2003: 209–240) use dedicated SVMs for tagging unknown words by considering the set of open categories as the set of allowed tags for them. These dedicated SVMs make use of features specially designed for tagging unknown words, some of which are similar to the rule templates used by Brill (1995a: 543–565) and described next.

In the transformation-based error-driven learning of rules for PoS tagging discussed above (Brill 1992: 152–155, 1995a: 543–565, 1995b: 1–13), tagging of unknown words is performed in a two-step procedure. The first step assigns a single PoS tag to each unknown word; this PoS tag is a proper noun if the word is capitalized, and a common noun otherwise. The second step learns contextual patching rules to reduce the number of unknown words wrongly tagged; these patching rules differ from those used for the known words because unknown words have a distributional behavior which is quite different from that of the known ones. In addition to the rule templates used for the known words, Brill (1995a: 543–565) uses rule templates that take into account lexical aspects of the unknown word such as prefix, suffix, or if the addition or removal of a suffix (or prefix) from the unknown word results in a known word.

Tagset design

The tagset to be used may differ depending on the natural language processing application in which the PoS tagger will be integrated. While for syntactic parsing it would be sufficient to differentiate a noun from a verb, for machine translation one might need to solve the ambiguity that may occur within the same lexical category, i.e. the ambiguity of words such as the Spanish word *canta* which may refer to the 3rd person of present tense of the verb *cantar* in indicative mood, or to that same verb in the imperative mood, 2nd person. Henceforth, we will refer as fine-grained PoS tags to those PoS tags that convey not only a lexical category, but also inflection information such verb tense, mood, person, number, definitiveness and case.

If fine-grained PoS tags are directly used for disambiguation, the number of parameters, or disambiguation rules, to learn becomes considerably high, making it harder to collect reliable evidences from corpora. For instance, in first-order HMMs, the number of parameters to estimate equals the square of the number of states (PoS tags). To avoid using such a large tagset, fine-grained PoS tags are usually manually grouped into coarse tags by following linguistic guidelines. In doing so one needs to avoid grouping tags having different syntactic roles because this would result in poor tagging results. In this regard it is important to bear in mind that, contrary to what one would expect, the relationship between tagging accuracy and tagset size is weak and is not consistent across language (Elworthy 1995: 1–9). The tagset must be carefully designed, having in mind the task for which the tagger will be used.

Automatic inference of the tagset

The manual definition of the tagset involves a human effort that it would be better to avoid. Moreover, linguistically motivated tagsets do not guarantee better PoS tagging performance because the underlying assumption, namely, that fine-grained PoS tags having the same lexical

category usually have similar probability distributions, does not necessarily hold for all lexical categories. Furthermore, not all the information provided by fine-grained PoS tags is useful for disambiguation, and the amount of information that is useful because it allows discrimination between different analyses may vary from one lexical category to another.

Brants (1995a: 1–10, 1995b: 287–289) automatically infers the grouping of fine-grained PoS tags into coarse ones by applying the HMM model merging method introduced by Stolcke and Omohundro (1994) and Omohundro (1992: 958–965), subject to some restrictions to guarantee that the information provided by the fine-grained PoS tags is preserved, i.e. that even though coarse tags are used, the inflection information provided by the fine-grained tags is not lost. Model merging is an iterative method that starts with an HMM that has as many states as fine-grained PoS tags. Then, in each iteration two states are selected for merging and combined into a single state, updating the transition and emission probabilities accordingly; the states to merge are chosen by using an error measure to compare the goodness of the various candidates for merging. This method assumes supervised training and has the advantage of finding the grouping of the fine-grained PoS tags into coarse ones at the same time as the HMM parameters are estimated. The main drawback is the computational cost of finding the pair of states to merge.

Another way of finding the best grouping of fine-grained PoS tags into coarse ones is to apply the model splitting strategy (Brants 1996: 893–896) which, in contrast to model merging, selects an HMM state to be divided into two new states, updating the transitions and emission probabilities accordingly. The state selected for splitting is the one that maximizes the divergence between the resulting probability distributions after splitting. The exponential growth of the number of possible splittings makes the computation of the global maximum infeasible, forcing the use of heuristics to find a local maximum.

Part-of-speech tagging in machine translation

Machine translation (MT) is one of the natural language processing applications in which PoS taggers are widely used, specially in rule-based MT (Hutchins and Somers 1992) where PoS ambiguities need to be resolved before doing any further analysis of the sentence to translate. The choice of the correct PoS tag may be crucial when translating to another language because the translation of a word may greatly differ depending on its PoS; e.g. the translation into Spanish of the English word *book* may be *libro* or *reservo,* depending on the PoS tag (noun or verb, respectively). However, not all words incorrectly tagged are wrongly translated since some of them may be involved in a *free-ride* phenomenon. A free-ride phenomenon happens when choosing the incorrect interpretation for an ambiguous word in a certain context does not cause a translation error. The more related two languages are, the more often this free-ride phenomenon may occur.

The fact that in MT what really counts is MT quality rather than tagging accuracy – one may not care whether a word is incorrectly tagged at a certain point as long as it gets correctly translated – may be exploited to train PoS taggers specially tuned to translation quality (Sánchez-Martínez *et al.* 2008: 29–66; Sánchez-Martínez 2008). Sánchez-Martínez *et al.* (2008) describe a method to train HMM-based PoS taggers that are specially tuned to translation quality by using information from the source language (as any other training method), and also information from the target language and the rule-based MT system in which the PoS tagger is to be embedded. The method is completely unsupervised and incorporates information from the target language by scoring the translation of each possible disambiguation of the source-language text segments in the untagged training corpus using a statistical model of the target language;[1] these scores are then renormalized and used as fractional counts to estimate the

HMM parameters in the same way the supervised training method does with counts collected from hand-tagged corpora. The resulting PoS tagger allows the MT system in which it is used to perform translations of the same quality as those produced when the PoS tagger is trained in a supervised way. PoS tagging performance is of better quality than that obtained with the unsupervised Baum-Welch expectation-maximization algorithm, but worse than that obtained when supervised training is used.

In statistical MT (Koehn 2010), PoS taggers have been used to reduce the problem caused by the fact that pure statistical MT systems treat inflected forms of the same word (e.g. *book* and *booked*) as if they were different words, which causes translation errors when the amount of parallel corpora available to learn the translation models is scarce or the languages involved in the translation show a rich morphology. PoS taggers have been used, among other things, to annotate each word in the training corpus with its PoS tag before learning *factored translation* models (Koehn and Hoang 2007: 868–876); to better model word reorderings and local agreements; to reorder the source-language side of the training corpus, as well as the source sentences to translate; to better match the word order of the target language (Popović and Ney 2006: 1278−1283); and to help the automatic alignment of words with their translations in the training parallel corpus (Ayan and Dorr 2006: 96–103), a common task when building statistical MT systems.

PoS tagging has also been used for the automatic evaluation of MT to devise linguistically motivated MT quality measures (Giménez and Màrquez 2010: 209–240), for the automatic categorization and analysis of translation errors (Popović and Ney 2011: 657–688) and for translation quality estimation (Felice and Specia 2012: 96–103; Popović 2012: 133–137).

Free/open-source tools for part-of-speech tagging

There are several free/open source PoS taggers freely available on the Internet; this section provides pointers to well-known implementations.

HMM-based PoS Taggers. The FreeLing suite of language analysers (http://nlp.lsi.upc.edu/freeling/; Padró and Stanilovsky 2012: 2473–2479; Carreras *et al.* 2004: 239–242) provides a classical second-order HMM-based PoS tagger.

HunPos (http://code.google.com/p/hunpos/; Halácsy *et al.* 2007: 209–212) implements a second-order HMM-based PoS tagger in which, contrary to what standard HMM implementations do, emission probabilities are based on the current state (PoS tag) and on the previous one.

The PoS tagger used by the Apertium rule-based MT platform (http://www.apertium.org/; Forcada *et al.* 2011: 127–144) implements a first-order HMM-based PoS tagger that can be trained using the method described in the previous section to get PoS taggers specially tuned to translation quality (Sánchez-Martínez *et al.* 2008: 29–66).

Maximum-entropy PoS Tagger. The Stanford *log-linear* part-of-speech tagger (http://nlp.stanford.edu/software/tagger.shtml; Toutanova and Manning 2000: 63–70; Toutanova *et al.* 2003: 252–259) includes, in addition to the baseline features described by Ratnaparkhi (1996: 133–142), features for the disambiguation of the tense forms of verbs, for disambiguating particles from prepositions and adverbs, and a more extensive treatment of capitalization for tagging unknown words.

SVM-based PoS Taggers. SVMTool (http://www.lsi.upc.edu/~nlp/SVMTool/) is a free/open-source implementation of SVM that can be used to train SVM-based PoS taggers (Giménez and Màrquez 2003: 153–163). It supports feature modelling (including lexicalization), and disambiguates thousands of words per second.

Rule-based PoS Tagger. GPoSTTL (http://gposttl.sourceforge.net/) is an enhanced version of the rule-based PoS tagger described by Brill (1992: 152–155, 1995a: 543–565, 1995b: 1–13). The enhancement includes a built-in (English) tokenizer and lemmatizer, and better handling on unknown numerals.

Note

1 A language model measures how likely it is that a given text segment represents a valid construction of the language.

References

Ayan, Necip Fazil and Bonnie J. Dorr (2006) 'A Maximum Entropy Approach to Combining Word Alignments', in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, New York, NY, 96–103.

Baum, Leonard E. and Ted Petrie (1966) 'Statistical Inference for Probabilistic Functions of Finite State Markov Chains', *The Annals of Mathematical Statistics* 37(6): 1554–1563.

Baum, Leonard E. (1972) 'An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process', *Inequalities* 3: 1–8.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra (1996) 'A Maximum Entropy Approach to Natural Language Processing', *Computational Linguistics* 22(1): 39–71.

Brants, Thorsten (1995a) 'Estimating HMM Topologies', in *Tbilisi Symposium on Language, Logic, and Computation*, 19–22 October 1995, Tbilisi, Republic of Georgia, 1–10.

Brants, Thorsten (1995b) 'Tagset Reduction without Information Loss', in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, 287–289.

Brants, Thorsten (1996) 'Estimating Markov Model Structures', in *Proceeding of the 4th International Conference on Spoken Language Processing*, Philadelphia, PA, 893–896.

Brants, Thorsten (2000) 'TnT – A Statistical Part-of-speech Tagger', in *Proceedings of the 6th Applied Natural Language Processing Conference and North American Chapter of the Association of Computational Linguistics Annual Meeting*, Seattle, WA, 224–231.

Brill, Eric (1992) 'A Simple Rule-based Part of Speech Tagger', in *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento, Italy, 152–155.

Brill, Eric (1995a) 'Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging', *Computational Linguistics* 21(4): 543–565.

Brill, Eric (1995b) 'Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging', in *Proceedings of the 3rd Workshop on Very Large Corpora*, Cambridge, MA, 1–13.

Carreras, Xavier, Isaac Chao, Lluis Padró, and Muntsa Padró (2004) 'FreeLing: An Open-source Suite of Language Analyzers', in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 26–28 May 2004, Lisbon, Portugal, 239–242.

Cristianini, Nello and John Shawe-Taylor (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge: Cambridge University Press.

Cutting, Doug, Julian Kupiec, Jan Pedersen, and Penelope Sibun (1992) 'A Practical Part-of-speech Tagger', in *Proceedings of the 3rd Conference on Applied Natural Language Processing*, Trento, Italy, 133–140.

Darroch, J.N. and D. Ratcliff (1972) 'Generalized Iterative Scaling for Log-linear Models', *Annals of Mathematical Statistics* 43: 1470–1480.

Elworthy David (1995) 'Tagset Design and Inflected Languages', in *Proceedings of the ACL SIGDAT Workshop from Texts to Tags: Issues in Multilingual Language Analysis*, 4 April 1995, Dublin, Ireland, 1–9.

Feldman, Anna and Jirka Hana (2010) *A Resource-light Approach to Morpho-syntactic Tagging*, Amsterdam and Atlanta: Rodopi.

Felice, Mariano and Lucia Specia (2012) 'Linguistic Features for Quality Estimation', in *Proceedings of the 7th Workshop on Statistical Machine Translation*, Montréal, Canada, 96–103.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers (2011) 'Apertium: A Free/Open-source Platform for Rule-based Machine Translation', *Machine Translation* 25(2): 127–144.

Gale, William A. and Kenneth W. Church (1990) 'Poor Estimates of Context Are Worse Than None', in *Proceedings of a Workshop on Speech and Natural Language*, Hidden Valley, PA, 283–287.

Giménez, Jesús and Llúis Màrquez (2003) 'Fast and Accurate Part-of-speech Tagging: The SVM Approach Revisited', in *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 153–163.

Giménez, Jesús and Llúis Màrquez (2010) 'Linguistic Measures for Automatic Machine Translation Evaluation', *Machine Translation* 24(3–4): 209–240.

Halácsy, Peter, András Kornai, and Csaba Oravecz (2007) 'HunPos – An Open Source Trigram Tagger', in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, 25–27 June 2007, Prague, Czech Republic, 209–212.

Hutchins, W. John and Harold L. Somers (1992) *An Introduction to Machine Translation*, London: Academic Press.

Kim, Jin-Dong, Sang-Zoo Lee, and Hae-Chang Rim (1999) 'HMM Specialization with Selective Lexicalization', in *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, Madison, WI, 121–127.

Koehn, Philipp and Hieu Hoang (2007) 'Factored Translation Models', in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, 868–876.

Koehn, Philipp (2010) *Statistical Machine Translation*, Cambridge: Cambridge University Press.

Kupiec, Julian (1992) 'Robust Part-of-speech Tagging Using a Hidden Markov Model', *Computer Speech and Language* 6(3): 225–242.

Manning, Christopher D. and Hinrich Schütze (1999) *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Merialdo, Bernard (1994) 'Tagging English Text with a Probabilistic Model', *Computational Linguistics* 20(2): 155–171.

Mikheev, Andrei (1996) 'Unsupervised Learning of Word-category Guessing Rules', in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, New York, NY, 327–333.

Nakagawa, Tetsuji, Taku Kudoh, and Yuji Matsumoto (2001) 'Unknown Word Guessing and Part-of-speech Tagging Using Support Vector Machines', in *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, 325–331.

Omohundro, Stephen M. (1992) 'Best-first Model Merging for Dynamic Learning and Recognition', *Neural Information Processing Systems* 4: 958–965.

Padró, Lluís and Evgeny Stanilovsky (2012) 'FreeLing 3.0: Towards Wider Multilinguality', in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, 2473–2479.

Pla, Ferran and Antonio Molina (2004) 'Improving Part-of-speech Tagging Using Lexicalized HMMs', *Journal of Natural Language Engineering* 10(2): 167–189.

Popović, Maja and Hermann Ney (2006) 'POS-based Reorderings for Statistical Machine Translation', in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 1278–1283.

Popović, Maja and Hermann Ney (2011) 'Towards Automatic Error Analysis of Machine Translation Output', *Computational Linguistics* 37(4): 657–688.

Popović, Maja (2012) 'Morpheme- and POS-based IBM1 Scores and Language Model Scores for Translation Quality Estimation', in *Proceedings of the 7th Workshop on Statistical Machine Translation*, Montréal, Canada, 133–137.

Rabiner, Lawrence R. (1989) 'A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition', in *Proceedings of the IEEE*, 77(2): 257–286.

Ratnaparkhi, Adwait (1996) 'A Maximum Entropy Part-of-speech Tagger', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, 133–142.

Roche, Emmanuel and Yves Schabes (1995) 'Deterministic Part-of-speech Tagging with Finite-state Transducers', *Computational Linguistics* 21(2): 227–253.

Samuelsson, Christer (1993) 'Morphological Tagging Based Entirely on Bayesian Inference', in *Proceedings of the 9th Nordic Conference on Computational Linguistics*, Stockholm, Sweden, 225–237.

Sánchez-Martínez, Felipe, Juan Antonio Pérez-Ortiz, and Mikel L. Forcada (2008) 'Using Target-language Information to Train Part-of-speech Taggers for Machine Translation', *Machine Translation* 22(1–2): 29–66.

Sánchez-Martínez, Felipe (2008) 'Using Unsupervised Corpus-based Methods to Build Rule-based Machine Translation Systems', PhD Thesis, Universitat d'Alacant, Spain.

Sánchez-Villamil, Enrique, Mikel L. Forcada, and Rafael C. Carrasco (2004) 'Unsupervised Training of a Finite-state Sliding-window Part-of-speech Tagger', in *Lecture Notes in Computer Science 3230* (Advances in Natural Language Processing), Alacant, Spain, 454–463.

Schmid, Helmut (1994) 'Part-of-speech Tagging with Neuronal Networks', in *Proceedings of the International Conference on Computational Linguistics*, Kyoto, Japan, 172–176.

Shawe-Taylor, John and Nello Cristianini (2004) *Kernel Methods for Pattern Analysis*, Cambridge: Cambridge University Press.

Stolcke, Andreas and Stephen M. Omohundro (1994) *Best-first Model Merging for Hidden Markov Model Induction*, Technical Report TR-94-003, University of California, Berkeley, CA.

Toutanova, Kristina and Christopher D. Manning (2000) 'Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-speech Tagger', in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, China, 63–70.

Toutanova, Kristina, Dan Klein, Christopher D. Manning, and Yoram Singer (2003) 'Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network', in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Edmonton, Canada, 252–259.

Viterbi, Andrew (1967) 'Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm', in *IEEE Translations on Information Theory*, 13: 260–269.

37
SEGMENTATION

Freddy Y.Y. Choi

UNIVERSITY OF MANCHESTER, THE UNITED KINGDOM

Introduction

The aim of segmentation is to partition a text into topically coherent parts. The result is a structure that resembles the table of contents of a book, where each chapter and section focuses on a specific topic within a story. This technology supports machine translation by limiting the size, scope and context of the input text. It improves translation speed by reducing the size of the input text from a complete story to a series of shorter independent text segments, thus reducing the search space and the number of candidate translations for selection. Text segments can be processed in parallel to boost speed performance in practical applications. The technology also improves translation accuracy by reducing the level of ambiguity (e.g. river *bank*, world *bank*) and the range of references (e.g. he, she, the president) in the input text, thus enabling the translation process to generate the most appropriate and specific output for the local context.

The level of improvement gained from incorporating segmentation into a machine translation process varies according to the nature of the input text. Topically fragmented input texts are common in real world applications; examples include long plain-text emails and documents about multiple topics, closed caption (subtitle) text from television broadcasts, meeting minutes and interview transcripts. These input texts rarely contain any mark up data about the topic boundaries, thus segmentation is applied to separate the continuous text into coherent parts prior to translation. In practical applications, the recommendation is to investigate the availability of pre-partitioned texts from the data provider where possible, rather than relying on automatic text segmentation as the default solution. Common alternative data sources that provide manually segmented data include RSS news feeds and on-demand television services.

Machine translation requires domain and language independent text segmentation, where the algorithm is able to partition texts about any topic in any language. These algorithms rely on a statistical analysis of word frequency, co-occurrence and word occurrence profile to measure cohesion (Halliday and Hasan 1976), and cluster analysis to identify the location of topic boundaries by merging the related parts. This chapter will provide an overview of domain and language independent text segmentation solutions for machine translation.

Background

Linguistic foundation

Cohesion (Halliday and Hasan 1976) is a linguistic phenomenon that is observed in well written and well structured texts. It enables the reader to follow a story from one part to the next, recognizing the change of scene and shift of focus between the different parts. Halliday and Hasan identified five observable linguistic characteristics that contribute towards cohesion: reference, ellipsis, substitution, conjunction and lexical cohesion.

Anaphora and cataphoric references (e.g. '*John* went to England.' … '*He* studied English.') indicate the text fragments are focusing on the same set of actors and objects, thus is likely to be about the same topic. An anaphora reference refers to a previously identified item (e.g. 'John' … 'He'), whereas a cataphoric reference refers to an item to be identified in the following text (e.g. 'The President' … 'Barack Obama').

Ellipsis is the omission of words in a cohesive text to reduce repetition and eliminate easily inferred fragments (e.g. 'What are you having for dinner?' … 'Burgers.', instead of '*I am having* burgers.'). Substitution is the replacement of a word with a more general word (e.g. 'Large shops sell lots of different products.' … 'Small *ones* tend to sell more interesting products.'). Conjunction is a word for specifying the association between sentences and clauses (e.g. 'and', 'however', 'therefore'). Ellipsis, substitution and conjunction give the reader a clear signal that the text fragments are parts of the same topic, as the interpretation of one is dependent on the other.

Lexical cohesion is the repetition of related words in a text (e.g. 'I like *fruits*.' … '*Bananas* are the best.'), thus indicating the text fragments are related to a common topic. This is a more subtle and less well defined phenomenon as the definition of 'related' is unclear. Examples of word relations that suggest cohesion include semantic similarity (e.g. 'Apples' … 'Bananas'), class and subclasses (e.g. 'mammals' … 'human'), synonyms (e.g. 'sofa' … 'couch'), and antonyms (e.g. 'expensive' … 'cheap'). Word relations are the least consistent but also the strongest signal for cohesion. A human reader can easily recognize a topic shift from the text content but the actual relation that signalled the shift may vary between texts.

Practical applications

Text segmentation is a relatively new linguistic research challenge that has gained interest and popularity through the US DARPA funded Topic Detection and Tracking (Wayne 1998) evaluations which have been running annual international competitions since 1998. The aim of the competitions is to accelerate the development of core technologies for enabling access to a growing collection of unstructured free text information.

As a practical example, applying topic detection and tracking technology to RSS news streams from multiple broadcasters (e.g. BBC, ITV and CNN) will enable the human reader to monitor the development of a specific story over time without the need to read all the reports about other irrelevant topics. This is achieved by partitioning the input streams into topically coherent fragments using a text segmentation algorithm and then grouping semantically similar fragments (e.g. about a specific story or theme) using a clustering algorithm, thus making it possible to monitor the development of an individual story over time without the need to read all the irrelevant fragments.

While topic detection and tracking will remain the main motivation for advancing text segmentation algorithms, the technology is becoming a key enabler for many large-scale

information management challenges, including machine translation. Natural language processing algorithms typically work at the character and sentence levels. Tokenization and sentence boundary disambiguation algorithms use minimal local context (e.g. less than 100 characters) to establish the token boundary and the purpose of punctuations. Part of speech tagging, shallow parsing, syntactic parsing and named entity recognition algorithms are applied to individual sentences (e.g. less than 100 tokens) to establish the grammatical structure of a sentence, extract semi-structured information (e.g. date, time, currency) and named references to world objects (e.g. people, places, events). These algorithms all work on small independent text fragments, thus making them scalable in practical applications by processing different parts of a longer text in parallel.

Natural language understanding algorithms, in contrast, tend to work at the document level. Reference resolution algorithms search the whole document to find the most probable actor or object (e.g. *John*, *sofa*) that is being referred to by an anaphoric or cataphoric reference (e.g. *he*, *it*). Summarization algorithms find the most salient information across the whole document to generate a more concise text. Machine translation algorithms find the most appropriate transformation of the whole document to produce the same text in a different language. These algorithms are applied to the whole story and tend to utilize more and more computational resources as the story length grows; scalability is a real concern in practical applications. Topic segmentation makes it possible to partition a text into independent parts for parallel processing, thus providing a linguistically sound basis for enabling divide and conquer.

Evaluation

Problem definition

The aim of text segmentation is to identify the existence of topic shifts in a text and the exact location of the topic boundaries. Given cohesion is a loosely defined linguistic phenomenon, the problem is defined by examples where a collection of texts is manually segmented by human readers (Hearst 1994). The challenge is to develop an automated process that finds all the segment boundaries and their exact locations according to human judgement.

The use of manually annotated example data to characterize a loosely defined linguistic phenomenon is common in linguistic research; however the production of an example data set is labour intensive and expensive, thus providing only a limited set of examples for investigation and testing purposes. An example-driven investigation requires a large data set to generate consistently useful findings that are applicable to real situations, especially when the inter-annotator agreement levels are variable across the example data set.

Given the limitations of manually annotated example data, an alternative approach was created and adopted (Reynar 1998; Allan *et al.* 1998; Choi 2000b: 26–33) to facilitate large scale testing. Rather than using manually annotated example data, an artificial data set is created by concatenating text fragments from different texts, thus ensuring the topic boundaries are well defined and adjacent segments are about different topics. The assumption is that a good segmentation solution must at least perform well on the artificial data set to be considered for application to real data. The use of artificial data makes it possible to conduct large-scale testing of segmentation algorithms under controlled conditions. The key variables for investigation are algorithm sensitivities associated with segment size and algorithm performance associated with document length.

Evaluation metrics

The accuracy of a segmentation algorithm is measured by its ability to accurately identify and locate all the segment boundaries. This is computed by comparing the expected segment boundaries with the algorithm output. Given a manually annotated or artificially generated test data set, where each example document is a sequence of word tokens with known segment boundaries, accuracy is measured by a variant of precision-recall (Hearst 1994; Reynar 1998), edit distance (Ponte and Croft 1997) or cluster membership (Beeferman *et al.* 1999: 177–210).

Precision-recall metrics consider segmentation as a retrieval task where the aim is to search and retrieve all the segment boundaries. A false positive is a segment boundary that did not exist in the reference segmentation. A false negative is a segment boundary that was missed by the algorithm. The basic metric does not consider near misses, i.e. the algorithm finds the boundary but not at the exact location.

Edit distance based metrics aim to remedy the near miss problem by considering segmentation as a transformation task where the aim is to generate the most similar segmentation to the reference example. Similarity is measured by the minimum number of elementary operations (insert, delete) required to transform the output segmentation to the reference segmentation. A perfect result will require no transformation, a near miss will require a few operations and a poor algorithm will require many operations. The metric offers a graded result according to the level of mismatch, taking into account both near misses and complete omissions.

Cluster membership based metrics follow a similar principle by considering segmentation as a clustering task where the aim is to group consecutive text fragments about the same topic together into a sequence of clusters. Accuracy is measured by the number of text fragments that have been placed in the correct and incorrect clusters; more specifically the Beeferman metric (Beeferman *et al.* 1999) considers all possible pairs of text fragments across the document and tests whether they belong to the same or different clusters as defined in the reference segmentation. Once again, the metric offers a graded result that considers near misses; however, the calculation considers all cluster membership tests as equal, thus it can generate unintuitive results. For instance, the separation of distant text fragments (e.g. start and end of a document) should be easier than the separation of adjacent fragments around a topic boundary, but the metric considers both decisions as equally valuable.

None of these earlier metrics were perfect and more recent works (e.g. Kazantseva and Szpakowicz 2011: 284–293) have gravitated towards the WindowDiff metric (Pevzner and Hearst 2002) which combines the concepts from the precision-recall and cluster membership metrics. It scans the text with a fixed size window and compares the number of boundaries in the output segmentation and reference segmentation. The metric penalizes the algorithm whenever the number of boundaries within the window does not match. Although the metric addresses the key issues associated with the previous evaluation approaches (i.e. near misses are ignored by precision-recall metrics; all cluster membership decisions are considered equal by the Beeferman metric), the WindowDiff metric is a parameterized metric (window size), thus making direct comparison of reported results impossible if the researchers used different window sizes in the evaluation (Lamprier *et al.* 2007).

All existing evaluation metrics have different issues and undesirable characteristics. Given the main purpose of an evaluation metric is to facilitate direct comparison of algorithms in literature, the recommendation is to adopt the Beeferman metric as the standard metric for reporting segmentation performance on a common publicly available test data set, e.g. the artificial test set presented in Choi (2000b).

Solution

Architecture

All text segmentation algorithms are based on a common architecture with three key components: normalization, cohesion metric and clustering. As an overview of the end-to-end processing chain that makes up a text segmentation algorithm, an input text is first normalized to provide the clean elementary parts for text analysis (e.g. lower cased paragraph with no punctuation). An elementary part is the smallest unit of text that may be considered as a complete topic segment in an application (e.g. a sentence or a paragraph). The cohesion metric is then applied to the elementary parts to estimate the level of cohesion between the different parts. The estimates are used by the clustering algorithm to determine the optimal segmentation (i.e. each segment is a cluster of elementary parts) that maximizes within cluster cohesion and minimizes intra-cluster cohesion.

The result of the analysis is either a list of topic segments for linear text segmentation, or a tree of segments for hierarchical text segmentation. The former is typically used in real world applications for topic detection and tracking across news streams, or to introduce scalability to natural language understanding applications such as summarization and machine translation. Hierarchical text segmentation provides the topic structure of a text by capturing the intermediate results of the divisive (i.e. top-down) or agglomerative (i.e. bottom-up) clustering algorithm used in linear text segmentation. The topic structure is typically used as one of several linguistic cues in a natural language understanding algorithm (e.g. summarization, table of contents generation).

Accuracy drivers

The accuracy of a text segmentation algorithm is largely determined by (a) the accuracy of the cohesion metric, (b) the level of noise eliminated and introduced by the normalization procedure and (c) the clustering algorithm's ability to establish the optimal segmentation granularity. The cohesion metric is fundamental to text segmentation algorithms, as it is responsible for recognizing the signals (e.g. references, similarity) that suggest two texts are part of the same topic segment. Step improvements to the accuracy of an algorithm are typically achieved by changing the linguistic basis of the cohesion metric and combining multiple linguistics cues in a single metric. Intuitively, the combination of multiple cues should provide the best results, but in practice a metric that combines just the most powerful cues tends to yield the best accuracy and speed performance over a large diverse corpus (i.e. real texts in a practical application). The reason is that no single linguistic cue gives the perfect result, thus the combination of multiple cues will introduce both positive improvements and additional sources of errors to the analysis. This is managed by introducing weighted cues, to ensure only the most useful cues are applied in each context and their contribution to the overall estimate is adjusted according to their accuracy in general. Both weight estimation and context recognition require a large corpus of training data to yield beneficial results; thus the recommendation is to use a simple and minimal combination of linguistic cues to boost accuracy, while ensuring the combination is applicable and stable across a wide range of texts.

The normalization procedure is responsible for filtering the input text for irrelevant signals (e.g. capitalization, punctuation), cleaning the input (e.g. remove invalid characters) and amplifying the relevant signals (e.g. morphological variants of the same word) to support the cohesion metric. Input data cleansing is a necessity in any practical applications for ensuring the

core algorithm components are not exposed to invalid input caused by common errors such as character encoding issues, carriage return and line feed characters on different operating systems and input file size limits. From an accuracy improvement perspective, the signal to noise ratio enhancement function offered by the normalization procedure is the main focus. The challenge is to eliminate all the noise without removing any useful signals. For instance, the removal of closed class words (e.g. determiners such as 'the', 'a', 'some' and prepositions such as 'on', 'over', 'with') is generally a positive improvement as these have little contribution to cohesion across multiple sentences, unless the text is about the usage and distinctions between determiners and prepositions. As for the amplification of relevant signals, the challenge is to avoid introducing errors and noise. For instance, the translation of all word tokens into word stems (e.g. 'run' instead 'running') is generally a good process for highlighting cohesion in a text to support the cohesion metric, unless the story is about the usage and distinctions between the morphological variants of a word. The examples presented here have been selected to highlight the subtle issues in normalization, rather than give concrete examples of specific challenges, as they will differ in every practical application. The recommendation is to test and refine a text segmentation algorithm iteratively by applying the algorithm to a large corpus of actual input texts. In general, the data cleansing function will need to be refined first to enable large-scale testing over a diverse text collection without errors. For accuracy improvements, the refinement of the normalization component is secondary to the cohesion metric, although the two components are closely coupled. Improvements to the cohesion metric tend to offer significant enhancements, whereas adjustments to the normalization component will provide relatively smaller benefits, thus it should be applied once the cohesion metric is stable and all avenues for enhancing the metrics have been exhausted.

The clustering algorithm has both a large and small contribution to the overall accuracy of the text segmentation algorithm. The component is used to translate the cohesion estimates into an optimal segmentation by merging the most cohesive parts and splitting the least cohesive parts, thus it has minimal contribution from a linguistic perspective, as it is simply acting upon the results of the cohesion metric. The function can be performed by a wide range of common clustering algorithms. The challenge here is in knowing when to stop in a specific application (i.e. when to terminate the clustering algorithm to avoid under or over segmentation), to ensure the right level of granularity is being achieved by the segmentation algorithm. What constitutes a complete topic segment frequently varies across applications. For instance, the task is clearly defined for broadcast news, where the aim is to find the series of different news reports about unrelated events; whereas the segmentation of a story book may generate many equally valid results, such as chapters, sections within a chapter, themes and scenes across chapters and dialogues amongst different set of characters in the story. This aspect of the clustering algorithm has a significant and obvious impact on the overall accuracy, especially when it is being tested on a gold standard data set for comparative analysis of algorithm performance. The recommendation for comparative studies is to perform the test with a known number of target segments first (i.e. the clustering algorithm knows how many segments are expected in the test text), thus making it possible to isolate the automatic termination problem and enabling the experimenter to focus on refining the cohesion metric to recognize the segments boundaries.

Normalization

The aim of normalization is to pre-process the input text to ensure the content is valid for processing by the cohesion metric and clustering algorithm, and to enhance the content for assessing cohesion. The processing chain for a typical normalization component comprises the

following processes: input validation, tokenization, surface feature normalization, stemming or morphological analysis, and stop word removal.

Input validation is typically implemented using a series of handcrafted regular expressions for text transformation and filtering. The process conducts basic checks on the input text to ensure the content does not contain any invalid characters (e.g. due to file corruption or encoding errors), and the size of the content is within design limits (i.e. not too long and not too short).

Tokenization in this instance covers both the partition of a string of characters into word tokens (i.e. a series of letters, a series of digits or punctuation) and the partition of the whole document into elementary parts. For most applications, a plain text document will contain line breaks for paragraphs or sentences. These are the elementary parts that can be combined to form topic segments; in other words, the assumption is that a sentence or a paragraph is unlikely to contain more than one topic, thus the segmentation algorithm will start the analysis at this level of granularity. The key challenge in tokenization is recognizing the different uses of punctuation in words (e.g. 'I.B.M.') and structured information (e.g. date, time, currency, reference numbers). This is solved using a combination of handcrafted regular expressions (e.g. for domain specific information such as credit card numbers) and pattern recognition algorithms for sentence boundary disambiguation (e.g. does a full stop mean the end of a sentence, or is it a part of an abbreviation). Tokenization is considered a solved problem as language-independent solutions are now achieving over 99.5 per cent accuracy (Schmid 2000) and even rule-based language specific methods have achieve over 99.7 per cent accuracy since 1994 (Grefenstette and Tapanainen 1994). For the development of text segmentation algorithms, or any complex natural language processing solutions, the recommendation is to use a simple rule-based tokenization solution in the first instance to generate consistent input to the cohesion metric to facilitate analysis and refinement; then replace it with a more sophisticated tokenization solution to boost accuracy once the cohesion metric is settled.

Surface feature normalization simplifies the input text by replacing all the characters with the lower (or upper) case equivalent, removing all punctuation and optionally replacing or removing structured information with keywords (e.g. replacing a specific date with just the keyword '[DATE]'). The aim is to eliminate all the obviously irrelevant content for assessing cohesion (e.g. punctuation marks are unhelpful as they are used throughout the text for other purposes) and to improve string matching for recurrences of the same word (e.g. 'President' at the beginning of a sentence and 'president' in the middle of a sentence are describing the same concept, thus they should be normalized to 'PRESIDENT' to offer basic improvements in recognizing cohesion). The replacement of structured information (e.g. date, time) with keywords (e.g. for describing the information type) is optional as an application may choose to simply remove or ignore these overly specific and infrequently occurring tokens that tend to play a small part in assessing cohesion.

Stemming or morphological analysis aims to replace every word in the document with the word stem (e.g. 'running' becomes 'run') such that semantically similar words are represented by the same string to eliminate false signals for topic shifts (i.e. the use of different words suggests a topic shift). Morphological analysis performs deep analysis of a word (e.g. 'restructuring') to identify its prefixes (i.e. 're-'), suffixes (i.e. '-ing') and stem (i.e. 'structure'). This is a relatively complex and generally language specific process, thus is less desirable for machine translation applications. Stemming (Porter 1980: 130–137) is a rough kind of morphological analysis that was created to support early information retrieval applications in matching user query keywords with the text document contents. The aim here is to generate a reasonable approximation of the stem (e.g. 'running' becomes 'runn') with minimal effort and maximum speed. The process involves using a series of fixed rules to strip out all known

prefixes and suffixes for a language to reveal the word stem. Although the solution is still language dependent, the linguistic resource required to construct a stemmer is simple and widely available for most languages. For the development of text segmentation algorithms, the recommendation is to use an existing stemmer where possible, or simply not to use a stemmer if one is unavailable for the target language, in the initial stages as stemming generally has a positive but small impact (Choi 2000c) on segmentation accuracy.

Stop word removal uses a lookup table or word frequency analysis to eliminate semantically irrelevant word tokens in the input text, such as frequently occurring closed class words (e.g. determiners 'a', 'an', 'the'). The aim is to remove all the obviously unhelpful signals for assessing cohesion. Lookup tables are popular in existing text segmentation algorithms but these are language specific, thus for machine translation applications, the recommendation is to use frequency analysis to filter the most frequently occurring word tokens in a text according to Zipf's law (Zipf 1949).

Cohesion metric

Lexical cohesion can be estimated by a range of linguistic cues. These can be broadly classified into four groups: lexical repetition, semantic similarity, collocation networks and keywords. The first assumes the reoccurrence of the same word token in different text fragments implies the segments are likely to be focusing on the same topic, and a change of vocabulary signals a topic shift. Semantic similarity uses semantically related words such as synonyms, as defined in a dictionary or thesaurus, to form lexical chains across a text for use as evidence of lexical cohesion and repulsion. Collocation networks are similar to semantic similarity in principle, except that related words are discovered by applying statistical analysis to a training corpus to find frequently co-occurring words within small text fragments, thus implying they are related to the same topic. Finally, keyword-based metrics use a collection of handcrafted or inferred words and phrases to detect cohesion and topic shift; for example the words 'however', 'as such', 'furthermore' implies the next sentence is a continuation of the same topic segment, whereas phrases such as 'over to you in the studio', 'and now for the weather' are frequently used in broadcast news to signal the change of topic to the viewer or listener.

Machine translation requires language independent metrics, thus the discussion here will focus on lexical repetition and collocation networks based metrics. Semantic similarity relies on language specific resources. Keyword-based solutions are developed and tuned for a specific language and domain. Even if they offer minor accuracy improvements over language independent metrics in specific contexts, the overhead and additional complexity of introducing a language detection component and the use of multiple context specific models make them unattractive for practical machine translation applications.

Lexical repetition

Lexical repetition based metrics are all based on a statistical analysis of word token distribution in the text (e.g. Hearst 1994; Reynar 1994: 331–333; Heinonen 1998; Choi 2000b: 26–33; Utiyama and Isahara 2001) which is fast (Choi 2000b: 26–33) and as accurate as language specific methods achieving over 95 per cent (Utiyama and Isahara 2001) on large test data sets. Repetition can be represented as lexical chains, vector space model or a language model. A lexical chain in this context is simply a list of word position references for describing the occurrence of a word token in the text. A typical algorithm will start by analysing the positions of each unique token in the text to create a complete chain for each token spanning the entire

text. A filter is then applied to fragment each chain according to distance, e.g. occurrence of the same word just at the beginning and end of the text is unlikely to be a useful linguistic cue. Cohesion between elementary text fragments is then estimated by counting the number of chains that cross the fragment boundaries. One of the key challenges here is in adjusting the filter to apply the right level of fragmentation to the chains, as there is little linguistic basis for setting the distance threshold.

Vector space model representations of lexical repetition consider each elementary text fragment as an unordered collection of word tokens. A text fragment is represented by a vector of word frequencies in the fragment. Each unique word token is a dimension in the vector space. Cohesion between two text fragments is estimated by the cosine metric which measures the angle between the corresponding vectors. In essence, text fragments containing similar words will have similar vectors, thus the angle will be small and they are considered to be cohesive; whereas fragments containing different words will have vectors that point towards very different directions, thus the angle will be large and they are not cohesive. The vector space model is a well-established solution for estimating semantic similarity in information retrieval. A simple enhancement to the basic metric adjusts the vector dimensions with weights to take into account the information value of each word in the text, for instance using the term frequency and inverse document frequency weighting (TF-IDF) scheme (Jones 1972). Another enhancement projects the vector space according to an inferred word similarity matrix that has been generated through statistical analysis of a large corpus, for instance using latent semantic analysis (Choi *et al.* 2001). This is in essence the same as the collocation networks approach. Finally, another avenue for enhancing the metric applies smoothing and scaling to the raw estimates to improve the stability and reliability of the analysis. For instance, minor differences in the similarity estimates are meaningless when they have been calculated from short texts; thus one can only rely on the rough order of magnitude in similarity as the basis for estimating cohesion, rather than the absolute value. This is the theoretical basis of the local ranking scheme (Choi 2000b: 26–33) that delivered significant improvements to segmentation accuracy.

Language model based representations of lexical repetition consider the likelihood of word distributions to detect topic shifts. In general, existing solutions frame the text segmentation problem as one of model optimization, to find the most likely word sequences that belong together according to a probabilistic word distribution model. A word distribution model describes the most likely collection of words that belong to each topic; this is generally derived from the text itself, thus making these solutions language independent. The most probable segmentation for a text is discovered by applying dynamic programming techniques to find the optimal solution. Experiment results (Utiyama and Isahara 2001) have shown the language model based approach delivers at least the same level of accuracy as vector space model approaches.

Collocation networks

Collocation network based metrics (Ponte and Croft 1997; Kaufmann 1999: 591–595; Choi *et al.* 2001; Ferret 2002) are not strictly language independent as they rely on a domain and language specific model of word co-occurrence to detect cohesion. However, the model is created automatically using raw text data, and can thus be applied to any language or domain; hence it is included in this discussion. From a linguistic perspective, collocation networks assume words that co-occur in a small text fragment (e.g. a paragraph) are likely to be about the same topic; thus a model of cohesion can be derived from a training corpus by analysing the word co-occurrence statistics. The result is a word distribution model for building lexical chains, enhancing the basic vector space model and providing prior probabilities for language model based metrics.

Existing solutions are all based on some variation of latent semantic analysis where word co-occurrence statistics are collected from a training corpus to produce a frequency matrix. Eigenanalysis is then applied to the matrix to perform principle component analysis, thus identifying the most discriminating dimensions for estimating similarity and enabling filtering of the matrix to reduce the impact of noisy data and its size to reduce the computational cost of runtime calculations. Experiment results (Choi *et al.* 2001) have shown that collocation networks based metric can outperform vector space model based metrics by about 3 per cent.

Discussion

Lexical repetition has established itself as one of the most reliable linguistic cues for detecting cohesion, especially in language and domain independent applications. The lexical chain approach is linguistically sound in principle, but it does not take into account the strength of each link and there is no reliable linguistic foundation for setting the threshold for fragmenting long chains. The vector space model has consistently delivered good results, especially when enhanced with smoothing techniques to make it more resilient against sparse data issues in short texts, false cues and local variations. Language model based approaches have delivered equally good and sometimes better results than vector space model approaches but the solution is more complex and the practical implementation usually requires more computational resources than vector space models. Collocation networks have been shown to deliver further improvements over vector space models and offer similar level of accuracy as language model based solutions.

For machine translation, the recommendation is to use vector space model based metrics as the default solution given the underlying technology, mathematical foundation and linguistic basis are well established and proven across a wide range of practical large-scale multilingual applications such as internet search engines. Its implementation is simple and computationally inexpensive relative to other methods. More recent works have shown that language model based metrics are emerging as the front runners for consistently delivering the best accuracy under test conditions. Future work in text segmentation should investigate the combination of collocation networks and language models for boosting accuracy, and the use of smoothing techniques to improve stability and resilience of the estimates; especially taking into account the statistical significance and linguistic basis of the estimates to ensure the interpretation is in line with what information is actually available in the source text, i.e. small differences in the cohesion estimates are meaningless when the source text fragments only contains a few words.

Clustering

The clustering component analyses the cohesion estimates to establish the optimal segmentation for a text. Clustering is a well-understood process with many proven solutions. For text segmentation, there are broadly three kinds of clustering solutions:, divisive clustering and agglomerative clustering. Moving window-based solutions (e.g. Hearst 1994) are not strictly clustering as it simply involves analysing the local cohesion estimates and dividing a text when the value drops below a predefined or calculated threshold. The challenge is setting the correct threshold value to prevent over- or under-segmentation, and making the value adapt all the texts in the target collection as cohesion estimates will vary across different documents. A popular approach to setting an adaptive threshold is to take the mean value for all the cohesion estimates for a text and use the standard deviation to compute the threshold (i.e. number of standard deviations from the mean) according to a sample of test examples (Hearst 1994).

Divisive clustering considers all cohesion estimates for the whole text and performs segmentation by dividing the text recursively, selecting the candidate boundary that maximizes the total cohesiveness of the resulting topic segments, and repeating the process until it reaches a termination condition. Overall cohesiveness of a text is computed by the sum of cohesive estimates for all pairs of elementary segments within each topic segment (i.e. attraction within each cluster), and optionally subtracted by the sum of estimates for elementary segments in different topic segments (i.e. repulsion between clusters). The termination condition is generally estimated by a variant of the adaptive threshold scheme produced in Hearst (1994), which in this instance assesses the rate of change in the overall cohesiveness value of the segmentation. Practical implementations of divisive clustering in text segmentation have used the Dot-plotting algorithm (e.g. Reynar 1994; Choi 2000b: 26–33) and Dynamic Programming algorithm (Utiyama and Isahara 2001). The former is generally faster but requires more working memory, whereas the latter is more computationally intensive but more memory efficient.

Agglomerative clustering is similar to divisive clustering, except that it starts bottom up, merging the most cohesive consecutive elementary segments at each step until it reaches the termination condition. Practical implementations of agglomerative clustering in text segmentation have achieved better results than moving window methods (Yaari 1997), but are less accurate than divisive clustering based methods.

As a direct comparison of the three clustering methods, moving window methods only consider the level of cohesion between consecutive elementary segments within a finite window. Agglomerative clustering only considers the level of cohesion between consecutive segments across the whole text. Divisive clustering considers all the estimates across the whole text between all pairs of elementary segments, thus making it possible to find the globally optimal solution, taking into account the local and global variations in cohesion. The recommendation is to use divisive clustering where possible to obtain the best segmentation to support machine translation. Although the method is the most computationally expensive and memory intensive, the resource requirement is relatively small in comparison to the key machine translation processes. The key limitation of divisive clustering is the input text length as the method needs to build a similarity matrix for all pairs of elementary segments. For translation applications that operate on long texts, the recommendation is to either use any available macro level segmentation (e.g. chapters) to partition the story into more manageable blocks for topic segmentation; or apply the moving window based solution with an adaptive threshold set according to maximum segment length to perform a coarse grain segmentation before detailed analysis and topic segmentation with a divisive clustering based algorithm.

Summary

Text segmentation enables machine translation algorithms to operate on long input texts that would otherwise be impractical due to computational constraints. It partitions a text into smaller topically independent segments for parallel processing by multiple instances of the same machine translation algorithm, thus improving throughput while ensuring the individual results are complete and the combined result is linguistically sound.

Machine translation requires language independent linear text segmentation. This implies all the underlying processes with a text segmentation algorithm must be adapted to eliminate the need for language specific resources. The general architecture of a text segmentation algorithm consists of three key components: normalization, cohesion metric and clustering. A conceptual language-independent text segmentation algorithm for machine translation that combines

proven parts of existing solutions will comprise a normalization component that uses regular expression for tokenization (Grefenstette and Tapanainen 1994) and frequency analysis for stop word removal (Zipf 1949); a cohesion metric that uses the vector space model (Reynar 1994; Choi 2000b: 26–33), TF-IDF weighting (Jones 1972: 11–21) and ranking filter (Choi 2000b: 26–33) to estimate the level of cohesion between the elementary segments; and a clustering algorithm that uses the moving window method to partition the text into more manageable sizes (Hearst 1994) and then a divisive clustering algorithm that performs the detailed analysis (Choi 2000b: 26–33).

Bibliography

Allan, James, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang (1998) 'Topic Detection and Tracking Pilot Study Final Report', in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Computer Science Department, Paper 341.

Beeferman, Doug, Adam Berger, and John Lafferty (1997a) 'A Model of Lexical Attraction and Repulsion', in *Proceedings of the 35th Annual Meeting of the ACL*, 373–380.

Beeferman, Doug, Adam Berger, and John Lafferty (1997b) 'Text Segmentation Using Exponential Models', in *Proceedings of EMNLP-2*, 35–46.

Beeferman, Doug, Adam Berger, and John Lafferty (1999) 'Statistical Models for Text Segmentation', in Claire Cardie and Raymond J. Mooney (eds) *Machine Learning: Special Issue on Natural Language Processing* 34(1–3): 177–210.

Cettolo, Mauro and Marcello Federico (2006) 'Text Segmentation Criteria for Statistical Machine Translation', in Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala (eds) *Advances in Natural Language Processing: Proceedings of the 5th International Conference, FinTAL 2006*, Turku, Finland, LNCS 4139, Berlin: Springer Verlag, 664–673.

Choi, Freddy Y.Y. (2000a) 'A Speech Interface for Rapid Reading', in *Proceedings of IEE Colloquium: Speech and Language Processing for Disabled and Elderly People*, April, London, England.

Choi, Freddy Y.Y. (2000b) 'Advances in Domain Independent Linear Text Segmentation', in *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, Association for Computational Linguistics, 26–33.

Choi, Freddy Y.Y. (2000c) 'Content-based Text Navigation', PhD Thesis, Department of Computer Science, University of Manchester, UK.

Choi, Freddy Y.Y., Peter Wiemer-Hastings, and Johanna Moore (2001) 'Latent Semantic Analysis for Text Segmentation', in *Proceedings of the Conference on Empirical Methods in Natural Lanaugage Processing 2001*.

Church, Kenneth W. (1993) 'Char_align: A Program for Aligning Parallel Texts at the Character Level', in *Proceedings of the 31st Annual Meeting of the ACL*.

Church, Kenneth W. and Jonathan I. Helfman (1993) 'Dotplot: A Program for Exploring Self-similarity in Millions of Lines of Text and Code', *The Journal of Computational and Graphical Statistics*.

Eichmann, David, Miguel Ruiz, and Padmini Srinivasan (1999) 'A Cluster-based Approach to Tracking, Detection and Segmentation of Broadcast News', in *Proceedings of the 1999 DARPA Broadcast News Workshop (TDT-2)*.

Ferret, Olivier (2002) 'Using Collocations for Topic Segmentation and Link Detection', in *Proceedings of the 19th International Conference on Computational linguistics Volume 1*, Association for Computational Linguistics.

Grefenstette, Gregory and Pasi Tapanainen (1994) 'What Is a Word, What Is a Sentence? Problems of Tokenization', in *Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX'94)*, July, Budapest, Hungary.

Hajime, Mochizuki, Honda Takeo, and Okumura Manabu (1998) 'Text Segmentation with Multiple Surface Linguistic Cues', in *Proceedings of COLING-ACL'98*, 881–885.

Halliday, Michael and Ruqaiya Hasan (1976) *Cohesion in English*, New York: Longman Group.

Hearst, Marti A. and Christian Plaunt (1993) 'Subtopic Structuring for Full-length Document Access', in *Proceedings of the 16th Annual International ACM/SIGIR Conference*, Pittsburgh, Philadelphia, PA.

Hearst, Marti A. (1994) 'Multi-paragraph Segmentation of Expository Text', in *Proceedings of the ACL'94*.

Heinonen, Oskari (1998) 'Optimal Multi-paragraph Text Segmentation by Dynamic Programming', in *Proceedings of COLING-ACL '98*.

Helfman, Jonathan I. (1996) 'Dotplot Patterns: A Literal Look at Pattern Languages', *Theory and Practice of Object Systems* 2(1): 31–41.

Jones, Karen Sparck (1972) 'A Statistical Interpretation of Term Specificity and Its Application in Retrieval', *Journal of Documentation* 28(1): 11–21.

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown (1998) 'Linear Segmentation and Segment Significance', in *Proceedings of the 6th International Workshop of Very Large Corpora (WVLC-6)*, August, Montreal, Quebec, Canada, 197–205.

Kaufmann, Stefan (1999) 'Cohesion and Collocation: Using Context Vectors in Text Segmentation', in *Proceedings of the 37th Annual Meeting of the Association of for Computational Linguistics (Student Session)*, June 1999, College Park, USA, 591–595.

Kazantseva, Anna and Stan Szpakowicz (2011) 'Linear Text Segmentation Using Affinity Propagation', in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 27–31 July 2011, Edinburgh, Scotland, 284–293.

Kozima, Hideki (1993) 'Text Segmentation Based on Similarity between Words', in *Proceedings of ACL'93*, 22–26 June 1993, Columbus, OH, 286–288.

Kurohashi, Sadao and Makoto Nagao (1994) 'Automatic Detection of Discourse Structure by Checking Surface Information in Sentences', in *Proceedings of COLING '94*, 2: 1123–1127.

Lamprier, Sylvain, Tassadit Amghar, Bernard Levrat, and Frederic Saubion (2007) 'On Evaluation Methodologies for Text Segmentation Algorithms', in *Proceedings of ICTAI '07*, Volume 2.

Litman, Diane J. and Rebecca J. Passonneau (1995) 'Combining Multiple Knowledge Sources for Discourse Segmentation', in *Proceedings of the 33rd Annual Meeting of the ACL*.

Miike, Seiji, Etsuo Itoh, Kenji Ono, and Kazuo Sumita (1994) 'A Full Text Retrieval System with Dynamic Abstract Generation Function', in *Proceedings of SIGIR '94*, Dublin, Ireland, 152–161.

Morris, Jane (1988) *Lexical Cohesion, the Thesaurus, and the Structure of Text*, Technical Report CSRI 219, Computer Systems Research Institute, University of Toronto.

Morris, Jane and Graeme Hirst (1991) 'Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text', *Computational Linguistics* 17: 21–48.

O'Neil, Mark A. and Mia I. Denos (1992) 'Practical Approach to the Stereo-matching of Urban Imagery', *Image and Vision Computing* 10(2): 89–98.

Palmer, David D. and Marti A. Hearst (1994) 'Adaptive Sentence Boundary Disambiguation', in *Proceedings of the 4th Conference on Applied Natural Language Processing*, 13–15 October 1994, Stuttgart, Germany/San Francisco, CA: Morgan Kaufmann, 78–83.

Pevzner, Lev and Marti A. Hearst (2002) 'A Critique and Improvement of an Evaluation Metric for Text Segmentation', *Computational Linguistics* 28(1): 19–36.

Ponte, Jay M. and Bruce W. Croft (1997) 'Text Segmentation by Topic', in *Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries*, University of Massachusetts, Computer Science Technical Report TR97-18.

Porter, M. (1980) 'An Algorithm for Suffix Stripping', *Program* 14(3): 130–137.

Press, William H., Saul A. Teukolsky, William T. Vettering, and Brian P. Flannery (1992) 'Numerical Recipes in C: The Art of Scientific Computing', Cambridge: Cambridge University Press, 2nd edition, 623–628.

Reynar, Jeffrey C. (1994) 'An Automatic Method of Finding Topic Boundaries', in *ACL'94: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (Student session)*, 331–333.

Reynar, Jeffrey C. (1998) 'Topic Segmentation: Algorithms and Applications', PhD thesis, Computer and Information Science, University of Pennsylvania.

Reynar, Jeffrey C. (1999) 'Statistical Models for Topic Segmentation', in *Proceedings of the 37th Annual Meeting of the ACL*, 20–26 June 1999, MD, 357–364.

Reynar, Jeffrey C. and Adwait Ratnaparkhi (1997) 'A Maximum Entropy Approach to Identifying Sentence Boundaries', in *Proceedings of the 5th Conference on Applied NLP*, Washington, DC.

Reynar, Jeffrey C., Breck Baldwin, Christine Doran, Michael Niv, B. Srinivas, and Mark Wasson (1997) 'Eagle: An Extensible Architecture for General Linguistic Engineering', in *Proceedings of RIAO '97*, June 1997, Montreal, Canada.

Schmid, Helmut (2000) *Unsupervised Learning of Period Disambiguation for Tokenisation*, Internal Report, IMS, University of Stuttgart.

Utiyama, Masao and Hitoshi Isahara (2001) 'A Statistical Model for Domain-independent Text Segmentation', in *Association for Computational Linguistics: 39th Annual Meeting and 10th Conference of the*

European Chapter: Workshop Proceedings: Data-driven Machine Translation, 6–11 July 2001, Toulouse, France.

van Rijsbergen, C. J. (1979) *Information Retrieval*, Newton, MA: Butterworth-Heinemann.

Wayne, Charles L. (1998) 'Topic Detection and Tracking (TDT) Overview and Perspective', in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 8–11 February 1998, Lansdowne, VA.

Yaari, Yaakov (1997) 'Segmentation of Expository Texts by Hierarchical Agglomerative Clustering', in *Proceedings of RANLP'97: Recent Advances in Natural Language Processing*, 11–13 September, Tzigov Chark, Bulgaria.

Youmans, Gilbert (1991) 'A New Tool for Discourse Analysis: The Vocabulary-management Profile', *Language* 67(4): 763–789.

Zipf, George K. (1949) *Human Behavior and the Principle of Least Effort*, London: Addison-Wesley.

38
SPEECH TRANSLATION

Lee Tan

THE CHINESE UNIVERSITY OF HONG KONG, HONG KONG, CHINA

What is speech translation?

Speech translation, or more precisely speech-to-speech translation (abbreviated as S2S), is a technology that converts a spoken utterance in one language into a spoken sentence in another language. It enables natural speech communication between two persons who speak different languages. A speech translation system is a computer system equipped with an audio interface which captures what the speaker says and plays back the output speech to the listener. The translation process is realized by specially designed software installed on the computer.

In the simplest case, a speech translation system could work like a speaking dictionary that performs word-by-word translation, without considering the grammatical relation and other linguistic properties of individual words. The technology level of such a system is relatively low and its practical applications are very limited. Today's technology is generally expected to perform whole-sentence translation and deal with continuous speech in a conversational or enquiry setting. The translated output is required to be natural speech with high intelligibility and human-like voice quality.

Applications of speech translation

Communication is fundamental societal behavior. With the trend of globalization, being able to communicate with people who speak different languages has become a basic and required ability for many individuals. Learning to speak a new language takes a lot of time and effort. Human translators may not be available all the time and are often very costly. Computer-based speech translation technology provides a feasible and efficient solution to address many practical needs.

In business communication, a telephone conversation between speakers at remote locations is often needed. A speech translation system can be integrated into the telephony system as an added-value service. This allows for interactive spoken dialogue, which is preferable for effective negotiation, lobbying and decision-making. If the translation system is able to support multiple languages, it will help small enterprises and organizations to develop international connections at lower costs.

Military and security applications have long been one of the major driving forces for the development of speech translation technology. For example, speech translation software was

developed for two-way conversation in Arabic and English to support the United States military operations in the Middle East. It enabled frontline soldiers and medics to communicate efficiently with civilians when human translators were not available. Similar systems are also useful to the United Nation peacekeeping forces, which need to execute missions in different countries.

The invention of the smartphone has made revolutionary changes in our experience with personal communication devices. The smartphone provides an ideal platform for deploying and popularizing speech-to-speech translation systems because of its user-friendliness, high portability, and large customer base. A smartphone with a speech translation function would empower the user with stronger communication ability and a broader range of information sources. For example, when people travel across different countries, they are able to use smartphones to 'speak' naturally to local people and 'listen' to their responses. Users may also use the same speech translation system as a convenient tool to learn to speak another language.

It has become part of our daily life to watch and share online video and audio recordings. Traditionally sharing of media files is for entertainment purpose and within a group of connected friends. Nowadays the applications extend widely to news broadcasting, commercial advertisements and promotions, education and self-learning, and many other areas. With speech translation technology, online spoken documents would be accessible to a much wider audience.

History of speech translation

In the early 1980s, the NEC Corporation developed a concept demonstration system for Spanish–English automated interpretation and later extended it to Japanese and French (http://www.nec.com/en/global/rd/innovative/speech/04.html). This system could handle only 500 words and the processing time for each utterance was as long as several seconds. A number of large-scale research projects on speech translation and related technologies were launched in early 1990. These projects were carried out at universities and research organizations in Japan, Germany and the United States. The major groups include the Advanced Telecommunications Research Institute International (ATR) in Japan, Carnegie Mellon University (CMU) and IBM Research in the United States, and the University of Karlsruhe in Germany. ATR was founded in 1986 to carry out systematic research on speech translation technology. The ATR-ASURA and ATR-MATRIX systems were developed for speech translation in a limited domain between Japanese, German and English (Takezawa *et al.* 1998: 2779–2782). In Germany, the Verbmobil project was a major initiative of speech translation technology development. Funded by the German Federal Ministry for Education and Research and multiple industrial partners, this large-scale project involved hundreds of researchers during 1992–2000 (http://verbmobil.dfki.de/overview-us.html). The Vermobil system was built to support verbal communication in mobile environments and handle spoken dialogues in three domains of discourse, including appointment scheduling, travel planning, and remote PC maintenance. JANUS was another domain-specific system developed at CMU for translating spoken dialogues between English, German, Spanish, Japanese, and Korean (Levin *et al.* 2000: 3–25).

Since the year 2000, research and development of speech translation technology have progressed gradually to deal with real-world scenarios. A wider range of application domains was explored and more realistic speaking style and acoustic conditions were assumed. The NESPOLE! (NEgotiation through SPOken Language in E-commerce) system was designed to allow novice users to make enquiries using English, French or German about winter-sports

possibilities in Italy via a video-conferencing connection (Metze *et al.* 2002: 378–383). In 2004, the European Commission funded a long-term project named TC-STAR, which targeted unconstrained conversational speech domains in English, Spanish and Chinese.

In the past decade, the Defense Advanced Research Projects Agency (DARPA) of the USA launched three influential programmes on technology advancement in speech translation. They are well known by the acronyms of GALE (Global Autonomous Language Exploitation), TRANSTAC (TRANSlation system for TACtical use), and BOLT (Broad Operational Language Translation). These projects were featured by wide international collaboration among academic institutions and industrial research laboratories from the United States and western European countries. Arabic languages were the major focus of the technologies, in order to support the US military operations and national security actions. The TRANSTAC systems were required to be installed on portable devices for tactical use without involving visual display (http://www.dtic.mil/dtic/pdf/success/TRANSTAC20080915.pdf). The MASTOR (Multilingual Automatic Speech to Speech Translator) system developed by IBM Research and the IraqComm system developed by Stanford Research Institute (SRI) were computer software running on laptop computers. These systems were deployed to various US military units to support their operations in Iraq.

Development of speech translation technology for Chinese started in the late 1990s at the National Laboratory of Pattern Recognition (NLPR), Chinese Academy of Sciences (CAS). The first system, named LodeStar, supported Chinese-to-English and Chinese-to-Japanese translation in the travel domain (Zong and Seligman 2005: 113–137). The 29th Summer Olympic Games was held in Beijing, the People's Republic of China. As part of the Digital Olympics initiative, a prototype speech-to-speech translation system was developed to assist foreign tourists in making travel arrangements. The project involved joint efforts from CAS-NLPR, Universität Karlsruhe and Carnegie Mellon University. The system supported Chinese, English and Spanish, and could run on laptop computers and PDAs with wireless connection (Stüker *et al.* 2006: 297–308).

The Asian Speech Translation Advanced Research (A-STAR) Consortium was formed in 2006 by a number of research groups in Asian countries, with the aims of creating infrastructure and standardizing communication protocols in the area of spoken language translation. In July 2009, A-STAR launched the first Asian network-based speech-to-speech translation system. In 2010, A-STAR was expanded to a worldwide organization, namely the Universal Speech Translation Advanced Research (U-STAR) consortium (http://www.ustar-consortium.com). The standardizing procedures for network-based speech-to-speech translations were adopted by ITU-T. The U-STAR consortium currently has 26 participating organizations.

Architecture of a speech translation system

Speech translation is made possible by three component technologies, namely speech recognition, spoken language translation, and speech synthesis. Figure 38.1 shows the basic architecture and operation of a bi-directional speech translation system for a pair of languages A and B. The input utterance spoken in language A is first recognized to produce a textual representation, e.g., an ordered sequence of words, in language A. The text in language A is then translated into an equivalent textual representation in language B, which is used to generate synthesized speech in language B. The same process is followed vice versa.

A speech translation system can be built from independently developed systems that perform speech recognition, machine translation and speech synthesis. These component systems are loosely coupled to operate in a sequential manner. There is no mechanism of information

feedback or error correction between the systems. In this approach, the performance deficiency of a preceding component may greatly affect the subsequent components, and hence degrade the performance of the entire system.

In an integrated approach, the component systems work coherently with each other with a unified goal of achieving optimal end-to-end performance. For example, the speech recognition system may produce more than one possible sentence and allow the machine translation system to choose the most suitable one according to the linguistic constraints of the target language. If there exist a few translation outputs that are equally good in meaning representation, the quality and fluency of the synthesized speech outputs can be used as the basis for selection (Hashimoto *et al.* 2012: 857–866).

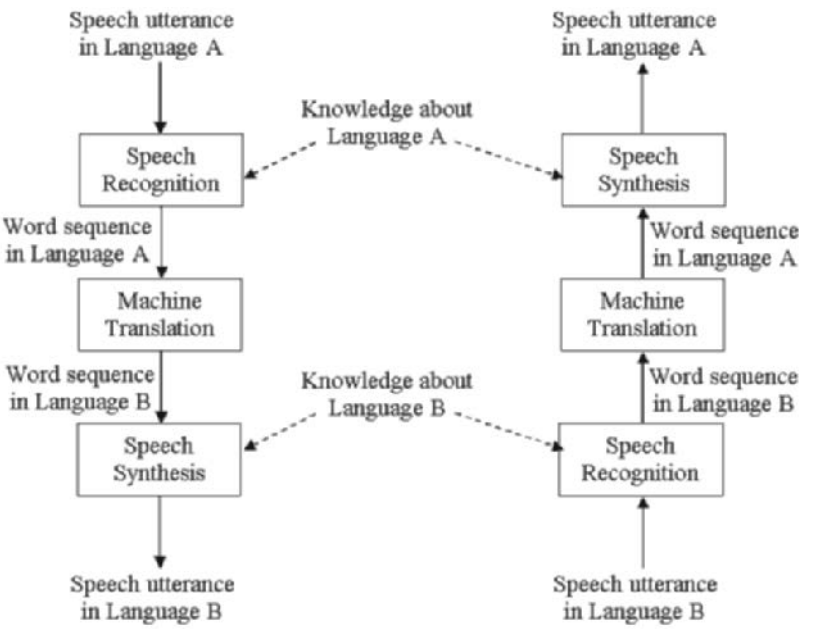


Figure 38.1 Architecture of bi-directional speech translation system

Automatic speech recognition

Automatic speech recognition (ASR) refers to the computational process of converting an acoustic speech signal into a sequence of words in the respective language. Statistical modeling and pattern recognition approach are widely adopted in today's ASR systems. The problem of speech recognition is formulated as a process of searching for the most probable word sequence from a large pool of candidate sequences. A general mathematical formulation is given as

$$\mathbf{W}^{\star} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{O}) ,$$

where $\mathbf{W}$ denotes a word sequence, $\mathbf{O}$ denotes a parameterized representation of the input signal, $P(\mathbf{W} \mid \mathbf{O})$ is known as the posterior probability of $\mathbf{W}$ given the observation $\mathbf{O}$. $\mathbf{W}^{\star}$ is the output of speech recognition, which corresponds to the word sequence with the highest posterior probability. By Bayes' rule, the above equation can be re-written as

$$\mathbf{W}^{\star} = \arg\max_{\mathbf{w}} \; P(\mathbf{O} \mid \mathbf{W})P(\mathbf{W}),$$

where $P(\mathbf{O} \mid \mathbf{W})$ is the probability of observation **O** given that **W** is spoken, and $P(\mathbf{W})$ is the prior probability of **W**. $P(\mathbf{O} \mid \mathbf{W})$ is referred to as the Acoustic Model (AM). It describes that when **W** is spoken, how probable **O** is observed in the produced speech. $P(\mathbf{W})$ is known as the Language Model (LM). It indicates how probable **W** is spoken in the language. AM and LM jointly represent the prior knowledge about the spoken language concerned as well as the intended domain of application. The models are obtained through a process of training, which requires a good amount of speech and text data. The basic architecture of an ASR system is depicted as in Figure 38.2.

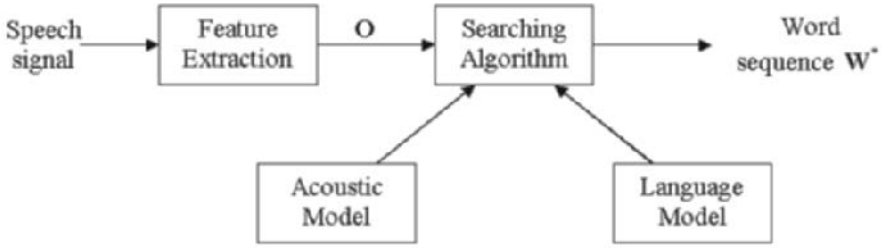


Figure 38.2 Architecture of a speech recognition system

Signal pre-processing and feature extraction

The acoustic signal is picked up by a microphone, which converts the signal into an electrical one. The electrical signal is amplified, sampled and digitized at the audio interface of a computer system. The digitized signal is a sequence of signed integer values that represent the signal amplitudes. There are typically 8,000–16,000 samples in each second of signal, depending on the application. The sample sequence is divided into short-time frames within each of which the signal properties are assumed to be homogeneous. The time advancement between successive frames is 0.01 second and each frame is 0.02 to 0.03 seconds long. In other words, neighboring frames overlap each other. A group of feature parameters is computed from each short-time frame by applying prescribed signal processing procedures. These parameters form a feature vector that consists of 20−40 elements. Thus a digitized speech signal is represented by a temporal sequence of feature vectors, which are used in speech recognition.

Mel-frequency cepstral coefficients, abbreviated as MFCC, are the most commonly used feature parameters. The computation of MFCC features starts with fast Fourier transform (FFT) of the signal samples, which results in a set of spectral coefficients that outline the frequency spectrum of the respective frame. A bank of nonlinearly spaced filters is applied to the log magnitude spectrum. The nonlinearity is designed in the way that the filter-bank simulates human auditory mechanism. The number of filters is 20−30, depending on the signal bandwidth. Subsequently discrete cosine transform (DCT) is applied to the filter-bank output, and the first 13 (low-order) DCT coefficients are used for speech recognition. The complete feature vector in a state-of-the-art ASR system also includes the first-order and second-order time differences between the MFCC parameters of successive frames, which characterize the temporal dynamics of the frame-based features. Other methods of feature extraction for speech recognition include perceptual linear prediction (PLP) and modulation spectrum.

Feature parameters may go through a transformation process, which maps the features into a new space of representation. The transformed feature vectors are desired as they are more effective in representing and discriminating speech sounds, and hence achieve a higher recognition accuracy. Examples of such transformation are the principal component analysis (PCA), linear discriminant analysis (LDA), vocal tract normalization (VTN), and cepstral mean normalization (CMN). Some of these transformations are particularly useful in dealing with a change of speaker or a change of microphone.

Acoustic modeling

Acoustic models are built for a set of speech units that are the basic constituents of the language concerned. The models are used to compute the likelihood of the input utterance with respect to a hypothesized sequence of speech units. In state-of-the-art ASR systems, the acoustic model of a speech unit $\mathbf{W}$ is in the form of a probability distribution function, denoted by $P(\mathbf{O}|\mathbf{W})$. $P(\mathbf{O}|\mathbf{W})$ is obtained via a process of statistical inference that involves a large number of incidences of $\mathbf{W}$. This process is known as model training.

The choice of modeling units is application-dependent. Word-level modeling is adequate and appropriate for small-vocabulary applications, e.g., recognition of digit strings. However, this approach is difficult to scale up to handle thousands of words, due to the scarcity of training data for each word. Sub-word modeling is a practical choice for large-vocabulary continuous speech recognition. Since the same sub-word unit may appear in different words, better sharing of training data can be achieved. This approach also offers the capability of recognizing a word that is not covered in the training data. The definitions of sub-word units are different from one language to another. For English, there are about 40 phonemes to be modeled. For Mandarin Chinese, it is a common practice to model about 60 sub-syllable units, namely Initials and Finals, which are defined based on traditional Chinese phonology. The exact choices and definitions of modeling units are not critical to the performance of a speech recognition system, provided that the chosen units completely cover the anticipated input speech.

Context-dependent phoneme models are commonly used in large-vocabulary systems. A basic phoneme may be represented with multiple models that cater for different phonetic contexts. In the approach of tri-phone modeling, the immediately neighboring phonemes on the left and right of a basic phoneme are taken into account. Other contextual factors that can be considered include stress, lexical tone, and sentential position. Effective clustering methods, e.g., decision tree based clustering, can be used to control the total number of context-dependent units.

Hidden Markov model (HMM) has been widely and successfully applied to acoustic modeling of speech. HMM is an extension of observable Markov process. Each state in an HMM is associated with a probability distribution function, which is typically represented by a Gaussian mixture model (GMM). The acoustic observations, i.e., feature vectors, are regarded as random variables generated by the HMM according to the state-level probability functions. As an example, the English phoneme /o/ can be modeled with an HMM with three states arranged in a time-ordered manner. Each of the states corresponds to a sub-segment of the phoneme. The state transitions reflect the temporal dynamics of speech. The training of HMMs for a large-vocabulary system typically requires hundreds of hours of transcribed speech data.

Other approaches to acoustic modeling for ASR include artificial neural networks (ANN) and segmental trajectory models. ANN is a powerful technique of pattern classification. It can be used to model the state-level probability functions in an HMM. In particular, recent studies showed that deep neural networks (DNN) with many hidden layers could achieve significantly better ASR performance than using GMM.

Lexical and language modeling

Speech is not a random sequence of phonemes and words. It is governed by the linguistic rules of the language. The lexical and language model of an ASR system reflect the system's knowledge of what constitutes a word, how individual words are arranged in order to form a sentence, etc. The lexical model is in the form of a pronunciation dictionary, which tabulates all legitimate words and their pronunciations in terms of a sequence of phonemes. If a word has multiple pronunciations, they will be listed as separate entries in the dictionary. The language model is a statistical characterization that attempts to encode multiple levels of linguistic knowledge: syntax, semantics, and pragmatics of the language. The lexical model can be constructed from published dictionaries of the language. The language model is usually developed via a computational process with a large amount of real-world language data.

N-gram language models are widely used in large-vocabulary continuous speech recognition. Let $W = w_1, w_2,\ldots,w_N$ be a sequence of N words. The probability P(W) can be expressed as

$$P(W)=P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)LP(w_N|w_1,w_2,\ldots,w_{N-1})$$
$$=\prod_{n=1}^{N} P(w_n|w_1,w_2,\ldots,w_{n-1})$$

$P(w_n|w_1,w_2,\ldots,w_{n-1})$ denotes the probability that the word w_n follows the sequence $w_1,w_2,\ldots,w_{n-1}$. It is referred to as the n-gram probability. $P(w_n)$ is called the uni-gram, $P(w_n|w_{n-1})$ the bi-gram, and $P(w_n|w_{n-2},w_{n-1})$ the tri-gram.

N-gram probabilities are estimated by counting word occurrences. For example, the tri-gram probability $P(w_n|w_{n-2},w_{n-1})$ is computed as

$$P(w_n|w_{n-2},w_{n-1})=\frac{C(w_{n-2},w_{n-1},w_n)}{C(w_{n-2},w_{n-1})},$$

where $C(\cdot)$ is the count of occurrences of the word sequence in a given corpus.

The uni-gram language model contains information about how frequently a word is used in the language, which does not help the recognition of a word sequence. Word bi-grams and tri-grams are commonly used in continuous speech recognition because they are able to capture local grammatical properties and are computationally manageable in practical applications. The process of estimating the n-gram probabilities is called language model training. If there are 10,000 words in the vocabulary, the total number of bi-grams and tri-grams will be 10^8 and 10^{12} respectively. Reliable estimation of these probabilities requires a huge amount of training data. For a word combination that does not appear in the training corpus, the respective n-gram probability will be assigned a zero value. This may lead to an undesirable generalization that a sentence containing this word combination will never be recognized correctly. To alleviate this problem, the technique of language model smoothing is applied to make the n-gram probabilities more robust to unseen data. Another way of handling the data sparseness problem is to use class n-grams. A relatively small number of classes are formed by grouping words with similar linguistic functions and grammatical properties. N-gram probabilities are estimated based on word classes instead of individual words.

Search/decoding

The goal of continuous speech recognition is to find the optimal word sequence, which has the highest value of $P(\mathbf{W})P(\mathbf{O}|\mathbf{W})$. This is done via a process of search over a structured space that contains many candidate sequences. The process is also called decoding because it aims at discovering the composition of an unknown signal. The acoustic models and the language models together define the search space, which is represented by a graph with many nodes and arcs. The nodes are HMM states and the arcs include HMM state transitions, cross-phoneme and cross-word transitions. Phoneme-level HMMs are connected to form word-level models according to the lexical model. The lexical model can be represented with a tree structure, in which words having the same partial pronunciation are merged instead of reproduced. With n-gram language models, the end state of each word is linked to many other words with probabilistic transitions. Higher-order language models require a longer word history. This makes the search space expand exponentially.

Each spoken sentence corresponds to a legitimate state-transition path in the search space. The likelihood of the sentence is computed from the state output probabilities, state transition probabilities, and word n-gram probabilities along the path. Exhaustive search over all possible paths is impractical and unnecessary. Many efficient search algorithms have been developed. Examples are the time-synchronous Viterbi search with pruning and the best-first A* stack decoder. The search algorithm is required to find not only a single best answer but also other alternatives that may rank just below the best. This is important because the single best output of speech recognition often contains errors. Since these errors may not be recoverable in the subsequent language translation process, inclusion of a broader range of hypotheses makes the whole system more robust. The most commonly used representations of multiple hypotheses are N-best list, word graph and word lattice.

Machine translation

A machine translation (MT) system performs text translation from one language to another. For speech translation applications, the input text to the MT system is derived from natural speech and the output text from the MT system is used to generate natural speech. The difference between spoken language and written language has to be well understood when applying general machine translation techniques to speech-to-speech translation.

Language translation is a knowledge-based process. A good human translator must have a profound understanding of both the source and the target languages, as well as their similarities and differences. Knowledge-based machine translation starts with linguistic analysis or parsing of the input sentence. The result of parsing is a structured representation, e.g., parse tree, which describes the syntactic relation between individual words in the sentence. A set of transformation rules are applied to change the syntactic structure of the source language to that of the target language. The translation of content words is done with a cross-language dictionary. This approach is referred to as rule-based translation. In the case where the parsing algorithm fails to analyse a sentence, the method of direct translation is used to produce a conservative result by performing word-for-word substitution.

The use of interlingua is an effective approach to domain-specific speech-to-speech translation in a multilingual scenario. Interlingua is a kind of meaning representation that is language-independent. In other words, sentences in different languages but having the same meaning are represented in the same way. Translation is formulated as a process of extracting the meaning of the input text and expressing it in the target language. Interlingua representations

are crafted manually based on both domain knowledge and linguistic knowledge. Since there is a need to determine the exact meaning of an input sentence, interlingua-based approaches require a deeper parsing than rule-based transformation.

Example-based machine translation is an empirical approach that does not require deep linguistic analysis. It is sometimes called or related to corpus-based or memory-based approach. An example-based system is built upon a bilingual corpus of translated examples. The translation is formulated as a process of matching fragments of the input sentence against this corpus to find appropriate examples, and recombining the translated fragments to generate the output sentence (Somers 1999: 113–157). Since examples are used directly in the translation process, the generalizability is limited unless the corpus can cover everything in the language.

Statistical machine translation has become a mainstream approach in the past decade. It leverages the availability of large-scale bilingual parallel corpora and statistical modeling techniques. In a bilingual parallel corpus, each sentence in the source language is aligned with a counterpart in the target language. Typically millions of sentence pairs are required for establishing a meaningful translation model. EUROPARL is one of the well-known parallel corpora for machine translation research (http://www.statmt.org/europarl). It contains a large collection of recordings of the European Parliament meetings in 11 different languages.

Statistical translation follows the same principle and mathematical framework as automatic speech recognition. Let $\mathbf{F}$ denote a string of words (sentence) in the source language. Given $\mathbf{F}$, the conditional probability of a translated word string $\mathbf{E}$ in the target language is denoted as $P(\mathbf{E}|\mathbf{F})$. The goal of translation is to find the optimal choice of $\mathbf{E}$, which has the largest $P(\mathbf{E}|\mathbf{F})$, i.e.,

$$\mathbf{E}^{\star} = \arg\max_{\mathbf{E}} P(\mathbf{F}|\mathbf{E})P(\mathbf{E}),$$

where $P(\mathbf{E})$ is the language model probability in the target language, and $P(\mathbf{F}|\mathbf{E})$ is the translation model probability (Brown *et al.* 1993).

The sentence-level probability $P(\mathbf{F}|\mathbf{E})$ can be computed from word-level probability $P(\mathbf{f}|\mathbf{e})$, in which $\mathbf{f}$ and $\mathbf{e}$ denote a pair of aligned words. Word alignment is a critical step in translation. It is done in the same way as HMM state alignment in speech recognition. To capture the dependencies between words, phrase-level alignment is performed in translation. Phrase is defined as a group of words. In the training of the phrase-level translation model, each phrase in the target sentence needs to be mapped to a phrase in the source sentence. The conditional probability for each pair of phrases is estimated from the training data.

With the translation model and the language model of target language, translation is a process of search for the optimal sequence of words in the target language. Different hypotheses of target sentences are generated in a bottom-up manner. Similar to speech recognition, the techniques of A* search and Viterbi beam search can be applied. However, the search algorithm has to be flexibly designed such that different word orders between the two languages are allowed. In speech recognition, the input feature vector and the corresponding phoneme sequence are aligned in the same temporal order.

Since statistical machine translation uses the same computational framework as speech recognition, an integrated approach can be developed to perform the conversion from speech input in the source language to text output in the target language. Stochastic finite-state transducers (SFTS) are commonly used to implement the integrated search.

Speech synthesis

Speech synthesis refers to the process of generating an audible speech signal to express a given message. Usually the message is in the form of written text, and the process is called text-to-speech (TTS) conversion. A text-to-speech system consists of three modules as shown in Figure 38.3.

The text processing module maps the textual input into a sequence of sound units. The acoustic synthesis module generates a continuous speech signal according to the sound unit sequence. The prosodic control module contributes to improve naturalness of the synthesized speech (Dutoit 1997: 25–36).

Similar to the consideration in speech recognition, the selection of sound units for speech synthesis is a trade-off between generalizability and accuracy. A small number of phoneme-level units are adequate to synthesize speech of arbitrary content. However, they may not be accurate enough to represent the contextual variation in natural speech. Use of word-level or even phrase-level units is effective in capturing local co-articulation effects. This is at the expense of a large number of distinctive units that need to be processed and stored. For general-domain TTS systems, context-dependent phonemic units are most commonly used. There are also systems using variable-length units.

For most languages, phonemic symbols cannot be straightforwardly observed from written text. Grapheme-to-phoneme conversion for TTS involves linguistic analysis at lexical, morphological and syntactic levels. The input sentence is first parsed into a list of words. Chinese written text does not have explicitly marked word boundaries. Thus word segmentation is an important problem in Chinese text processing. A pronunciation lexicon is used to map each word into a phoneme sequence. If a word has alternative pronunciations, the most appropriate one is chosen according to its linguistic context. The pronunciation of proper names, abbreviations, idiomatic expressions, numbers, date and time, etc., cannot be covered by the lexicon, because there are too many variations of them. These items are handled specially by heuristic rules, which are application-dependent.

Prosody refers to the temporal variation of rhythm, pitch, and loudness along a spoken utterance. It plays an important role in human speech communication. Prosodic phenomena in natural speech include focus, stress, accentuation, sentential intonation, pause, and many others. They are realized in acoustic signals through the variation of fundamental frequency (F0), duration, and signal intensity. The prosodic control module specifies the target values of these parameters based on the results of text analysis. Prosodic control in TTS can be rule-based, model-based or corpus-based. Rule-based methods generate target prosody from a set of pre-determined rules that are derived by linguistic observations. Model-based approaches assume that prosody production is governed by an underlying model, usually a parametric one.

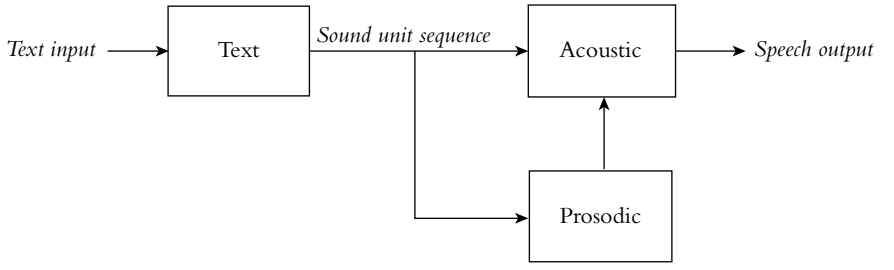


Figure 38.3 Architecture of a text-to-speech system

Corpus-based approaches use a large amount of natural speech data to train a generative prosody model in a statistical sense.

Waveform concatenation is a predominant approach for acoustic synthesis in commercial TTS systems. It is an engineering approach that leverages the availability of low-cost computer storage. An acoustic inventory is designed to cover all basic sound units in the language. For each of these sound units, multiple waveform templates are stored to represent its contextual variations. A continuous speech utterance is produced by selecting appropriate waveform templates from the acoustic inventory and concatenating them in the time domain. The selection of waveform templates can be formulated as an optimization process. The objective is to minimize phonetic and prosodic mismatches, and signal discontinuities that are incurred by concatenation. If the acoustic inventory is well designed and comprehensive in coverage, the concatenated speech would be able to reach a high level of smoothness and naturalness. Further modification on the prosody of concatenated speech utterances can be done using pitch-synchronous overlap and add (PSOLA) technique.

HMM-based speech synthesis has been investigated extensively in recent years. It makes use of a set of context-dependent HMMs that are trained from natural speech data in the same way as in a speech recognition system. For speech synthesis, the HMMs that correspond to the desired phoneme sequence are concatenated. A temporal sequence of spectral and prosodic parameters are generated from the HMM parameters and used to synthesize the output speech. Without the need for storing original speech waveforms, an HMM-based speech synthesis system has a much lower memory requirement than a concatenation system, making it a better choice for portable and personalized applications. An appealing advantage of HMM-based speech synthesis is that it provides a good mechanism for flexible change, modification or customization of voice characteristics. This is done by retraining and adaptation of the HMMs with a relatively small amount of new training data. It is even possible to develop a multilingual system that can speak several different languages using the same voice. This is particularly useful in speech translation.

Compared to speech recognition and language translation, speech synthesis technology is considered to be mature and ready for real-world applications. Multilingual text-to-speech capabilities are now provided in standard computer systems running Windows 8 and iOS.

Examples of speech translation systems

There were many speech-to-speech translation systems developed for research demonstration purposes. Most of them worked in specific application domains. Very few general-purpose systems are available in the commercial market. One of them is the Compadre® product suite of SpeechGear, Inc. It consists of a series of software modules that are designed for different modes of communication. The latest version supports bi-directional translation between English and 40 other languages. The speech recognition engine in the Compadre® modules is the Dragon Naturally Speaking™ provided by Nuance Communications, Inc.™

Recently a number of smartphone Apps for multilingual speech-to-speech translation have become available. This makes the technology more accessible and portable for general users. Jibbigo was initially developed by a start-up company founded by Dr. Alex Waibel, who is one of the pioneer researchers in speech translation technologies. The Jibbigo App supports about 20 languages. The online version is available for free download at Apple's AppStore and Android. The offline version charges per language pair. It can be used without internet connection.

Simutalk is a speech-to-speech translation App developed by ZTspeech in Beijing, China. The technologies in the software are backed up by speech translation research at the Institution

of Automation, Chinese Academy of Sciences. Simutalk supports bi-directional Chinese–English translation with a large vocabulary and requires internet connection. Another multilingual speech-to-speech translation App is named VoiceTra4U. It is developed by the U-STAR consortium. A major feature of VoiceTra4U is that it supports many Asian languages. The software is available for free download at Apple's AppStore.

Further reading

Casacuberta, Francisco, Marcello Federico, Hermann Ney, and Enrique Vidal (2008) 'Recent Efforts in Spoken Language Translation', *IEEE Signal Processing Magazine* 25(3): 80–88.

Nakamura, Satoshi (2009) 'Overcoming the Language Barrier with Speech Translation Technology', *NISTEP Quarterly Review* (31): 35–48.

Weinstein, Clifford J. (2002) 'Speech-to-Speech Translation: Technology and Applications Study', *MIT Lincoln Laboratory Technical Report*.

References

Brown, Peter, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer (1993) 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational linguistics* 19(2): 263–311.

Dutoit, Thierry (1997) 'High-Quality Text-to-Speech Synthesis: An Overview', *Journal of Electrical and Electronic Engineering Australia* 17: 25–36.

Federico, Marcello (2003) 'Evaluation Frameworks for Speech Translation Technologies', in *Proceedings of EUROSPEECH 2003 – INTERSPEECH 2003: 8th European Conference on Speech Communication and Technology*, 1–4 September 2003, Geneva Switzerland, 377–380.

Gao, Yuqing, Bowen Zhou, Ruhi Sarikaya, Mohammed Afify, Hong-Kwang Kuo, Wei-zhong Zhu, Yong-gang Deng, Charles Prosser, Wei Zhang, and Laurent Besacier (2006) 'IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-Speech Translator', in *Proceedings of the Workshop on Medical Speech Translation*, 9 June 2006, New York, 53–56.

Hashimoto, Kei, Junichi Yamagishi, William Bnyrne, Simon King, and Keiichi Tokuda (2012) 'Impacts of Machine Translation and Speech Synthesis on Speech-to-Speech Translation', *Speech Communication* 54(7): 857–866.

http://verbmobil.dfki.de/overview-us.html.

http://www.dtic.mil/dtic/pdf/success/TRANSTAC20080915.pdf.

http://www.nec.com/en/global/rd/innovative/speech/04.html.

http://www.statmt.org/europarl.

http://www.ustar-consortium.com.

Levin, Lori, Alon Lavie, Monika Woszczyna, Donna Gates, Marsal Gavalda, Detlef Koll, and Alex Waibel (2000) 'The Janus-III Translation System: Speech-to-Speech Translation in Multiple Domains', *Machine translation* 15(1–2): 3–25.

Metze, Florion, John McDonough, Hagen Soltau, Chad Langley, Alon Lavie, Tanja Schultz, Alex Waibel, Roldano Cattoni, Gianni Lazzari, and Fabio Pianesi (2002) 'The NESPOLE! Speech-to-Speech Translation System', in Mitchell Marcus (ed.) *Proceedings of the 2nd International Conference on Human Language Technology Research*, 24–27 March 2002, San Diego, CA, 378–383.

Papineni, Kishore, Salim Roukos, Todd Ward, and Zhu Wei-Jing (2002) 'DFHJLBLEU: A Method for Automatic Evaluation of Machine Translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-2002,* 7–12 July 2002, University of Pennsylvania, Philadelphia, PA, 311–318.

Somers, Harold L. (1999) 'Review Article: Example-based Machine Translation', *Machine Translation* 14(2): 113–157.

Stüker, Sebastian, Chengqing Zong, Jürgen Reichert , Wenjie Cao, Guodong Xie, Kay Peterson, Peng Ding, Victoria Arranz, and Alex Waibel (2006) 'Speech-to-Speech Translation Services for the Olympic Games 2008', in Andrei Popescu-Belis, Steve Renals, and Hervé Bourlard (eds) *Machine Learning for Multimodal Interaction: Proceedings of the 4th International Workshop, MLMI 2007*, 28–30 June 2007, Brno, Czech Republic, 297–308.

Takezawa, Toshiyuki, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Lida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto (1998) 'A Japanese-to-English Speech Translation System: ATR-MATRIX', in *Proceedings of the 5th International Conference on Spoken Language Processing, Incorporating the 7th Australian International Speech Science and Technology Conference*, 30 November–4 December 1998, Sydney Convention Centre, Sydney, Australia, 2779–2782.

Zhang, Ying (2003) 'Survey of Current Speech Translation Research'. Available at: http://projectile.is.cs.cmu.edu/research/public/talks/speechTranslation/sst-survey-joy.pdf.

Zong, Chengqing and Mark Seligman (2005) 'Toward Practical Spoken Language Translation', *Machine Translation* 19(2): 113–137.

39

TECHNOLOGICAL STRIDES IN SUBTITLING

Jorge Díaz Cintas

UNIVERSITY COLLEGE LONDON, THE UNITED KINGDOM

The audiovisualization and internetization of communication

The production and exchange of material in which written texts, images and speech are integrated and exploited through visual and auditory channels is an everyday reality in our society, which increasingly relies on these audiovisual programmes for information, entertainment, education, and commerce. This type of format is appealing not only for its alluring semiotic complexity, but also because thanks to today's technology these messages can travel nearly instantly and have the potential to reach large audiences anywhere in the world. Traditionally, the flow of communication was unidirectional, through the cinema and the television, but nowadays it takes the form of bidirectional, dynamic exchanges increasingly through the World Wide Web.

Audiovisual exchanges are appealing because they can communicate complex messages in a ludic way. Their composite, audio and visual nature gives them the edge over written communication and has triggered the audiovisualization of our communicative environment, where sounds and visuals coalesce in a winning combination over other formats, particularly among younger generations. A situation like this is a fertile ground for the blossoming of subtitling, which has grown exponentially in the profession, has gained much deserved visibility in the academe, and has become the international voice of millions of bloggers and netizens.

The catalyst for this moulding of our habits towards greater audiovisual communication can be traced back to cinema in the first instance and television some decades later, though the real impact came about with the start of the digital revolution in the 1980s. Boosted by vast improvements in computing technology, it marked the beginning of the information age and the globalization trends.

The phasing out of analogue technology and the advent of digitization opened up new avenues not only for the production but also for the distribution, commercialization and enjoyment of subtitles. This transition is best symbolized in the death of VHS and the upsurge of the DVD at the end of the last millennium, followed by the switch-off of the analogue signal and the switch-over to digital broadcasting in the early years of the twenty-first century. From a linguistic point of view, the cohabitation of several languages and translations on the very same digital versatile disc has provided consumers with a different viewing experience altogether, allowing them a greater degree of interactivity and more control over the language combination(s) they want to follow.

All these changes have favoured the audiovisualization of translation, which has been taken to new levels thanks to the omnipresent and omni-powerful World Wide Web. Without a doubt, the biggest catalyst of changes in audiovisual communication and translation has been, and continues to be, the internet. Since its launch in the early 1990s, it has known a phenomenal growth and has had an enormous impact on culture, commerce, and education. The potential unleashed by the technology has meant that video material, once too heavy to travel through the ether, can now be transmitted and received with surprising ease virtually anywhere. This, together with the consolidation of Web 2.0 – associated with applications that facilitate participatory and collaborative activities among netizens of virtual communities as well as the production of user-generated content – have made possible that the viewing, exchange and circulation of audiovisual materials is just a keystroke away for nearly everybody. Passive viewers of the first static websites have now become prosumers and bloggers of the cyberspace, with the power of creating and distributing their own material.

This internetization of communication, aimed at reaching commercial success and visibility on a global scale, has found its best ally in subtitling, an easy, economical and fast way of breaking language and sensory barriers by making audiovisual programmes linguistically available and accessible to the rest of the world and potential clients. Indeed, unless this material comes with translations into other languages, it risks capping its potential exposure and its reach across countries and cultures; and without subtitles most videos will be equally inaccessible to audiences with hearing impairments.

Although subtitling took its first steps soon after the invention of cinema over a century ago, its technical evolution was rather slow for many decades focussing primarily in modernizing the various methods of engraving subtitles on the celluloid (Ivarsson and Carroll 1998: 12–19). In more recent decades, the efforts of the technology manufacturers have been directed to the development of powerful software packages specifically designed for subtitling. Yet, and perhaps rather surprisingly when compared with other areas in translation (O'Hagan 2013), little attention has been paid so far to the role that computer-aided translation (CAT) tools can play in subtitling or to the potential that translation memories and machine translation can yield in this field, although the situation is changing rapidly.

When talking about CAT tools, Chan (2013: 1) states that 'the history of translation technology is short, but its development is fast'; an affirmation that in the field of subtitling is particularly apparent. In fact, it could be argued that developments in subtitling are taking place at a faster pace than in any other areas of translation because of, among other reasons, the ubiquitous presence of subtitles in the cyberspace and the magnetism they seem to exert on netizens. Researchers, software developers, subtitling companies and even amateurs are finally paying closer attention to the technical intricacies of this translation practice, as they have realized that subtitling is much more than just adding two lines at the bottom of a film and that technology holds the key for companies (and individuals) to be able to cope with the vast amounts of audiovisual material, both commercial and user-generated, that needs translating.

The audiovisualization of communication is having a great impact not only on the nature of the translation practice – with traditional ones being reassessed (dubbing, voiceover, subtitling) and new ones entering the market (subtitling for the deaf and the hard-of-hearing, audio description for the blind and the partially sighted, audio subtitling) – but also on the working flows of companies, the technology being used, the role of the translator, the nature of the subtitling job, the formal conventions being applied in the subtitles (Díaz Cintas 2010: 105–130), and the multifarious audiovisual genres that get subtitled these days. The following sections concentrate mainly on the technical dimension.

The commoditization and globalization of subtitling

Of the various modes of translating audiovisual programmes, subtitling is arguably the most widely used in commercial and social environments for two main reasons: it is cheap and it can be done fast.

Subtitles are used in all distribution channels – cinema, television, DVD, Blu-ray, internet – both intra and interlingually. The entertainment industry, and increasingly the corporate world, has been quick to take advantage of the potential offered by digital technology to distribute the same audiovisual programme with numerous subtitled tracks in different languages. On the internet, the presence of subtitles has been boosted by the development and distribution of specialist subtitling freeware, which allows fansubbers and amateur subtitlers to create their own translations and distribute them around the globe. On a more domestic note, one of the most symbolic ways in which (intralingual) subtitles have been propelled to the media centre stage has been the inclusion of a subtitle button on most TV remote controls, which takes viewers to the subtitles in an easy and straightforward manner. Legislation in many countries is also having a great impact on the total number of subtitled hours that TV stations must broadcast to satisfy the needs of deaf and hard-of-hearing viewers, with some corporations like the BBC subtitling 100 per cent of their output.

As subtitling projects have become bigger in the number of hours and languages that need to be translated, their budgets have also risen, making the whole operation an attractive field for many old and new companies setting up innovative businesses or expanding the portfolio of services they provide to their clients. In this highly competitive commercial environment, the role of new technologies aimed at boosting productivity is being keenly explored by many of the stakeholders.

The technology turn

As in many other professions, technical advancements have had a profound impact on the subtitling praxis. The profile expected of subtitlers has changed substantially and linguistic competence, cultural awareness and subject knowledge are no longer sufficient to operate effectively and successfully in this profession. Would-be subtitlers are expected to demonstrate high technical know-how and familiarity with increasingly more powerful subtitling software.

The first programs designed exclusively for subtitling started being commercialized in the mid 1970s. At the time, subtitlers needed a computer, an external video player in which to play the VHS tapes with the material to be translated, and a television monitor to watch the audiovisual programmes. The computer would have a word processor with a special subtitling program which made it possible to simulate the subtitles against the images on screen. Some subtitlers would also need a stopwatch to perform a more or less accurate timing of the dialogue.

The situation has changed significantly and these days, with a PC, a digital copy of the video, and a subtitling program, subtitlers can perform all pertinent tasks in front of a single screen; they can watch the video and type their translation, decide the in and out times of each of their subtitles, take due care of shot changes, monitor the reading speed and length of their subtitles, decide on the positioning and colour of the text, spell check their translation, and simulate their subtitles against the images.

The capability and functionality of most professional subtitling programs have been improved at an incredibly fast pace in recent decades, with some of the leading commercial manufacturers being EZTitles (www.eztitles.com), FAB (www.fab-online.com), Miranda Softel (www.miranda.com/softel), Spot (www.spotsoftware.nl), and Screen Systems (www.screensystems.tv), the latter developers of the program WinCAPS (Figure 39.1):

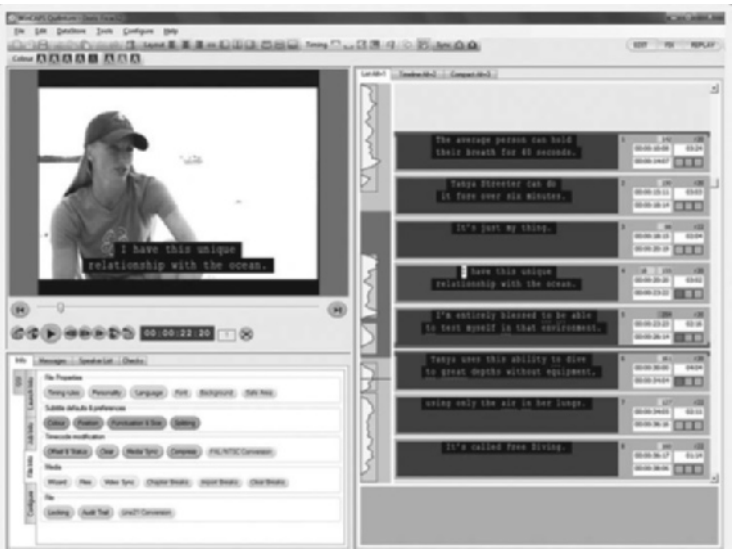


Figure 39.1 Interface of the professional subtitling program WinCAPS Qu4ntum

The fact that professional subtitling software has traditionally been rather expensive and out of reach for many translators has encouraged some to take advantage of the potential offered by technology and come up with their own creative solutions, favouring the development of a vast array of free subtitling programs, of which some of the best known are: Subtitling Workshop (http://subworkshop.sourceforge.net), DivXL and Media Subtitler (www.divxland.org/en/media-subtitler), Aegisub (www.aegisub.org) and Subtitle Edit (www.nikse.dk/SubtitleEdit).

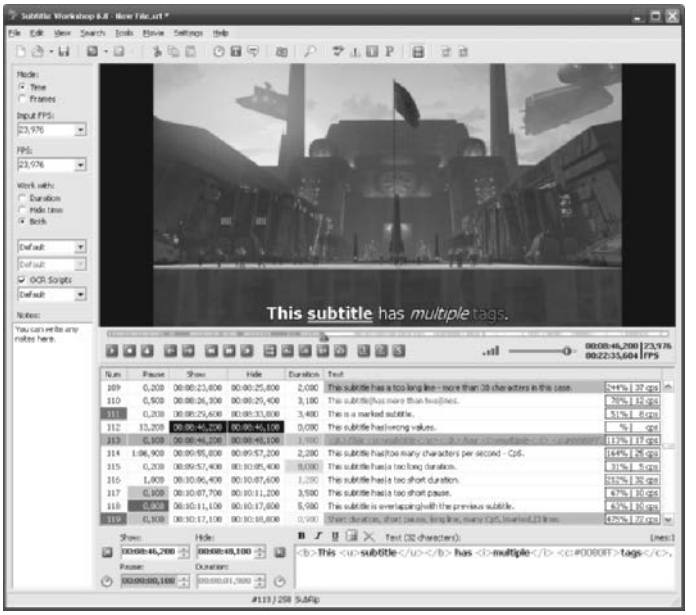


Figure 39.2 Interface of Subtitle Workshop 6.0b, subtitling freeware developed by URUWorks

New software functionality

As time equates to money, professional (but also amateur) subtitling programs are being constantly updated with a view to maximizing productivity and, hence, reducing costs. Improved user interfaces and the automation of certain subtitling tasks, particularly at the technical level, have always been the favoured remit of software engineers, though experiments have been conducted in recent years into the potentiality of automating some steps in the linguistic transfer.

To speed up the spotting process (i.e. the synchronization of the subtitles with the soundtrack) whilst respecting shot changes, some subtitling software applications detect shot changes in the audiovisual program automatically, displaying a timeline in which the video track and the shot change boundaries are shown, thus making it easier and quicker to set the subtitles' in and out times. Another improved feature of most programs is the provision of an audio level indication waveform, whereby changes in soundtrack volume are shown and speech presence can be detected and distinguished from music or background effects. The main benefits of these efficiency tools are twofold. First, subtitlers can skip the scenes with no speech, saving time especially during the final preview or quality check. Second, by assisting them in identifying the timing of speech points, it helps making spotting a lot easier, faster and more accurate.

Technology can further assist subtitlers by simplifying the tasks of text input and timecode synchronization. The automatic timing of subtitles is achieved by means of speech alignment technology: the script or transcript of the dialogue is fed to the subtitling program which, equipped with a speech recognition system, synchronizes it with the soundtrack of the video and assigns it a given timecode, taking account of parameters such as timing rules for shot changes, reading speeds, minimum gaps between subtitles, and minimum and maximum duration of subtitles. If the script contains more textual information than just the dialogue exchanges, the latter can still be imported into the software, with a script extractor that is capable of parsing the script layout to extract dialogue or any information deemed relevant, such as speaker cues. When subtitling for the deaf and the hard-of-hearing (SDH), this information can be used in order to automatically colour the interventions of the different actors, for instance.

In the case of live subtitling, automatic speech recognition (ASR) has been instrumental in the growth of respeaking:

> a technique in which a respeaker listens to the original sound of a live programme or event and respeaks it, including punctuation marks and some specific features for the deaf and hard of hearing audience, to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay.
>
> (Romero-Fresco 2011: 1)

Although this is a new, cost-effective alternative to conventional keyboard-based methods for live subtitling using stenotype or velotype, its very own survival is already being challenged by experiments that look into using speech recognition for subtitling directly from the voice of the TV presenter, thus doing away with the figure of the respeaker.

Reaching for the cloud

When it comes to the production of subtitles, the traditional model of a translation company that commissions a project from professional subtitlers and pays them for their work has ceased to be the only one in existence. In today's global world, viewers are also bound to come across

subtitles that nobody has commissioned or paid for (fansubs) as well as subtitles that organizations have requested from volunteers but not reimbursed (crowdsourced subtitling or crowdsubtitling).

The latter usually refers to collaborative, nonprofit subtitles powered by specific organizations or teams of volunteers. From a technical perspective, they often use applications or platforms built for the specific purpose of this task and which are very easy to learn and use, as is the case of dotSub (dotsub.com) or Amara (www.amara.org), since they usually do not allow the participants to decide the timing of the subtitles and ask them to concentrate on the linguistic transfer. The process of adding subtitles is fast and easy and no software needs to be downloaded or installed. The final output, clips and subtitles, is shared on open websites like TED (www.ted.com), Khan Academy (www.khanacademy.org) or Viki (www.viki.com).

Fansubbers or amateur subtitlers, on the other hand, tend to operate within their own *ad hoc* groups, motivated by their ultimate belief in the free distribution on the net of subtitles made by fans for the consumption of other fans. The first fansubs date from the early 1990s and their exponential rise in recent years has been made possible thanks to the availability of free subtitling and video editing software. If in the early years fansubbers' drive was confined to the popularization of Japanese anime, the reality these days is that most audiovisual programmes, including the latest US films and most popular TV series, find their way into the fansubbing circuit, raising thorny ethical considerations. As opposed to crowdsubtitling, in which both clips and subtitles are distributed with the consent of the interested parties, fansubs are technically illegal as they are not officially licensed and, therefore, infringe the copyright of the owners of the audiovisual programme. On occasions, fansub sites have been closed by copyright enforcement agencies, as in Sweden, and in Poland nine fansubbers were arrested by police in 2007 (Briggs 2013).

Another difference with crowdsubtitling is that fansubbers tend to work with free subtitling programs that they download from the web, such as the ones mentioned earlier in this chapter, whereas crowdsubtitling is usually done through online platforms without the need of downloading any software. In this sense, fansubbing can be said to be closer to professional subtitling since the fansubber tends to be in charge of the technical as well as the linguistic tasks.

A recent trend, cloud subtitling refers to the notion of subtitling on the cloud through collaboration among people based in different geographical locations. On the surface, the only common characteristics it has with fansubbing and crowdsubtitling are their delivery on the internet, the use of teams of subtitlers for different tasks and the relative ease in the preparation of subtitles as opposed to conventional professional subtitling. But, essentially, cloud subtitling adopts a different working model overall and resembles closely the typical chain of subtitling preparation followed by subtitling companies. The final product is no longer considered user-generated content as it is prepared by subtitlers rather than volunteers. It is a solution mostly adopted by translation companies who act as mediators between clients and vendors. The entire subtitling project is managed online, through a cloud-based platform that usually incorporates a project management environment as well as a subtitling editor with a user-friendly application that operates as subtitlers' workspace. One of the advantages of working in this way is that subtitlers can manage their projects without having to buy or download software themselves. What is more, cloud-based subtitling is often provided in different formats, supporting most of the current technologies and devices in the market as well as internet applications, and deliverables are forwarded automatically to clients without any additional effort by subtitlers.

Although a very new development, cloud subtitling has made rapid inroads into the industry, opening new avenues in the provision of subtitles. Among the most prominent examples of

cloud subtitling platforms are ZOOsubs (www.zoosubs.com) by Zoo Digital and iMediaTrans (www.imediatrans.com) by the i-Yuno Media Group. The former was launched in 2012 and currently offers services for subtitling and post-production in more than 40 languages, including fully visible monitoring of the subtitling process, archiving and reviewing of the content, while the final content as well as subtitles files can be converted in several formats and clients have the opportunity to actively participate in the workflow. iMediaTrans also replicates all the tasks involved in the subtitling industry chain, while in-house teams coordinate projects to make sure that the outcome quality within the cloud is of the standard requested by clients. These cloud platforms, including also Übertitles (www.ubertitles.com), tend to work on the basis of automatic alignment of text with audio, whilst still allowing for subtitle editing with options on positioning and use of colours as well as various other technical attributes that can be set by the client or the vendor. These subtitling providers usually select their collaborators online but, unlike crowdsubtitling, they claim to employ professional subtitlers rather than volunteers, and offer clients the possibility of choosing a particular subtitler to take care of their project.

This streamlining of labour management makes cloud subtitling a unique solution for saving time, money and space in the production, editing, post-production and delivery of subtitles. When compared with the practices followed by traditional subtitling companies, a certain degree of harmonization can be detected, with the latter relying more than ever on freelance subtitlers and orders and project management being conducted online. What cloud subtitling notably brings is the potential for closer monitoring on the part of the clients themselves, the possibility of delivering the final product in different formats with greater ease, and the use of cloud-based applications and platforms that lower the cost of subtitling and post-production overall.

Machine translation and subtitling

Whilst developments in the technical dimension of subtitling have been numerous as regards, for example, spotting, shot changes, audio and speaker recognition and automatic colouring of text, the advances on the linguistic front have been much more modest. Although some programs can facilitate text segmentation by automatically dividing the text of a script into subtitles based on linguistic rules that are set up for a specific language, the results can be rather disappointing and the participation of the translator is crucially required.

Translation memory tools, which store previously translated sentences and allow the user to retrieve them as a base for a new translation, have had a great impact in translation, particularly in the fields of specialized and technical translation. However, their worth has been called into question in the case of subtitling because of the fictional and literary nature of audiovisual programmes. Though this might have been true in the past, when most of the materials being subtitled belonged to the entertainment genre, the situation is rapidly evolving. The fact that companies and institutions involved in selling, marketing, education and science, to name but a few areas, are discovering the virtues of communicating audiovisually, mainly through the internet, is clearly bringing changes to this state of affairs. DVD bonus material, scientific and technical documentaries, edutainment programmes, and corporate videos tend to contain the high level of lexical repetition that makes it worthwhile for translation companies to employ assisted translation and memory tools in the subtitling process.

As one of the pioneers in this area, the Taiwanese company Webtrans Digital (www.webtrans.com.tw) has been working with a computer assisted tool called Wados for many years now, and claim that it enhances their efficiency and subtitling consistency.

A step further from computer-assisted translation in the form of memory tools is machine translation (MT). Subtitling has only recently been recognized as an area that could benefit from the introduction of statistical machine translation (SMT) technology to increase translator productivity (Etchegoyhen *et al.* 2013). Two of the first funded research projects to look into its feasibility were MUSA (MUltilingual Subtitling of multimediA content, http://sifnos.ilsp.gr/musa) and eTITLE (Web 1). MUSA ran from 2002 until 2004 and had English, French and Greek as the working languages. The team's seemingly straightforward but highly ambitious goal was (1) to create a multimedia system that would convert the audio stream of audiovisual programmes into text transcriptions with the help of a speech recognition system; (2) to use this output to condense the sentences into subtitles in the same language, by performing an automatic analysis of the linguistic structure of the sentence; and finally (3) to translate automatically these first generation subtitles into other languages by combining a machine translation engine with a translation memory and a term substitution module.

Despite such high hopes, no tangible results ever materialized from either of these two projects. One of the downfalls of such a utopian approach was the over-reliance of the projects on technology that at that point was not developed enough, the other one being the lack at the time of professional quality parallel subtitle data, without which it is difficult to adequately train SMT systems for the creation of subtitles.

More recently, the European Commission funded the project SUMAT, an online service for SUbtitling by MAchine Translation, under its Information and Communication Technologies Policy Support Programme (Web 2). Run by a consortium of four subtitling companies and five technical partners from 2011 until 2014, one of its aims was to use the archives of subtitle files owned by the subtitling companies in the consortium to build a large corpus of aligned subtitles and use this corpus to train SMT systems in various language pairs. Its ultimate objective was to benefit from the introduction of SMT in the field of subtitling, followed by human post-editing in order to increase the productivity of subtitle translation procedures, reduce costs and turnaround times while keeping a watchful eye on the quality of the translation results. To this end, the consortium has built a cloud-based service for the MT of subtitles in nine languages and seven bidirectional language pairs. The service offers users, from individual freelancers to multinational companies, the ability to upload subtitle files in a number of industry-standard subtitle formats as well as in plain text format and to download a machine translated file in the same format, preserving all original time codes and formatting information where possible (Georgakopoulou and Bywood, 2014).

Although the switch from rule-based approaches to statistical translation methods has the potential to improve the accuracy of the translation output, the reality is that no current system provides the holy grail of fully automatic high-quality MT. Indeed, as foregrounded by Hunter (2010: online):

> There is scope for machine translation technology to be used in the creation of translated subtitle files, but as this is not yet a perfect science, there is a fine line between the time taken to check and edit automated content and the time taken to translate each subtitle in turn.

In the toolbox of automatic translation undertaken within the context of subtitling, TranslateTV™ (www.translatetv.com) has been translating English closed captions into Spanish subtitles in real time as a commercial venture in the USA since 2003. Taking advantage of the high volume of intralingual subtitles (English into English) for the deaf and the hard-of-hearing being done in the USA, Vox Frontera, Inc. offers an automatic translation service of

those subtitles into Spanish, aimed primarily at the Hispanic and Latino community, who see and hear exactly what English-speaking viewers see and hear with the only difference of the added real-time Spanish subtitles.

A bolder approach in the automation of subtitling has been taken by Google and YouTube. In an attempt to boost accessibility to audiovisual programmes, primarily to people with hearing impairments, they introduced in 2006 a new feature allowing the playback of captions and subtitles (Harrenstien 2006). In 2009, they announced the launch of machine-generated automatic captions, with the firm belief that "captions not only help the deaf and hearing impaired, but with machine translation, they also enable people around the world to access video content in any of 51 languages" (Harrenstien 2009: online). Their philosophy is summarized in the following quote:

> Twenty hours of video is uploaded to YouTube every minute. Making some of these videos more accessible to people who have hearing disabilities or who speak different languages, not only represents a significant advancement in the democratization of information, it can also help foster greater collaboration and understanding.
>
> (YouTube 2010)

Automatic captioning, based on Google's automatic speech recognition technology and YouTube caption system, is only available for user-generated videos where English is spoken (ibid.). For the system to work best a clearly spoken audio track is essential and videos with background noise or a muffled voice cannot be auto-captioned. The video owner can download the auto-generated captions, improve them, and upload the new version; and all viewers are offered the option to translate those captions into a different language by means of machine-translated subtitles (Cutts 2009), with various degrees of success.[1]

The second subtitling feature launched by the two internet giants allows for a higher degree of accuracy in the linguistic make-up of the captions. Called *automatic timing*, it permits video owners to add manually created captions to their videos by automatically cueing the words uttered in the video. All the user needs is a transcript of the dialogue and, using speech-to-text technology, Google does the rest, matching the words with the time when they are said in the audio and chunking the text into subtitles. The owner of the video can download the timecoded subtitles to modify or to use somewhere else, and the subtitles can also be automatically translated into other languages. As pointed out by Lambourne (2011: 37), 'Look at Google AutoCaps. Submit your media file and see it create automatic captions. The quality and accuracy varies from the sublime to the ridiculous but if you're deaf you may not be able to determine which is which.'

Other developments

Assistive technology and audiovisual translation have started to combine as a successful tandem to foster access services in online education, with the aim of making educational material on the web accessible for people with sensory impairments (Patiniotaki 2013). With regard to live distribution on web-based media – broadcasts, webinars or web-supported conferences – one of the upcoming needs is that of real-time captioning, both for audiences with hearing impairments but also for interlingual transfer. The use of speech recognition and speech-to-text services has been explored by research groups which have developed their own platforms, like eScribe (www.escribe.cz) or Legion Scribe (www.cs.rochester.edu/hci/pastprojects.php?proj=scb), both relying on crowdsourcing human transcribers (Bumbalek *et al.* 2012).

Automatic speech recognition (ASR) is also being tested as a potential solution, though the quality of real-time captions created in this way is still problematic.

Other hardware developments that prove this thirst for subtitles in everyday life include Will Powell's glasses, which, inspired by Google Glass, 'provide (almost) real-time subtitles for conversations being held in foreign languages' (Gold 2012: online).

On a different note, *A Christmas Carol* (Robert Zemeckis, 2009) marked a milestone in UK cinema as the first movie ever in the UK to become truly accessible in 3D to deaf and hard-of-hearing viewers (Web 3), and hence, as the first film to show 3D intralingual subtitles. The release of *Avatar* (James Cameron, 2009) a month later in December saw the birth of interlingual 3D subtitles and set the trend of the changes to come. With the surge in interest for 3D stereographic movies, more pressure is being applied to the broadcast and entertainment industry to provide 3D content for the array of 3D media players, fifth generation video games consoles, televisions and cinemas. This migration to high definition and 3D is bringing along new job profiles – like the *3D subtitle mapper*, responsible for the positioning of the subtitles – as well as fresh challenges and novel ways of working in subtitling and is bound to have an impact on the workflows and the skills required of the translators.

The need for 3D subtitles in multiple languages has become a commercial necessity since the use of traditional subtitles in a 3D environment risks destroying the 3D illusion (Screen Subtitling 2010). The main challenges derive particularly from the way the 3D subtitles are positioned on screen and how they interact with the objects and people being depicted. Any apparent conflict between an onscreen object and the subtitle text will destroy the 3D illusion and can lead to physiological side effects in the form of headaches and nausea. To address the issues raised by 3D subtitling, the British company Screen Subtitling have been pioneers in the development of Poliscript 3DITOR, a subtitle preparation package that helps design, display and deliver 3D subtitles (Web 4).

Final remarks

Though technology has been a defining feature of subtitling ever since its origins, the linguistic transfer has somewhat been forgotten when it comes to the use of CAT tools, perhaps because originally subtitling was used to translate audiovisual genres (i.e. films) that did not feature high levels of lexical repetition, as opposed to technical manuals for instance. Given the commercial importance of subtitling, it is intriguing that software engineers do not seem to have made any serious attempts to develop tools, beyond the inclusion of spell checkers, that would help subtitlers with the linguistic dimension and not only with the technical tasks. For example, by integrating a search function in the interface of the subtitling software, time could be saved as subtitlers will not have to exit the program every time they need to document themselves. In addition, some of these tools could help improve consistency in terminology, especially when dealing with team translations or TV series consisting of numerous episodes; facilitate the consultation and creation of glossaries when working in specific projects; include thesauri and suggest synonyms when space restrictions are at a premium; and propose to reuse (parts of) subtitles that have been previously translated to give account of the same or very similar expressions.

This status quo may soon change though, as the visibility of subtitling has grown exponentially around the world, including in the so-called dubbing countries where until recently this practice was rather marginal. The output has multiplied quantitatively, the outlets and screens where subtitles are displayed have proliferated and diversified, and the demand for subtitles has never been so high. Their attraction for learning and maintaining a foreign language and the

ease and speed at which they are produced are part of subtitling's strongest appeal. In an audiovisualized world, subtitles have become a commodity expected by most viewers and a translation field worth of further exploration from a technical (and linguistic) perspective.

Despite this promising outlook, subtitling also faces important challenges such as the deprofessionalization of this activity and the downward price spiral of recent years. On the industry's side, the mantra of the subtitling companies can be summarized in three key concepts: (low) costs, (speedy) turnovers and (high) quality. The first two are being clearly addressed by the various technological advancements mentioned in the previous pages. The latter, not so much, leaving quality as one of the unresolved questions that needs to be addressed, with some professionals advocating the formation of a subtitling trade body by the industry for the industry (Lambourne 2011). The high demand for subtitles to translate both user-generated content and commercial programmes is the driving force behind most technical developments taking place in the field. Reconciling costs, time, quality and professional satisfaction is not an easy task and, to date, there is no technology that adequately fills the gap.

Instead of looking for ways to do away with the human translator, technology should concentrate more on how subtitlers can be assisted in their work. Ultimately, the solution to the conundrum has to be the development of technology that finds synergies with the individual and relies on the participation and *savoir faire* of professional subtitlers. The key to success may not be so much in the technology itself, but rather in the innovative use the industry and the subtitlers make of it.

Note

1 More information on viewing videos with automatically generated captions can be found on <www.google.com/support/youtube/bin/answer.py?hl=en&answer=100078>, and a video singing the virtues of the system is available on <www.youtube.com/watch?v=QRS8MkLhQmM>.

References

Briggs, John (2013) Swedish Fan-made Subtitle Site Is Shut Down by Copyright Police, *TechCrunch*, 10 July. Available at: http://techcrunch.com/2013/07/10/swedish-fan-made-subtitle-site-is-shut-down-by-copyright-police.

Bumbalek, Zdenek, Jan Zelenka, and Lukas Kencl (2012) 'Cloud-based Assistive Speech-transcription Services', in Klaus Miesenberger, Arthur Karshmer, Petr Penaz and Wolfgang Zagler (eds) *Computers Helping People with Special Needs, ICCHP 2012, Part II* (pp. 113–116). Heidelberg: Springer. Available at: www.rdc.cz/download/publications/bumbalek12cloudescribe.pdf.

Chan, Sin-wai (2013) Translation Technology on the Fast Track: Computer-aided Translation in the Last Five Decades. In Rokiah Awang, Aniswal Abd. Ghani and Leelany Ayob (eds) *Translator and Interpreter Education and Training: Innovation, Assessment and Recognition*, Kuala Lumpur: Malaysian Translators Association, 1–11.

Cutts, Matt (2009) *Show and Translate YouTube Captions*. Available at: www.mattcutts.com/blog/youtube-subtitle-captions

Díaz Cintas, Jorge (2010) The Highs and Lows of Digital Subtitles. In Lew N. Zybatow (ed.) *Translationswissenschaft – Stand und Perspektiven, Innsbrucker Ringvorlesungen zur Translationswissenschaft VI*, (pp. 105–130). Frankfurt am Main: Peter Lang.

Etchegoyhen, Thierry, Mark Fishel, Jie Jiang, and Mirjam Sepesy Maučec (2013) SMT Approaches for Commercial Translation of Subtitles. In K. Sima'an, M. L. Forcada, D. Grasmick, H. Depraetere and A.Way (eds) *Proceedings of the XIV Machine Translation Summit*, (pp. 369–370). Nice, 2–6 September 2013. www.mtsummit2013.info/files/proceedings/main/mt-summit-2013-etchegoyhen-et-al.pdf.

Georgakopoulou, Panayota and Lindsay Bywood (2014) Machine Translation in Subtitling and the Rising Profile of the Post-editor. *MultiLingual* January / February 2014, 24–28.

Gold, John (2012) See Real-time Subtitles through Google Glass-like Apparatus. *Network World* 24 July 2012. www.networkworld.com/news/2012/072412-powell-subtitles-261112.html.

Harrenstien, Ken (2006) *Finally, Caption Playback*, 19 September 2006. http://googlevideo.blogspot.com/2006/09/finally-caption-playback.html.

Harrenstien, Ken (2009) *Automatic Captions in YouTube*, 19 November 2009. http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html.

Hunter, Gordon (2010) Services for Impaired and Disabled Users. *CSI, Cable and Satellite International* September/October 2010. www.csimagazine.com/csi/Services-for-impaired-and-disabled-users.php.

Ivarsson, Jan and Mary Carroll (1998) *Subtitling*, Simrishamn: TransEdit.

Lambourne, Andrew (2011) Substandard Subtitles: Who's Bothered? *TVB Europe* January, 37. www.sysmedia.com/downloads/pdf/clippings/TVB_Europe_%202011_Jan.pdf

O'Hagan, Minako (2013) The Impact of New Technologies on Translation Studies: A Technological Turn? In Carmen Millán and Francesca Bartrina (eds) *The Routledge Handbook of Translation Studies* (pp. 503–518), London and New York: Routledge.

Patiniotaki, Emmanouela (2013) Assistive Technology and Audiovisual Translation: A Key Combination for Access Services in Online Education. *A Global Village* 11: 52–55. http://aglobalvillage.org/journal/issue11/e-democracy/emmanuela-patiniotaki.

Romero-Fresco, Pablo (2011) *Subtitling through Speech Recognition: Respeaking*, Manchester: St. Jerome Publishing.

Screen Subtitling (2010) *Subtitling for Stereographic Media*, Screen Subtitling Systems Ltd. Available at: www.screen-and-media.com/downloads/Subtitling%20for%20Stereographic%20Media.pdf.

Web 1: 'E-TITLE'. www.upf.edu/glicom/en/proyectos/proyectos_finalizados/e_title.html.

Web 2: Europe's Information Society – SUMAT: An Online Service for SUbtitling by MAchine Translation. http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=270919.

Web 3: Your Local Cinema – A Christmas Carol. www.yourlocalcinema.com/christmascarol.PR.html

Web 4: Poliscript 3DITOR Brochure. www.screensystems.tv/download/poliscript-3ditor-brochure

YouTube (2010) *The Future Will Be Captioned: Improving Accessibility on YouTube*. http://youtube-global.blogspot.com/2010/03/future-will-be-captioned-improving.html.

40

TERMINOLOGY MANAGEMENT

Kara Warburton

INDEPENDENT SCHOLAR

Introduction

A frequent question raised by professionals in the information industry, such as writers and translators, is how we can optimize information for the computer medium that is so ubiquitous today. The effectiveness of an information product largely depends on how easily it can be found in online searches. The cost of producing information is related to how clear, concise and effective the information is, and how often it can be reused in different delivery media and for different purposes and audiences. There are increasing demands for information to be suitable for automated processes such as machine translation or content classification. New natural language processing (NLP) technologies such as controlled authoring software, content management systems, and computer-assisted translation (CAT) tools are becoming commonplace in documentation and translation departments. These and other technology-driven changes are transforming information into tangible digital assets that can be organized, structured, managed, and repurposed.

One of these digital information assets is *terminology*, and it can help address these new challenges. Methodologies have been developed to create, manage, and use *terminology* – or *terminological resources* – for specific aims such as to improve communication within and across disciplines, to strengthen minority languages, or to create and manage knowledge resources that are increasingly in highly structured electronic form.

The purpose of this chapter is to raise awareness among information professionals about terminology as a discipline and as a valuable language resource, and to describe the work known as *terminology management*. We will cover both theoretical and practical aspects.

Terminology as a discipline

Definition

Within the broader field of linguistics, Terminology[1] is the scholarly discipline that is concerned with understanding and managing terminologies, that is, words and expressions carrying special meaning. There are various definitions of this discipline, reflecting different theoretical views. The definition adopted by the ISO Technical Committee 37, which sets international standards in the area of terminology management, is as follows:

> (Terminology is) the science studying the structure, formation, development, usage and management of terminologies in various subject fields.
>
> (ISO 1087-1, 2000)

where *terminologies* are sets of *terms* belonging to special language (sometimes called languages for special purposes, or LSP), and *subject fields* are fields of special knowledge.

Subject fields, sometimes called domains, are what differentiates LSPs from general language (Rondeau 1981: 30; Dubuc 1992; Sager 1990: 18). LSPs are the language used in specific subject fields. Examples of subject fields are classic disciplines such as law, medicine, economics, science and engineering, but also applied fields such as sports, cooking, and travel, or commercial activities such as product development, shipping, business administration, manufacturing, and so forth.

Terms, concepts, and objects

In the field of Terminology, the relationship between terms, concepts, and objects is fundamental. A *term* is the name or *designation* of a *concept* in a particular *subject field*. A concept is a mental representation of a class of objects that share the same properties or characteristics, which are known as *semantic features*. For example, the concept of 'pencil' would comprise the following properties: graphite core, wood casing, used for writing, and usually yellow, with an eraser at one end, and sharpened to a point at the other end. Essential and delimiting characteristics help in the crafting of definitions of concepts such that each concept is distinguished from all other related concepts. Figure 40.1 is known as the semantic triangle in linguistics.

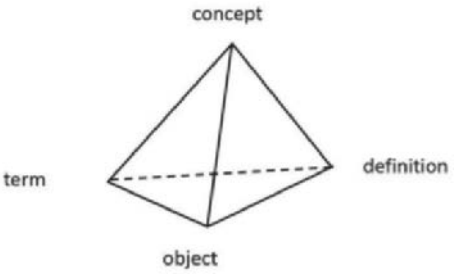


Figure 40.1 The semantic triangle

The triangle shows the relationship between an object of the real world, our mental conceptualization of that object, the term which at the same time represents the object of the real world and our conceptualization of it, and finally, the definition which is a verbal description of the concept and the object. The term and definition denote the object. The object and an individual's conceptualization of that object may not be perfectly equivalent; for instance, most people visualize pencils as yellow, yet they can come in any colour. Likewise, two individuals' conceptualizations of the same object may differ slightly; for instance, they may have different mental images of a tree or a boat, depending on their personal experiences.

In LSPs, the relationship between the term and the concept is more stable than in general language, that is, the concept is less subject to individual interpretation or variation; it is more objective.

Terms versus words

Association with a subject field is what differentiates terms from so-called common words, which are members of the general lexicon (Pearson 1998: 36) that do not refer to a specialized activity (Rondeau 1981: 26). Terms are lexical units belonging to an LSP; they are *subject-field-specific lexical units*.

Terms are also distinguished from common words by their single-meaning relationship (called monosemy) with the specialized concepts that they designate and by the stability of that relationship (Pavel 2011).

Words are generally understood to be a sequence of characters bounded by a white space at both ends (or by punctuation). However, terms often comprise more than one word, for example *climate change* or *sports utility vehicle*; these types of terms are referred to as multi-word units (MWU). In contrast, the lexical units one finds in a general language dictionary are usually single words. A sports utility vehicle is also known by the acronym *SUV*; such abbreviated forms, a phenomenon known as lexical variation, is very common in terminology. Thus, both semantically and morphologically, terms exhibit certain properties that distinguish them from so-called common words.

Relation to other disciplines

While Terminology is related to many disciplines in linguistics and in information technology, the two main ones are mentioned here: lexicography and translation.

Terminology is concerned with the language used in distinct subject fields, whereas lexicography studies the general lexicon of a language. In Terminology, the *concept* is the focus of study and the central structure for organizing data, whereas in lexicology it is the *word*.

Terminologists typically create and manage a multilingual concept-oriented database, a kind of 'knowledge base', whereas lexicologists develop dictionaries, and typically monolingual ones. The two professionals thus focus on different parts of the lexicon, look at the lexicon from different perspectives, and structure the data differently. They also differ in their methodologies, but this will be explained later. Both, however, perform some of the same tasks such as preparing definitions and describing the usage and grammatical properties of words and terms.

The field of Terminology traces its origins to the need for speakers of different languages to communicate clearly with each other in various subject fields. Therefore, it almost always takes a multilingual approach. As such, it is closely related to the field of translation, and most terminologists have a translation background. Many training programmes for translators include modules about Terminology, and software programs used by translators often include functions for managing terminology.

Theoretical evolution

The original and still predominant theory of Terminology was developed by Eugen Wüster and colleagues in Vienna in the middle of the last century (Wüster 1979). This theory is referred to as the Traditional Theory, the General Theory, the Wusterian Theory, or the Vienna School. An engineer, Wüster developed his theory while preparing a multilingual dictionary of machine tools (Wüster 1967). According to this theory, clear communication is achieved by fixing the relationship between terms and concepts. The objective is *biunivocity*, whereby a linguistic form corresponds to one and only one concept, and a concept is expressed

by one and only one linguistic form (L'homme 2004: 27). The focus of study is the concept, to which terms are secondarily assigned as designators (Cabré Castellvi 2003: 166-167). Concepts occupy fixed positions in a language-independent concept system, where they are hierarchically related to other concepts. The preferred methodology is *onomasiological*, that is, concepts are delimited *before* any of their corresponding terms are even considered, and the goal is normalization, or standardization. Works based on the General Theory of Terminology include Cabré Castellví (1999), Felber (1984), Rondeau (1981), Dubuc (1992) and Picht and Draskau (1985).

The onomasiological approach is one of the basic tenets of the General Theory of Terminology. It contrasts with the *semasiological* approach, which is used in lexicography. These two approaches are fundamental for distinguishing terminology from lexicography. With the semasiological approach, a word is described in all its possible meanings. In contrast, terminologists study and describe concepts and then, only secondarily, determine how these concepts are expressed verbally by terms. The two perspectives are shown in Figures 40.1 and 40.2:

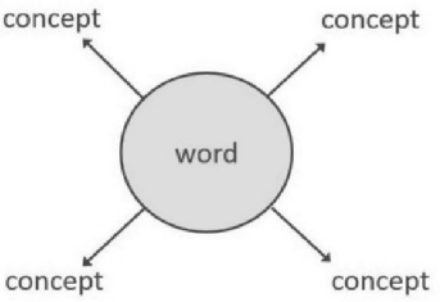


Figure 40.2 Lexicology – semasiological

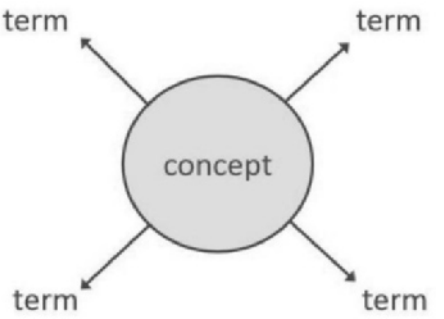


Figure 40.3 Terminology – onomasiological

Parallel to a shift from structuralist linguistics to corpus linguistics and cognitive linguistics, since the mid 1990s, the General Theory of Terminology has been subject to some criticism. The main critique is that it does not take into account language in use (Pearson 1998; Temmerman 1997: 51–90 and 2000; L'Homme 2004 and 2005: 1112–1132; Cabre 1999; Kageura 2002). Concepts are studied outside of their use in communicative settings. Terms are considered at the level of *langue*, and not of *parole*, in de Saussurian terms. Consequently, several new theories emerged in recent decades that emphasize communicative, cognitive, and

lexical aspects (see for example Cabré Castellví, Temmerman, Castellvi, Sager, and L'Homme). The socio-cognitive theory and the lexico-semantic theory come to mind, exemplified by Temmerman and L'Homme respectively. The emergence of these theories was facilitated by advances in NLP technologies and the availability of large machine-readable corpora, which opened up new opportunities to study terms in their natural context.

The Socio-Cognitive Theory views terms as expressions of meaning that are dependent on the context of communication. The Lexico-Semantic Theory considers terms first and foremost as lexical units. The focus is on lexical structures rather than conceptual ones.

The three aforementioned theories (General, Socio-Cognitive, and Lexico-Semantic) diverge considerably in their definition of what constitutes a term (sometimes called *termhood*), emphasizing respectively, the concept, cognitive aspects and communicative context, and lexical behavior. The General Theory of Terminology considers membership in an objectivist, structured system of concepts as a criterion of termhood and it determines this membership status on the basis of the concept, not of the term. In contrast, subsequent theories place more emphasis on a range of linguistic properties (morphological, syntactic, paradigmatic, etc.) of terms, properties that can be determined from the corpus. Indeed, in these more recent theories, the notion of term is intrinsically linked to the text in which it occurs.

Terminology management

General explanation

The act of managing terminology refers to a wide range of tasks focused on terminology data, i.e. terms and information about terms such as definitions, context sentences, and grammatical information. These tasks include collecting, developing, storing, reviewing, harmonizing, enhancing, and distributing terminology data. Today terminology is always managed by using computers, and terminology data is stored in a terminology database, or *termbase*. The person who manages terminology is referred to as a *terminologist*.

Spreadsheets are commonly used to record terminology in the initial stage, such as by a translator. However, this activity would not be considered 'terminology management' and ultimately, in order to be properly utilized, the terminology in the spreadsheet would need to be imported into a termbase that the terminologist manages with the aid of a terminology management software. This is why terminology management software programs usually provide an import function for spreadsheets.

Terminologists create and manage termbases, which are composed of terminological *entries*. They work in the language services of governments, where they play a key role in supporting the national languages, in the private sector, where they support corporate communications, and in supra-national non-governmental organizations, such as the United Nations, where they facilitate multilingual communication.

When an organization needs to communicate clearly, it examines its terminology and decides which terms to use and which terms to avoid. This decision-making process results in a prescribed set of terms, which need to then be distributed to members of the organization, often with definitions to ensure that everyone using the terms knows exactly what they mean.

Terminology resources are often created to support multilingual communication in commercial sectors and in specialized subject fields or domains, such as law, science, and medicine. Clarity and precision are paramount. Providing semantic information about the source language terms, such as definitions, can greatly assist translators to determine the correct target language equivalents. In many production settings, target language equivalents are determined

before the document is translated, and the target language terms are provided to the translators working on the document. This sequence of events – translate the key terms before the document – helps to reduce terminology errors and raise terminology consistency.

Terminology resources can also be developed in monolingual settings, such as to provide sets of pre-approved terminology for writers to use when they prepare documentation in a specialized field. A case in point is the aeronautical industry, which was an early adopter of controlled terminology in technical writing.

The need for terminology management

Consistency of language (terms and expressions) is frequently cited as one of the key factors in information clarity and usability. Terminology consistency in an information set also has an impact on the reader's ability to find this information, through search engines or online indices. When they occur in a text that will be translated, terminology inconsistencies often increase in frequency in the translated version compared to the original, due to the fact that there can be several ways to translate a given term or expression. When a document or a collection of documents is divided into smaller parts which are translated by several translators, terminology in the target language will be more inconsistent than when only one translator is involved. In spite of this risk, this approach is often adopted by companies, under pressure to get their products to market as quickly as possible. In industries with highly specialized terminology, such as the automotive industry, terminology inconsistencies and other terminology errors are among the most frequently occurring translation errors (Woyde 2005). In high-risk fields such as health sciences, engineering, national defense, and law, problems of ambiguity, inconsistency, or imprecision can have serious consequences. Prescribed terminology can be provided to writers and translators to help them avoid these problems. This will be explained further later.

Aside from improving the quality of information content, the benefits of terminology management can also be demonstrated from a business perspective, that is, in terms of cost, time, and productivity gains. The return-on-investment (ROI) needs to be separately measured for each organization, since the gains depend on its specific production environment. Nevertheless, several generic ROI evaluations have been produced; see for instance Champagne (2004), Warburton (2013a, 2013b) and Schmitz and Straub (2010). These studies show that costs are saved by reducing wasteful duplication of work, such as when two translators research and work through the process of translating the same term, or when two technical writers create definitions for the same term in different company documents. Another area of cost savings is reducing the effort of editors and revisers to correct terminology mistakes by reducing the occurrence of such mistakes.

The purpose of managing terminology in an organization, such as a company or an NGO, is to improve the use of terminology across that organization. In a language planning environment, such as a branch of government responsible for protecting the national language, the mission is to strengthen the language as a whole. The latter has a social dimension, and indeed, the term *socioterminology* has been coined for this type of terminology work. There is no specific term yet for the former type of terminology management, but we could call it *institutional terminology*.

In summary, establishing and using consistent and appropriate terminology helps increase the quality of information, which in turn improves the usability of related products, makes information easier to find, and lowers translation costs.

Types of terminology management

Onomasiological versus semasiological

As previously noted, the classical approach to terminology management is onomasiological, whereby the concept is the central focus. This distinguishes terminology from lexicology, which adopts a semasiological (word-based) approach. However, in many situations, terminology work is actually semasiological. Translators, for instance, often identify a handful of key terms in a document that they are translating, and determine the correct translations after doing a bit of research. To save their work for future reference, they add the terms to a termbase.[2] Although they have taken care to ensure that the target language terms they choose have the same meanings as the source language terms, they spend little if any time analysing the concepts and rarely record concept information such as definitions or subject-fields, much less produce diagrams of hierarchical concept systems. In practice, the dividing line between terminography and lexicography has become quite fuzzy.

Descriptive, prescriptive and normative

Terminology management methods also vary depending on whether the goal is *descriptive*, *prescriptive* or *normative*. In descriptive terminology, the terminologist 'describes' the current and actual behavior and usage of terms without making any value judgments about them. This approach is adopted for instance to record the vocabulary used in so-called minority languages or languages at risk. Normative terminology seeks to develop a 'standardized' terminology in specific subject fields; an example is the terminology found in ISO standards. Prescriptive terminology adopts some aspects of both the former: it documents terms in use but at the same time it is concerned about consistency and quality of terminology and therefore it 'prescribes' terms to use and terms to avoid in cases of synonymy. The prescriptive approach is common in institutional terminology management.

Different methodologies and types of information are needed to achieve these different aims. Descriptive terminology emphasizes recording the sources of terms and context sentences. Normative terminology is the most likely type to adopt an onomasiological approach, where significant time and resources are spent on concept analysis, synonym ranking, and the crafting of definitions. In fact, definitions are often mandatory and adhere to strict rules of style. Prescriptive terminology adopts the normative approach only for difficult cases of problematic synonyms or conflicting terms and for the remainder settles on basic description. Definitions are less rigorous than in normative settings and will only be present in a small proportion of the total number of entries.

When providing an aid to translators is the primary purpose of terminology work, the descriptive approach is most common. The terminologist finds source language terms that are used and puts them into the termbase, and target language equivalents are then added to the entries if and when they are needed. Thus, most translation companies and their clients accumulate their terminology data using descriptive methods over a period of years or even decades. Problems arise, however, if the needs for terminology move beyond the translation activity, since the nature and structure of terminology data required for other purposes can vary considerably. The organization may find that its termbase needs to be modified to handle these new uses. This issue will be further discussed in the section *Repurposing and interoperability*.

Methodology

Thematic versus ad-hoc

According to the General Theory, concepts can only be studied systematically, that is, as members of a logical and coherent concept system (Rondeau 1981). This onomasiological approach to terminology work is also referred to as *thematic* (L'Homme 2004). The terminologist studies and then defines key concepts from a subject field by applying the rigor of Aristotelian logic, that is, naming a genus (i.e. the superordinate class that the concept belongs to), and differentiae (i.e. the semantic properties that differentiate this concept from other concepts belonging to the same class). The concepts are then represented in a diagram showing the hierarchical relations between them. Only then are terms selected to denote the concepts. Figure 40.3 is a concept system elaborated according to the thematic method:

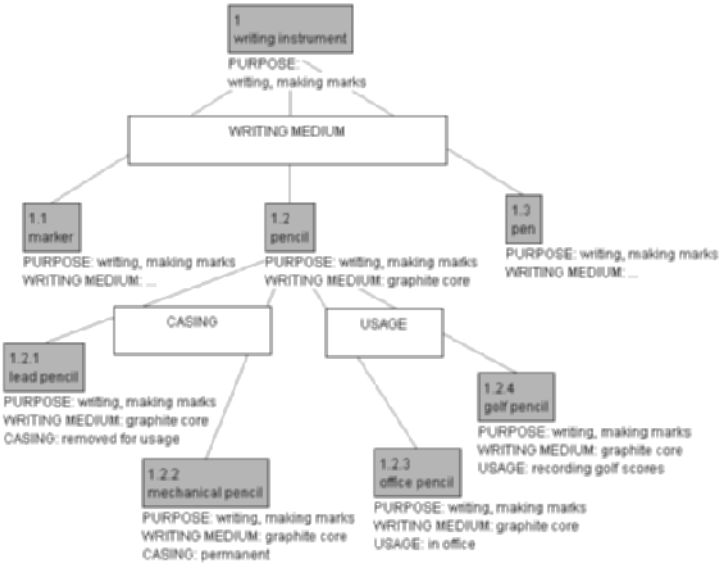


Figure 40.4 Concept system for writing instruments

Source: Copenhagen Business School

Concepts, represented by the grey boxes, are denoted by terms (the box labels). Concepts are numbered and arranged hierarchically to show their relations. Concepts at the same level of the hierarchy, called coordinate concepts, share the same parent, their superordinate concept. For instance, *marker*, *pencil*, and *pen* are coordinate concepts to the superordinate *writing instrument*. The white boxes represent criteria of subdivision, i.e. on what basis the subordinate concepts are differentiated from each other. For instance, what makes a *lead pencil* different from a *mechanical pencil* is the casing. The text below the grey boxes lists the essential characteristics of the concept; one can write a definition using these characteristics. For instance:

> mechanical pencil: pencil with a fixed[3] casing

Note that it is not necessary to state 'writing instrument used for making marks that has a graphic core and a fixed casing' since all the properties except 'fixed casing' are present in the

concept of *pencil*. This inheritance principle allows definitions to be quite concise. However, in order for this definition to be valid, the term *pencil* used as the genus must also be defined in the terminology resource in question.

The thematic approach is widely recognized as characteristic to the field of Terminology. It is particularly well-suited for developing standardized terminology, such as for nomenclatures. However, in more practical situations, such as when a translator or a writer working on a document quickly needs help deciding what term to use, a task- and text-driven approach is usually adopted. The terminologist starts with a source language term, and looks for target language equivalents by searching in various sources. Finding instances of the target language term in authentic contexts with the same meaning as the source language term is the only evidence required. Definitions and concept diagrams are almost never prepared. This approach is referred to as *ad-hoc* (Wright and Wright 1997: 13–24; Wright 1997) or *punctual* (Cabré Castellvi 2003; Picht and Draskau 1985) because the aim is to fulfil an immediate need for a translation and then move on to other tasks.

These different methodologies have different interpretations about the notion of what constitutes a term. In the thematic approach, a term is a designator of a conceptual node in a structured concept system. Real contexts may be studied to confirm the existence of these terms, but their final determination is based solely on the concept system. The *ad-hoc* approach, in contrast, accepts the existence of a term based solely on observations of text.

Corpora, concordances and term extraction

As noted earlier, terminologists frequently adopt a semasiological approach to term research and identification. Due to advances in text processing capabilities made possible through the use of computers, terminologists can now use large bodies of text, called corpora, in their research. Using corpora as empirical evidence of terms is essential in any medium to large scale terminology project. Due to the large size of most corpora, the task of identifying terms is usually carried out with the assistance of technologies such as concordancing software and automatic term extraction (ATE) tools. Not only do these technologies allow terminologists to identify more terms than would be possible manually, their use also raises the correspondence between the terms in the termbase and the corpus that the terms are supposed to reflect. In other words, terms for the termbase are pulled from the corpus directly rather than resulting from a decision made by the terminologist with potentially no reference to the corpus at all. The latter results in some terms being entered in the termbase that are not useful as a support to writers and translators, simply because they occur rarely in the texts that the writers and translators are dealing with. Since entering a term and associated information such as definitions in a termbase typically triggers a downstream process of adding translations, the wasted cost in terms of manpower caused by adding source language terms that have little value is multiplied, in some cases tenfold or higher. Deciding which terms to include in a termbase to ensure its value is a major challenge for terminologists, and basing this decision on corpus evidence is highly recommended.

Workflow

The workflow of identifying terms, recording them in a termbase, adding information and obtaining the necessary approvals varies considerably depending on whether the approach is thematic or ad-hoc, and the aim is descriptive, prescriptive, or normative. ISO TC37 has published several standards that describe workflows,[4] such as ISO 15188 and ISO 10241. Figure 40.4 shows a workflow for prescriptive terminology that was elaborated by TerminOrgs,[5] a group of professional terminologists.

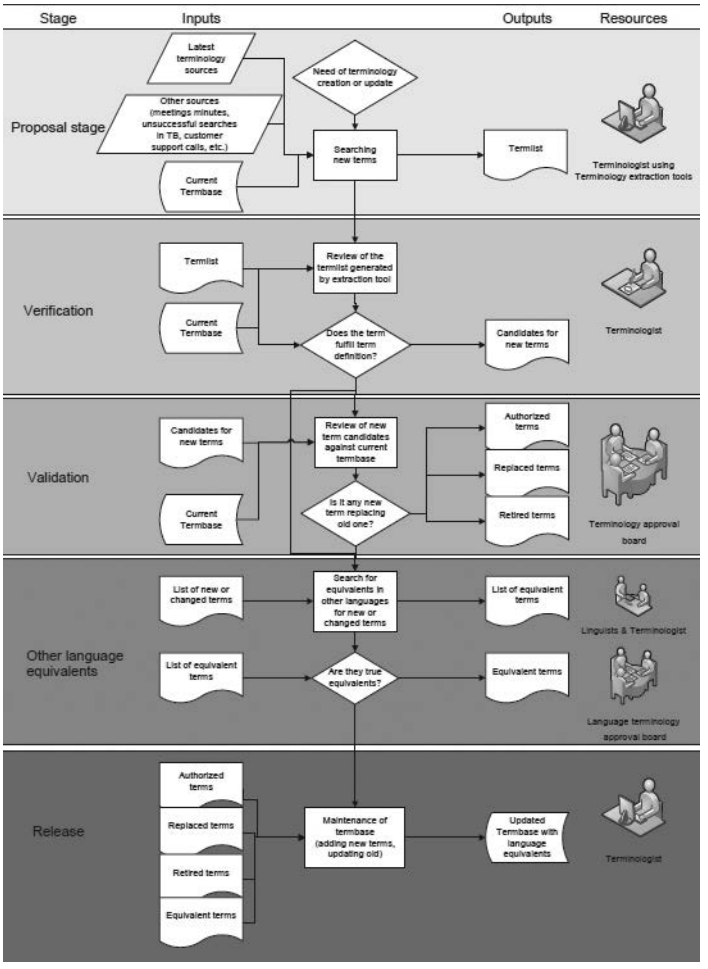


Figure 40.5 A workflow for prescriptive terminology.

Source: TerminOrgs

Terminology databases

General explanation

A terminology database (termbase) is a collection of terminology in electronic form, similar to yet different from a dictionary that one can access on the Web. Termbases are almost always multilingual.

Organizations of various sorts develop a termbase to serve their communication needs. Governments of countries that have more than one official language or significant populations speaking different languages may develop a termbase to store terms needed to express concepts for their various programmes, services, and industries. Examples are Canada and Sweden. Supra-national organizations such as the European Union develop termbases to support interlingual communication which facilitates cross-border trade and collaboration. Non-

governmental organizations, such as the World Health Organization, develop termbases in order to help implement their programmes effectively in different linguistic communities. Increasingly, commercial enterprises are using termbases to store multilingual terminology about their products and services; these termbases are leveraged in the authoring and translation process to increase quality, improve productivity, and save time.

Data categories

Many termbases are quite simple, containing primarily just terms. Some, however, contain a wide range of other types of information such as definitions, usage notes, and grammatical descriptors. Some also include links between entries which can range from simple pointers between related terms to hierarchical relations of various sorts, such as to link broader and narrower terms. These bits of information are called *data categories* by terminologists. There are hundreds of different data categories possible for a termbase. In 1999, ISO TC37 published an inventory of terminology data categories as an international standard (12620). In 2009 this standard was revised and the inventory was moved into an electronic database which is available on the Internet (www.isocat.org).

Broadly speaking, data categories are organized into three groups: conceptual, terminological, and administrative. Conceptual data categories describe concepts; they provide semantic information. Examples are subject field values and definitions. Terminological data categories describe terms, for example, usage notes, part-of-speech and context sentences. Administrative data categories include, for example, the name of the person who added some information or the date that it was added.

Structure

The structure of termbases has also been standardized by ISO TC37, through ISO 16642: *Terminological Markup Framework*. According to this standard, terminological entries are structured in three hierarchical sections: concept, language, and term. Information at the concept level describes the concept as a whole and thus is shared by all the terms in all languages in the entry, such as a subject field value and a definition. All the information pertaining to a given language is organized in a dedicated language section, which is sub-divided into term sections for information about individual terms. Information that is shared by all the terms in a given language occurs at the language level, such as a language-specific definition or a comment, and information about a specific term is inserted at the term level, such as a usage note or a context sentence. A term section can contain only one term, but a language section can contain multiple term sections and a concept section can contain multiple language sections. All the terms in the entry are synonyms. This principle is referred to as *concept orientation*.

Examples of termbases

Some large termbases are publicly available today. The following are a few examples:

- United Nations – UNTERM: http://unterm.un.org/
- European Union – IATE: http://iate.europa.eu
- Government of Canada – TERMIUM: http://www.btb.termiumplus.gc.ca/
- Microsoft - https://www.microsoft.com/language/en-us/search.aspx
- Eurotermbank - http://www.eurotermbank.com/

Terminology management systems

Single-purpose versus multi-purpose

A terminology management system (TMS) is a software program specifically designed for managing terminology. When the first terminology databases were developed in the 1970s, none existed, and so the organizations responsible for these termbases developed their own in-house systems. In the 1980s and 1990s, TMSs began to emerge as part of desktop software programs for translators, known as computer-assisted translation (CAT) tools. Today, virtually all CAT tools have functions for collecting and storing terminology.

It should be pointed out that the terminology components of CAT tools and controlled authoring tools tend to lack features that may be needed for extended applications of terminology data, since they are designed specifically for use by translators and writers respectively. One should be careful not to assume that a TMS designed for a single purpose will be suitable for developing and managing terminology resources for other purposes.

Using a Wiki application can be effective for collecting terminology from grass roots users on a large scale. Wikis are designed for open collaboration. If the institutional setting requires any level of control over the terminology data, such as how it is entered, by whom, and what types of terms are allowed in the system, a Wiki-based TMS may not be appropriate. In addition, Wikis tend to lack some of the more advanced functions available in a more robust TMS, such as the ability to create relations between concept entries.

Today, the majority of TMS available on the market continue to be single-purpose, that is, designed for the needs of one type of user. As developers of termbases begin to want to use their terminology data for purposes in addition to translation, the need for a TMS that supports a wider range of users has increased, and a few robust products exist today. The ISO standard 26162, *Design, Implementation and Maintenance of Terminology Management Systems*, is an excellent resource.

Key features

It is beyond the scope of this chapter to describe the features of terminology management systems. For that purpose, refer to the standard mentioned above. It is, however, essential that a TMS adhere to the international standards mentioned in this chapter. Some of the most important principles and features include:

- Web interface
- concept orientation
- term autonomy
- a variety of import and export formats, minimally including spreadsheets and TBX
- views and layouts customizable for different user types
- ability to record relations between terms and between concepts.

Some of these topics are described in later sections of this chapter.

Push and pull approach

Terminology can be shared among all translators working on a project, which obviously helps to increase consistency. However, expecting translators to look up terms and obtain the prescribed translations is not effective; if they already know how to translate a term in *their* way,

it will not occur to them to look for a different translation. Working under time pressures, they are not likely to check their terminology frequently.

CAT tools address this problem by providing functionality whereby if a sentence to be translated contains a term that is in the termbase, the corresponding entry is automatically displayed to the translator. This function, commonly known as *autolookup*, reflects a 'push' approach. It 'pushes' information to the user at the moment it is needed. In contrast, terminology in separate files such as spreadsheets reflects a 'pull' approach, where the user decides if and when to access the information.

This push approach applies in other user scenarios, such as in controlled authoring. A controlled authoring software normally has a terminology component for writers just as a CAT tool does for translators, only in this case, it is monolingual. The terminology function contains source language terms with usage information; in particular, terms that should be avoided are clearly indicated. When a writer uses one of these deprecated terms, the function automatically highlights the term and displays another that should be used in its place. It works just like a spell checker, which highlights spelling errors and makes suggestions.

Repurposing and interoperability

Explanation

Terminology data is a language resource that can be used for various purposes, as described in the next section. The term *repurposing* refers to the activities of using terminology data in different applications. In order to repurpose terminology, the data has to be interoperable, meaning that it must be possible to exchange the data between the termbase and other systems easily and without loss of information. This can be achieved by an import/export procedure, or by direct real-time links between the termbase and other applications that are implemented programmatically such as through application programming interfaces. Repurposing requires a standard interchange format: TermBase eXchange (TBX).

Applications of terminology resources

The performance of many NLP applications can be improved through the use of richly structured terminological resources (Castellvi *et al.* 2007). Nearly 20 years ago, Meyer (1993) predicted that machines would become a primary user of terminological data.

> It is predicted that machines may become a category of user for terminology banks; machine translation tools, expert systems, natural-language interfaces to databases, and spelling checkers are just a few of the most obvious applications. … Machines will need very large quantities of explicitly represented conceptual information since they do not possess much of the basic real-world information that humans know implicitly.

Ibekwe-SanJuan *et al.* (2007) consider commercial applications of what they call 'terminology engineering'.

> Applications of terminology engineering include information retrieval, question-answering systems, information extraction, text mining, machine translation, science and technology watch, all are external fields that can benefit from the incorporation of terminological knowledge.

They further note that terminology is useful for building other types of language resources such as ontologies and aligned corpora. Numerous works describe the role of controlled terminologies for indexing (Strehlow 2001: 419–425; Buchan 1993: 69–78). Strehlow notes that a sound strategy for the use of terminology in strategic areas of content (such as titles, abstracts and keywords) can lead to significant improvements in information retrieval. Wettengl *et al.* (2001: 445–466) describe how terminology data can help build product classification systems.

There are numerous references in the literature to applications of terminology resources: automatic back-of-the-book indexing, indexing for search engines, ontology building, content classification, contact record analysis, search engine optimization (query expansion and document filtering), federation of heterogeneous document collections, cross-lingual information retrieval, document summarization/abstraction and keyword extraction, product classifications, automated construction of domain taxonomies and ontologies, and so forth (Park *et al.* 2002: 1–7; Jacquemin 2001; Oakes *et al.* 2001: 353–370; Cabre 1999).

Repurposing considerations

As shown in the previous section, the use of terminology data in an organization can change over time. We mentioned that repurposing requires a standard interchange format. But that in itself is not sufficient to prepare the data for its different potential applications; the data also needs to be designed so that it is usable in these systems. For instance, to reduce translation costs, an organization may deploy controlled authoring software to help raise the consistency, quality, and translatability of its source content. Controlled authoring software requires terminology data of a nature and structure that is different than that required for translation purposes. For instance, it requires information about synonyms in the source language, which is frequently lacking in translation-oriented terminology resources. But even more fundamental is the fact that terminology resources developed for translators are frequently not even concept-oriented: synonyms are not stored in the same entry. This is a problem for any organization wishing to use its termbase for controlled authoring.

Another problem occurs when manually prepared glossaries are imported into a TMS. Typically these glossaries lack a part-of-speech value for the terms. This particular data category is essential for any NLP-oriented application of the data. It is therefore recommended that the part of speech value be added when importing glossaries into a TMS.

Terminology data developed for translation purposes may not be usable 'out of the box' for other applications. Re-engineering the terminology resources to meet new requirements can be very difficult and cost prohibitive. If an organization anticipates that it may want to repurpose its terminology, consulting a professional terminologist prior to purchasing a TMS or creating a database of any significant size can help to protect its investment.

The following are two general guidelines that can help to ensure that a termbase is repurposable:

- comply with the industry standards listed at the end of this chapter;
- be aware of the potential limitations of a TMS that is locked into an application-specific software.

Standards and best practices

Adhering to standards ensures that the terminology data is developed according to long-standing best practices and sound principles, some of which are described in this section.

Concept orientation

Concept orientation is the fundamental principle whereby a terminological entry describes one and only one concept. This principle originates from the General Theory of Terminology and is still extremely relevant today even for the most practical types of terminology work. It distinguishes terminology from lexicology, where an entry describes an individual word or expression, which can be polysemic. Terminology databases are usually multilingual; each language term in an entry is equivalent in meaning to the others. However, because entries are meaning-based, they can also contain multiple terms in a given language. Synonyms, abbreviations, and spelling variants of a term must all be placed in the same entry. One can say that terminology resources are structured more like a thesaurus than a conventional dictionary.

Term autonomy

Term autonomy is the principle whereby each term in a terminological entry can be documented or described with the same level of detail. In other words, the TMS should not have an unbalanced structure whereby there are more fields available to describe a so-called 'main' term than for other terms. This unbalanced structure has been observed in some TMSs where there are a number of fields available to document the first term, such as **Part-of-speech** and **Definition**, but fewer fields for subsequent terms, such as **Synonym** and **Abbreviation**. This reflects a poor design. Instead, the TMS should use a Term type data category. For example, instead of:

Term:
  Part-of-speech:
  Definition:
  Synonym:
  Abbreviation:

One should have:

Term:
  Part-of-speech:
  Definition:
  Term type:

where the **Term type** field allows values such as *abbreviation* and *acronym*. The value *synonym* should not be necessary since all terms in a given entry are synonyms of each other. So long as the term section shown above can be repeated in the entry, all terms can be equally described.

Data elementarity

Data elementarity is a best practice in database management in general, and it also applies to termbases. According to this principle, there can only be one type of information in a database record or field. This means that in the termbase design separate fields are required for different types of information, which takes careful planning to account for all types of information that may be required by the users. An example where this principle has been violated is when the **Definition** field contains not only a definition, but also the source of that definition. Another

example is when both the part-of-speech and gender is inserted into a field, such as 'n f' for a feminine noun. Possibly the worst case is when the **Term** field contains two terms, as this also violates term autonomy. Unfortunately this seems to occur fairly frequently, such as when a term field contains both a full form and an abbreviation, for example:

Term: access control list (ACL)

These are two different terms. The following arrangement is better:

Term: access control list
  Term type: full form
Term: ACL
  Term type: acronym

Summary

Terminology is a field in applied linguistics that is experiencing rapid growth due to advances in information technology and natural language processing, economic globalization, and linguistic diversity. While Terminology shares some methodology with lexicography, the two have different foci, namely concept description and language description, and a different scope, i.e. special language and general language, respectively. Terminology resources help translators, writers, and other content producers use clear and consistent terms so that communication is most effective. These resources are created and managed in databases that reflect fundamental principles unique to the field of Terminology. If properly developed, they can improve the performance of many natural language processing applications that are driving innovation in information technology.

Industry standards

ISO 704: Terminology Work – Principles and Methods
ISO 30042: TermBase eXchange (TBX)
ISO 16642: Terminological Markup Framework (TMF)
ISO 26162: Design, Implementation and Maintenance of Terminology Management Systems
ISO TC37 data categories (www.isocat.org)

Notes

1 The word 'terminology' is polysemous as it can mean (1) the field of terminology as a discipline, (2) a set of terms, and (3) the practice of managing terminology. To address this ambiguity, the word Terminology will be capitalized when referring to the discipline.
2 Many translators use spreadsheets rather than a proper termbase.
3 Here, we have chosen 'fixed' instead of 'permanent' as it seems more accurate. The term *permanent* frequently has a temporal meaning.
4 See: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=48104&published=on&includesc=true
5 http://www.terminorgs.net.

References

Buchan, Ronald (1993) 'Quality Indexing with Computer-aided Lexicography', in Helmi B. Sonneveld and Kurt L. Loening (eds) *Terminology – Applications in Interdisciplinary Communication*, Amsterdam and Philadelphia: John Benjamins, 69–78.

Cabré Castellví, M. Teresa (1999) *Terminology – Theory, Methods and Applications*, Amsterdam and Philadelphia: John Benjamins.

Cabré Castellví, M. Teresa (2003) 'Theories of Terminology', *Terminology* 9(2): 163–199.

Champagne, Guy (2004) *The Economic Value of Terminology – An Exploratory Study*. (Report commissioned by the Translation Bureau of Canada).

Dubuc, Robert (1992) *Manuel pratique de terminologie*, Quebec, Canada: Linguatech.

Felber, Helmut (1984) *Terminology Manual*, UNESCO.

Ibekwe-SanJuan, Fidelia, Anne Condamines, and M. Teresa Cabre Castellvi (2007) *Application-driven Terminology Engineering*, Amsterdam and Philadelphia: John Benjamins.

ISO TC37, SC1. 2000. ISO 1087–1 (2000) *Terminology Work – Vocabulary – Part 1: Theory and Application*, Geneva: International Organization for Standardization.

Jacquemin, Christian (2001) Spotting and Discovering Terms through Natural Language Processing, Cambridge, MA: MIT Press.

Kageura, Kyo (2002) The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth, Amsterdam and Philadelphia: John Benjamins.

L'Homme, Marie-Claude (2004) *La terminologie: principes et techniques*, Montreal: Les Presses de l'Université de Montréal.

L'Homme, Marie-Claude (2005) 'Sur la notion de terme', *Meta* 50(4): 1112–1132.

Meyer, Ingrid (1993) 'Concept Management for Terminology: A Knowledge Engineering Approach', in Richard Alan Strehlow and Sue Ellen Wright (eds) *Standardizing Terminology for Better Communication*, ANSI.

Oakes, Michael and Chris Paice (2001) 'Term Extraction for Automatic Abstracting', in Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme (eds) *Recent Advances in Computational Terminology*, Amsterdam and Philadelphia: John Benjamins, 353–370.

Park, Youngja, Roy J. Byrd, and Branimir K. Boguraev (2002) 'Automatic Glossary Extraction: Beyond Terminology Identification', in *Proceedings of the 19th International Conference on Computational linguistics*, Volume 1, Association for Computational Linguistics, 1–7.

Pavel, Sylvia (2011) *The Pavel Tutorial*. Available at: http://www.btb.gc.ca/btb.php?lang=eng&cont=308.

Pearson, Jennifer (1998) *Terms in Context – Studies in Corpus Linguistics*, Amsterdam and Philadelphia: John Benjamins.

Picht, Heribert and Jennifer Draskau (1985) *Terminology: An Introduction*, Denmark: LSP Centre, Copenhagen Business School.

Rondeau, Guy (1981) *Introduction à la terminologie*, Montreal: Centre educatif et culturel Inc.

Sager, Juan (1990) *A Practical Course in Terminology Processing*, Amsterdam and Philadelphia: John Benjamins.

Schmitz, Klaus-Dirk and Daniela Straub (2010) *Successful Terminology Management in Companies: Practical Tips and Guidelines: Basic Principles, Implementation, Cost-benefit Analysis and System Overview*, Stuttgart: Stuttgart TC and More GmbH.

Strehlow, Richard (2001) 'Terminology and Indexing', in Sue Ellen Wright and Gerhard Budin (eds) *Handbook of Terminology Management*, Volume 2, Amsterdam and Philadelphia: John Benjamins, 419–425.

Temmerman, Rita (1997) 'Questioning the Univocity Ideal: The Difference between Socio-cognitive Terminology and Traditional Terminology', *Hermes Journal of Linguistics* 18: 51–90.

Temmerman, Rita (2000) *Towards New Ways of Terminology Description*, Amsterdam and Philadelphia: John Benjamins.

Warburton, Kara (2013a) 'Developing a Business Case for Managing Terminology'. Available at: http://termologic.com/?page_id=56.

Warburton, Kara (2013b) 'The ROI of Managing Terminology: A Case Study'. Available at: http://termologic.com/?page_id=56.

Wettengl, Tanguy and Aidan van de Weyer (2001) 'Terminology in Technical Writing', in Sue Ellen Wright and Gerhard Budin (eds) *Handbook of Terminology Management*, Volume 2, Amsterdam and Philadelphia: John Benjamins, 445–466.

Woyde, Rick (2005) 'Introduction to SAE J1930: Bridging the Disconnect Between the Engineering, Authoring and Translation Communities', *Globalization Insider*, Geneva: Localization Industry Standards Association. Available at: http://www.translationdirectory.com/article903.htm.

Wright, Sue Ellen (1997) 'Term Selection: The Initial Phase of Terminology Management', in Sue Ellen Wright and Gerhard Budin (eds) *Handbook of Terminology Management*, Volume 1, Amsterdam and Philadelphia: John Benjamins, 13–24.

Wright, Sue Ellen and Leland Wright (1997) 'Terminology Management for Technical Translation', in Sue Ellen Wright and Gerhard Budin (eds) *Handbook of Terminology Management*, Volume 1, Amsterdam and Philadelphia: John Benjamins Publishing Company, 147–159.

Wüster, Eugen (1967) *Grundbegriffe bei Werkzeugmaschinen*, London: Technical Press.

Wüster, Eugen (1979) *Introduction to the General Theory of Terminology and Terminological Lexicography* (translation), Vienna: Springer.

41
TRANSLATION MEMORY

Alan K. Melby

BRIGHAM YOUNG UNIVERSITY, THE UNITED STATES

Sue Ellen Wright

KENT STATE UNIVERSITY, THE UNITED STATES

Introduction

A 'translation memory' (TM) is a database of paired text segments, where Segment B is a translation of Segment A. Translators use TMs to 'remember' the content of past translations. TM programs comprise the prototypical function associated with so-called 'CAT' systems (Computer Assisted Translation).

The term 'translation memory' is ambiguous. As noted in Macklovitch and Russell (2000: 137–146) it is sometimes used to designate a database containing a collection of paired source-language (SL)/target-language (TL) text units, but in common parlance the term is also inaccurately used to refer to one of the software programs used to create, store, access, retrieve and process the units contained in the TM database. Somers and Diaz (2004: 5–33) try to circumvent this ambiguity by referring to *translation memory systems* as *TMS*, although care must be taken to ensure the context is clear, since *TMS* can also stand for *terminology management system* or *translation management system*. Another ambiguous term that needs to be clarified before discussing TM technology is 'translation unit.' Early in the development of TMs, it was defined as 'an entry [as in a database entry] consisting of aligned segments of text in two or more languages' (TMX Standard 1998, where it is assigned the XML tag *<tu>*, as shown in Figure 41.2). We will call this the *formal translation unit*, and the process of dividing text into segments, logically enough, is called 'segmentation.' Formal <tu>s have traditionally been automatically identified based on elements of primary punctuation, such as periods (full stops), question marks, exclamation points, and (optionally) colons and semi-colons, as well as end-of-paragraph markers. Consequently, although they may be fragments, they usually equate to full sentences or sometimes paragraphs.

The term *translation unit* is also common in translation studies, where it has undergone several shifts. Early on Vinay and Dalbelnet (1958/1995) focused on the smallest (atomic) units of thought, essentially terminological and collocational units. Nevertheless, for many translators the translation unit (unit of translation) is instead 'a stretch of source language text' on which the translator focuses during the cognitive translation *process*, or, viewed conversely, the corresponding target-text chunk that is the *product* of that process and that can be 'mapped onto the source-text unit' (Malmkjaer 1998: 373).

Kenny notes that the translation unit undergoes constant transformation, depending on changes in a translator's cognitive processing (2009: 304). These *cognitive translation units* may

vary dynamically in rank (unit level) from single terms to collocations to clauses, even to whole sentences (McTait *et al*. 1999), but do probably tend to occur at the clause level, which is confirmed in research involving *think-aloud protocols (TAP)*.

If formal <tu>s are set at the sentence or paragraph level, and cognitive <tu>s are frequently sub-sentential, there can be a cognitive disconnect between the human translator and the TM. Consequently, Reinke even asserts that the notion of the translation unit used to designate pairs of SL and TL texts is a complete misnomer. He observes that the translator's cognitive unit (regardless of rank) is actually the *expression* (*Äußerung*) of a conceptual (semantic) content, whereas the chunk the TM retrieves is just a string of matchable characters (*Satz*). Therefore, he proposes that instead of formal *translation units*, we should speak of *retrieval units* (2003: 186). Regardless of how they are defined, these retrieved segments have a coercive effect, prompting the translator to use them as such, even if the conceptual reality of the TL might dictate otherwise. Our discussion will include attempts to resolve this anomaly by integrating methods for including sub-segment identification and processing in the TM workflow.

Translation memory and the CAT in the TEnT

CAT tool environments feature a variety of individual, often interactive, functions, including TMs, terminology management, alignment, analysis, conversion, and knowledge management. These tools may also offer project management features, spellchecking, word and line-counting, machine translation input, and the ability to output a variety of file types. Jost Zetzsche has called this combination of features and resources a *Translation Environment Tool (TEnT)* and adds code protection, batch processing, and code page conversion, among others, to the list of features (2007). Although TMs, termbases, and concordances may be separate programming functions within a TEnT, they are typically integrated at the interface level.

Traditional TM databases are created in three different modes: Interactive generation of <tu>s during the translation of texts, alignment of existing parallel translations, and subsetting of existing aligned resources (Somers and Diaz). <tu>s created interactively become immediately available for intratextual matching during the ensuing translation of the SL text in question. <tu>s are commonly stored together with a set of metadata, such as client, subject matter, and location in the source text. This means that although the full ST is not saved as such, it can be regenerated at any time. These metadata can be used to create subsets of larger TMs, which enables the creation of highly specialized, specific job-related TMs and to validate the appropriateness of specific SL/TL matches. Subordinate portions of a <tu> cannot generally be manipulated or combined, although they do contribute to so-called fuzzy matches.

An alternative approach that stores whole texts with their complete translations is called 'bitext'. When Brian Harris introduced the term in 1988, his description sounded very much like our TMs, which leads many people to use the terms as synonyms. Bowker, for instance, speaks of individual paired segments as bitexts (2002). In the larger sphere of computational linguistics, however, bitexts are complete bilingual or multilingual parallel texts that are aligned with one another. Tiedemann provides a very broad definition: 'A bitext $B=(B_{src}, B_{trg})$ is a pair of texts B_{src} and B_{trg} that correspond to each other in one way or another … Correspondence is treated as a symmetric relation between both halves of a bitext' (Tiedemann 2011: 7).

Sub-segment identification

At the interface, users may not be clearly aware of the differences between TM and bitext, but instead of saving matched, pre-defined text chunks in a database as individual frozen segments,

bitext or full-text, bitext systems employ pattern-matching algorithms to identify matched similar text chunks of any length, which introduces the possibility of working with sub-sentential 'coupled pairs' (Toury 1995) of SL-TL segments. Even in standard TMs, not all segments consist of sentences. Section headings and items in a list are examples of non-sentence segments, as are the isolated text strings that often appear in computer program interfaces. However, short segments are not at all the same thing as sub-segments. Sub-segments are automatically extracted from segments, long or short, without modifying the actual translation memory database.

Sub-segmenting can be thought of as an automatic concordance lookup in which the software automatically chooses which sub-segments to look up within a source segment and presents the <tu>s from the translation-memory database that contain each sub-segment, sometimes highlighting the probable translation(s) of the sub-segment. Clearly, if too many sub-segments are chosen within a source-text segment and presented to the translator, or if there are multiple variations in sub-segment target language matches, the translator can be overwhelmed by the amount of information presented. Another aspect of sub-segmenting is that potentially incorrect <tu>s that are rarely or never displayed using full-segment lookup can often be retrieved during sub-segment lookup, requiring increased maintenance to ensure substantially higher quality translation memory.

Macklovitch *et al.* (2000) describe sub-sentential segmentation that is based essentially on bilingual concordancing within the framework of the Canadian RALI project, while Benito, McTait *et al.* as well as Somers and Diaz, observe that traditional TMs cannot combine fragments from different translation units (<tu>s) to build new target language options based on analysed patterns. A variety of strategies to achieve sub-sentential segmentation have been suggested in addition to concordancing, including morphological analysers associated with example-based MT (EBMT), part-of-speech taggers used in corpus analysis, and closed class word lists designed to enhance the effectiveness of corpus-based bitexts (ibid., Rapp 2002). Interestingly, although many proponents assert the increased recall potentially afforded by sub-sentential segmentation, Gow (2003) shows that efforts to establish statistical comparison between the two styles produce inconclusive results – with standard TM benefiting from greater certainty (matches are often confirmed during creation or editing), but shorter segments providing greater recall, but also additional noise that may actually slow the translation process. Gow concludes that implementers need to weigh their options and choose the system type that best meets their needs. Nevertheless, recent upgrades in sub-segment algorithms and the enthusiasm of some current users may well inspire researchers like Gow to reassess their results.

Advantages of a TM

The ability to retrieve previously translated text enables individual translators to quickly and efficiently create the translation for a revised source text, even when the original was completed long ago. It also enables pairs or groups of translators working in networked environments or over the Internet to collaborate using the same translation memory and terminology resources. New translators can take over in the absence of initial translators and produce consistent results. Furthermore, networking among multiple translators rapidly builds the size of the TM.

An even more dramatic example of the power of translation memory is when a product is delivered simultaneously to multiple markets at the same time that the product is released in its domestic market, in each case with localized – that is, regionally adapted – documentation. This approach to localization is called *simship*, short for *simultaneous shipment*. In order to achieve simship without delaying release of the domestic version, translation into multiple

languages must begin before completing the final version. Based on translation memory, versions with minor changes can be translated very quickly. In this regard, Bowker (2002) distinguishes between *revisions* (texts revised over time) and *updates* (ongoing changes during a production phase).

Translation memory is also useful even when the source text is not a revised version of a previously translated source text, so long as it is part of a collection of related documents that contain segments that are identical or very similar to previously translated segments. Applied appropriately in a translation workflow, TM can significantly enhance efficiency, time to market, and quality (O'Hagan 2009: 48–51). Used without effective quality assurance, however, TM enables the rapid and efficient replication of bad translations, giving rise to a special kind of multilingual 'garbage-in, garbage-out' exercise (GIGO), reflecting the fact that TM promotes *consistency* without necessarily having any effect on *accuracy* (Zerfass 2002).

Bowker (2002: 92–128) and O'Hagan (2009: 48–51) distinguish between *exact matches* and *full matches*. With exact (100 percent) matches, stored segments are identical to the new SL segment not only in string content, but also with regard to any layout features. Context matches (alternatively called *perfect matches* and *in context exact (ICE) matches* by various vendors) also ensure that the immediately preceding segment (or even both the preceding and the following segments) in the new SL is identical to that for the original document, which a growing number of database TMs and bitext systems in general can provide. *Full matches,* in contrast, are almost 100 percent, but may feature minor differences such as punctuation, capitalization, or dates.

The identification of *partial* equivalents supports the retrieval of so-called 'fuzzy matches' in traditional systems. Fuzzy matches are typically assigned a value indicating what percentage of a source segment matches the source segment in a retrieved <tu>. The screenshot in Figure 41.1 shows an example of a 74 percent fuzzy match, which resulted when the phrase '*for example our protagonist Bishop Otto von Bamberg*' was modified in a second version of a previously translated text. This capability is especially powerful for frequently modified texts, such as software manuals and machine instructions. Freigang and Reinke call this value an *Ähnlichkeitswert,* something like a 'coefficient of similarity' (2005; *similarity coefficient* in Macklovitch and Russell), which corresponds to McTait *et al.*'s 'similarity measure' (1999). Typically, the translator can assign a threshold (such as 70 percent) below which a fuzzy match is not displayed, since excessive fuzziness results in noise that slows translation down.

In the translation/localization industry, translator compensation is often influenced by how closely a segment retrieved from translation memory matches the current segment of source text. However, as of this writing, there is no standard method of computing the percentage of match. Some methods are word based and some are character based. This makes it difficult to use the same threshold across multiple tools.

Two other factors can affect fuzzy-match values. One is in-line markup, which consists of tags used to identify layout features such as boldface, underline, or italics. There may also be hypertext links, footnotes, or other embedded controls or operative features. These items may not end up in the same place in the target text, or in some cases it may be desirable to delete them, or possibly to add something new in the target text that is not there in the original. They need to be protected on the one hand, and manipulated on the other so that they will be in the right place in the finished translation. It is not necessarily uncommon for inline markup to vary from one version of a document to the next, in which case just a change in, say, a boldface command, will cause an otherwise identical passage to register as a fuzzy match.

The second kind of item that is often treated using similar strategies is the presence of so-called 'named entities.' These could be people's names or the names of institutions, countries,

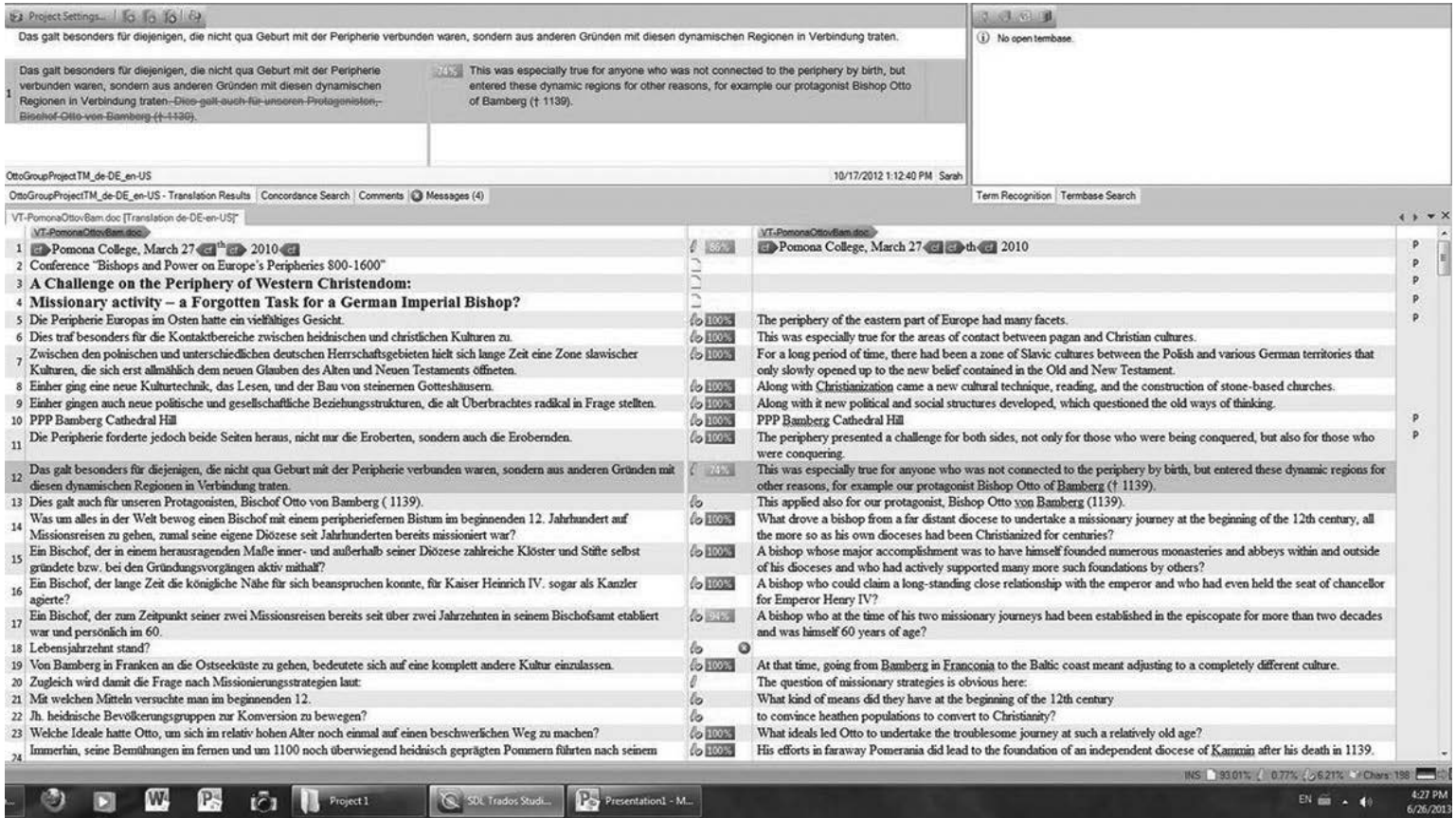


Figure 41.1 Side-by-side segment display

Source: © SDL™ Suite 2011

and the like, but in localization environments, named entities include dates, numbers, currency units, proper names, identifiers, URLs, inline graphics or other region-specific values that can be simply transferred or automatically generated according to set rules. Translators do not usually have to 'translate' these items because they will either not change (in the case of people's names) or the TM will generate the proper TL form, for instance: American English 8/2/2013 (for August 2, 2013) becomes 2.8.2013 in German (for the 2nd of August 2013), which is the standard German order and notation.

Different TM applications and writers use different terms for these variables. Sometimes they are both called 'placeables' because systems typically convert native code to visual markers that appear in the TM editor and can be manipulated (placed) by the translator without breaking the original markup. (See the small pointed icons in Figure 41.1, Line 1.) Azzano calls these elements 'placeables' and 'localizables,' respectively: 'Placeable elements are portions of a document which – in translation – do not change in content. Localizable elements are portions of a document which – in translation – are adapted in content to a target locale according to standard or given rules' (Azzano 2011: 38). Bowker, however, uses the term *placeable*s for Azzano's *localizables,* and refers to them as 'named entities', while Kenny talks about *formatting tags* and *placeables* (Bowker 2004; Kenny 2009), and Macklovitch and Russell dub them *non-translatables*.

Creating, using and maintaining a TM

Although TMs may be used in multilingual computing environments, each individual TM is in most cases bilingual and directional. TMs can be created with general language pairs in mind (for instance, English → Spanish), but many are based on locale-specific SL and TL language code values, such as: en-US → es-ES (US English to Spanish (Iberian) Spanish). This choice has relative advantages and disadvantages. More general resources are more reusable if a translator or translation team encounters texts from many different locales (such as with Spanish from countries all over the Western Hemisphere).

Such TMs may also end up being larger overall because of the range of input, and TM size can accelerate return on investment (ROI) up to a certain point. However, more specific locale codes support the automatic generation of useful localizable entities, and overly large TMs may produce excess noise, require extra maintenance, and waste time. The notion of so-called 'Big Mamma' TMs has inspired such efforts as the TAUS project, where individuals and companies are encouraged to pool their TMs in a single global resource; Linguee, which harvests identifiable <tu>s from across the Web; and MyMemory, which claims to house 'the World's Biggest Translation Memory.' The latter is perhaps unique in that it offers interface capability with a wide range of TM tools, in addition to offering its own API (*All references* 2013). Many experienced translators prefer a mixed solution, maintaining their own universal TM, but actively subsetting it as needed using metadata-based filters for client, product, subject field, and text type classifications.

Regardless of how the <tu>s will be added to the database, the first step in creating a TM is to create the empty database itself. This involves making various choices, depending on the tool used, such as designating segmentation rules and stop lists, which are lists of lexical items or possibly punctuation that may cause problems during segmentation. A typical example might be abbreviations or German ordinal numbers (German *1.* = English *1st*) that end in periods, which may trigger unwanted mid-sentence segmentation.

In typical 'interactive mode,' aligned translation segments are stored in the TM during the initial translation process. One or many texts for translation can be associated with the working

TM, provided that the language direction of the project is also aligned with that of the TM. The greater the repetition factor, the greater the ROI from the TM. Needless to say, interactive mode requires considerable time and effort before reaching a desirable break-even point. In some cases, however, large amounts of legacy translation are available from documents that have been translated without using a TM. Users may even harvest 'Rosetta Stone' translations (existing multilingual versions of authoritative texts) from the Internet to increase recall by leveraging these texts to populate the TM. The segments contained in such legacy documents can be matched up in *batch mode* using alignment tools.

Alignment involves a series of stages: an initial automatic first-pass alignment, followed by a second pass in which a human translator 'corrects' any errors in the machine alignment. Such errors can result when, for instance, one long, possibly complicated sentence in the SL has been split into two, or when two short or repetitive sentences have been combined into one in the TL. Flexible alignment tools allow human editors to move and recombine segments (including whole sentences or paragraphs) as needed to accommodate for segments that may have been moved around during translation. New material can also have been introduced that does not directly correspond to any SL segment (explicitation), or conversely, an SL segment may be intentionally omitted (implicitation) because it is inappropriate to the target audience. Where such null segments occur, matches are most likely to be lost.

The concordance feature, if turned on during TM setup, allows users to search for a specified string wherever it occurs and provides a quick reference to its equivalent (or equivalents if it is not translated the same way in every case). By retrieving information that does not meet the validation criteria for fuzzy matching, concordance search supports translators if they have a hunch that they have seen a word or string before. Concordance lookup is not satisfying, however, for text varieties such as patents, where clauses (but not full sentences) are used intratextually with a high degree of repetition. These cases offer a good argument for sub-sentential matches as discussed earlier.

As the TM grows, the power of TM technology supports a pre-translation phase whereby the system-generated indexing function quickly finds paired translation segments and populates the TL with possible matches even before starting the follow-up human processing stage. Pre-translation can also predict expansion or compression rates based on average relationships for a given language pair or other calculable factors that may occur.

In this way, human knowledge and translation competence have been captured in machine-processible format, but it is important to remember that neither traditional translation memory nor bitext lookup software in any way 'understands' the translation process or the subject matter of the translated text. Nor does the software attempt to parse sentences or negotiate grammatical equivalents as some machine translation systems do, or attempt to do. Furthermore, translators need to carefully review suggestions presented by the TM to ensure they are still valid and that they fit the new micro- and macro-context, a process de Vries calls a 'sanity check' (2002). At this point the translator both enjoys the efficiency of retrieval, but is also obliged to bear the above-cited GIGO principle in mind.

Layout and interface issues

Editor designs vary. In the past, some editors worked on the premise that most texts for translation are created in Microsoft Word™. These systems enabled users to display TM functionality on the same screen while editing the actual translation as a Word document. This approach was attractive insofar as users knew Word and how to use it.

By the same token, however, the interface between the TM software and Word can cause problems, especially if there are even slight differences in software versioning. File corruption is an ever present danger. Furthermore, there is a wide range of other formats that also need to be translated (for instance, PowerPoint™, InDesign™, HTML, as well as standard formats such as DITA and XLIFF). As a result, the current trend is for TMs to provide their own editing environment where most major file formats can be processed by converting native mode text, often to the XLIFF (XML Localization Interchange File Format), which is designed 'to standardize the way localizable data are passed between tools during a localization process' (XLIFF).

Such programs process the global formatting codes accompanying the original file, strip them out and save them as 'stand-off' markup. The translator sees more or less plain text, and placeables appear as placeholders that can be positioned correctly in the target text. This approach helps prevent broken codes, but manipulating placeable markers can itself be challenging and requires additional knowledge from translators, revisers, and reviewers. Other files with special layout requirements, such as PowerPoint slides, can demand careful editing and reformatting in the TL export file because expansion and contraction in text volume can significantly affect slide design.

Separation of data and software

Translators become accustomed to individual TM features and interfaces, such as whether source and target language text are presented in side-by-side tables (as shown in Figure 41.1) or whether they are stacked vertically. Supplemental keyboard and mouse commands used to control program functions require a certain learning curve before users become truly proficient. Furthermore, it is often difficult to get several programs to run correctly on the same machine. Consequently, translators generally work with a limited number of applications. This leads to the need to allow multiple translation team members to use different programs. In such cases, the demand for a universal file exchange format is critical.

Soon after the appearance of multiple TM tools, it became apparent to the owners of translation memory databases that it was desirable to separate the data from the software, so that a TM can be used in a different tool and so that multiple TMs, created using different tools, can be combined. This business need led to the development of the Translation Memory eXchange format (TMX) format, which was developed by the OSCAR special interest group of the now defunct Localisation Industry Standards Association (LISA). TMX is an XML markup formalism that 'allows any tool using translation memories to import and export databases between their own native formats and a common format' as shown in Figure 41.2. This interchange format not only allows people to work together despite different applications and platforms; it also frees up assets so that users can change systems without losing data.

Another obstacle to maximizing the re-use of information in a TM is differing methods of segmentation. If a TM is set up to segment at the sentence or sub-sentence level, and a revised version of the source text is to be translated using a different tool that segments in even just a slightly different way, it will be impossible to fully leverage the data from the original TM. These segmentation differences can keep relevant translation units from being retrieved, or exact matches can be categorized as fuzzy matches. This problem inspired the development of the *Segmentation Rules Exchange Standard* (SRX) for recording and transmitting information on the segmentation method used when creating either TMs or bitexts.

```
<?xml version="1.0" ?>
<tmx version="1.4">
    <header creationtool="XYZTool" creationtoolversion="1.01-023"
    datatype="html" segtype="sentence" adminlang="en-US" srclang="en" o-
    tmf="ABCTransMem">
    </header>
    <body>
        <tu>
            <tuv xml:lang="en">
            <seg>Text in <bpt i="1">&lt;B></bpt>bold
            <ept i="1">&lt;/B></ept>.</seg>
            </tuv>
            <tuv xml:lang="fr"> <seg>Texte en
            <bpt i="1">&lt;B></bpt>gras<ept i="1">&lt;/B></ept>.</seg>
            </tuv>
        </tu>
    </body>
</tmx>
```

Figure 41.2 Sample TMX markup

Special language support

Most TM software was created to support languages that use white space to separate words, that have relatively stable noun and verb forms with easy to manage inflections, and that track from left to right. Arabic, for instance, poses special problems in this regard. If a system is not carefully configured, bidirectional text sometimes results in bizarre behaviors, such as the actual inversion of text so that it reads in mirror image. Arabic morphology, where roots are embedded inside words, as well as orthographic variation, make it very difficult to ensure that real matches will be recognized. These problems have led many Arabic translators to eschew the use of tools.

In similar fashion, some Russian translators argue that terminology components and concordance searches are less useful because they still have a great deal of typing to do in order to account for the many inflectional forms in Russian, although many users have developed compensatory ways of dealing with these languages, such as creating multiple index terms for different forms. Other languages (Chinese and Thai, for instance) pose issues because they lack white space between 'words.' What is sorely needed in this regard is the development of special resources designed to facilitate more comfortable processing in a variety of different languages. Historical discussions often mention problems involving character sets, but these issues have been widely solved by near-universal implementation of Unicode character codes. What has not been widely implemented in TM systems is language-specific morphological processing, including word-boundary identification.

Professional considerations

We have already discussed the advantages and disadvantages of building large global TMs, along with the merits of carefully maintained specialized smaller resources. We have not discussed related copyright issues, client confidentiality, and various knowledge management considerations, all of which frequently dictate careful segregation of resources. Ownership of TMs and termbases is a critical factor in managing CAT tools. Technically, if

translators contract with clients simply to translate text, without mentioning TM or terminology management in the job order for the project, then they may feel they have a claim on resulting resources, even though the translation produced under the agreement belongs to the client as a work made for hire. The argument has been made that translations and TMs are different works. If in doubt, it is wise to contact a copyright attorney for specific advice.

Nevertheless, in today's translation and localization markets, many language service providers (LSPs) control the copyright to TMs by asserting ownership and include statements to this effect in agreements. Many end clients in today's market, however, are increasingly realizing that they need to retain the rights to these resources as valuable intellectual property documenting enterprise knowledge assets. In any event, confidentiality issues may play a bigger role with regard to a translator's relationship to the client than actual ownership of the TM. With the advent of online TM service, where the translator never really has the TM on his or her machine, the question is rendered moot. Nevertheless, other TM owners have bought into a sharing principle as referenced early in the TAUS project, which designers hope will benefit both human and machine translation.

A further consideration that has significantly altered the translation market concerns discounted payment for full and fuzzy matches. Assuming that previously translated text segments should not be double charged, many LSPs and end clients have demanded percentage reductions for leveraged text. This may make sense provided translators enjoy improved productivity adequate to compensate for lost income. The practice is not, however, without its problems. As de Vries points out, TM maintenance itself (such as TM clean-up after subsequent editing outside the TM editor) requires effort, which is traditionally incorporated into hidden overhead, but which is exposed if discounts are imposed.

Fuzzy matches can require significant editing, and even full matches may suggest inappropriate solutions if the overall context has changed. Failure to review the entire translation for macro-contextual coherence results in what Heyn (1998) has called the 'peep-hole' phenomenon, where translators see the text through the aperture of the individual segment. In worst case scenarios, translators only receive compiled lists of new segments for processing, without any view of the greater context. Technology aside, the business practices surrounding implementation of the tools can create environments where pressure to increase productivity at the cost of quality adversely affects professionalism and even erodes translator skills (LeBlanc 2013).

This situation contributes to the commodification of translation services and serves neither the well-being of the translator nor the quality of the final text. It also has produced a two-tier system where 'premium' translators focus on demanding jobs that require a high level of stylistic skill, while others produce 'bulk' work. The former generally turn down jobs involving discounts, while the latter work at bulk rates, producing high volumes. Durban's solution to this apparent payment issue is for translators to charge for their time instead of using traditional by-the-word rates (Durban, unpublished).

History of translation memory

During the early phase of research and development in translation technology (from the early 1950s until the mid-1960s) the focus was on fully automatic translation intended to replace human translators rather than increase their productivity. Probably the most influential early document on machine translation was a 1949 memo by Warren Weaver, who held an important position at the Rockefeller Foundation. The memo was later reprinted (Locke and Booth 1955), but by then it had already had an influence on machine translation. W. John

Hutchins (1999), on the occasion of the fiftieth anniversary of the Weaver memo, published an article about its historical context and significance. According to Hutchins:

> perhaps the most significant outcome of the Weaver memorandum was the decision in 1951 at the Massachusetts Institute of Technology to appoint the logician Yehoshua Bar-Hillel to a research position. Bar-Hillel wrote the first report on the state of the art (Bar-Hillel 1951) and convened the first conference on machine translation in June 1952.

By 1960, Bar-Hillel had soured on the prospects of what he termed FAHQT (fully-automatic high-quality machine translation). In an important (1960) article he concluded that humans would need to be involved but only suggested two possible roles: pre-editor or post-editor of machine translation. The idea of translation memory would come later.

Six years later, the ALPAC report (1966) brought a halt to US government funding for translation technology, even though the report listed 'means for speeding up the human translation process' as its second recommendation (34). The ALPAC report was probably the first widely read publication recommending the use of technology to increase the productivity of human translators. This recommendation marks an important shift from an HAMT (human-assisted machine translation) perspective to an MAHT (machine-assisted human translation) perspective. Appendix 13 of the report shows what looks, at first glance, like translation memory (Figure 41.3*;* note the limitation to uppercase lower ASCII characters).

Despite the apparent TM 'look' in the reproduced printout, these source-target pairs were manually entered in a database, one pair at a time, by human translators to provide context for terms. A search of the database using keywords retrieved information for the purpose of terminology look-up. Although the database was not derived by processing existing source texts and their translations, it can be considered to be a precursor to translation memory. This system was developed for the European Coal and Steel Community in Luxembourg between 1960 and 1965 and thus may be the first hint at modern translation memory.

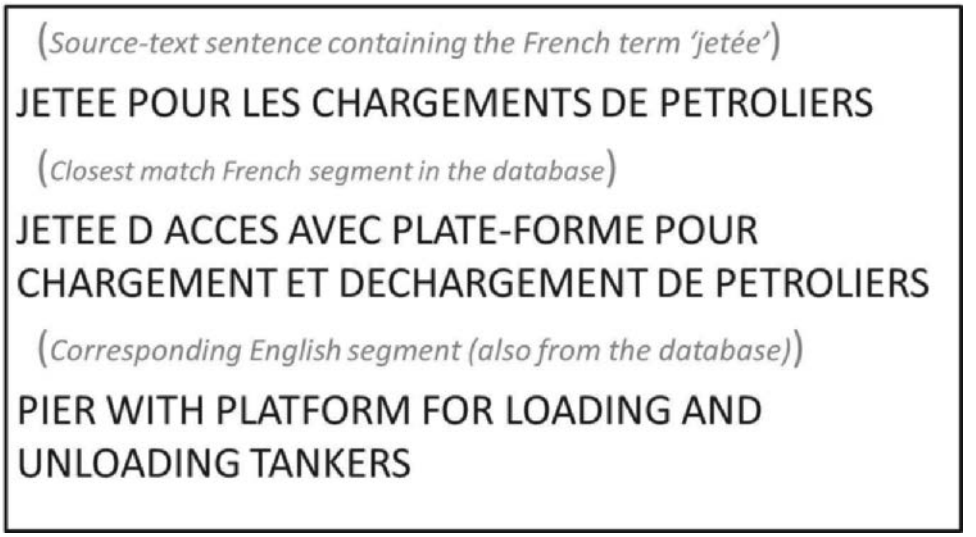


Figure 41.3 Pseudo TM segment from the ALPAC Report

Until the late 1970s, the emphasis for the involvement of human translators with translation technology was on

(a) humans assisting machine translation systems through pre-editing, post-editing, or interactive resolution of ambiguity;
(b) computers assisting humans by providing access to a terminology database;
(c) humans examining a database of previously translated documents to avoid re-translating the same text.

As an example of (c), Friedrich Krollmann (1971) cited the recommendations of the ALPAC report and made several suggestions for supporting translators using technology. One suggestion, a database of entire translations to avoid translating an entire text more than once in a large organization, could be considered to be another precursor to translation memory.

The first published description of translation memory as it has come to be understood was by Peter Arthern. In a 1978 presentation at the ASLIB conference on translating and the computer (published in 1979), he discussed the impact of translation technology on the Council of the European Commission, beginning with the status of machine translation. He then writes that since submitting an abstract for his ASLIB presentation, he realized that 'it is the advent of text-processing systems, not machine translation or even terminology data banks, which is the application of computers which is going to affect professional translators most directly – all of us, freelances and staff translators alike' (1979: 82).

Arthern points out that text-processing systems (currently better known as word processing systems) must be able to transmit and receive files through central computers rather than function as stand-alone machines. The first word processing system for a microcomputer, Electric Pencil (Bergin 2006) had been released in 1976 shortly before Arthern was preparing his presentation, and it was word processing on a general-purpose microcomputer with communication capabilities that led to TM tools for translators.

After a description of terminology management, Arthern presents his newly conceived proposal:

> The pre-requisite for implementing my proposal is that the text-processing system should have a large enough central memory store. If this is available, the proposal is simply that the organization in question should store all the texts it produces in the system's memory, together with their translations into however many languages are required.
>
> This information would have to be stored in such a way that any given portion of text in any of the languages involved can be located immediately, simply from the configuration of the words, without any intermediate coding, together with its translation into any or all of the other languages which the organization employs.
>
> This would mean that, simply by entering the final version of a [source] text for printing, as prepared on the screen at the keyboard terminal, and indicating in which languages translations were required, the system would be instructed to compare the new text, probably sentence by sentence, with all the previously recorded texts prepared in the organization in that language …
>
> Depending on how much of the new original was already in store, the subsequent work on the target language texts would range from the insertion of names and dates in standard letters, through light welding at the seams between discrete passages, to

> the translation of large passages of new text with the aid of a term bank based on the organization's past usage.
>
> (Arthern 1979: 94)

Arthern's proposal is a remarkably accurate description of a modern translation tool with a translation memory function and a terminology look-up function, integrated with a text editor or word processing software.

In 1980, another visionary, Martin Kay, proposed similar features in an internal report to colleagues at Xerox PARC, Palo Alto [California] Research Center, titled 'The Proper Place of Men and Machines in Language Translation.' Kay was probably not aware of Arthern's paper, which had been published only months before. Kay's report was not officially published until 1999, but had a substantial impact years earlier in the 1980s on the community of translation developers who were sympathetic to the idea that post-editing raw machine translation was certainly not the only and probably not the best use of the talents of a professional human translator.

The result of proposals from Arthern, Kay, and others is what is known as a translator workstation, consisting of a personal computer and various software functions, including translation memory, that have been integrated in some way with text processor or word processing. Considerably more detail on the history of the translator workstation is available in Hutchins (1998).

Software development work on commercial implementation of translation memory began in the early 1980s, with the first commercial translation memory system being the ALPS system mentioned by Hutchins (1998). An examination of ALPS corporate documents has determined that an ALPS translator support system with a translation memory function (called 'repetitions processing') was commercially released in 1986. Shortly thereafter (1988), a translation editor called TED was developed for internal use within Trados, then a translation service provider, but not commercialized. It was not until the 1990s that other translation-memory systems, such as Trados and STAR Transit, would be offered to the public.

Most of the features of an integrated translator workstation foreseen in the early 1980s have been implemented in today's translator tools, with a few notable exceptions, such as morphological processing in terminology lookup and automatic 'quality estimation' of machine translation.

Kay (1980) foresaw translator productivity software that would be able to match on the base form of a word. This is, of course, particularly important when doing automatic terminology lookup on highly inflected languages where the form of the term in the source text is not the same as the form of the term in the termbase. As noted above, most current translator tools try to work for all languages and thus have limitations when applied to morphologically complex languages such as Finnish and Arabic.

Melby (1982) foresaw the current trend in translator tools which provide multiple resources for a translator. For each segment of source text, various resources would be consulted, including terminology lookup and a database of previously translated texts. In addition, a segment of raw machine translation would be presented, but only if the machine translation system's self-assessment of quality exceeded a translator-set threshold. Automatic quality estimation (NAACL 2012) of raw machine translation did not achieve a really useful status until the early 2010s, about 30 years after Melby's proposal.

In summary, translation memory emerged from a combination of the availability of word processing and a change in perspective on the role of humans, from assistant to master of a translation technology system.

Future developments and industry impact

When translation memory was introduced to the commercial translation market in the 1980s, all machine translation systems were rule based. In the first decade of the twenty-first century, statistical machine translation (SMT) and example-based machine translation (EBMT) rose to prominence. This led to an unanticipated connection between TM and these new forms of MT. A large TM can provide the training data for SMT and EBMT. This raises the question of where to obtain a very large TM for general-purpose machine translation. The technology currently used in the Internet was around in the 1980s, but it was used primarily by the military and selected universities. Since the 1990s, as the Internet and, in particular, multilingual websites and other on-line multilingual resources have blossomed, it has gradually become more feasible to automatically harvest gigantic TMs from the Internet, at least for those languages with the most online content. All this led to the improvement of the SMT version of Google Translate, for instance. It has also led to additional privacy concerns, since material submitted to Google Translate becomes available to Google for other purposes. Nevertheless, examination of online resources such as Linguee reveals the great danger in simply accepting translation segments taken out of context. Indeed, as more and more online content is produced through machine translation in the first place, automatically generated <tu>s create a kind of questionable vicious circle.

There is a further tension between massive distributed translation memories and careful terminology management. The broader the set of bitexts that are used to create a large TM, the more likely it is that terminological variation will be introduced, originating from multiple text types and domains, as well as from the proliferation of styles representing multiple authors. Especially in combination with sub-segment look-up, such TMs can result in inconsistent terminology across various TM suggestions to a translator as a text is translated. One possibility for addressing this problem is to combine TM look-up and terminology lookup in a more integrated fashion. This kind of integration has not yet appeared in commercial translation tools.

Melby (2006) proposes that data-driven MT may in the future converge with TM technology to improve recall and cut down on fuzziness. EBMT incorporates existing rule-based architectures (Somers) or draws on statistical probabilities that parse linguistic coherence factors in determining probable matches (Simard). This vision of the future implies that coordinating linguistic awareness through MT tools into the current relatively blind look-up capabilities of TM could result in increased coherence, consistency, and accuracy (ibid., also Rapp, Koehn and Senellart).

Creative document writing is the enemy of TM efficiency. The advent of single-source controlled authoring enables not just the reuse of translated segments, but of source language chunks as well, which can be used to generate a number of text varieties, such as advertising brochures, maintenance manuals, product literature, and web pages. Here the same concerns apply as for TM leveraging in general: assembling disparate units to create a new text may not always meet audience needs.

It may be that in the future translation memory will become less visible as a separate function in a translation tool. Information from translation memory, terminology databases, and machine translation may be integrated and presented to the translator as a single suggestion in some cases, refined by automatic or explicit domain identification, with the sources of the integrated suggestion only visible to the translator upon request.

References

ALPAC (1966) *Languages and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC: National Academy of Sciences, National Research Council, 1966.

Arthern, Peter J. (1979) 'Machine Translation and Computerized Terminology Systems: A Translator's Viewpoint', in Mary Snell (ed.) *Translating and the Computer: Proceedings of a Seminar*, London, 14 November 1978, Amsterdam: North Holland, 77–108.

Bar-Hillel, Yehoshua (1960) 'The Present Status of Automatic Translation of Languages', *Advances in Computers* 1: 91–163.

Benito, Daniel (2009) 'Future Trends in Translation Memory', *revista tradumàtica: Traducció I Technologies de a Informació I la Comunicació*, 07 (desembre 2009).

Bergin, Thomas J. (October–December 2006) 'The Origins of Word Processing Software for Personal Computers: 1976–1985', *IEEE Annals of the History of Computing* 28 (4): 32–47. Available at: doi:10.1109/MAHC.2006.76.

Bowker, Lynne (2002) 'Translation-memory Systems', in *Computer-Aided Translation Technology: A Practical Introduction*, Ottawa: University of Ottawa Press, 92–128.

de Vries, Arjen-Sjoerd (2002) 'Getting Full or Fuzzy? The payment issue with full matches generated by translation memory systems', *Language International* (44). http://www.language-international.com.

Dunne, Keiran (2012b) 'Translation Tools', in Carol A. Chapelle (ed.) *The Encyclopedia of Applied Linguistics*, Hoboken, NJ: Wiley-Blackwell Publishing Ltd.

Freigang, Karlheinz and Uwe Reinke (2005) 'Translation-Memory-Systeme in der Softwarelokalisierung', in Detlef Reineke und Klaus-Dirk Schmitz (eds) *Einführung in die Softwarelokalisierung*, Tübingen: Gunter Narr Verlag, 55–72.

Gow, Francie (2003) *Metrics for Evaluating Translation Memory Software*, Ottawa: University of Ottawa, Canada.

Harris, Brian (1988) 'Bi-text, A New Concept in Translation Theory', *Language Monthly* 54: 8–10.

Heyn, Matthias (1998) 'Translation memories: Insights and prospects', *Unity in diversity*, 123–136.

Hutchins, W. John (1998) 'The Origins of the Translator's Workstation', *Machine Translation* 13(4): 287–307.

Hutchins, W. John (1999) 'Warren Weaver Memorandum', July 1949, *MT News International* 22: 5–6, 15.

Kay, Martin (1998/1980) 'The Proper Place of Men and Machines', *Language Translation* 12(1–2): 3–23, Hingham, MA: Kluwer Academic Publishers.

Kenny, Dorothy (2009) 'Unit of Translation', in Mona Baker and Gabriela Saldanha (eds) *Routledge Encyclopedia of Translation Studies* (2nd edn), London and New York: Routledge, 304–306.

Koehn, Philipp and Jean Senellart (2010) 'Convergence of Translation Memory and Statistical Machine Translation', in *AMTA Workshop on MT Research and the Translation Industry*.

Krollmann, Friedrich (1971) 'Linguistic Data Banks and the Technical Translator', *Meta* 16(1–2): 117–124.

LeBlanc, Matthieu (2013) 'Translators on Translation Memory (TM): Results of an Ethnographic Study in Three Translation Services and Agencies', *Translation and Interpreting* 5(2): 1–13.

Linguee: 'Dictionary and Search Engine for 100 Million Translations'. Available at: http://linguee.com.

Macklovitch, Elliott, Michael Simard, and Philippe Langlais (2000) 'TransSearch: A Translation Memory on the World Wide Web', in *Proceedings of LREC 2000*.

Macklovitch, Elliott and Graham Russell (2000) 'What's Been Forgotten in Translation Memory', in *AMTA '00, Proceedings of the 4th Conference of the Association for Machine Translation in the Americas on Envisioning Machine Translation in the Information Future*, London: Springer, 137–146.

Malmkjaer, Kristen (1998) 'Unit of Translation', in Mona Baker and Kirsten Malmkjaer (eds) *Routledge Encyclopedia of Translation Studies* (1st edn), London and New York: Routledge, 286–288.

McTait, Keven, Maive Olohan, and Arturo Trujillo (1999) 'A Building Bocks Approach to Translation Memory', in *Translating and the Computer 21: Proceedings from the ASLIB Conference*, London.

Melby, Alan (1982) 'Multi-level Translation Aids in a Distributed System', in Jan Horecký (ed.) *Proceedings of Coling 1982*, Amsterdam: North Holland.

Melby, Alan (2006) 'MT +TM+QA: The Future Is Yours', *Traducció I Tenologies de la informació I a Comunicació* Numero 4.

MyMemory. Available at: http://mymemory.translated.net/doc.

NAACL (2012) 7th Workshop on Statistical Machine Translation, Shared Task: Quality Estimation, Montreal, Quebec, Canada.

O'Hagan, Minako (2009) 'Computer-aided Translation (CAT)', in Mona Baker and Gabriela Saldanha (eds) *Routledge Encyclopedia of Translation Studies*, (2nd edn), London and New York: Routledge, 48–51.

Rapp, Reinhard (2002) 'A Part-of-speech-based Search Algorithm for Translation Memories', in *Proceedings of LREC 2002*.

Reinke, Uwe (2003) *Translation Memories – Systeme – Konzepte – Linguistische Optimierung (Systems, Concepts, Linguistic Optimization)*, Frankfurt am Main: Peter Lang.

Simard, Michel (2003) 'Translation Spotting for Translation Memories', HLT-NAACL-PARALLEL '03: Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond, 27 May – 1 June 2003, Edmonton, Canada, 3: 65–72.

Somers, Harold L. and Gabriela Fernandez Diaz (2004) 'Translation Memory vs. Example-based MT – What's the Difference?' *International Journal of Translation* 16(2): 5–33.

Tiedemann, Jörg (2011) *Bitext Alignment*, San Rafael, CA: Morgen and Claypool.

Toury, Gideon (1995) *Descriptive Translation Studies and Beyond,* Amsterdam and Philadelphia: John Benjamins.

TMX 1.4b Specification: Translation Memory eXchange format. 2005-04-26. Available at: http://www.gala-global.org/oscarStandards/tmx/tmx14b.html (Original: 1998).

Vinay, Jean-Paul and Jean Darbelnet (1995) *Comparative Stylistics of French and English: A Methodology for Translation* (1958/1995), Juan Sager and Marie-Jo Hamel (trans.), Amsterdam and Philadelphia: John Benjamins.

Weaver, Warren (1949/1955) 'Translation', in William N. Locke and Donald Booth (eds) *Machine Translation of Languages: Fourteen Essays*, Cambridge, MA: Technology Press of the Massachusetts Institute of Technology, 15–23.

XLIFF (XML Localization Interchange File Format). Available at: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xliff.

Zerfass, Angelika (2002) 'Evaluating Translation Memory Systems', in *Proceedings of LREC 2002*.

Zetzsche, Jost (2007) (first edition 2003) *The Translator's Tool Box: A Computer Primer for Translators*, Winchester Bay Oregon: The International Writers' Group.

42

TRANSLATION MANAGEMENT SYSTEMS

Mark Shuttleworth

UNIVERSITY COLLEGE LONDON, THE UNITED KINGDOM

Introduction and definition

Perhaps less well known than some other types of translation technology, computerized Translation Management Systems have been in existence since the late 1990s. This technology, which was introduced in order to enable translation companies or individual translators to remain in control of the ever-increasing volumes of content that they need to process, was designed to facilitate the monitoring of the business, process and language aspects of their translation and localization projects.

Through the reduction or almost complete elimination of manual tasks, using a TMS should permit translation professionals to work more efficiently and to focus on the aspects of their job that require some exercise of judgment. Given that products and services are being constantly developed and improved and new systems continue to appear, this chapter represents a snapshot of the technology as it appeared in summer and autumn 2012. It should also be pointed out that, while the terminology used to refer to different features and functions varies widely from system to system, this chapter seeks to use a nomenclature that is neutral and does not follow the terminology of any one product systematically.

Translation management is defined (as 'translation workflow management') by Chan (2004: 261) in the following terms:

> This is to track and manage the progress of translation projects by the use of workflow management tools. These tools help to keep track of the location of outsourced translations and their due dates, text modifications, translation priorities, and revision dates. The larger the size of the translation project, the more important it is to know the status of the variables.

The following alternative definition is provided by Budin (2005: 103):

> Translation management is a comprehensive concept covering all procedures of the computational management of translation processes by using a broad spectrum of computer tools (including machine translation systems, translation memory systems, term bases, etc.), modeling these processes into operational work flow models, and including economic and human resource management aspects.

While both these definitions envisage a situation whereby translation management functions are performed by a suite of programs, as we shall see, in practice it is usual for many or all of them to be combined within a single application or service that permits the user (who will generally be a senior executive, co-ordinator or project manager within a translation or localization company or language service provider) to control the company's translation assets (i.e. translation memories, terminology and machine translation) and monitor every aspect of workflow:

> These applications orchestrate the business functions, project tasks, process workflows, and language technologies that underpin large-scale translation activity, coordinating the work of many participants in the communications value chain, working inside, outside, and across organizations.
>
> (Sargent and DePalma 2009: 1)

The key advantages brought by this technology include the ability to handle an increased workflow, to accomplish more with less and to manage language service vendors better (ibid.: 20). When combined with the company's content management system (CMS), the result is a seriously enhanced capability in at least the first two of these three areas, as a CMS will help manage the whole content life cycle, from the writing of original material, through editing work and localization to publication to multiple output formats, including, for example, PDF, HTML and RTF. Many of the functions offered by a TMS are made available to a freelance translator within a desktop CAT (or Computer-Aided Translation) tool, there being perhaps something of an overlap between the former and the latter in terms of functionality.

Workflow is described by Rico (2002) as including stages for commissioning, planning, groundwork (term extraction and research, text segmentation and alignment, and text preparation), translation (using translation memory, machine translation and/or localization tools) and wind-up (including consistency check, detection of missing elements, grammar check and testing). However, one might sometimes wish to include other procedures as well, such as desktop publishing engineering, review feedback and so forth (see for example Shaw and Holland, 2009: 112 for an alternative workflow scheme). Great importance is ascribed to workflow simply because defining the step-by-step procedures involved in a complex translation project helps prevent possible errors occurring. It should be added that some workflows are highly complex in nature and involve loops, parallel steps, the skipping of individual steps and the inclusion of multiple transitions in and out of each step (Peris 2012).

TMSs, which can also be referred to as globalization management systems, global content management systems or project management systems, can be commercial (desktop, server-based or 'software-as-a-service'), open source or formed on an *ad hoc* basis by combining a number of different applications if a particular company decides that none of the commercial solutions available suit their particular requirements.

Scope and main functionalities

By way of providing a rationale for this technology, Chamsi (2011: 51–68) tracks the various stages in the development of a new language service provider (LSP), from start-up and growth phase to a greater maturity, in terms of the kind of translation management procedures it is likely to have in place. In the first phase, the system used is likely to be *ad hoc*, and may be even largely paper-based or may consist of a single Excel spreadsheet, although given the relatively low level of workflow such a simple system is likely to be flexible enough for the demands

placed on it (ibid.: 53). Once the LSP starts to experience significant growth, in many cases the original basic translation management procedures will continue to be used, although by this time the system's simplicity and flexibility will both have been seriously compromised (ibid.: 54). If the LSP has not already taken this step, then under pressure from heavy workloads, customer or invoicing issues or reduced profitability (ibid.: 55), sooner or later it will be forced to make the transition to a more robust set of procedures or adopt an off-the-shelf TMS solution. Sargent and DePalma similarly argue that the alternative to adopting a TMS is to 'reach a choke-point in communication with key constituencies' (2009: 1), and observe that it is companies with a high degree of 'digital saturation' that will be likely to opt for an IT-based TMS solution (ibid.: 28; see also Lawlor 2007).

Translation management comes in a variety of forms. Some basic TMS functions are provided by many desktop CAT systems (e.g. SDL Trados, Déjà Vu, MemoQ and OmegaT), while there are now also a number of software-as-a-service (SaaS – or in other words cloud-based) systems that combine translation memory (TM) functionality with some more sophisticated TMS capabilities (e.g. Wordbee and XTM Cloud). Systems that follow the SaaS model have the advantage of requiring no installation and being available in the latest version from anywhere on a 24/7 basis using browser-based access; the implications of the phrase 'cloud-based' are that users access the cluster of software services (known as a 'cloud' because of the cloud-shaped symbol that is normally used to represent them in diagrams) held on the provider's server in the manner described above, and indeed entrust to them their private data, which is held securely on their behalf (see Muegge 2012). Such products are generally licensed via a subscription model, different payment plans being typically available for freelancers and translation companies of different sizes. In addition, a few systems are what are termed as 'captive', or in other words are only available to a company's clients as part of a language service contract (DePalma 2007). There are also more dedicated TMSs that do not offer all the features of a typical CAT tool but that provide heavy-duty support for a wide range of business and project-related tasks. For those who choose to implement their own solution (whether that be by creating a new product or configuring an existing one), this will greatly enhance the system's flexibility while not leading to a significant increase in costs (Stejskal 2011).

Chamsi (2011: 61–62) identifies the following list of basic characteristics that a TMS should possess. Some of these would be typical of any information system and others specific to the activity of project management:

> *simplicity*: the system should not require users to undergo extensive training and should be as intuitive as possible, as regards most areas of functionality at least;
>
> *adaptability/flexibility*: it should be possible for the system to handle a number of different processes in parallel;
>
> *scalability*: the system should be able to grow as the organization also experiences expansion;
>
> *ease and security of access*: since many translation projects are performed by people living in different parts of the world the system should be securely accessible from any location;
>
> *automation of repetitive tasks*: the system should enable project managers to focus on the 'value-added' (ibid.: 62) aspects of their work by freeing them from repetitive tasks as well as helping them to improve their productivity by allowing them to spend less time on each task; by acting as a 'productivity multiplier' (ibid.: 62) in this way it will facilitate the growth of the LSP;

reduction in file management and file transfer overhead: the task of storing and transferring files should be managed automatically so that files can be made available to those who need them with a minimum of effort;

reduction of risks of mistakes: any risk, such as those intrinsic to understanding a project specification or transferring files, should be reduced through the implementation of specific pre-defined processes;

access to relevant data for decision-making: practical information such as details of suppliers, information on clients, project specifications and suchlike should be easily accessible within the system.

Depending on the precise needs of a particular user, some of these characteristics will be essential and others maybe less so.

According to Sargent and DePalma (2009: 6–8; see also Sargent 2012), there are four types of TMS: language-centric, business, enterprise and house. These will be considered below; as will be seen, they are distinguished to a large extent by the sophistication of the workflow management that is built in.

Language-centric

A typical language-centric system consists of tools for project managers, translators, editors and reviewers. While these are likely to be web or server based, some will also offer a desktop translation client. The workflow management will typically only include the translation process itself rather than more peripheral functions such as desktop publishing and testing. Terminology management is also frequently included (Sargent and DePalma 2009: 6–7). This type of system is relatively widespread. It is frequently adopted because of its centralized translation memory capability and its ability to pre-process files for translation (ibid.: 7–8). A TM tool with TMS functionality is only considered a TMS if it is web rather than desktop-based (ibid.: 7). Typically, a language-centric tool can be deployed very quickly (ibid.: 33).

Business

The second type, business systems, comprise project, resource and finance modules and are frequently used in conjunction with a language-centric system. Once again, the workflow capabilities are relatively limited (Sargent and DePalma 2009: 7). Such systems tend to be favoured by LSPs (ibid.) because of the business management capabilities that they offer.

Enterprise

Enterprise systems, the third type that the authors identify, place the emphasis firmly on workflow. They combine the functionalities of language-centric and business types although they can be weaker than business-type systems in the areas of project, resource and finance management. Workflow management is sophisticated, and allows for collaborative multi-vendor scenarios and for non-core processes such as desktop publishing and testing (Sargent and DePalma 2009: 7).

House

Finally, house systems also focus on workflow and collaboration management. In addition to this, they provide clients with logins to enable them to access status reports, and permit online job submission and retrieval (Sargent and DePalma 2009: 7).

Unfortunately, Sargent and DePalma do not provide examples of each system, although in 'Examples of different types of TMS' below the attempt is made to assign each of the three systems discussed to one or other of their categories.

To date, a certain amount of academic research has been carried out into different approaches to translation management. Budin (2005), for example, discusses the possible use of ontologies for translation management purposes (although within the context of his study his understanding of the notion of translation management is in some ways different from that assumed in the remainder of this article). Ontologies are 'formal and explicit specifications of shared conceptualizations of a domain' (ibid.: 103) or in other words conceptual maps, designed for computers and expressed in formal language, of the concepts that exist within a particular subject domain and the interrelations that exist between them. They can be used to provide structure for domain-specific knowledge and texts, and they are already employed in this manner in terminology management. Budin observes that techniques of terminological knowledge modelling and ontology engineering are increasingly being used for the development of translation assets (ibid.: 112). In TMS the use of ontologies would entail the following stages and/or processes (ibid.: 114–118): structuring of knowledge system of a given domain; creation of the term base; using the terminology database as a knowledge base from which to develop an ontology that can subsequently be employed for further terminology management processes in large organizations, including source text analysis, target text production, checking TM consistency, term extraction from corpora and so on (117–118).

History

Individual translators and translation companies have always had to use some method for managing their various procedures, while two kinds of application dedicated to this area, referred to as 'infrastructure' and 'translation workflow and billing management', were envisaged by Melby as early as 1998 as two of eight different categories of translation technology (Melby 1998: 1–2). (The former was to deal with document creation and management, the terminology database and telecommunications, while the latter handled the non-translation-specific logistics of processing translation projects.) Even today, the TMS sector can still be thought of as a 'still-young category of business software' (Sargent and DePalma 2009: 20). A number of systems have a claim to be the first TMS. GlobalSight was one of the first TMSs, having been released also in 1998 (see DePalma 2011). Other early systems include the LTC Organiser and Projetex. SaaS systems first appeared around 2002 (Muegge 2012: 18).

According to Sargent and DePalma (2007), by early 2006 a shift in emphasis was occurring amongst the software vendors, who were repositioning themselves as translation workflow providers rather than competing with CMS producers. In 2007 Lingotek made their system available free of charge (Muegge 2012: 18), while in 2009 Google launched the Translator Toolkit, although this was chiefly intended to improve the quality of output from Google Translate (ibid.).

Common features

Systems vary greatly in terms of the functions that they offer their users. In particular, features offered in translation memory systems, dedicated desktop and server-based systems, and SaaS

systems will differ from each other as in each case a different type of user and corporate environment is generally envisaged. Similarly, many vendors offer different versions of their software for specific groups of user (e.g. freelancer, LSP, large enterprise, etc.). Here differences between versions are likely to involve the availability of a supplier module, possible connectivity to a CMS, the potential for being networked and the provision of access to the system for clients and suppliers. The price differential between the most basic and most sophisticated versions will normally be huge. In addition, it should be pointed out that there are tools that are designed to accomplish a single translation management task (e.g. TMbuilder for TM export and import, SDL Studio GroupShare for group collaboration, globalReview for reviewing and approving translations performed with SDL Trados, one2edit for managing, editing and translating InDesign documents online, and ICanLocalize for running multilingual websites).

There are a wide range of features that different TMSs possess, depending on their level of sophistication. However, all TMSs will include a number of functions that are typically available in desktop CAT tools. These would include items such as file format conversion, text segmentation, formatting tag handling, text alignment, a built-in editing environment and, possibly, pre-translation. In some cases, vendors of TMSs will ensure a close integration with their own CAT tool (as is the case with SDL Wordserver and SDL Trados Studio). In addition, there will be at least a basic kind of workflow management, including some deadline management and mechanisms for controlling translation assets. The system will also allow for connection to an SQL database or other CMS via an API (Application Programming Interface) in order to raise the level of automation in the translation process. It will assign roles within a project (e.g. translator, reviser, desktop publishing expert) to service providers, and the project itself to a project manager, and will handle the submission of translated content electronically. Reports on status, deadlines, costs, workload, quality parameters, etc. can be generated for clients to inspect. The system will analyse files to be translated against project TMs in order to determine the leverage that it will be possible to achieve and where possible will 'pretranslate' – or in other words compare entire source texts with TMs and automatically insert all exact and fuzzy matches that it finds – so that translators are only presented with material that needs to be processed. Some systems also have procedures to manage version control issues.

In addition to what might be termed these 'core' functionalities, each system will typically offer many further features. These will be discussed in the following paragraphs. (It should be noted that not all these features are available in all systems.)

Statistics Statistical information relating to revenues, costs, quality control, etc. can often be generated by the system.

Email Many systems provide email templates and automatic sending of job offers, deadline reminders, payment advices and so forth. Similarly, incoming emails that are identified with a project number can be automatically filed under the appropriate project.

Business Many systems can store templates for purchase orders, job assignments, invoices, quotes, etc. and help to control the budget. Automatic numbering of projects, currency calculations and setting of credit limits for customers are also possible features.

Project management A complete overview of the structure of a project (e.g. in tree-like view) can often be provided. Source language content can be monitored for changes. Freelancers in the database of service providers can be searched according to a range of different criteria (e.g. by job details, tool use, area of specialization or previous performance). Time management support can be offered in the form of different job statuses (e.g. Done, Due today and Overdue).

Customer front end Customers can be offered a log-in in order to enable them to track their project and documentation, upload and download files and access quotes and invoices (e.g. Cloudwords, OTM, GlobalSight).

Freelancer front end Via their own log-ins freelancers can receive job documentation, can upload and download documents, etc.

Crowd-sourcing management Some tools offer support for this new area of translation activity via integration with an open-source CMS, as well as automatic publication of translations. Greater management flexibility is required here as translators need to be able to collect a job that they would like to work on and submit it once they are ready to do so.

Sales Systems can provide sales representatives with a simple means to ensure consistency of pricing, both in general and on a customer-specific level.

Translation quality assessment Tools are usually offered to permit linguistic and formal quality assessment, and job quality and sizing assessment. The tracking of freelancers' performance via indicators enables the ranking when selecting resources for a job. Finally, complaints can be managed and documented.

Standards Most tools should conform to the usual industry standards (e.g. XLIFF, TMX, TBX, SRX and GMX) as well as open standards.

Collaboration Various collaborative tools may be offered, such as collaborative post-editing.

Smartphone version Some systems (e.g. TPBox and LSP.net's OTM) permit project managers to track projects via mobile devices (e.g. iPhone, iPod touch, iPad and Android) in order to help ensure fast responses to customer requests.

The idea of a company connecting a TMS to their CMS via an API has been mentioned above. There are various advantages to this. First, the documents remain secure throughout the translation process. Second, files can be transferred automatically by the system. Third, various different workflows are unified within a single process managed by the software. Fourth, in this way it is possible to exercise strong version control, thus eliminating the risk of selecting an out-of-date version of a file (see MultiCorpora 2011).

Examples of different types of TMS

In this section, as representatives of different kinds of TMS we will be looking at three different systems: Déjà Vu X2 Workgroup (as opposed to Déjà Vu X2 TEAMserver, which possesses a greater range of TMS functions) as an example of a desktop TM tool that offers limited translation management functionality; XTM 7.0 as a typical SaaS-type TMS-cum-TM-tool; and OTM 5.6.6 as a representative of dedicated industry-strength TMSs.

Déjà Vu X2 Workgroup

Like many other TM tools, Déjà Vu X2 Workgroup offers the freelance translator a range of analysis tools. As explained before, as a desktop TM tool with TMS functionality it is not considered to be a TMS.

Déjà Vu X2 Workgroup's main TMS functionalities involve three different areas: project management, file analysis and QA. Regarding the first of these, as is the case with comparable systems, the tool's functionality is relatively limited and is restricted to the process of setting up a new project. Here, besides being able to manage the translation resources, the software user can specify a particular client (from a predefined but editable list) and a particular subject area (from a non-editable list), as shown in Figure 42.1:

Figure 42.1 Specifying client and subject area in Déjà Vu X2 Workgroup

On the other hand, other project management functions such as job allocation or time management cannot be performed.

The second area is that of file analysis, and once again Déjà Vu X2 Workgroup's set of available functions is broadly representative of other similar CAT tools. Here the system will provide the user with statistics regarding repetition within the file ('Duplicates') and matches from the TM ('Guaranteed Matches,' 'Exact Matches,' '95% – 99%,' '85% – 94%,' etc., all the way down to 'No Match').

Finally, regarding QA, Déjà Vu X2 Workgroup will, among other functions, check terminology, numerals and embedded codes, add missing spaces between segments, check warnings and allow the user to process segments by type (e.g. Multiple Matches, Unfinished Fuzzy Matches, Commented Segments, etc.).

By way of brief contrast, OmegaT, as an open-source tool, allows the users to conduct file analysis, but its functionality in other areas is strictly limited.

XTM 7.0

As stated above, this system has been selected as a typical SaaS-type application that combines CAT and TMS functionality. It can be classified as a language-centric system according to the description presented in the section on 'Scope and main functionalities'.

When creating a project the user is presented with the same kinds of choice as was the case with Déjà Vu X2 Workgroup. However, the system not only permits the specifying of a customer, the user can also assign work to freelancers. In addition, while setting up a project XTM 7.0 permits the user to select from a range of pre-defined workflows, as shown in Figure 42.2:

Figure 42.2 Selecting a workflow in XTM 7.0

As would be expected of a language-centric system, the workflow management stops short of peripheral activities, such as testing and desktop publishing.

The system will also produce cost estimates on the basis of word counts. It will also generate different file types – not only TMX and XLIFF, but also PDF and HTML – once the project is completed.

OTM 5.6.6

Finally, we turn our attention to this dedicated TMS, which can be classified as a business system. As we might expect, a wide range of functionality is offered to the user. First of all, the system provides five different interfaces: administration, resource service pages, project management, customer service pages and OTM website.

The administration interface permits the user to create user accounts, to select or modify the settings for corporate identity (e.g. logos, stationery and signature files), to manage resources, to create or edit standard texts (e.g. terms and conditions) and to set a wide range of preferences.

Preferences can be selected for language (for the user interface, for communicating with customers and service providers, and for the website), for project management, and for payments, credit limits, currencies and taxes. In addition, the system can be configured to send automatically generated emails for operations such as quote requests, order confirmations, payment requests and invoice submissions.

The resource service pages interface gives access to job offers and allows users to set specialization parameters. Log-ins are created by the administrator for service providers once their application to work for the company has been approved.

In the project management interface there are a very wide range of options. The user can search for service providers against a range of criteria such as skills (e.g. translation, localization, editorial work and DTP/Typesetting), language pair and area of specialization. The new project creation process allows an entire customer profile to be entered (rather than just a name); the type of service, the volume (in lines, pages, hours, etc.) and the deadline can also be specified. There are also bookkeeping, administration and report-generation sections.

The OTM website allows the user to create a website that acts as a front end for the system, providing an inquiry form for customers, an application form for potential service providers and, if desired, pages for company profile, customer feedback and business terms.

Finally, as stated in the section 'Common features', although this is not a part of the version discussed in this section, a version of this system is available for mobile devices. A sample iPhone screenshot is shown in Figure 42.3:

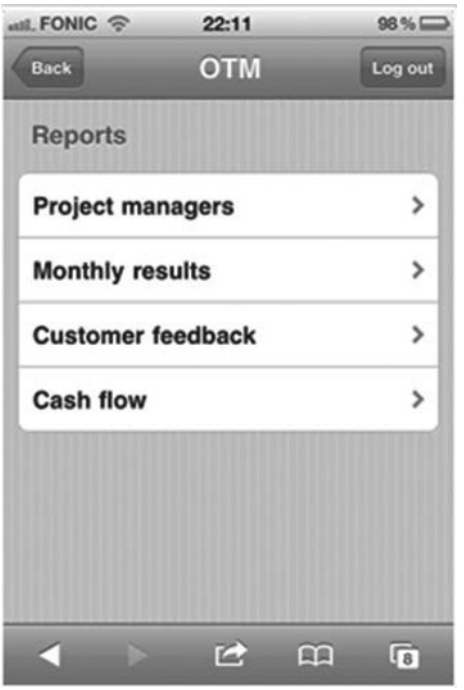


Figure 42.3 Screenshot of the OTM 5.6.6 iPhone interface

Source: image taken from http://www.lsp.net/otm-mobile-devices.html

This interface includes a wide range of functionalities that permit a project manager to stay fully in touch at all times.

Future of the technology

It is important not to underestimate the potential significance of the role that TMSs might play as receptacles of considerable linguistic assets. As pointed out by Sargent and DePalma (2009: 39), it is essential for companies to formulate a well-thought-out corporate policy for capturing and managing terminology and TM assets that can be re-used – 'today, tomorrow, and for many years to come'. As a side effect of this, organizations that succeed in pursuing such a policy and gathering language assets within a TMS 'will also achieve useful MT output sooner than companies that do not' (ibid.). The authors predict that in 5 or 50 years' time it will be impossible to calculate 'how much the accumulation of knowledge and fixed language assets could mean to a commercial or governmental entity' (ibid.: 29).

A few years ago, Sargent (2007: 36) made a number of predictions about the future of the technology, most of which have so far been only partially implemented in the technology:

> a greater level of interoperation with CMSs via a plug-in interface, allowing these systems to order translations within their own environment;
>
> a higher level of tie-in with authoring environments, needed in order to permit authors to select existing segments that have equivalent meanings when there are translated matches within the TM;

more sophisticated resource management functions, to permit 'push-button' generation of purchase orders and invoicing of other cost centres within the company for a wide range of services;

more long term, the need for integration with other enterprise systems that provide process, resource and financial management.

Besides these, it is possible to speculate as follows:

Already of great significance for the translation industry, the role of SaaS and cloud computing in the translation industry is likely to grow considerably over the next few years. How will this new technology affect more traditional workflow structures?

It is possible that crowdsourcing or community translation functionality will become more prominent in systems.

Likewise, it is highly probable that more and more systems will start to offer versions of their software for different kinds of mobile device.

One possible direction of development is for the technology to move towards conformity with the most important project management standards, such as Prince2, TenStep, Six Sigma, PMP, PMI, PMBoK and Agile.

Conclusion

With new systems appearing all the time, in view of all the parameters that have been discussed above it is a potentially confusing matter for the customer to select the one that is most suited to his or her company's requirements. However, the core functions of the technology have been well defined for some years now, even though, as already stated, it is still a relatively young category of business software. As a technology it has a number of clear strong points. First, it is designed to manage complexity within translation projects, greatly facilitating the task of project managers and enabling them to focus on the more creative aspects of their work. Second, it will automate many tasks for project managers and control workflow, with much the same result. Finally, as a repository of linguistic resources it will preserve and accumulate knowledge. On the other hand, there are clear lines of development that are still in progress, with the result that much is still promised by this exciting area of translation technology.

References

Budin, Gerhard (2005) 'Ontology-driven Translation Management', in Helle V. Dam, Jan Engberg, and Heidrun Gerzymisch-Arbogast (eds) *Knowledge Systems and Translation* (Text, Translation, Computational Processing 7), Berlin: Mouton de Gruyter, 103–123.

Chamsi, Alain (2011) 'Selecting Enterprise Project Management Software: More Than Just a Build-or-buy Decision?' in Keiran J. Dunne and Elena S. Dunne (eds) *Translation and Localization Project Management: The Art of the Possible*, Amsterdam and Philadelphia: John Benjamins, 51–68.

Chan, Sin-wai (2004) *A Dictionary of Translation Technology*, Hong Kong: The Chinese University Press.

DePalma, Donald A. (2007) 'Managing Translation for Global Applications', *Galaxy Newsletter* (3). Available at: http://www.gala-global.org/articles/managing-translation-global-applications-0.

DePalma, Donald A. (2011) 'The Attack of the TMS Patents', Lowell, Massachusetts, Common Sense Advisory, Inc. Abstract available at: http://www.commonsenseadvisory.com/AbstractView.aspx?ArticleID=2171

Lawlor, Terry (2007) *Globalization Management Systems: Building a Better Business Case*, White Paper. Available at: http://www.sdl.com/products/translation-management

Melby, Alan (1998) 'Eight Types of Translation Technology'. Available at: http://www.ttt.org/technology/8types.pdf.

Muegge, Uwe (2012) 'The Silent Revolution: Cloud-based Translation Management Systems', *TC World*, July 2012. Available at: http://www.csoftintl.com/Muegge_The_silent_systems.pdf.

MultiCorpora (2011) 'Other Available Products'. Available at: http://www.multicorpora.com/en/products/other-available-products.

Peris, Nick (2012) 'SDL WorldServer: Getting Started with Custom Reports'. Available at: http://localizationlocalisation.wordpress.com/category/translation-management-systems.

Rico, Celia (2002) 'Translation and Project Management', *Translation Journal* 6(4). Available at: http://translationjournal.net/journal/22project.htm.

Sargent, Benjamin B. and Donald A. DePalma (2007) 'How TMS Developers Pitch Their Wares to LSPs', *Galaxy Newsletter* 2007 (4). Available at: http://www.gala-global.org/articles/how-tms-developers-pitch-their-wares-lsps-0.

Sargent, Benjamin B. (2007) 'What's Next for TMS?' *Multilingual* September 2007, 36. Available at: http://www.multilingual.com/articleDetail.php?id=979.

Sargent, Benjamin B. and Donald A. DePalma (2009) *Translation Management Takes Flight: Early Adopters Demonstrate Promise of TMS*, Lowell, Massachusetts: Common Sense Advisory, Inc. Preview available at: http://www.commonsenseadvisory.com/Portals/_default/Knowledgebase/ArticleImages/090331_R_tms_takes_flight_Preview.pdf.

Sargent, Benjamin B. (2012) 'What's a TMS? And Why Isn't It a CMS?' Available at: http://www.softwareceo.com/blog/entry/46120/What-s-a-TMS-And-Why-Isn-t-It-a-CMS.

Shaw, Duncan R. and Christopher P. Holland (2009) 'Strategy, Networks and Systems in the Global Translation Services Market', in Peter H. M. Vervest, Diederik W. van Liere, and Zheng Li (eds) *The Network Experience: New Value from Smart Business Networks*, Berlin, Springer-Verlag, 99–118.

Stejskal, Jiri (2011) 'Translation Management System: Buy It or Build It?' *Translating and the Computer 33*, 17–18 November 2011, London, UK. Available at: http://mt-archive.info/Aslib-2011-Stejskal.pdf.

Appendix: List of translation management systems

This list is not exhaustive but is believed to include at least a representative cross section of systems available during the autumn of 2012. Note that the list does not include desktop CAT tools that offer some limited TMS functionality.

(1) Desktop or server-based

(a) PROPRIETARY

Across Language Server (http://www.across.net) A general-purpose TM-cum-TMS tool available in versions for companies of different sizes and also as a personal edition.

Advanced International Translations Projetex (http://projetex.com) Workstation and Server versions of this tool ensure that both freelance and corporate workflow management are both are catered for.

Asia Online Language Studio (http://www.languagestudio.com/Default.aspx) Lite, Post Editor and Pro versions available for this project management, editing and machine translation environment.

Déjà Vu TEAMserver (http://www.atril.com/en/node/343) A general-purpose tool intended for companies.

Heartsome Translation Studio (http://www.heartsome.net/EN/home.html) A CAT tool that integrates project management functions.

Memsource Server (http://www.memsource.com/translation-server) Solution combining translation memory, termbase and workflow modules.

QuaHill (http://www.quahill.com) A tool available in Basic, Professional and Enterprise versions.

TOM (http://www.jovo-soft.de) Agency, Team and Solito versions available for this tool.

(b) OPEN SOURCE

GlobalSight Translation Management System (http://www.globalsight.com) An open-source system designed to enable companies to manage and localize global content.

(2) SaaS

(a) PROPRIETARY

Cloudwords (http://www.cloudwords.com) A tool for business users.

Google Translator Toolkit (http://translate.google.com/toolkit) An extension of Google Translate, originally launched to improve the output of that online machine translation engine.

iTrac TMS (http://www.merrillbrink.com/itrac.htm) A system for companies and freelancers.

Lionbridge Translation Workspace (https://en-gb.lionbridge.com/LanguageTech.aspx?pageid=1287&LangType=2057) A project and language management product available through the Lionbridge business unit GeoWorkz and based on the concept of 'Live Assets.'

LTC Worx (http://www.langtech.co.uk/en/products/ltc-worx.html) A company-oriented system.

ONTRAM (http://www.andrae-ag.de/index.php) For enterprise translation management; available in Enterprise and Premium versions.

OTM – Online Translation Manager (http://www.lsp.net/online-translation-management.html) A system that controls and manages workflows for translation, editing, localization and desktop publishing.

Plunet BusinessManager (http://www.plunet.com) Available in Team, Corporate and Enterprise editions.

Smartling Translation Management System (http://www.smartling.com/platform) Translation management tool for crowdsourcing and professional translation of websites and mobile apps.

Synble Get Localization (http://about.synble.com/?page_id=28) A localization and crowdsourcing management tool.

Synble Lingodesk (http://www.synble.com) Versions available for higher volume Mobile, .NET, Web and Collaborative projects.

Text United (http://www.textunited.com) A tool aimed at both corporate and individual users.

Transifex (http://www.transifex.com) A workflow and crowdsourcing tool for individual translators and small businesses.

Wordbee (http://www.wordbee.com) A general-purpose TM-cum-TMS tool intended for freelancers, LSPs and businesses.

XTM Cloud (http://xtm-intl.com) A versatile TM-cum-TMS tool that is available in Freelance, Small Group and LSP versions.

YOOmanage (http://www.crosslang.com/en/translation-management/translation-management-systems/yoomanage) A corporate workflow and linguistic assets management system.

Zanata (http://zanata.org) A system for managing localization projects.

(b) OPEN SOURCE

]project-open[(http://www.project-open.com) A general-purpose enterprise resource planning-cum-project management solution that has a module for translation.

(3) Both Desktop/server-based and SaaS versions available

Lingotek Collaborative Translation Platform (http://www.lingotek.com/collaborative_translation_platform) A service that is based on three tiers of translation: automatic, community and professional.

MultiCorpora MultiTrans Prism (http://www.multicorpora.com/en/multitrans-prism) A system that offers management of projects, workflow and translation assets.

Promax (http://www.promax-vpm.com) Project, business and database management system.

SDL TMS (http://www.sdl.com/products/sdl-translation-management-system) One of SDL's two TMS products.

SDL Worldserver (http://www.sdl.com/products/sdl-worldserver) The other of SDL's two TMS products, which is likely to be the one that will be developed in the future.

TPBox (http://www.tpbox.com/cgi-bin/aec.cgi) A general-purpose TMS, available in English, French and German versions.
XTRF (http://www.xtrf.eu/page/overview) A company-oriented system.
YOOprocess (http://www.crosslang.com/en/translation-management/translation-management-systems/yooprocess) An LSP-oriented translation project management system.

(4) 'CAPTIVE' SYSTEMS THAT ARE ONLY AVAILABLE AS PART OF A LANGUAGE SERVICE CONTRACT. EACH OF THESE SYSTEMS IS A GENERAL-PURPOSE TMS THAT OFFERS CLIENTS A WIDE RANGE OF FUNCTIONALITIES.

i plus (http://www.translateplus.com/i-plus.aspx)
Lionbridge Freeway (http://en-gb.lionbridge.com/language-technology/products-technologies/freeway.htm)
Sajan GCMS Translation Management System (http://www.sajan.com)
The bigword Language Director (http://www.thebigword.com/client/ukgov/en/language-director.html)
translations.com GlobalLink (http://www.translations.com/products/technology_globallink_globalization_mananagement.html) [*sic*]

(5) TOOLS THAT PERFORM A SINGLE OR A LIMITED NUMBER OF TRANSLATION MANAGEMENT TASKS

Clay Tablet Platform (http://www.clay-tablet.com/page.asp?intNodeID=910&intPageID=917) A tool that allows any TMS to be linked with any CMS through a range of connectors.
eTranslate Translation Management System (http://www.etranslate.com.au/technology/translation-management-system.asp) A tool that enables web developers to maintain multilingual and international websites.
globalReview (http://www.kaleidoscope.at/English/Software/GlobalReview/globalreview.php) A tool for managing and reviewing SDL Trados projects.
ICanLocalize (http://www.icanlocalize.com/site/services/website-translation) A tool for managing multilingual websites, intended for small businesses.
one2edit (http://www.1io.com) A system that manages, edits and translates InDesign documents.
+Tools (http://www.wordfast.com/products_plustools.html) Includes some TMS functions; intended to work alongside WordFast.
Pootle (http://www.ohloh.net/p/pootle) Manages community-based localization of PO and XLIFF files.
SDL Studio GroupShare (http://www.sdl.com/products/sdl-studio-groupshare) A tool that facilitates collaboration between localization team members.
String (http://ads.gengo.com/string-landing-page) A free tool for creating and managing multilingual websites.
TMbuilder (http://ankudowicz.com/tmbuilder) A tool for creating TM export and import files, free for non-commercial use.
YOOsource (http://www.crosslang.com/en/translation-management/crowdsourcing-platforms/yoo source-translation-crowd-sourcing-platform) A system focusing on crowdsourcing projects.

Acknowledgement

I wish to express my gratitude to Dr Elina Lagoudaki for sharing her voluminous collection of relevant bookmarks with me early on in the project, and also to Common Sense Advisory, Inc., for making the full text of Sargent and DePalma (2009) available to me while I was preparing this chapter.

INDEX