Continuum Advances
in Translation

Corpus-Based Translation Studies

Research and Applications

Edited by
Alet Kruger, Kim Wallmach
and Jeremy Munday

continuum

Corpus-Based Translation Studies

Continuum Advances in Translation Studies
Series Editor: Jeremy Munday is a Reader in Translation Studies at the School of Modern Languages and Cultures, University of Leeds, UK.

Continuum Advances in Translation Studies publishes cutting-edge research in the fields of translation studies. This field has grown in importance in the modern, globalized world, with international translation between languages a daily occurrence. Research into the practices, processes and theory of translation is essential and this series aims to showcase the best in international academic and professional output.

Titles in the Series:

Adaptation, Translation and Transformation
Edited by Laurence Raw and Joanne Collie

Music, Text and Translation
Edited by Helen Julia Minors

Quality In Professional Translation
Jo Drugan

Translation, Humour and Literature: Volume 1
Edited by Delia Chiaro

Translation, Humour and the Media: Volume 2
Edited by Delia Chiaro

Corpus-Based Translation Studies

Research and Applications

Edited by

Alet Kruger, Kim Wallmach and Jeremy Munday

Continuum Advances in Translation Studies

continuum

Continuum International Publishing Group

The Tower Building	80 Maiden Lane
11 York Road	Suite 704
London SE1 7NX	New York NY 10038

www.continuumbooks.com

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

EISBN: 978-1-4411-8919-6

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

Typeset by Newgen Imaging Systems Pvt Ltd, Chennai, India
Printed and bound in Great Britain

Contents

Part III: Studies in Specific Sub-Fields

General Editor's Preface for Advances in Translation Studies Series

The aim of this new series is to provide an outlet for advanced research in the broad interdisciplinary field of translation studies. Consisting of monographs and edited themed collections of the latest work, it should be of particular interest to academics and postgraduate students researching in translation studies and related fields, and also to advanced students studying translation and interpreting modules.

Translation studies has enjoyed huge international growth over recent decades in tandem with the expansion in both the practice of translation globally and in related academic programmes. The understanding of the concept of translation itself has broadened to include not only interlingual but also various forms of intralingual translation. Specialized branches or sub-disciplines have developed for the study of interpreting, audiovisual translation and sign language, among others. Translation studies has also come to embrace a wide range of types of intercultural encounter and transfer, interfacing with disciplines as varied as applied linguistics, comparative literature, computational linguistics, creative writing, cultural studies, gender studies, philosophy, postcolonial studies, sociology, and so on. Each provides a different and valid perspective on translation, and each has its place in this series.

This is an exciting time for translation studies, and the new Continuum Advances in Translation Studies series promises to be an important new plank in the development of the discipline. As General Editor, I look forward to overseeing the publication of important new work that will provide insights into all aspects of the field.

Jeremy Munday
General Editor
University of Leeds, UK

Notes on Contributors

Claudio Bendazzoli has a PhD in Languages, Cultures and Intercultural Communication (Interpreting Studies) from the University of Bologna at Forlì, Department of Interdisciplinary Studies in Translation, Languages and Cultures (SITLeC); his thesis (2010) concerned the creation of a machine-readable simultaneous interpreting corpus of Italian/English speeches from real conferences held in the Italian market. During 2004–2006 he was a Grant Holder of the Scuola Superiore di Studi Umanistici of the University of Bologna for the European Parliament Interpreting Corpus (EPIC) project. He has a degree in Conference Interpreting from the Advanced School of Modern Languages for Interpreters and Translators (SSLMIT) of the University of Bologna at Forlì and works as a freelance interpreter, translator and trainer. His main research interests are corpus-based interpreting studies, theatrical training and interpreter training.

Lynne Bowker holds both a BA and MA in translation from the University of Ottawa (Canada) and a PhD in Language Engineering from the University of Manchester (United Kingdom). She previously taught translation and computational linguistics at Dublin City University (Ireland), and she is currently Associate Professor at the School of Translation and Interpretation at the University of Ottawa (Canada), where her teaching and research interests focus on translation technology, corpus-based studies and terminology. She is author of *Computer-Aided Translation Technology* (University of Ottawa Press, 2002) and co-author of *Working with Specialized Language: A Practical Guide to Using Corpora* (Routledge, 2002). She is also the editor of *Lexicography, Terminology and Translation: Text-Based Studies in Honour of Ingrid Meyer* (University of Ottawa Press, 2006).

Michael Hoey is currently Pro-Vice Chancellor for Internationalisation at the University of Liverpool (United Kingdom) where he also holds the Baines Chair of English Language. He is a member of the Academy of Social Sciences and his books include *On the Surface of Discourse* (1983), *Patterns of Lexis in Text* (1991) (winner of the Duke of Edinburgh English Speaking Union Award for best book in Applied Linguistics), *Textual Interaction* (2001) and *Lexical Priming: A New Theory of Words and Text* (2005) (shortlisted for the British Association of Applied Linguistics prize for best book on Applied Linguistics). He is Chief Consultant to Macmillan Publishers on dictionaries, one of which, *The Macmillan English Dictionary,* won the 2002 Duke of Edinburgh English Speaking Union Award and a 2004 British Council Innovation Award.

Juliane House received her first degree in English, Spanish, translation and international law from Heidelberg University, her B.Ed, MA and PhD in applied linguistics from the University of Toronto, Canada and an honorary doctorate from the University of Jyväskylä, Finland. She is Professor emerita of Applied Linguistics at Hamburg University, and a senior member of the Hamburg Research Centre on Multilingualism. She is also Chair of Applied Linguistics at the Hellenic American University in Athens, Greece and President of the International Association for Translation and Intercultural Studies (IATIS). Her research interests include contrastive pragmatics, discourse analysis, politeness theory, English as a lingua franca, intercultural communication, and translation. Among her latest publications are *Translation* (Oxford University Press, 2009) and *Convergence and Divergence in Language Contact Situations* (John Benjamins, 2009, with K. Braunmueller).

Dorothy Kenny is Senior Lecturer at Dublin City University (Ireland), where she lectures in translation studies, specializing in translation technology and corpus linguistics. Her publications include: *Lexis and Creativity in Translation: A Corpus-Based Study* (St. Jerome, 2001) and the co-edited volumes *Unity in Diversity: Current Trends in Translation Studies* (St. Jerome, 1998) and *Across Boundaries: International Perspectives on Translation Studies* (CSP, 2007). She was co-editor of the annual *Bibliography of Translation Studies* (St. Jerome) between 1998 and 2004. She has also authored numerous refereed articles and book chapters on corpus-based translation studies, computer-aided translation, translator training and translation theory, and sits on the editorial boards of journals such as *Machine Translation* and *The Interpreter and Translator Trainer.*

Alet Kruger taught and trained postgraduate translation students at the University of South Africa in Pretoria for 26 years before taking early retirement at the end of 2007. She is the author of various scholarly articles on translation and corpus-based translation studies and was the editor of a special issue of *Language Matters* – a journal devoted to Studies in the Languages of Africa on 'Corpus-Based Translation Studies: Research and Applications' (vol. 35(1), 2004). She is translation consultant to the Southern African Bible Society, which is involved in a new translation of the Bible into Afrikaans for 2015. She is the co-owner and currently full-time Office Manager of Multilingua Translation/Interpreting Services, a translation and interpreting agency that specializes in the 11 official South African languages. She is also a practising professional translator and among others has translated into Afrikaans *The Coastal Guide of South Africa* (Jacana, 2007), *Find It: Your Guide to Kruger* (Jacana, 2008) and *Exploring Our Provinces* (Jacana, 2009).

Sara Laviosa is Lecturer in English and translation at the University of Bari 'Aldo Moro' and Visiting Lecturer at the University of Rome 'Tor Vergata', Italy. From 1999 to 2002 she was Head of the Italian Section at the School of Languages, University of Salford, UK. She is the author of *Corpus-Based Translation Studies: Theory, Findings, Applications* (Rodopi, 2002) and *Linking Wor(l)ds: Lexis and Grammar for Translation* (Liguori, 2005, 2008). She is editor of the first volume of *Translation Studies Abstracts* (St. Jerome, 1998), *L'Approche Basée sur le Corpus/The Corpus-Based Approach* (special issue of *Meta*, vol. 43(4), 1998) and *Assessment and Accreditation for Languages: The Emerging Consensus?* (CILT, 2000, co-edited with Anke Hübner and Toni Ibarz). Her research interests are in corpus-based translation studies and translation pedagogy.

Saturnino Luz is Lecturer in Computer Science at Trinity College, Dublin. He has collaborated in the Translational English Corpus (TEC) project and in developing and maintaining the TEC software. His research interests include computational linguistics, interaction design, collaborative technologies and information visualization. He is currently a principal investigator of the Centre for Next Generation Localisation (CNGL), coordinating the Systems Framework area, and the principal investigator of a Research Frontiers project aimed at enhancing technological support for collaboration by medical teams at interdisciplinary meetings. He has served in the programme committee of several international conferences and the editorial boards of international journals. He has been a member of the Association for Computing Machinery (ACM) since 1994 and contributes regularly to the *ACM Computing Reviews* journal.

Koliswa Moropa is an Associate Professor in the Department of Linguistics at the University of South Africa, Pretoria and is currently Chair of the Department. She completed the first doctoral study in Southern Africa using corpus-based translation methodology in Xhosa. She has translated over 45 children's books into Xhosa, among others Walt Disney classics such as *The Lion King, The Jungle Book* and *101 Dalmatians*. She has also translated various Bible story books from Afrikaans into Xhosa. In 2003, her translation of *Plays by Zakes Mda* published by Unisa Press was nominated for the South African Translators' Institute's Award for Outstanding Fiction. In 2007, she translated *The Prophet* by Kahlil Gibran – a project sponsored by the National Department of Arts and Culture.

Jeremy Munday is Reader in Translation Studies at the University of Leeds, UK. He is the author of *Introducing Translation Studies* (Routledge, 2001, 2008), *Translation: An Advanced Resource Book* (Routledge, 2004, with Basil Hatim) and *Style and Ideology in Translation* (Routledge, 2007). He is editor of *Translation as Intervention* (Continuum and IATIS, 2007), *Translation and Ideology* (Special issue of *The Translator*, vol. 13(2), 2007, co-edited with Sonia Cunico) and the

Routledge Companion to Translation Studies (Routledge, 2009). His main research interests are in discourse analysis and translation, corpus-based translation studies and Spanish and Latin American writing in translation. He is also a qualified translator and has experience in bilingual lexicography.

Mariachiara Russo graduated in Conference Interpreting from the Advanced School of Modern Languages for Interpreters and Translators (SSLMIT) of the University of Trieste in 1987 and has been a freelance conference interpreter ever since. In 1993, she became Associate Professor at the SSLMIT of Trieste where she taught simultaneous and consecutive interpreting from Spanish into Italian. In 2001, she moved to the SSLMIT of the University of Bologna at Forlì where she also teaches interpreting theory. Since November 2005 she has been Director of the MA Programme in Conference Interpreting. She coordinated the EPIC (European Parliament Interpreting Corpus) project and is on the Scientific Board of *PUENTES* (University of Granada) and *TRANS. Revista de Traductología* (University of Malaga). Her main research fields are aptitude testing, corpus-based interpreting studies and community interpreting.

Gabriela Saldanha is Lecturer in Translation Studies at the Centre for English Language Studies, University of Birmingham, where she teaches on two MA programmes in Translation Studies (distance and campus-programme). Her research has focused on gender-related and personal stylistic features in translation, using corpus linguistics. She is co-editor, together with Mona Baker, of the revised edition of the *Routledge Encyclopedia of Translation Studies* (2010). She is also co-editor of *Translation Studies Abstracts*, an online database published by St Jerome Ltd. (www.stjerome.co.uk/tsaonline). She is a Member of the ECPC (European Parliamentary Comparable and Parallel Corpora) research group (http://www.ecpc.uji.es/).

Annalisa Sandrelli has a degree in Conference Interpreting (English and Spanish) from the Advanced School of Modern Languages for Interpreters and Translators (SSLMIT) of the University of Trieste, Italy. After teaching Italian and liaison interpreting at the University of Hull (1996–2002), she returned to Italy to teach English and Spanish Interpreting at the University of Trieste and the University of Bologna at Forlì (2002–2006). She is currently a Research Fellow in English Language and Translation at LUSPIO University in Rome. She is a member of the International Advisory Board of *The Interpreter and Translator Trainer*. Her main research interests lie in corpus-based interpreting studies, interpreter training and new technologies, legal interpreting and translation, and audiovisual translation (dubbing, subtitling, respeaking and audiodescription). She is also a practising interpreter, translator and subtitler.

Robin Setton has been a professional conference interpreter since 1979, with active English and French and passive German and Chinese and is currently on staff at the Organisation for Economic Co-operation and Development (OECD). He has trained interpreters and designed and led training courses in schools in Europe and the Far East, including Paris (ESIT, Sorbonne), Geneva (ETI), Taipei and currently Shanghai (GIIT/SISU), where he has been responsible for developing a doctoral programme in interpreting studies. He holds a PhD in Applied Linguistics from the Chinese University of Hong Kong and separate postgraduate degrees in conference interpreting, translation, Chinese Studies and Linguistics from the University of Paris (Sorbonne). He is the author of a monograph (*Simultaneous Interpretation: A Cognitive-Pragmatic Analysis*, John Benjamins, 1999) and articles on cognitive, linguistic, cultural and pedagogical aspects of interpreting.

Kim Wallmach is a Senior Lecturer at the University of the Witwatersrand and is currently head of Wits Language School's newest unit, Translation and Interpreting. She has over 15 years' experience in teaching translation and interpreting at university level, as well as over ten years' practical experience in running a translation and interpreting agency. She regularly works as a project manager for simultaneous and consecutive interpreting and translation in the 11 official languages of South Africa and the major European languages. Her current research interests include interpreting/translation and nation-building, translation/interpreting and ideology, sign language interpreting and corpus-based interpreting studies. She holds an MA and PhD in translation studies from the University of the Witwatersrand.

Federico Zanettin is Associate Professor of English Language and Translation at the University of Perugia, Italy. His research interests range from comics in translation to corpus-based translation studies and intercultural communication. His publications include the volumes *Comics in Translation* (2008, editor) and *Corpora in Translator Education* (2003, co-editor) and articles in various journals and edited volumes. He is co-editor of the online database *Translation Studies Abstracts* and the *Bibliography of Translation Studies* and of the journal *inTRAlinea* (University of Bologna), and is a member of the advisory board of the journals *The Translator* and *Entreculturas* (University of Malaga).

Introduction

Alet Kruger, Kim Wallmach and Jeremy Munday

Corpus-based translation studies has come a long way since the ground-breaking work of the early 1990s that began to build on the methods and findings of monolingual corpus linguistics of the 1980s (e.g. Sinclair 1991). John Laffling's research brought together machine translation (MT) and contrastive textology in an analysis of political manifestoes which was remarkably prescient, identifying naturally occurring translation equivalents in (non-translated) parallel texts in German and English (Laffling 1991); two years later, Mona Baker's paper 'Corpus Linguistics and Translation Studies: Implications and Applications' proposed the benefits of corpus-based methods for the emerging discipline of translation studies in order to research central questions such as universals of translated language and procedures such as explicitation (Baker 1993). The late Stig Johansson was also for many years a key figure in the promotion of corpus-based cross-linguistic research (see Johansson 1998).

The intervening years have seen huge advances in corpus linguistics and in the use of corpora not only to create a new paradigm of statistical MT systems (starting from the seminal work of Brown et al. 1993) but also for wider purposes, such as identifying gaps in the coverage of rule-based MT systems (Babych and Hartley 2009). Furthermore, a keen interest has emerged in the use of corpora for language learning and translator training (e.g. Bowker and Pearson 2002; Zanettin et al. 2003). Like Olohan (2004), however, the main focus of the current volume is firmly on a third strand: on how corpus-based studies have assisted, and may continue to assist, translation studies researchers in the investigation of key questions of translation and interpreting.

The articles in this volume are written by many of the leading international figures in the field. They provide an overall view of developments in corpus-based translation (and interpreting) studies and also specific case studies of how the methodology is employed in specific scenarios, such as contrastive studies, terminology research and stylistics. A corpus here is understood as a collection of texts in electronic form, enabling speedy interrogation by software specially designed to analyse linguistic and other features (cf. Kenny 2009: 59). These corpora may be monolingual or multilingual and may amount to millions (even billions) of words. Indeed, the whole World Wide Web may be considered a corpus (see Luz, Chapter 5, below). Corpus-based translation (and interpreting) studies therefore are studies of translation (and interpreting) that are based on the analysis of electronically prepared corpora of texts. What is striking in the diversity of these studies is a common methodology and the number of interrelated issues and observations.

The first section of the book presents core concepts and tools. The first two chapters are particularly salient. **Sara Laviosa** begins by exploring the antecedents of corpus-based translation studies as well as its future. This critical survey extends and updates the earlier state of the art covered in her book *Corpus-based Translation Studies: Theory, Findings, Applications* (Laviosa 2002). When the idea of investigating translation and interpreting through corpora was first put forward by Baker (1993), it was envisaged that in this new partnership corpus linguistics would provide the methodology for carrying out empirical investigations while translation theory would identify the areas of enquiry and elaborate operational hypotheses. The two partners would work in harmony mainly for the benefit of the advancement of the descriptive branch of the discipline. Since then corpus-based translation studies (CTS) has embraced descriptive and applied studies in a wide range of different languages. In this chapter Laviosa examines, in the light of recent developments, what type of relationship holds between CTS and descriptive translation studies (DTS), on the one hand, and CTS and corpus linguistics on the other. She seeks to establish which claims and predictions put forward in the past still hold true and which are the most promising areas for long-term CTS research.

On the other hand, research into interpreting is considered either a subdiscipline of the umbrella term ‘translation studies’ or a (semi-)autonomous discipline in its own right (Pöchhacker 2009). **Robin Setton**’s contribution in this volume fills an important gap in research by providing a very detailed survey of the field of corpus-based interpreting studies (CIS). Perhaps naturally, the number of such studies is fewer because of the inherent difficulties associated with accessing recordings and preparing transcriptions or video material. However, Setton’s chapter stands out by the comprehensiveness of its coverage, describing both the evolution and latest developments in the field. Its length, exceeding all others, contributes to the new and welcome prominence of interpreting studies and usefully prepares the way for Bendazzoli et al.’s analysis of the rich corpus of European Parliament interpreting recordings (see Chapter 12, and below).

Chapter 3, by **Dorothy Kenny**, centres on the key concept of the **unit of translation**. Kenny points out that, although it is a basic concept, it has received only scant attention in corpus-based translation studies to date, despite the fact that parallel corpora are likely to contain a wealth of data on translation units. In her view, corpus linguistics may offer new, as yet barely tapped, ways of looking at translation units. She begins by reviewing existing approaches to translation units, focusing in turn on comparative stylistics, process and product-oriented descriptive translation studies, and natural language processing, before making her own suggestions for identifying translation units in parallel corpora, illustrated by her own corpus of German literary translations.

Although this volume focuses on the qualitative translation perspective, the stretching of theoretical boundaries in recent years has naturally been accompanied by a corresponding 'hardwiring' of corpus-based translation studies. Thus, in their chapters Federico Zanettin and Saturnino Luz examine the different stages in the creation of a corpus and the benefits of different types of encoding, vital in order to make the most of corpus-based resources in descriptive translation studies. **Federico Zanettin** describes the multiple layers of annotation based on XML/TEI standards. His examples are taken from CEXI, a bidirectional parallel English-Italian corpus hosted by the School for Translators and Interpreters of the University of Bologna. **Saturnino Luz** similarly presents and discusses recent advances in tools, technologies and standards that aim to form an infrastructure for creating and sharing dynamic, and widely accessible, corpora. Luz points out that researchers will necessarily have to focus on corpus software and tools that use the Web as a communication medium. As an example, Luz describes the model adopted by the Translational English Corpus (TEC) at the University of Manchester.

This is clearly an area that is undergoing rapid and exciting change. The many applied benefits of technological advances in corpus-based methods are demonstrated, amongst others, within the University of Leeds Centre for Translation Studies, where several strands of CTS research are being pursued in individual and collaborative research projects across many languages. New work with European partners uses parallel and comparable corpora to create MT systems for under-resourced languages, such as Croatian and Latvian. This builds on ASSIST, which uses very large comparable (rather than parallel) corpora of English and Russian news texts to enable human translators to find translations for rare or novel general-language expressions not yet in a dictionary (Sharoff et al. 2009). In addition, the automatic identification of the genre of texts harvested automatically from the Web in various languages has far-reaching benefits for work on human and machine translation and language-learning (Sharoff 2010). Another current project, IntelliText, provides a simple, integrated interface for carrying out a range of searches and statistical analyses on large, publicly available corpora in several languages (Wilson et al. 2010). This promises to be of great future benefit to CTS and other humanities researchers. Furthermore, in closely focused studies, Peng (2009) created an annotated corpus of interpreting to compare the relative coherence of expert and trainee interpreters working into English and Chinese; and coherence between text and image in multimodal, multilingual documents is captured by Thomas (2009a, 2009b), who provides both a theoretical framework and annotation software to analyse packaging designed for English- and Chinese-speaking consumers.

The second part of the volume, entitled 'Methods for the qualitative analysis of contrastive patterns in large corpora', presents methods for the analysis of key pragmatic and discourse features. **Michael Hoey**'s chapter extends his own concept of lexical priming for the first time explicitly to translation. According to Hoey (2005), as words are acquired through encounters with them in talk and writing, they become loaded with the collocational, colligational, semantic, pragmatic and textual contexts in which they are encountered, which can be called their 'primings'. This means that the description of a word is a much more complex matter than simple definition and usage notes. To discover to what extent it might be a problem in translation, Hoey analyses the translation into Portuguese of the first sentence of a book by Bill Bryson, translated by three Portuguese postgraduate students. The English original and the Portuguese translations are analysed in terms of the primings (colligations, etc.) of the component vocabulary of each text, making use of reference corpora of English and Portuguese. The comparison highlights a number of important differences in the typical primings of specific lexical items in the source and target languages that may cause shifts of emphasis and connotation in translation. The implication is that there is a whole new class of 'false friends' – words whose primings differ radically even though they appear to have the same denotation.

Jeremy Munday's chapter is linked to Hoey's concept of lexical priming. It seeks to explore the phenomenon of semantic prosody in English and Spanish. Semantic prosody is defined by Louw (1993: 157) as 'a consistent aura of meaning with which a form is imbued by its collocates'. That is, certain words, such as *cause*, which might normally be thought of as neutral, tend to occur in negative or positive contexts and seem to accrue such values from surrounding collocates (Stubbs 1996). Munday uses as an example the English *loom large* and one of its dictionary equivalents, the Spanish *cernerse.* Comparable reference corpora, the *British National Corpus for English* and the *Real Academia Corpus of Spanish*, are used to test the methodology and to examine the kinds of questions which such an analysis raises. The results again encompass collocational and colligational patterns, as well as authorial evaluation and stance.

Juliane House substantiates her distinction between two main types of translation: 'overt', or source-text oriented translation, and 'covert' translation, culturally filtered to fit target text conventions (House 1977/1981 and 1997). These are some of the foundational concepts on which modern translation studies is based. She shows how corpus-based methods can be fruitfully used to investigate – both qualitatively and quantitatively – the role that global English has come to play today in initiating and propelling language change via language contact in translation and multilingual text production.

House also addresses some more general methodological issues arising from such corpus-based work, suggesting that corpus-based translation studies can fruitfully be combined with qualitative and quantitative analysis, thus averting the danger of treating text/discourse as context-free objects. In her view, a translation studies corpus should be seen only as *one* of many tools of scientific inquiry, a methodological basis for pursuing translation research. Regardless of frequency and representativeness, corpus data are useful because they are often better data than those derived from accidental introspections. But if the use of corpora is to maximally fulfil its potential, it should be used in conjunction with other tools, i.e. introspection, observation, textual and ethnographic analysis.

The chapters by Hoey, Munday and House all show a common concern with pushing the boundaries of corpus-based studies to include context. This trend points the way to future work into the issues of phraseology and discourse in CTS.

The third and final section of the book comprises studies in specific subfields, namely terminology studies, stylistics, translation universals and simultaneous interpreting. In her chapter, **Lynne Bowker** assesses the impact that corpus-based resources and tools have had on terminological research. For Bowker, corpus-based resources and tools include not only collections of electronic texts that can be interrogated with the help of tools such as concordancers and word listers, but also pairs of texts that have been aligned and stored in specially designed databases for processing by tools such as translation memory systems or automatic term extraction tools. She makes the point that it is difficult to know what the long-term effects of new technologies will be when they are first introduced into a profession. Electronic corpora and associated tools for corpus processing have now been in fairly widespread use in the translation profession for over a decade, and it would seem to be a good time to look back at the use of these tools, and to reflect on the changes that they have brought about. By assessing and understanding these changes, Bowker believes that we will be in a better position to develop strategies for the future, such as designing new types of resources for translators or making modifications to the curricula of translator training programs to reflect the new reality of the translation profession.

Translational stylistics has evolved into its own specialism in CTS. **Gabriela Saldanha**'s contribution in this volume points out that the notion of style has traditionally been associated exclusively with 'original', i.e. non-translated, texts. However, recent work in translation studies has started to look at style from the perspective of the target text, particularly Baker (2000), Malmkjær (2003), Bosseaux (2007) and Munday (2008). Saldanha draws on Baker's (2000) argument that it is possible to conduct a corpus-based study to identify the translators' own stylistic preferences in presenting her own case study of the use of foreign lexical items in literary translations from Spanish and

Portuguese into English by Margaret Jull Costa and Peter Bush. Her results suggest that the way in which the two translators treat source culture lexical items is a distinctive characteristic of their translation style.

Koliswa Moropa's chapter reflects one of the first attempts to apply corpus-based translation research to an African language. Xhosa is an agglutinating language spoken in South Africa, which makes corpus research particularly challenging, as phrases and even whole sentences are written as one word. With the advent of a multilingual language policy in post-apartheid South Africa, there has been a significant increase in the demand for the translation of hegemonic languages (English and Afrikaans) into the previously marginalized languages (Xhosa, Zulu, Ndebele, Swati, Northern Sotho, Southern Sotho, Tswana, Tsonga and Venda) and vice versa. Although the volume of translation into African languages has increased substantially, translated texts, due to a lack of terminology, are not of a uniformly acceptable standard. This has prompted a number of scholars to use parallel corpora as part of their research on terminology and standardization, as well as on typical translation strategies such as simplification and explicitation. The interest of Moropa's chapter is the link between simplification and explicitation and how far this may be affected by the concordial system of Xhosa.

The difficulties associated with developing spoken corpora large enough to yield stable analytical results have meant that corpus linguistics has tended to focus on the analysis of written discourse. However, as Setton emphasizes in Chapter 2, spoken corpora can provide a particularly valuable resource for both quantitative and qualitative analyses of specific pragmatic functions and thus help generate new empirical insights into patterns of language usage. Of course, one key difference between written and spoken corpus analysis is that a transcribed spoken corpus is a mediated record, a textual rendering of an event which is multimodal in nature, and thus the transcribed spoken corpus may capture only a limited and limiting aspect of the reality of that event (Knight and Adolphs 2007). This is changing with the advent of multimodal corpora, which allow for a fuller representation of the context and co-text in which human communication takes place. A multimodal environment which includes both gesture and prosodic features of language makes it possible to study pragmatic functions in a more complete light (see Baldry and Thibault 2006; Baldry 2007).

The chapter by **Bendazzoli**, **Sandrelli** and **Russo** demonstrates some of the opportunities of this type of research, in their case in simultaneous interpreting. They conduct a specific case study, presenting an analysis of disfluencies in the European Parliament Interpreting Corpus (EPIC), an electronic corpus of speeches in English, Italian and Spanish and their simultaneous interpretations into the same languages, in all possible combinations and directions. The focus of their chapter is on two specific types of disfluencies, truncated words and mispronounced words, produced by both speakers and interpreters. After automatically extracting and calculating the relevant occurrences in the sub-corpora, the main features of these disfluencies are described and their possible causes are discussed. The specific nature of EPIC has enabled the researchers to carry out both comparable and parallel analyses of the material. That is, they are able to compare disfluencies occurring in the sub-corpora of original speeches delivered in Italian, English and Spanish with disfluencies in the interpreted speeches in the same language (comparable analysis); and to compare disfluencies in each sub-corpus of original speeches with those occurring in the corresponding two interpreted versions available in EPIC (parallel analysis).

The EPIC corpus is noteworthy not only because it is one of the first representative collections of authentic interpreted text, but also because it is a multidimensional tool comprising video, audio and machine-readable written transcripts. The multimedia archive comprises digital video and audio clips of both original and interpreted speeches and transcriptions of the recorded material. The EPIC corpus transcripts are fully machine-readable, which means that it is possible to exploit its potential as a parallel or a comparable corpus. The creation of a multilingual parallel corpus of interpreted speeches and their corresponding source speeches also offers the opportunity of comparing more interpretings of the same text, something which is not often possible, as pointed out by Kalina (1994: 227). Future alignment of the corpus will allow more refined studies on quality features, specific interpreting strategies and possible differences depending on language pair and language direction.

In conclusion, all these studies on corpus-based translation studies lead away from a narrow source-text/target-text comparison and towards a broader approach, into areas such as prosody, phraseology, pragmatics, stylistics, discourse and contrastive linguistics. House (this volume) aptly sums up this direction as follows:

> In the last analysis, the object of corpus-based translation studies should not be the explanation of what is present in the corpus, but the understanding of translation. The aim of a corpus is not to limit the data to an allegedly representative sample, but to provide a framework for finding out what sort of questions should be asked about translation and about language used in different ways.

References

Babych, Bogdan and Anthony Hartley (2009) 'Automated Error Analysis for Multiword Expressions', *Linguistica Antverpiensia New Series* 8: 81–104.

Baker, Mona (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications', in Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair,* Amsterdam: John Benjamins, 233–50.

— (2000) 'Towards a Methodology for Investigating the Style of a Literary Translator', *Target* 12(2): 241–66.

Baldry, Anthony (2007) 'The Role of Multimodal Concordancers in Multimodal Corpus Linguistics', in Terry Royce and Wendy Bowcher (eds) *New Directions In The Analysis Of Multimodal Discourse,* New Jersey: Erlbaum, 173–93.

Baldry, Anthony and Paul J. Thibault (2006): *Multimodal Transcription and Text Analysis,* London: Equinox.

Bosseaux, Charlotte (2007) *How Does It Feel? Point of View in Translation*, Amsterdam and Philadelphia: Rodopi.

Bowker, Lynne and Jennifer Pearson (2002) *Working With Specialized Language: A Practical Guide to Using Corpora*, London and New York: Routledge.

Brown, Peter, Stephen Della Pietra, Vincent Della Pietra and Robert Mercer (1993) 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational Linguistics*, 19(2): 263–311.

Hoey, Michael (2005) *Lexical Priming: A New Theory of Words and Language*, London: Routledge.

House, Juliane (1977/1981) *A Model for Translation Quality Assessment,* Tübingen: Gunther Narr.

— (1997) *Translation Quality Assessment*: *A Model Revisited,* Tübingen: Gunther Narr.

Johansson, Stig (1998) 'On the Role of Corpora in Cross-linguistic Research', in Stig Johansson and Signe Oksefjell (eds) *Corpora and Cross-Linguistic Research: Theory, Method and Case Studies,* Amsterdam: Rodopi, 3–24.

Kalina, Sylvia (1994) 'Analyzing Interpreters' Performance: Methods and Problems', in Cay Dollerup and Annette Lindegaard (eds) *Teaching Translation and Interpreting 2: Insights, Aims, Visions*, Amsterdam: John Benjamins, 225–32.

Kenny, Dorothy (2009) 'Corpora', in Mona Baker and Gabriela Saldanha (eds) *Routledge Encyclopedia of Translation Studies,* 2nd edn, London and New York: Routledge, 59–62.

Knight, Dawn and Svenja Adophs (2007) **'**Multi-Modal Corpus Pragmatics: The Case of Active Listenership', in J. Romeo (ed.), *Corpus and Pragmatics*, Berlin: Mouton de Gruyter.

Laffling, John (1991) *Towards High Precision Machine Translation: Based on Contrastive Textology*, Berlin and New York: Foris Publications.

Laviosa, Sara (2002) *Corpus-Based Translation Studies: Theory, Findings, Applications,* Amsterdam: Rodopi.

Louw, Bill (1993) 'Irony in the Text or Insincerity in the Writer: The Diagnostic Potential of Semantic Prosodies', in Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds) *Text and Technology*: *In Honour of John Sinclair,* Amsterdam: John Benjamins, 157–76.

Malmkjær, Kirsten (2003) 'What Happened to God and the Angels: An Exercise in Translational Stylistics', *Target* 15(1): 37–58.

Munday, Jeremy (2008) *Style and Ideology in Translation: Latin American Writing in English,* New York: Routledge.

Olohan, Maeve (2004) *Introducing Corpora in Translation Studies*, London and New York: Routledge.

Peng, Gracie (2009) 'Using Rhetorical Structure Theory (RST) to Describe the Development of Coherence in Interpreting Trainees', *Interpreting* 11(2): 216–43.

Pöchhacker, Franz (2009) 'Interpreting Studies', in Jeremy Munday (ed.) *The Routledge Companion to Translation Studies*, Abingdon and New York: Routledge, 128–40.

Sharoff, Serge (2010) 'In the Garden and in the Jungle: Comparing Genres in the BNC and Internet', in Alexander Meheler, Serge Sharoff and Maria Santini (eds) *Genres on the Web: Computational Models and Empirical Studies,* Berlin and New York: Springer, 149–66.

Sharoff, Serge Bogdan Babych and Anthony Hartley (2009) '"Irrefragable Answers": Using Comparable Corpora to Retrieve Translation Equivalents', *Language Resources and Evaluation* 43: 15–25.

Sinclair, John (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Stubbs, Michael (1996) *Text and Corpus Analysis*, Oxford: Blackwell.

Thomas, Martin (2009a) 'Localizing Pack Messages: A Framework for Corpus-Based Cross-Cultural Multimodal Analysis'. Unpublished PhD Thesis, University of Leeds.

— (2009b) 'Developing Multimodal Texture', in Eija Ventola and Moya Guijarro (eds) *The World Told and the World Shown: Multisemiotic Issues*, London: Palgrave Macmillan, 39–55.

Wilson, James, Anthony Hartley, Serge Sharoff and Paul Stephenson (2010) 'Advanced Corpus Solutions for Humanities Researchers', *Proceedings of the Workshop on Advanced Corpus Solutions, Pacific Asia Conference on Language, Information and Computation PACLIC24,* Sendai, Japan, November 2010.

Zanettin, Federico, Silvia Bernardini and Dominic Stewart (eds) (2003) *Corpora in Translation Education*, Manchester: St. Jerome.

Part I

Core Concepts and Tools

Chapter 1

Corpus-Based Translation Studies: Where Does It Come From? Where Is It Going?

Sara Laviosa

In this chapter, a map of corpus-based translation studies (CTS) is drawn, followed by a critical reflection on its achievements and the identification of new avenues of enquiry for the future. The chapter is organized into three chronological sections, each corresponding to a salient moment in the evolution of CTS: the dawn of corpus-based translation studies (1993–1995); the establishment of corpora in translation studies (1996–1999); the spread of corpora across languages and cultures (from 2000 onwards).

1.1 The Dawn of Corpus-Based Translation Studies

Mona Baker published her seminal paper, 'Corpus Linguistics and Translation Studies: Implications and Applications' in 1993 as part of a collection of research articles in honour of John Sinclair (Baker et al. 1993). 'The availability of large corpora of both original and translated text, together with the development of a corpus-driven methodology', Baker predicted, 'will enable translation scholars to uncover the nature of translated text as a mediated communicative event' (Baker 1993: 243). Two years later, in 'Corpora in Translation Studies: An Overview and Some Suggestions for Future Research', published in *Target,* that original idea was further developed by suggesting specific research projects involving the design and analysis of parallel, bi/multilingual and, above all, monolingual comparable corpora. It was the dawn of a new partnership that led 'to new ways of looking at translation' (Kenny 2009: 53) during the 1990s. This decade was characterized by a myriad of competing and complementary theoretical approaches and methodologies grown out of the cross-fertilization with new fields of studies as varied as pragmatics, critical linguistics, post-colonialism, gender studies and globalization. In the meantime, well-established conceptual paradigms, such as polysystem theory, skopos theory, and poststructuralist and feminist approaches to translation theory, continued to enliven translation research (Venuti 2000/2004).

This novel research agenda was launched at a time of unprecedented growth in corpus linguistics, whose innovative techniques for language observation, analysis and recording had given rise to 'a new perspective on description' (Sinclair 1991: 2) and 'a new way of thinking about language' (Leech 1992: 106). Mega corpora of English of no fewer than 100 million words were being compiled, such as the *British National Corpus* (BNC), *Cambridge Language Survey*, *Longman Corpus Network* and *Bank of English*, the last-mentioned counting over 300 million words at the time. New corpus types were being designed, the interactive *Corpus of Spoken American English*, for example, allowed the simultaneous presentation of visual and auditory information in each concordance line (Chafe et al. 1991). Also of significance was the fact that the Council of Europe was planning large-scale projects aimed at creating reference corpora for the languages of the Union.[1]

Many areas of study in applied linguistics were being influenced by the insights and methodology of corpus linguistics: lexicography first of all, then educational linguistics, natural language processing (NLP), machine translation (MT), computer-assisted translation (CAT), contrastive analysis, terminology, forensic linguistics and critical linguistics, to name just the principal ones. With such an impressive record of achievements the timing was right for predicting that corpus linguistics would make a triumphant entry also into translation studies. Of course, corpora were not unknown to the discipline when Baker put forward her proposals. In fact, at the University of Lund, Gellerstam (1986) had already compiled the first monolingual comparable corpus of Swedish novels to study translationese and Lindquist (1989) had investigated the Swedish renderings of English adverbials with a parallel language database. Their research intended using corpora as aids to improve the practice of translation; it therefore found its place within the applied branch of the discipline. Instead, what Baker proposed in the early 1990s was a composite programme of research conceived within Descriptive Translation Studies (DTS). The strong links forged in those years between corpus linguistics and DTS, which where underlain by a set of common concerns stemming from a descriptive, functional and empirical perspective, is, in my view, one of the keys, if not *the* key, to the success story of CTS. What are these shared issues? First, the object of study consists of authentic samples of language use rather than idealized entities; linguistic regularities are regarded as probabilistic norms of behaviour rather than prescriptive rules; language patterns reflect and reproduce culture. Moreover, both corpus linguistics and DTS adopt a comparative research model in which descriptive hypotheses that make claims about the probabilistic generality of a given phenomenon are put forward, and texts are examined across corpora representing different language varieties, for example, translated versus non-translated language, original texts and their translations, different text types or different modalities within the same language, and so on. The same empirical paradigm therefore embraces the target-oriented, historical-descriptive approach developed by Toury from polysystem theory and the corpus-linguistic descriptive approach advocated by Baker.

Furthermore, corpus-linguistic analytical procedures together with corpus-design principles were largely compatible with Toury's (1995: 36–9) discovery and justification procedures involving an inductive and helical progression from observable translational phenomena to the non-observable and culturally determined norms that govern translators' choices. In the mid-1990s, the methodologies foreseen by Toury for DTS and Baker for CTS were at the stage of being developed and tested. Both scholars stressed the importance of developing a coherent descriptive methodology that would allow researchers to compare results, replicate studies and systematically widen the scope of research into the nature of translation. In the years that followed, CTS developed and fruitfully utilized a methodology not only for the advancement of DTS but also in applied translation studies.

1.2 The Establishment of Corpora in Descriptive Translation Studies

In 1996, the first CTS analysis was carried out at the University of Manchester (Laviosa-Braithwaite 1996). As part of that study a multi-source-language, monolingual comparable corpus of English was created. It offered a synthesis between a corpus linguistic methodology and the investigation of simplification, a line of research pursued within DTS in the 1980s. The findings revealed core patterns of lexical use in narrative prose and newspaper articles which were largely independent of the influence of the source language and could be regarded as aspects of simplification in translational English. Almost at the same time and right up to the end of the decade other novel syntheses were proposed.

Munday (1997) combined systemic functional linguistics, corpora, cultural studies and reception theory to analyse translation norms in a parallel corpus of Spanish short stories by Gabriel García Márquez and their English translations. The findings obtained from comparative analyses of the target and source texts *vis-à-vis* English and Spanish reference corpora, suggested that the initial norm characterizing the translator's choices was orientated towards acceptability. Later on, the exploration of the 'third code' (Frawley 1984) inspired two studies of normalization. The first is Scott's analysis (1998) of the novel *A hora da estrela* by Clarice Lispector and its translation, *The Hour of the Star,* carried out by Giovanni Pontiero. The second study, by Kenny (1999), examined lexical norms and creativity in a two-million-word parallel corpus of contemporary German literary texts and their English translations (GEPCOLT). Scott (1998) looked in particular at how the repetition of the negative type *não* had been translated and discovered two kinds of normalization, one linked to the systemic differences between the source and target language, the other resulting from the translator's stylistic preferences. However, although normalization was found to be a feature of translation in GEPCOLT, occurring in 44 per cent of cases where translators had to deal with creative hapax legomena, most of the time, creative lexis was not normalized, which shows that 'normalization is far from an automatic response to lexical creativity in source texts' (Kenny 2001: 210). The co-occurrence of regularities and instances of counter-examples to prevailing patterns in the investigation of universals is also a feature of Laviosa-Braithwaite's study where simplification was by no means uniform in all text types represented in the *English Comparable Corpus* or for all the parameters considered.

This is also the case with Øverås' (1998) research, which tested Blum-Kulka's (1986) explicitation hypothesis in a sub-corpus of literary translations taken from the *English Norwegian Parallel Corpus* (ENPC). Her interpretive hypothesis was that a rise in the level of cohesion is an aspect of explicitation in the process of translation; the descriptive hypothesis was that English and Norwegian target texts are more cohesive than their source texts. Both predictions were largely confirmed since the explicitation shifts outnumbered the implicitation strategies, notwithstanding a lower level of explicitation in Norwegian–English translations. In addition to the constraints inherent in the mediating process of translation, a variety of factors were hypothesized as good candidates for explaining this feature of translation in follow-up studies. These are the stylistic preferences of the source and target language, their systemic differences as well as culture-bound translation norms, which, as Weissbrod (1992) points out, are amenable to change with historical circumstances and according to the relative position of translation in different literary systems. Moreover, explicitation was found to be associated with the tendency to prefer typical rather than unusual collocations, which suggests that the universals of translation may present distinctive features but may also overlap with one another to some extent.

It seems that for every question answered, many more are being thrown up, thus giving rise to a serendipity process of continual quest, discovery and explanatory hypotheses. The question that comes to mind at this point is: how shall we unravel the intricate maze of what is norm-dependent and what is universal, what is specific and what is general, what is central and what is peripheral? Shall we follow Toury's (1995: 259) lead and aim to uncover the laws of translational behaviour through the gradual refinement of compatible methods of enquiry, theoretical speculation and the accumulation of data yielded by studies relating to different languages, sociocultural milieus and periods of history? Or shall we go along with Tymoczko's (1998: 656) view that 'the primary purpose of CTS is neither to be objective nor to uncover universal laws' but to build 'many different corpora for specialized, multifarious purposes, making room for the interests, inquiries and perspectives of a diverse world'? Perhaps this is a false dilemma since the two positions are in all probability much less apart than they appear to be if we examine them in greater depth. If we regard the notion of translation universal not as an absolute principle that can explain translation strategies in every single circumstance, but as a descriptive construct, an open-ended working hypothesis about "similarities, regularities, patterns that are shared between particular cases or groups of cases" (Chesterman 2004: 33), it will, I believe, unveil more and more aspects of the variegated nature of translation products and processes and their complex relationships with culture. It will also allow us 'to see both similarities and differences in a perspective that increases our understanding of the whole picture, and also of how this picture relates to other pictures' (Chesterman 2004: 33).

The translational laws put forward by Toury (1995: 259–79) constitute an eminent example of such hypotheses. They are not deterministic propositions, but conditioned, probabilistic explanations intended to tie together particular modes of translational behaviour and the vast array of variables that impinge on them to a lesser or greater extent (or not at all) in different conditions – linguistic, sociocultural or cognitive (Toury 2004). For example, according to Toury's (1995: 268–74) law of growing standardization, the special textual relations created in the source text, such as creative collocations, are often replaced with conventional relations in the target text (such as habitual or typical collocations), which leads to the dissolution of the original set of textual relations. The operation of the law is influenced by factors such as age, extent of bilingualism, the knowledge and experience of the translator as well as the status of translation within the target culture, so that the more peripheral the status of translation, the more it will accommodate itself to established models and repertoires in the target language.

The law of growing standardization coexists with the law of interference, whereby 'phenomena pertaining to the make-up of the source text tend to be transferred to the target text' (Toury 1995: 275). Interference, "which is a kind of *default*" (Toury 1995: 275), is in turn conditioned by factors such as the professional experience of the translators, the status of the source language or the prestige value assigned to different text types in the target language, so that technical translation, for instance, may be less affected compared with literary translations (Toury 1995: 274–9). The two laws are not totally unconnected; they are part of an intricate system (Toury 1995: 274), which systematic diachronic and synchronic research that is methodologically sound and firmly grounded in theory can gradually unravel, thus unveiling the specificity and regularities, the diversity and uniformity of translational phenomena across languages and cultures.

Corpus research into translation universals aimed to play a part in this wide-ranging and long-term plan by drawing on the insights of previous studies and moving forward with small-scale endeavours that had the potential to be followed up and extended. Therefore, their significance lies largely in the fact that they built upon, refined and diversified the work of scholars such as Blum-Kulka and Levenston (1983), Klaudy (1996), Shlesinger (1989, 1991, 1995), Toury (1985, 1991) and Vanderauwera (1985). They were able to do so thanks to the development of a coherent methodology, clear, explicit descriptive and interpretive hypotheses as well as a consistent comparative research model. In keeping with Baker's agenda, CTS contributed to bringing forward the state of the art in DTS through the study of universals, which can be considered, in line with Toury, as 'one of the most powerful tools we have had so far for going beyond the individual and the norm-governed' (Toury 2004: 29), without denying 'the existence or importance of that which is unique in each particular case' (Chesterman 2004: 33).

Right up to the end of the 1990s, Baker continued to be at the forefront of lively theoretical debates. From 1996 to 1999, she wrote three articles which had an influential role in strengthening the collaboration between CTS and DTS (Baker 1996, 1998, 1999). She provided guidelines on how to refine hypotheses and methodology so as to render operational and verifiable abstract concepts such as the notion of translation universals. Moreover, the search for the patterns that identify translation qua translation, argued Baker, should go hand in hand with the assessment of the relative status of source and target languages, that is, prestigious versus less prestigious ones. Also, given that translated texts are distinctive communicative events, shaped by their own goals, pressures and contexts of production, descriptive scholars were encouraged to focus on the interplay of three elements: readership expectations, theoretical pronouncements and professional practice. These can be studied by complementing textual analyses with the investigation of extralinguistic sources of data such as historical information, book reviews, interviews of authors and translators, trends revealed by the output of publishing companies and decisions taken by funding bodies.

Still from a theoretical stance, other novel syntheses were proposed: Halverson (1998) adopted 'prototypical categories' for defining the object of study in corpus investigations and resolving the impasse created by two contradicting statements. On the one hand, the legitimate data for empirical and theoretical research was claimed to consist of any translation that is 'presented or regarded as such within the target culture, on whatever grounds' (Toury 1985: 20) and, on the other, professional translations were assumed to enjoy a higher status, mainly on the basis of evidence from psycholinguistic studies. She suggested that the target parent population of translated works be regarded as a prototype category whose centre is taken up, but only for cultures of industrialized Western countries, by professional translations, whereas in the periphery there are clusters of different types of translation, for example, those carried out by translator trainees, or those performed in the foreign language. This means that a corpus intended to be representative of the population of translated texts would consist of an array of sub-corpora presenting differing degrees of relevance but all being regarded as legitimate objects of investigation. Just as prototypes are culture-bound so are the corpora designed to represent a given parent population.

This raises the thorny issue of the comparability of the object of study and the consequent generalization of individual research findings. It is a recurrent problem in CTS, one that has come up time and time again in the compilation of bilingual and monolingual corpora. In a country such as Brazil, for example, where about 90 per cent of all published literature is translated literature, it would be problematic to design a representative and balanced monolingual comparable corpus of narrative texts using the same criteria adopted for the creation of the *English Comparable Corpus* (ECC) (Magalhães 2001). In less common languages this difficulty is not restricted to literary genres but concerns general language use too. Many non-literary text types in Irish Gaelic, for instance, are translations, mainly from English, as Kenny (1998) points out. The influence of translation policies affected the design of the *Corpus of Translated Finnish* (CTF) compiled at Savonlinna School of Translation Studies under the direction of Anna Mauranen. Academic texts in the natural sciences were excluded since this text category is not translated into Finnish. The problem interests also bidirectional parallel corpora. In designing the *English-Norwegian Parallel Corpus* (ENPC), the choice of texts was limited by the fact that many text types were translated into Norwegian, but significantly fewer ones into English (Johansson and Hofland 1994; Johansson 1998).

This type of imbalance was also encountered during the design stage of the CEXI project, which envisaged the creation of a bidirectional parallel corpus of English and Italian, under the direction of Guy Aston at the University of Bologna in Forlì. The problem, explained in detail by Zanettin (2002), derived from the very different composition of the English and Italian parent populations of translated narrative and non-fiction works. This mismatch impacted (a) the representativeness of the original sub-corpora; and (b) the level of comparability of the translational and the original components of the corpus as a whole. This, in turn, would have limited the types of comparative analyses that were intended to be carried out. There is therefore always a trade-off between balance and comparability, on the one hand, and representativeness on the other. The problem may be solved by means of a compromise: balance may be obtained in a core corpus, while representativeness may be reached by flanking the core corpus with unidirectional parallel sub-corpora which better mirror the composition of the translation parent populations. So internal balance, representativeness and comparability can be obtained to a reasonable extent in corpus design with a bit of manoeuvring between what is given and what is taken as objects of study in a particular sociocultural environment. Comparability across cultures is a much more complicated matter, of course, but one can at least achieve consensus on the design principles and make them explicit, so that different research communities are able to mutually understand the rationale for particular decisions and the implications of each other's findings.

Still on the subject of corpus design, Shlesinger and Malmkjær put forward proposals and reflected on this aspect of the methodology. Shlesinger (1998) suggested unveiling the distinctive features of interpreting *vis-à-vis* written translation and original spoken discourse through a new type of monolingual comparable corpus which would include interpreted speeches from a variety of source languages, original spoken texts produced in comparable settings and written translations of oral source texts produced in similar settings. Malmkjær (1998) recommended a particular type of parallel corpus, one which would include as many different translations of the same source text as possible to yield data useful to the scholars interested in the study of equivalences and those who focus on the phenomenon of translation per se. This is because the traditional unidirectional parallel corpus may hide an important aspect of the translational process, namely the specific choices and strategies adopted by different translators.

1.3 The Establishment of Corpora in Applied Translation Studies

Applied CTS took off slightly later compared with descriptive studies and then grew fairly rapidly. Its beginning can be traced back to Gellerstam (1986) and Lindquist (1989), as stated earlier, but it is really from the end of the 1990s onwards that we can really talk of a growing body of research in this area. Applied corpus studies of translation forged strong links with contrastive analysis, language for specific purposes (LSP), foreign language teaching, terminology, lexicography and computational linguistics. At the core of corpus-based pedagogy were the design and navigation of corpora created not only as sources for the retrieval of translation equivalents or as aids for improving the quality and efficiency of the final translation product, but also as repositories of data used to better understand translation processes and language behaviour, from a monolingual and a contrastive perspective. Corpus-based teaching methods drew on Data-Driven Learning, which was developed by Johns (1991) in foreign language pedagogy. Within this student-centred perspective tutees act as researchers in as far as they identify problem areas, suggest descriptive hypotheses and then test them in cooperation with their tutor who assumes the role of facilitator in the learning process. Gavioli and Zanettin applied this methodology in the undergraduate translation classroom, where learners designed, compiled and analysed a comparable Italian–English corpus of medical research articles in order to acquire: (a) content knowledge about a specific subject field; (b) textual knowledge concerning the overall structure of medical research articles; (c) linguistic knowledge about the typical use of specialized terms (e.g. acronyms, names of tests or viruses) and words in specialized contexts, such as contrastive markers, modal verbs and tentative verbs. The analytical tools were frequency lists and keyword in context (KWIC) concordance lines through which the selected lexical items were investigated semantically, syntactically, collocationally and pragmatically in order to improve the understanding of medical terms obscure to non-expert readers as well as to identify the most accurate equivalents at the level of lexis and discourse units (Gavioli, 1997, 1999; Gavioli and Zanettin 2000). Similar techniques were used by Bowker (1998) in an experimental study which tested the effectiveness of using a target language monolingual corpus of specialized texts as an aid to improve subject field understanding, correct term choice and fluency.

So at the end of the 1990s the overall picture looks like this: within the empirical paradigm whose development in the early 1990s can be regarded, in line with Chesterman (1998), as the most important trend that characterizes translation studies, a number of novel syntheses in the pure and applied branches of the discipline were proposed and realized with corpus linguistic methods. On the whole, they were received with interest by the scholarly community. Here are some notable comments made about the new trend: Tymoczko (1998: 652, 658) regards CTS as 'central to the way that Translation Studies as a discipline will remain vital and move forward' and 'an opportunity to reengage the theoretical and pragmatic branches of Translation Studies, branches which over and over again tend to disassociate, developing slippage and even gulfs'. Hatim (1999) also praises CTS when he claims that corpus-based translation studies is a truly new wave of research providing it does not limit itself to studying only what is *in* translated text but also what is *of* translation, that is, its ideological impact.

At the end of this initial period of intense scholarly work, three main areas of development for the new millennium can be identified. These are: first, the search for common ground between linguistics and the rapidly developing interdisciplinary field of cultural studies; second, an awareness of ideology as a factor indissolubly intertwined with text, context, translation as well as the theory, practice and pedagogy of translation; and, finally, keeping pace with the development of modern technologies in order continually to update, refine and diversify the methodologies adopted in descriptive and applied studies (Tymoczko 1998: 657).

1.4 The Spread of Corpora across Languages and Cultures

We can share Venuti's (2000: 334, 2004: 326) perception that at the start of the new millennium, translation studies 'is an international network of scholarly communities who construct research and debate across conceptual and disciplinary divisions'. This insight is consistent with the intention expressed in the same year by Mona Baker, Theo Hermans and Maeve Olohan to open and focus the scholarly debate on three important issues: (i) comparing and contrasting the variety of research models elaborated by the different approaches and theories of translation; (ii) their relationship with existing paradigms; (iii) the extent to which they can be applied across the wide range of phenomena considered to be legitimate data for the discipline. These were the main themes of the first international conference devoted to 'Research Models in Translation Studies', held in Manchester, in April 2000. It brought to light not only the spread of methods of testing or developing theories or producing or exploring new data – the very definition of research models put forward by Chesterman (2000) – but the conference also revealed some important developments that were taking place in corpus studies of translation. These were later enhanced by two international conferences entirely devoted to corpus-based translation studies. The first one, 'Corpus-Based Translation Studies: Research and Applications', was held in Pretoria, in July 2003. The second one, 'Conference and Workshop on Corpora and Translation Studies', took place in Shanghai, in March 2007. To what extent do these advances substantiate Tymoczko's (1998: 657) claim that the appeal of CTS lies in its potential 'to illuminate both similarity and difference and to investigate in a manageable form the particulars of language-specific phenomena of many different languages and cultures'? First of all, let us examine descriptive CTS, which has grown considerably over the last decade or so, thanks to the creation of new corpora in many different countries (cf. Xiao 2006; Anderman and Rogers 2008). The availability of these multilingual resources has enabled descriptive translation scholars to interface with neighbouring disciplines, most notably, contrastive linguistics and lexicography, as testified by the themes addressed by the biennial international conference series, *Using Corpora in Contrastive and Translation Studies* (UCCTS). This initiative is intended to provide an international forum for the exploration of theoretical and practical concerns in contrastive and translation studies. The first conference was held at Zhejiang University in Hangzhou, China, on 25–27 September 2008 (see Xiao 2010), the second, jointly organized by Edge Hill University, the University of Bologna and Beijing Foreign Studies University, took place at Edge Hill University in Ormskirk, UK, on 27–29 July 2010.

Thanks to the spread of corpora, the quest for universals is now being pursued in multilingual settings, despite ongoing debate on the tenability of this notion as a valid concept for describing and explaining the specificity of translational language (cf. Mauranen and Kujamäki 2004; Klaudy 2009; Laviosa 2009; Xiao 2010). These studies demonstrate, through empirical and interdisciplinary research that 'translations are texts of a particular, specific kind, which reflect the complex cognitive processes and the particular social contexts from which they arise' (Mauranen 2008: 45). They sometimes show lexical and structural trace of language contact, while at other times they may under-represent features unique to the TL, over-represent elements that are less common in the TL or display a tendency towards conservative or conventional target language use (ibid.).

In addition to the study of regularities in translational behaviour, scholars have also focused in recent years on the particulars of culture-specific phenomena. For example, Baker (2000) examined the style of two literary translators represented in the TEC, namely Peter Bush and Peter Clark, where 'style' comprises the translator's choices regarding the type of works to translate, the consistent use of specific strategies as well as the use of prefaces, footnotes or glossaries. With this investigation, Baker outlines a framework within which the linguistic choices made by the translator are linked to extralinguistic aspects of the process of translation such as the relative status of source and target language, the distance between source and target culture, the translator's choices regarding themes and literary genres as well as the professional status of the translator. Stylistic variation is also at the heart of Kruger's research (2000). Drawing on Biber's (1988) approach to spoken and written register variation, she examined a cluster of 12 co-occurring linguistic features indicative of 'involved production' in a diachronic parallel corpus including Afrikaans translations of *The Merchant of Venice*. The study revealed a tendency towards a more oral, more involved and more situated style in a recent stage translation versus an older page translation. Other studies of the translator's voice in literature were carried out by Bosseaux (2007), Kenny (2001), Saldanha (2004, 2005) and Winters (2005).

Ideology is emerging as a distinctive theme in descriptive CTS. Kemppanen's (2000, 2001, 2004) study, for example, is based on a *Comparable Corpus of History Texts*, which consists of Russian–Finnish translations and original Finnish texts on Finnish political history. Most of the collected works were published in the 1970s when the Marxist-oriented approach to the scholarly study of Finnish history was highly valued in a political climate which favoured and encouraged Finnish–Soviet relations. Kemppanen applies A. J. Greimas' model of narrative structures to investigate two semantic fields identified in the translational sub-corpus: 'friendship and co-operation' and 'class consciousness'. The word *ystävyys* (friendship) in particular displays a positive semantic prosody in translated texts versus a negative one in original works. Moreover, in translation, friendship is portrayed as an aim to be achieved by Finland and the Soviet Union working in harmony, while in original Finnish, the Soviet Union is portrayed as an opponent. Another example of how the analysis of the sociocultural context can be fruitfully integrated with corpus data is Munday's (2002) investigation of an article by Gabriel García Márquez about a 6-year-old Cuban boy who was rescued while trying to reach the USA in 1999. The three English translations published in the *Guardian* and *New York Times* newspapers as well as by a Cuban group, *Gramma International* were analysed within a systemic functional linguistic framework. This revealed important shifts in the translation process, which were accounted for by linking the metafunctional profiles of source and target texts to their relative sociocultural contexts. For instance, the anti-USA feelings expressed in the original were relayed differently in the target texts and in the *New York Times*, they were largely omitted.

The investigation of Anglicisms through corpus linguistic methods combined with other tools, such as textual and ethnographic analysis (e.g. Baumgarten et al. 2004; House 2007; Laviosa 2010a, 2010b; House, this volume, Chapter 8) has recently attracted the attention of translation scholars, particularly in Europe, where the harmonization of a national and a transnational identity is closely linked to the issues of multilingualism and mutual comprehensibility. Owing to the status of English as a global lingua franca, these studies largely assume that translation is a mediator of language change induced by English source texts, as a result of the process known as 'negative transfer', which is more readily tolerated 'when translation is carried out from a "major" or highly prestigious language/culture, especially if the target language/culture is "minor", or "weak" in any other sense' (Toury 1995: 278). Yet, the empirical evidence is far from consistent, since there seems to be considerable variation across target languages, domain-specific discourses, text types and even different types of Anglicisms at different levels of linguistic analysis (cf. Anderman and Rogers 2005). It is, therefore, still an open issue whether translation plays a significant role in the process of Anglicization of the European languages and may be legitimately regarded as the main means of importing new linguistic trends vis-à-vis other forms of language contact.[2]

These recent lines of enquiry demonstrate that corpus scholars are increasingly becoming aware that '[d]escription is not enough. It has to serve a purpose, such as explanation' (Hermans 1999: 103). In order to achieve this objective they are beginning to contextualize the phenomena investigated and becoming more alert to the wealth of ideas and research tools offered by neighbouring disciplines, as recommended by several scholars (Tymoczko 1998; Olohan 2004; House, this volume, Chapter 8).

Finally, in applied CTS, corpora are being widely and systematically used in translator training, particularly in the 'translation praxis classroom' (Kiraly 2000, 2003), where they constitute valuable resources for retrieving and examining lexical, terminological, phraseological, syntactical and stylistic equivalents (Bowker and Pearson 2002; Bowker 2003; Koby and Baer 2003). They are also employed to acquire subject-specific knowledge, to evaluate translation quality, and as essential components of Computer-Aided Translation Technology (Bowker 2000; 2001; Koby and Baer 2003). At the same time, corpora are making inroads into translator education (cf. Zanettin et al. 2003), which comprises language teaching at an advanced level for trainee translators (Bernardini 2000, 2002, 2004a, 2004b) and translation-based teaching methods for ESP learners (Gavioli 2005; Zanettin 2009).

1.5 Conclusion

Revisiting Tymoczko's envisaged steps in the development of corpus-based translation studies, we can say: *yes*, there is growing awareness of the role played by ideology in shaping text, context and translation, and this can be investigated by integrating corpus techniques with other methods; *yes*, language- and culture-specific phenomena are being investigated alongside the quest for similarities across languages and cultures, and *yes*, CTS is keeping pace with modern technologies, applying them effectively to research and pedagogy. CTS has not disappointed our expectations, at least to a large extent, but how much progress has really been made towards bridging the gap between linguistics and cultural studies? This too was one of the pursuits of CTS, and at present there are tangible signs that this may become one of the most fruitful and groundbreaking venues of enquiry. If one considers that more and more diversified resources are continually being created around the world, ranging from large reference corpora to small, handpicked specialized corpora, from synchronic to diachronic repositories of linguistic data, from unidirectional to bidirectional parallel corpora and from monolingual to multilingual comparable resources, it becomes plausible to conceive and carry out interdisciplinary work that harmonizes history with critical linguistics and sociocultural and literary investigations.

At the level of theory, the general trend seems to be favourable to starting a dialogue between post-modern cultural studies and textual theories on the one hand, and empirical descriptive studies on the other. Here, in particular, reference is made to the set of 30 theses identified by Chesterman and Arrojo (2000) as representing the shared ground between essentialism and non-essentialism on three main issues in translation studies, namely the definition, nature and effects of its object of study. There is, of course, a connection between these endeavours to engage in constructive exchanges of views and perspectives and Baker's (2002) past and recent recommendations to corpus builders and analysts to document and study the extralinguistic factors that come into play when texts are produced so as to go beyond the text as a formal structure and explore the relationship between linguistic patterns and text users. In conclusion, it seems neither unreasonable nor too far-fetched to envisage that what the future holds for corpus-based translation studies is the promotion of rich, varied, multilingual and interdisciplinary work, which will lead the way towards greater unity in the field of translation studies, fully respecting the diversity of each perspective involved.

Notes

[1] PAROLE (Preparatory Action for Linguistic Resources Organization for Language Engineering) represents a large-scale harmonized effort to create comparable text corpora and lexica for EU languages. Fourteen languages are involved on the PAROLE project, including Belgian French, Catalan, Danish, Dutch, English, French, Finnish, German, Greek, Irish, Italian, Norwegian, Portuguese and Swedish. Corpora containing 20 million words and lexica containing 20,000 entries were constructed for each of these languages using the same design and composition principles during 1996–1998. These corpora all include specific proportions of texts from the categories book (20 per cent), newspaper (65 per cent), periodical (5 per cent) and miscellaneous (10 per cent) within a settled range. The PAROLE corpora that are currently available are distributed by ELRA http://cw.routledge.com/textbooks/0415286239/resources/corpa3.htm#_Toc9229895

[2] See also Mauranen (2008: 45) who contends that, even though translations reflect their source languages through interference, they are not the only form of frequent language contact in today's globalized world and cannot be regarded as the enemy within pure, isolated, self-contained languages.

References

Anderman, Gunilla and Margaret Rogers (eds) (2005) *In and out of English: For Better, for Worse?*, Clevedon: Multilingual Matters.

— (2008) *Incorporating Corpora: The Linguist and the Translator*, Clevedon: Multilingual Matters.

Baker, Mona (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications', in Mona Baker, Gill Francis and Elena Tognini-Bonellli (eds) *Text and Technology: In Honour of John Sinclair,* Amsterdam: John Benjamins, 233–50.

— (1995) 'Corpora in Translation Studies: An Overview and some Suggestions for Future Research', *Target* 7(2): 223–43.

— (1996) 'Corpus-Based Translation Studies: The Challenges that Lie Ahead', in Harold Somers (ed.) *LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager,* Amsterdam: John Benjamins, 175–86.

— (1998) 'Réexplorer la Langue de Ia Traduction: Une Approche par Corpus', *Meta* 43(4): 480–5. Available online at: http://www.erudit.org/revue/meta/1998/v43/n4/ (accessed 21 May 2010).

— (1999) 'The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators', *International Journal of Corpus Linguistics* 4(2): 281–98.

— (2000) 'Towards a Methodology for Investigating the Style of a Literary Translator', *Target* 12(2): 241–66.

— (2002) 'Corpus-Based Studies within the Larger Context of Translation Studies', *Genesis: Revista Cientifica do ISAI* 2: 7–16.

Baker, Mona, Gill Francis and Elena Tognini-Bonellli (eds) (1993) *Text and Technology: In Honour of John Sinclair,* Amsterdam: John Benjamins.

Baumgarten, Nicole, Juliane House and Julia Probst (2004) 'English as *Lingua Franca* in Covert Translation Processes', *The Translator* 10(1): 83–108.

Bernardini, Silvia (2000) *Competence, Capacity, Corpora. A Study in Corpus-aided Language Learning*, Bologna: CLUEB.

— (2002) 'Exploring New Directions in Discovery Learning', in Bernhard Kettemann and Georg Marko (eds) *Teaching and Learning by Doing Corpus Analysis*, Amsterdam and New York: Rodopi, 165–82.

— (2004a) 'The Theory behind the Practice: Translator Training or Translator Education?', in Kirsten Malmkjær (ed.) *Translation in Undergraduate Degree Programmes*, Amsterdam and Philadelphia: John Benjamins, 17–30.

— (2004b) 'Corpus-aided Language Pedagogy for Translator Education', in Kirsten Malmkjær (ed.) *Translation in Undergraduate Degree Programmes,* Amsterdam and Philadelphia: John Benjamins, 97–112.

Biber, Douglas (1988) *Variation across Speech and Writing,* Cambridge: Cambridge University Press.

Blum-Kulka, Shoshana (1986) 'Shifts of Cohesion and Coherence in Translation', in Juliane House and Shoshana Blum-Kulka (eds) *Inter-lingual and Inter-cultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies,* Tübingen: Gunter Narr, 17–35.

Blum-Kulka, Shoshana and Eddie A. Levenston (1983) 'Universals of Lexical Simplification', in Claus Faerch and Gabriele Kasper (eds) *Strategies in Interlanguage Communication,* London: Longman, 119–39.

Bosseaux, Charlotte (2007) *How Does it Feel? Point of View in Translation. The Case of Virginia Woolf into French,* Amsterdam and New York: Rodopi.

Bowker, Lynne (1998) 'Using Specialized Monolingual Native-language Corpora as a Translation Resource: A Pilot Study', *Meta* 43(4): 631–51. Available online at: http://www.erudit.org/revue/meta/1998/v43/n4/ (accessed 21 May 2010).

— (2000) 'A Corpus-Based Approach to Evaluating Student Translations', *The Translator* 6(2): 183–210.

— (2001) 'Towards a Methodology for a Corpus-Based Approach to Translation Evaluation', *Meta* 46(2): 345–64. Available online at: http://www.erudit.org/revue/meta/2001/v46/n2/ (accessed 28 May 2010).

— (2003) 'Towards a Collaborative Approach to Corpus Building in the Translation Classroom', in Brian James Baer and Geoffrey S. Koby (eds) *Beyond the Ivory Tower: Rethinking Translation Pedagogy*, Amsterdam and Philadelphia: John Benjamins, 193–210.

Bowker, Lynne and Jennifer Pearson (2002) *Working with Specialized Language. A Practical Guide to Using Corpora,* London and New York: Routledge.

Chafe, Wallace L., John W. Du Bois and Sandra A. Thompson (1991) 'Towards a New Corpus of Spoken American English', in Karin Aijmer and Bengt Altenberg (eds) *English Corpus Linguistics in Honour* of *Jan Svartvik,* London and New York: Longman, 64–82.

Chesterman, Andrew (1998) 'Causes, Translations, Effects', *Target* 12(1): 200–30.

— (2000) 'A Causal Model for Translation Studies', in Maeve Olohan (ed.) *Intercultural Faultlines. Research Models in Translation Studies: Textual and Cognitive Aspects,* Manchester: St. Jerome, 15–27.

— (2004) 'Beyond the Particular', in Anna Mauranen and Pekka Kujamäki (eds) *Translation Universals: Do they Exist?,* Amsterdam and Philadelphia: John Benjamins, 33–50.

Chesterman, Andrew and Rosemary Arrojo (2000) 'Shared Grounds in Translation Studies', *Target* 12(1): 151–60.

Frawley, William (1984) 'Prolegomenon to a Theory of Translation', in William Frawley (ed.) *Translation: Literary, Linguistic and Philosophical Perspectives,* London: Associated University Presses.

Gavioli, Laura (1997) 'Corpora di Testi Elettronici e Concordanze: Un' Esperienza in un Corso Universitario per Traduttori', in *Atti del Simposio su Didattica e Informatica, Livorno 9–11 Ottobre 1997,* Livorno: Accademia Navale, 131–4.

— (1999) 'Corpora and the Concordancer in Learning ESP. An Experiment in a Course for Interpreters and Translators', in Gabriele Azzaro and Margherita Ulrych (eds) *Anglistica e i Metodi* e *Percorsi Comparatistici nelle Lingue, Cultura* e *Letterature di Origine Europea, Vol. II Transiti Linguistici* e *Culturali,* Trieste: EUT, 331–44.

— (2005) *Exploring Corpora for ESP Learning,* Amsterdam and Philadelphia: John Benjamins.

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: CLUEB, 61–80.

Gellerstam, Martin (1986) 'Translationese in Swedish Novels Translated from English', in Lars Wollin and Hans Lindquist (eds) *Translation studies in Scandinavia, Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II Lund 14–15 June 1985, Lund Studies in English,* Lund: CWK Gleerup, 75, 88–95.

Halverson, Sandra (1998) 'Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and the Conception of the Object of Study', *Meta* 43(4): 494–514. Available online at: http://www.erudit.org/revue/meta/1998/v43/n4/ (accessed 21 May 2010).

Hatim, Basil (1999) 'The Cultural and the Textual in the Way Translation Studies has Evolved', Paper presented at The University of Salford, UK, ESRI Research Seminars, 24 March.

Hermans, Theo (1999) *Translation in Systems: Descriptive and System-oriented Approaches Explained,* Manchester: St. Jerome.

House, Juliane (2007) 'Language Change through Language Contact in Translation: Evidence from Diachronic Corpus Studies', Paper presented at

Conference and Workshop on Corpora and Translation Studies, Shanghai Jiao Tong University, Shanghai, 31 March–3 April.

Johansson, Stig (1998) 'On the Role of Corpora in Cross-Linguistic Research', in Stig Johansson and Signe Oksefjell (eds) *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies,* Amsterdam: Rodopi, 3–24.

Johansson, Stig and Knut Hofland (1994) 'Towards an English-Norwegian Parallel Corpus', in Udo Fries, Gunnel Tottie and Peter Schneider (eds) *Creating and Using English Language Corpora,* Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zurich, 19–23 May 1993*,* Amsterdam: Rodopi, 25–37.

Johns, Tim (1991) 'From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-driven Learning', in Tim Johns and Philip King (eds) *Classroom Concordancing, ELR Journal,* 4 (Special Issue), 27–46.

Kemppanen, Hannu (2000) 'Looking for Evaluative Keywords in Authentic and Translated Finnish: Corpus Research on Finnish History Texts', Paper presented at Research Models in Translation Studies Conference, UMIST and UCL, Manchester, 28–30 April.

— (2001) 'Ideology in Translation: Challenging Universals?', Paper presented at Translation Universals: Do they Exist? Conference, Savonlinna School of Translation, University of Joensuu, Finland, 19–20 October.

— (2004) 'Keywords and Ideology in Translated History Texts: A Corpus-Based Analysis', *Across Languages and Cultures* 5(1): 89–106.

Kenny, Dorothy (1998) 'Corpora in Translation Studies', in Mona Baker (ed.) *Routledge Encyclopedia of Translation Studies,* London: Routledge, 50–3.

— (1999) 'Norms and Creativity: Lexis in Translated Text'. Unpublished PhD Thesis, Centre for Translation and Intercultural Studies (CTIS), University of Manchester.

— (2001) *Lexis and Creativity in Translation. A Corpus-Based Study*. Manchester: St. Jerome.

— (2009) 'Corpora', in Mona Baker and Gabriela Saldanha (eds) *Routledge Encyclopedia of Translation Studies,* 2nd edn, London and New York: Routledge, 59–62.

Kiraly, Donald C. (2000) *A Social Constructivist Approach to Translator Education. Empowerment from Theory to Practice,* Manchester: St. Jerome.

— (2003) 'From Instruction to Collaborative Construction: A Passing Fad or the Promise of a Paradigm Shift in Translator Education?', in Brian James Baer and Geoffrey S. Koby (eds) *Beyond the Ivory Tower: Rethinking Translation Pedagogy*. American Translators Association Scholarly Monograph Series, Vol. 12, Amsterdam and Philadelphia: John Benjamins, 3–27.

Klaudy, Kinga (1996) 'Concretization and Generalization of Meaning in Translation', in Marcel Thelen and Barbara Lewandoska-Tomaszczyk (eds) *Translation and Meaning Part 3*, *Proceedings of the Maastricht Session of the 2nd International Maastricht-Łódź Duo Colloquium on 'Translation and Meaning'*, held in Maastricht, The Netherlands, 19–22 April 1995, Maastricht: Hogeschool Maastricht, 141–63.

— (2009) 'Explicitation', in Mona Baker and Gabriela Saldanha (eds) *Routledge Encyclopedia of Translation Studies*, 2nd edn, London and New York: Routledge, 104–8.

Koby, Geoffrey S. and Brian James Baer (2003) ‘Task-Based Instruction and the New Technology. Training Translators for the Modern Language Industry’, in Brian James Baer and Geoffrey S. Koby (eds) *Beyond the Ivory Tower: Rethinking Translation Pedagogy,* Amsterdam and Philadelphia: John Benjamins, 211–27.

Kruger, Alet (2000) ‘Lexical Cohesion and Register Variation in Translation: The Merchant of Venice in Afrikaans’. Unpublished DLitt et Phil Thesis, University of South Africa, Pretoria.

Laviosa, Sara (2002) *Corpus-Based Translation Studies. Theory, Findings, Applications,* Amsterdam: Rodopi.

— (2009) ‘Universals’, in Mona Baker and Gabriela Saldanha (eds) *Routledge Encyclopedia of Translation Studies,* 2nd edn, London and New York: Routledge, 306–10.

— (2010a, in press) ‘Towards the Study of Drifts in the Priming of Anglicisms in Business Communication’, in Paola Evangelisti Allori and Vijay Bhatia (eds) *Identity in the Professions. Legal and Corporate Citizenship,* Bern: Peter Lang.

— (2010b) ‘Corpus-Based Translation Studies 15 Years on: Theory, Findings, Applications’, *SYNAPS* 24.

Laviosa-Braithwaite, Sara (1996) ‘The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation’. Unpublished PhD Thesis, Centre for Translation and Intercultural Studies (CTIS), University of Manchester.

Leech, Geoffrey (1992) ‘Corpora and Theories of Linguistic Performance’, in Jan Svartvik (ed.) *Directions in Corpus Linguistics,* Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991, Trends in Linguistics: Studies and Monographs, Berlin: Mouton de Gruyter, 65, 105–22.

Lindquist, Hans (1989) *English Adverbials in Translation: A Corpus Study* of *Swedish Renderings,* Lund Studies in English 80, Lund: Lund University Press.

Magalhães, Célia Maria (2001) ‘Corpora-Based Translation Studies in Brazil: Towards Universals of Translation?’, Paper presented at Claims, Changes and Challenges in Translation Studies, Third International Congress of the European Society for Translation Studies, Copenhagen Business School, 30 August–1 September.

Malmkjær, Kirsten (1998) ‘Love thy Neighbour: Will Parallel Corpora Endear Linguists to Translators?’, *Meta* 43(4): 534–41. Available online at: http://www.erudit.org/revue/meta/1998/v43/n4/ (accessed 21 May 2010).

Mauranen, Anna (2000) ‘Strange Strings in Translated Language. A Study on Corpora’, in Maeve Olohan (ed.) *Intercultural Faultlines. Research Models in Translation Studies: Textual and Cognitive Aspects,* Manchester: St Jerome, 119–41.

— (2008) ‘Universal Tendencies in Translation’, in Gunilla Anderman and Margaret Rogers (eds) *Incorporating Corpora: The Linguist and the Translator,* Clevedon: Multilingual Matters, 32–48.

Mauranen, Anna and Pekka Kujamäki (eds) (2004) *Translation Universals. Do They Exist?,* Amsterdam and Philadelphia: John Benjamins.

Munday, Jeremy (1997) ‘Systems in Translation: A Computer-Assisted Systemic Approach to the Analysis of the Translations of García Márquez’. Unpublished PhD Thesis, Department of Modern Languages, University of Bradford.

— (2002) 'Systems in Translation: A Systematic Model for Descriptive Translation Studies', in Theo Hermans (ed.) *Crosscultural Transgressions. Research Models in Translation II: Historical and Ideological Issues*, Manchester: St. Jerome, 76–92.

Olohan, Maeve (2004) *Introducing Corpora in Translation Studies,* Manchester: St. Jerome.

Olohan, Maeve (ed.) (2000) *Intercultural Faultlines. Research Models in Translation Studies: Textual and Cognitive Aspects,* Manchester: St. Jerome.

Øverås, Linn (1998) 'In search of the Third Code: An Investigation of Norms in Literary Translation', *Meta* 43(4): 571–88. Available online at: http://www.erudit.org/revue/meta/1998/v43/n4/ (accessed 21 May 2010).

Saldanha, Gabriela (2004) 'Accounting for the Exception to the Norm: A Study of Split Infinitives in Translated English', *Language Matters, Studies in the Languages of Africa* 35(1): 39–53.

— (2005) 'Style of Translation: An Exploration of Stylistic Patterns in the Translations of Margaret Jull Costa and Peter Bush'. Unpublished PhD Thesis, School of Applied Language and Intercultural Studies, Dublin City University.

Scott, Maria Nélia (1998) 'Normalisation and Readers' Expectations: A Study of Literary Translation with Reference to Lispector's "A Hora da Estrela" '. Unpublished PhD Thesis, AELSU, University of Liverpool.

Shlesinger, Miriam (1989) 'Simultaneous Interpretation as a Factor in Affecting Shifts in the Position of Texts in the Oral-Literate Continuum'. Unpublished MA Dissertation, Tel Aviv, Tel Aviv University.

— (1991) 'Interpreter Latitude vs. Due Process. Simultaneous and Consecutive Interpretation in Multilingual Trials', in Sonja Tirkkonen-Condit (ed.) *Empirical Research in Translation and Intercultural* Studies*,* Selected papers of the TRANSIF Seminar, Savonlinna 1988, Tübingen: Gunter Narr, 47–155.

— (1995) 'Shifts in Cohesion in Simultaneous Interpreting', *The Translator* 1(2): 193–214.

— (1998) 'Corpus-Based Interpreting Studies as an Offshoot of Corpus-Based Translation Studies', *Meta* 43(4): 486–93. Available online at: http://www.erudit.org/revue/meta/1998/v43/n4/ (accessed 21 May 2010).

Sinclair, John M. (1991) *Corpus,* C*oncordance, Collocation,* Oxford: Oxford University Press.

Toury, Gideon (1985) 'A Rationale for Descriptive Translation Studies', in Theo Hermans (ed.) *The Manipulation of Literature: Studies in Literary Translation,* London: Croom Helm, 16–41.

— (1991) 'What are Descriptive Studies into Translation Likely to Yield apart from Isolated Descriptions?', in Kitty M. van Leuven-Zwart and Ton Naaijkens (eds) *Translation Studies: The State of the Art: Proceedings from the First James S. Holmes Symposium on Translation Studies*, Amsterdam and Atlanta, GA: Rodopi. 179–92.

— (1995) *Descriptive Translation Studies and Beyond,* Amsterdam: John Benjamins.

— (2004) 'Probabilistic Explanations in Translation Studies: Welcome as they are, Would they Qualify as Universals?', in Anna Mauranen and Pekka Kujamäki (eds) *Translation Universals. Do they Exist?*, Amsterdam and Philadelphia: John Benjamins, 15–32.

Tymoczko, Maria (1998) 'Computerized Corpora and the Future of Translation Studies', *Meta* 43(4): 652–60. Available online at: http://www.erudit.org/revue/meta/1998/v43/n4/ (accessed 21 May 2010).

Vanderauwera, Ria (1985) *Dutch Novels Translated into English: The Transformation of a 'Minority' Literature,* Amsterdam: Rodopi.

Venuti, Lawrence (2000/2004) *The Translation Studies Reader,* London: Routledge.

Weissbrod, Rachel (1992) 'Explicitation in Translations of Prose-fiction from English to Hebrew as a Function of Norms', *Multilingua* II(2): 153–71.

Winters, Marion (2005) 'A Corpus-Based Study of Translator Style: Oeser's and Orth-Guttmann's German Translations of F. Scott Fitzgerald's *The Beautiful and Damned*'. Unpublished PhD Thesis, School of Applied Language and Intercultural Studies, Dublin City University.

Xiao, Richard (2006) 'Corpora Survey'. Available online at: http://cw.routledge.com/textbooks/0415286239/resources/corpa3.htm (accessed 27 May 2010).

Xiao, Richard (ed.) (2010) *Using Corpora in Contrastive and Translation Studies,* Newcastle upon Tyne: Cambridge Scholars Publishers.

Zanettin, Federico (1998) 'Bilingual Comparable Corpora and the Training of Translators', *Meta* 43(4): 616–30. Available online at: http://www.erudit.org/revue/meta/1998/v43/n4/ (accessed 21 May 2010).

— (2002) 'CEXI: Designing an English Italian Translational Corpus', *Language and Computers* 42(1): 329–43.

— (2009) 'Corpus-Based Translation Activities for Language Learners', *The Interpreter and Translator Trainer,* (3)2: 209–24.

Zanettin, Federico, Silvia Bernardini and Dominic Stewart (eds) (2003) *Corpora in Translator Education,* Manchester: St. Jerome.

Chapter 2

Corpus-Based Interpreting Studies (CIS): Overview and Prospects

Robin Setton

The confusions that occupy us arise when language is like an engine idling, not when it is doing work.

Ludwig Wittgenstein

2.1 Introduction

In the last thirty years, corpus-based translation studies (CTS) has piggybacked the fast-growing field of corpus linguistics (CL) to yield a rich harvest of insights into translation. Some researchers studying interpreting had also been examining performance data, in which the challenges soon appeared to be of a different order. In interpreting, the conditions of reception and production of the translation create special difficulties for data collection, transcription and interpretation, but also invite closer attention to a new dimension: process. As well as revealing interesting features of the product, as in mainstream CL and CTS, authentic interpreting corpora open a window onto the peculiar cognitive processes of this activity, and may even contribute to our understanding of language and cognition.

This chapter reviews the history of corpus-based interpreting studies (CIS), focusing on recent developments – new software, access to larger, quality corpora and new techniques for transcription, analysis, presentation and sharing – that hold significant future promise for this paradigm as an alternative and complement to intuition, surveys or laboratory experiments.

Technology and better access to corpora are vastly enhancing the data, but as the maxim has it, if theory without data is empty, data without theory is sterile. The interpreting studies community is still struggling to analyse and understand this unique activity, interweaving language, cognition and communication, with some relatively primitive theoretical baggage consisting of categories, assumptions and simple heuristic models inherited, almost intact, from early descriptive linguistics or cognitive psychology. More and better data cannot of themselves compensate for this theoretical lag: authentic corpora will demand a refined theoretical prism reflecting up-to-date models of human communication and cognition, adapted to the peculiar and complex conditions of interpreting. However, the necessary adaptations in moving from CL and CTS to CIS, and from theories of cognition and speech processing to models of interpreting, are by no means straightforward, as the discussion will show.

2.2 Object and Aims of CL, CTS and CIS

Information technology transformed philology by enabling the broad-based empirical study of language in use on the basis of large collections of real linguistic productions. Corpus linguists have been logging occurrences and distributions of language forms in search of usage patterns that identify different registers and help to understand the contextual factors which influence its variability (Biber et al. 1998). In a more militant vein, some 'empirical linguists' are using the new data on usage to challenge the tenets of a hitherto dominant linguistics seen as based on invented utterances and preoccupied only with inferring abstract universals (Sampson 2001).

The flowering of corpus linguistics inspired a similar empirical turn in translation studies, in which a strong impetus was given by Mona Baker and her colleagues in Manchester (Baker 1995). In its first two computer-aided decades, CTS has initially focused on product rather than process, exploring the characteristics of translations as texts, or the patterning specific to translational language. Text features are quantitatively logged and compared as in CL, but CTS also ventures to treat certain features, like lexical use patterns (proportion of lexical to grammatical words, high- to low-frequency words, word repetition, type-token ratios) as indicators of such putative universals of translation (or at least of translated text) as simplification or explicitation. Another mainstream research topic in CTS is the impact on translation of cultural, ideological, political and gender pressures and norms, or conversely, and more ambitiously, the effect of translation in these dimensions of intercultural communication.

The prospects seem particularly exciting for the study of interpretation, where corpora are potentially richer by one or two dimensions than both monolingual and translation data. The peculiar conditions of production, and the challenge of tracking the intensive use of local context which interpreters need to manage these conditions, make interpreting corpora a rich undeveloped resource for the study of psycholinguistic and pragmatic processes.

Alongside the relatively organized and coordinated landscape of CTS, with its substantial paradigmatic and methodological unity and the resolute adoption and integration of natural language processing (NLP) software, CIS is still a cottage industry. This is also due to some significant differences of orientation and emphasis. Interpreting research is more concerned with cognitive and psycholinguistic processes, and the conditions of production make it rather pointless to attempt any realistic model of the process without taking into account factors such as the live context and elusive features of live speech like prosody.

Fruitful research has already been done, along CTS lines, to identify distinctive characteristics of interpreting output as compared to spontaneous sovereign speech (see Section 2.4.5.1 below). But the interpreting community, insofar as it expects anything from research at all, is especially interested in discovering the factors determining quality and expertise (for training), a demand which CIS can only meet by exploring and inferring from the contrastive temporal and linguistic details of multilingual corpora.

Research in support of interpreter training and quality aims to identify the main factors in the success or failure of interpretation and the acquisition of interpreting expertise. Corpus analysis and controlled experiments offer alternative routes to this goal. While natural interpretation corpora obviously hold valuable clues to expertise, experimental research in a psycholinguistics paradigm is currently more influential. However, the yield of robust findings has been disappointing (see Gile 2001; Pöchhacker 2004), due in part to problems of experiment design, but also to a weakness or dispersion of the theoretical basis for making sense of experimentally generated data. Both paradigms urgently need consensus on agreed methods and procedures – in CIS, at least for corpus description and presentation, to allow sharing and comparability, and perhaps more elusively, on a theoretical framework and procedures for analysis. This will probably entail theoretical enrichment and updating beyond the often simplistic cognitive models and assumptions about language processing, often dating back decades, that still underlie much research.

In addition to interpreting-specific applications, some researchers also still believe that conference interpretation can serve as a laboratory for the study of language, cognition and communication, in both sociological and psychological dimensions. Given an understanding of the conditions of production, interpreting corpora might support research into the communication of implicit meaning, and thus help to test the predictions of cognitive and pragmatic models of speech communication (e.g. by comparing interpreting performance from discourse with and without cohesive links), or to explore how external knowledge or context is used in comprehension and translation, and how and when such knowledge is activated. CIS might also contribute to the study of multilingual communication as a social, historical or ideological vector (filter, amplifier, refractor . . .), in showing what 'goes through' in linguistically mediated discourse, and how, as compared to communication in a superficially common language.

This potentially rich dividend from the study of special types of language use like translation and interpretation will depend on the techniques that can be developed for extracting robust findings from corpora produced in natural or ‘ecological’ conditions, bypassing the noise in experimental studies from participants’ (mis-)interpreting the researcher’s instructions, or deviating from their usual task norms and behaviour, in ways that may be undetectable. Simultaneous interpreting (SI) data promise to be particularly rich, on condition of finding reliable methods to fix, display and interpret them. In short, the hopes placed in this new ‘ecological’ data are balanced by technical difficulties and other possible sources of distortion or misinterpretation.

CL and CTS have built an indispensable foundation for the study of interpreting corpora, but they are not motivated as CIS is – it can hardly proceed otherwise – to break into three further dimensions of description and representation: *multilingualism, orality* and *situated synchronicity*. Corpus linguists are still tempted to fall back on the relative safety of written corpora even when testing hypotheses about aural comprehension (Sampson 2001: 42 on syntactic complexity), or to give priority to deepening the detail for a single familiar language over searching for cross-linguistically valid coding categories (ibid.: 91). Research on written translation can also remain focused on features of the written product, and indeed has little evidence from which to make inferences about process. When translation is at its most oral, interpersonal and interactive, as in ‘dialogue’ interpreting, or at its most cognitively and temporally constrained, as in conference interpreting, help must be sought from other disciplinary quarters: models of the dynamics of human verbal communication, and models of cognition (especially different kinds of memory, attention and knowledge mobilization).

2.3 Dialogue and Conference Interpreting

Oral translation has been roughly divided into conference interpreting and ‘dialogue interpreting’ (Mason 1999), in which the prototype settings and conditions are different enough to determine quite distinct research orientations, methods and problems. Dialogue interpreting is usually face-to-face and more immediately interpersonal, involving the interpreter directly in mediating across power gradients and cultural differences. Research in this area has accordingly highlighted features which reveal this more intimate dynamic, studying indicators of politeness or power, such as forms of address or patterns of turn-taking, interruption or overlap, which are readily accessible through discourse analysis. Discourse analysts search natural corpora for “regularities [. . . to be] described relative to environments in order to illuminate language use and its articulation with various social and psychological factors” (Brown and Yule 1983: 22). Accordingly they work mostly on interactive communicative events with high participant involvement, such as conversations, interviews, and patient–doctor or lawyer–witness encounters.

In the more formalized and detached situation of conference interpreting, the interpreter’s interactive and mediating skills are less critical, but the task is probably more challenging cognitively and linguistically, and it is these skills that have attracted the spotlight of research. The conditions of production are highly task-specific. Of the four modes of translation shown in Figure 2.1 (written text translation and three variants on conference interpreting), only one, 'free' simultaneous interpretation from improvised speech, is done blind, that is, in a single pass with no preparation or access to the text or content in advance, other than what can be inferred from the general subject-matter of the event. The shape of each unfolding utterance remains unpredictable, and transcripts show that most of the time the interpreter must usually begin to produce well before each source utterance is complete. But unlike the text translator, she shares the live communicative situation with her speaker, and can see and hear him move, gesture, pause, slow down or accelerate and appeal to his audience in various immediate ways.

In interpreting, immediate speech processing is critical, and consequently so is whatever can be mobilized to support it from immediate context and prior knowledge. This means that unlike in CL and CTS, no sense can be made of an interpreting corpus without reference not only to these specific processing *constraints* – errors, for example, may be due to misunderstandings or wrong strategic choices, or to attention or memory overload – but also to the *flexibilities,* especially from the use of available context at the time of production, which plays a key role in making these constraints manageable. The theoretical prism for analysis must therefore be adjustable to take into account the variation in both these factors – processing and access to context – according to mode (Figure 2.1) and situation.

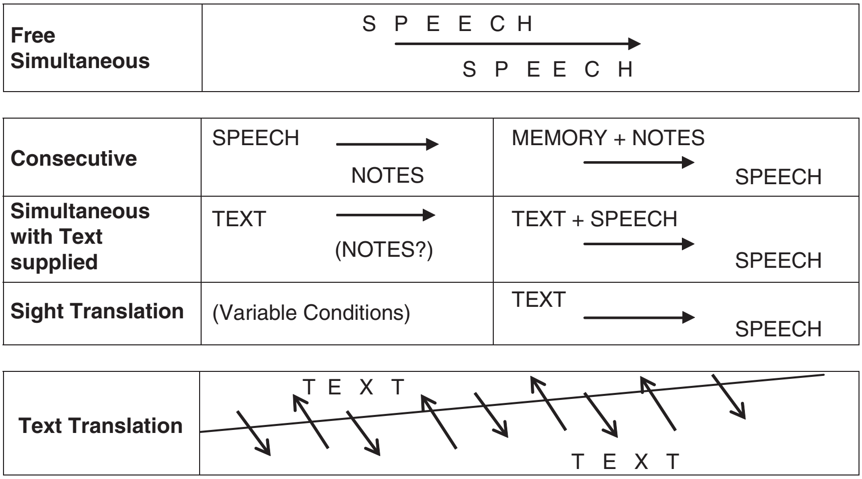


Figure 2.1 Dynamics and input sources in different translation modes

Corpus-based research on simultaneous interpretation combines the challenges inherent in spoken, multilingual and parallel corpora at each of the following stages: selection, data collection, alignment, transcription, presentation and of course analysis and the extraction of scientific conclusions. After reviewing the modest but no doubt heroic beginnings of corpus interpreting studies, we will look at each of these challenges in turn and explore prospects for developing the corpus approach into a more manageable and productive paradigm.

2.4 CIS: Still a Cottage Industry

The total volume of samples of authentic conference interpretation (from real events organized for other than interpreting research or training purposes) recorded and/or partially transcribed and analysed in the first fifty-odd years of interpreting research has been modest. Many small samples have, of course, been generated in experimental conditions, and several researchers have studied student productions, which are far more easily obtained. Table 2.1 lists some landmark studies of conference interpreting corpora. They reflect different strategies for data selection, presentation and analysis.

Enthusiasm for CIS may have fluctuated with the debate on research methods and doctrine within interpreting studies. The Paris school (Seleskovitch 1975, Lederer 1981, Dejean le Féal 1982, Donovan 1990, Laplace 1994) has generally insisted that the only valid basis for theorizing about the interpreting process is the experience and observations of professionals backed by illustrations from their performance at real-life events. However, this doctrine was vigorously challenged from the 1980s onwards on epistemological and methodological grounds, and corpus-intensive research lost ground in favour of controlled experimental studies. CIS survived through the 1990s mostly in another, more sociological paradigm pursued by researchers based largely in German-speaking countries, Central Europe and Scandinavia who have also preferred to work on authentic corpora (Pöchhacker 1994; Kalina 1998, Vuorikoski 2004), often modelling interpreting within action theory (*Handlungstheorie*) and the *skopos* school of Translation Studies, with input from functional and text linguistics and rhetorics.

The relative dormancy of CIS through this period also reflected the continuing difficulty of obtaining and analysing substantial authentic data as much as any doctrinal preference for controlled laboratory studies. In the new century, with the recent availability of large multilingual corpora from the European Parliament (EP) and steadily improving software tools, CIS is enjoying a revival, notably in Italy (see this volume, Chapter 12) and Spain; while in Japan (Nagoya) a very different research goal – the ongoing quest for automated interpreting – has prompted the analysis of a substantial corpus of professional interpreting in controlled conditions.

Table 2.1 Studies of interpreting based on authentic corpora

Languages: EN English, ES Spanish, DE German, FI Finnish, FR French, HEB Hebrew, IT Italian, RU Russian, SW Swedish, TR Turkish, ZH (Standard) Chinese

Researcher	Languages	Event	Mode	Subjects	Length	Transcription published or available	Sound files availability	Analysis
Oléron and Nanpon 1965	EN, FR, DE, ES	UNESCO impromptu (non-tech discussion)	SI	pros	~ 7 minutes	Unknown	Unknown	Time lag, speed, fidelity
Dejean Le Feal 1978	FR>DE	Various speeches	SI	pros		77 pages, speakers and interpreters	Text on micro fiches, AIIC*	Recited vs. impromptu input (école du sens)
Chernov 1978	EN, FR, ES, RU	UN 1968	SI	pros	'~40 hours'	'Parallel transcripts' extracts published	Probably N. A.	Illustrate theory (redundancy & prediction)
	EN>RU, ES, FR	1978 UN satellite interpreting experiment	SI	pros		(ditto)	Probably no longer available	(ditto)
Lederer 1981	DE > FR	Railway Consortium and lab (2nd versions)	SI	2 pros	3 hours taped (original+ 2 interpreters)	63 minutes original DE, FR; some extracts synchronized (interlinear)		Illustrate theory (école du sens)
Shlesinger 1989	HEB><EN	Courtroom testimony	SI	pros	4 hours			
	FR><EN	Extracts from 2 meetings 1986–88	SI	pros	~ 4 hours	50 pp SI (2 events), fluent text speaker + interpreter, some extracts synchronized (interlinear)	No longer available	Fidelity examples (école du sens)

Continued

Table 2.1 Continued

Researcher	Languages	Event	Mode	Subjects	Length	Transcription published or available	Sound files availability	Analysis
Pöchhacker 1994	EN><DE, FR >DE	Vienna small business conference *ICSB*	SI	pros	14 hours (original + interpreter)	Available as vol. 3 of doctoral diss. from U. of Vienna library; parts published in Pöchhacker 94	Tapes available from author	Intertextuality, situational & delivery factors (speed, slips/shifts, hesitation EN><DE)
Kalina 1998	EN><DE<>FR	*Bertell*: 1989 public lecture (anti nuclear)	SI	6 students	70 minutes	5 versions tiered. 5s per line.	Consult, loan at Heidlbrg	Choose examples, identify errors, strategies
	DE, FR, EN	*Würzburg*: 1992 law symposium	SI	6 pros (2 per booth)		1-track audio (simulation: 2 track)	(ditto)	(ditto)
Setton 1997, 99	DE>EN	Extracts from Kalina *Würzburg* corpus	SI	1 pro + 2 in mock	14/30 minutes microanalysis	1-track audio (simulation: 2 track)	Available from author	Linguistic, cognitive-pragmatic analysis for process modelling
Wallmach 2000	EN, Zulu, Afrikaans, Sepedi	Parliament speeches Gauteng (S. Africa) Provincial legislature	SI	16 pros	Pilot: 6–8 hours EN, Afrikaans, Zulu	Pilot material transcribed (6–8 hrs)	110 hours on tape (3 languages)	Norms/strategies vs. user expectations, effect of speed, technicality on performance, language-specific strategies
Cencini 2000	EN><IT	Television Interpreting Corpus (TIC),			36,000 words	Not available for outside use		TEI Standard (computer query-able)
Fumagalli 1999–2000	EN>IT	Comparable EN(18) and IT source speeches;	CI			Not available for outside use		Features of 'interpretese' *MultiConcord-Parallel Concordancer*
Diriker 2001	EN><TR	Conference on Metaphysics and Politics (2 days)	SI	pros	150 pp transcripts	Available as Annex to Ph.D.	Bogazici Univ. Library	Interactional and sociological

Vuorikoski 2004	mostly DE, EN ><FI, SW	European Parliament debates	SI	Ca. 70 pros	120 speeches, 65 analysed	Selected transcripts appended to PhD	CD and website	Difficulty, quality in rhetorical (political) speech genre
Beaton 2007		European Parliament debates via EBS	SI		7 hours			Text-discourse-ideology link through cohesive devices
Monacelli 2005, 2009	IT><EN	10 speeches from 4 events conferences	SI	10 pros	2 hours originals + interpretation	Selected transcripts in PhD	CD ST, TT and synchronized (from audio cues)	Interactional politeness, 'face', through deixis, mood, transitivity
Straniero 2003, 2007	IT><EN	TV talk shows	SI and CI	11 TV pros	>80 samples (1997–2002)			Emergency strategies, TV-specific quality norms
ECIS Group (Univ. of Granada)	EN>DE, ES, FR, DE, FR>ES	European Parliament debates via EbS	SI	pros	43 speeches, 73 interpretations	Linked files in MS Access, tool developed for multivariate visualization	Available from authors	Quality minimizers and maximizers, verbal and non-verbal
CIAIR (Nagoya University)	EN><JA	Monologue SI (simple lectures) and travel dialogues	SI and CI	SI: 4 pros per speech	182 hrs (1m words)	Web access/interface	Not shared	Strategies, lag, etc. for machine SI research. Own software.
EPIC**: Bologna-Forli group	EN-IT-ES (9 sub-corpora)	European Parliament sessions	SI	Multiple pros	Open-ended; 280 hrs usable in archive; ca. 280, 000 words transcribed as of mid-2010	.Available for query on line (POS-tagged and indexed database)	As separate files, on request	As of 2010: directionality, lexis, disfluencies; Ongoing (+ grad theses). Semi-automatic transcription technique

Continued

Table 2.1 Continued

Researcher	Languages	Event	Mode	Subjects	Length	Transcription published or available	Sound files availability	Analysis
Meyer (2008) 'K6'(CoSi)	PT>DE	Lecture tour	SI, CI		6 hours, 35,000 words transcribed (HIAT)	Web access, available at http://www.exmaralda.org/en_index.html		Proper names in CI, SI. Recording & transcription: EXMARaLDA
Meyer K2 (DiK) corpus	DE, TR, PT, ES	Doctor–patient dialogues in hospitals	Mono-lingual and inter-preted		25 hours recorded, 160,000 words	Web access, available at http://www.exmaralda.org/en_index.html		
Shlesinger 2008, Shlesinger and Ordan (forthcoming)	EN	EPIC	SI, written transla-tion					Features of 'interpretese' vs. original speech and written translation
DIRSI (Bendazzoli 2010a)	EN><IT	International conferences. (health) monologues	SI	5 pros (one EN-A)	2010: 20 hours, 130,000 words transcribed, tagged and indexed; text-to-sound and ST-TT aligned	Accessible via dedicated online web interface		Directionality
FOOTIE (Sandrelli forthcoming)	EN><IT	Football press conferences: interactive Q&A	SI			In progress		Work begun in 2010
CorIt (Straniero 2007: 20–1)	multilingual	TV interpreting in Italy (longitudinal)	CI, SI			In progress		In progress

*The Research Committee of AIIC (International Association of Conference Interpreters) provides copies of microfiches of some theses to researchers on request.

**European Parliament Interpreting Corpus.

2.4.1 Consecutive Interpretation

Very few corpora of *consecutive* (conference) interpretation have been published. The first in-depth analysis of consecutive interpretation in a substantial – though not authentic – corpus was also the first PhD in interpreting studies (Seleskovitch 1975). Seleskovitch analysed 12 French versions of two recorded English speeches by professional interpreters, publishing most of her corpus in orthographic transcription, along with some of the interpreters' notes (photocopied) and interviews in which the subjects explained them. Seleskovitch found formal independence between the three stages – input, notes and output – to support her theory of the 'triangular' process of interpreting with extensive deverbalization of the message in the intermediate stage.

Dollerup and Ceelen (1996) published a very large corpus (300 pages of transcripts) of speeches and consecutive student renditions involving six languages, from a one-month EEC training course held in 1976, in standard orthographic transcription, but virtually without commentary.[1] The data are 'deliberately raw: it is up to the individual researcher to decide what (s)he wants to study'. The authors are protective of the subjects, citing a 'moral obligation' always to keep in mind that oral presentation is never intended for the written page, so needs editing 'to be fair'. This illustrates a major quandary of CIS: lacking agreed and familiar conventions for marking rhythm and prosody on transcripts – like the musical notation which is now so familiar that it is 'visually audible' for musicians – publishers of interpreting corpora have no option but to exhort readers to 'read with their ears'. Similar caveats are given by Donovan who published consecutive transcripts from two live events along with SI samples for her own doctoral thesis dissertation (Donovan 1990: 25–8).

2.4.2 Simultaneous Interpretation

With the exception of an early study of a very small sample of UNESCO speeches by two professional psychologists (Oléron and Nanpon 1965), all the SI studies listed have been the work of practising interpreters for PhD or MA theses. Prior to the most recent Italian and Japanese activity, the most substantial studies were those of Chernov (1978, 1994, 2004), Lederer (1981), Pöchhacker (1994), Kalina (1998), Diriker (2001), Vuorikoski (2004), some of which were used by other researchers: Setton (1997/1999) analysed part of Kalina's corpus; Lamberger-Felber (2001) drew on Pöchhacker's.

Chernov (1978, 1994 and forthcoming) published extracts from recorded tapes of interpreted sessions at the United Nations in 1968 (in English, Russian, French and Spanish), and in a remote satellite interpreting experiment between New York and Buenos Aires in 1978, in the latter case showing interesting differences in interpreters' performance in the remote versus *in situ* conditions. Lederer (1981) timed and transcribed an hour-long trilingual meeting of a European railway consortium, of which she selected and synchronized about 18 minutes of German-to-French SI interpreters for closer analysis, illustrating the interpretive theory with examples and a few quantitative measurements such as the lag between the original and the interpretation, measured at 20 or more points. Pöchhacker (1994) recorded a 3-day conference on small businesses in Vienna, where his role as recruiter and coordinator of the interpreting team gave him a unique overview of the whole conference as a multilingual interpreted event. His analysis stresses the interplay between environmental factors and the macrocontext (intertextuality, advance documentation or mode of presentation) and the interpreters' local performance. Kalina (1998) recorded a 3-day conference on European law, with interpretation in English, French and German, which she used with various other smaller experimental and student samples to study interpreters' strategies, but published only a very short plain interlinear transcript with commentary to illustrate various strategic and error phenomena. Setton (1999) studied German–English extracts from Kalina's Wurzburg corpus, versions of the same speeches interpreted later in simulated conditions by two more professional volunteers, and several speeches and discussions interpreted from Chinese into English, deriving a theoretical model relating cognitive processing in SI to available context.

2.4.3 Experimental and Training Corpora

It would be well beyond the scope of this chapter to attempt to list the numerous studies of interpreting based on corpora generated experimentally or in simulated conditions. Some have produced abundant quantitative data (e.g. Barik 1973, 1975) or presented transcripts in novel ways (e.g. Goldman-Eisler 1972, who provided some graphics of temporally aligned parallel acoustic tracings). For the most part, however, the authors of experimental studies have not published substantial transcripts.

Most recently, a very large English–Japanese SI corpus (182 hours) of lectures on non-technical topics and their interpretations by professionals in booths (in a classroom) has been assembled and manually transcribed at the Center for Integrated Acoustic Information Research (CIAIR) of Nagoya University in Japan, as part of ongoing research towards developing an automatic SI system (Tohyama et al. 2005). This is a novelty for the machine interpreting community, which elsewhere has largely ignored human interpreting. In addition to design proposals for an automatic system, the project has already generated several papers analysing aspects of the human SI process, including patterns of segmentation or chunking (Ding et al. 2005), word-level variations in lag (EVS) (Ohno et al. 2008), strategies to reduce lag or increase output quality (Tohyama and Matsubara 2006a), and pausing and its effect on listeners

(Tohyama and Matsubara 2006c), as well as proposed pedagogical uses of the corpus in interpreter training (Tohyama and Matsubara 2006b).

2.4.4 Ready-Made Authentic Corpora: The European Parliament Corpora

In the last few years, CIS has been revolutionized by the cornucopia of authentic speeches and professional SI versions in multiple languages (now 27) released into the public domain by the European Parliament. This corpus is 'ready-made' only up to a point: it has taken intensive work by research groups at Granada in Spain (ECIS: Evaluación de la Calidad en Interpretación Simultanea) and Bologna (-Forli) in Italy (EPIC: European Parliament Interpreting Corpus) to compile samples from EP debates into indexed, searchable databases accessible via the internet (see Table 2.1). Details of the compilation, transcription, annotation, presentation and preliminary analysis of these corpora can be found in Monti et al. (2005), for EPIC, and Barranco-Droege et al. (forthcoming) for ECIS.

2.4.5 Recent Research Orientations and Topics in CIS

Analyses of these large, multilingual machine-readable interpreting corpora, albeit focused on a single institutional setting have already yielded some low-hanging fruit. The ECIS group had hitherto specialized in experimental work on factors in interpreting quality and have been using their corpus as an observational complement, generating researcher judgements on as many as 45 different (ST and TT) variables, and using sophisticated visualization tools to explore their interactions. EPIC researchers have so far done more quantitative and statistical analyses of both 'comparable' texts (originals vs. translations in the same language) and 'parallel texts' (STs vs. their TTs), on whole speeches (e.g. for lexical density and variety, Russo et al. 2006), but also some local correspondences (e.g. disfluencies), pending the full time-alignment that will open up a '*parallel-synchronised*' dimension in which to study technically more challenging issues like strategies, or test theories of SI processes. This is already underway in Nagoya, in the large heterogeneous, though only semi-authentic CIAIR corpus.

2.4.5.1 Features of the Product: 'Interpretese'

Large-scale CTS analyses of features such as collocations, part-of-speech distributions and frequencies and lexical richness (type-token ratios) in comparable texts had already shown evidence of a simplifying, levelling (or normalizing) 'universal' of 'translationese' (Baker 1995; Laviosa 1996, 2000). Similar evidence has been found for 'interpretese', or interpreters' speech (Fumagalli 2000), and recent studies on the EPIC corpus have positioned interpreting closer to original speech linguistically than to written translation (Shlesinger

2008; Shlesinger and Ordan forthcoming). In other words, orality is a more salient factor in the phenomenology of interpreting than what it may owe to being a form of translation.

2.4.5.2 Factors in Performance and Quality

Vuorikoski (2004) had already used European Parliament speeches to study interpreting quality as a function of rhetoric. The ECIS and EPIC groups have been exploring the impact of a wide range of input speech variables on interpreting performance, including topic, speed, disfluency patterns (Pradas Macías 2009; Bendazzoli, this volume), accent, 'problem triggers', and at a meta-level, patterns of quality evaluation itself (Collados Aís 2009; García Becerra, in press). In terms of topic, for example, interpreters in EPIC are significantly more fluent from impromptu speech and on procedural or 'housekeeping' subject-matters than on political, or 'specialised' discourse in all language combinations and directions (Sandrelli et al. 2007). As to the impact of speed on quality (all EP input speech is fast by normal professional standards), first results from EPIC are mixed, while initial ECIS studies seem to confirm previous intuitions that it can only be meaningfully assessed in combination with pitch patterns (Iglesias Fernández 2010). Meanwhile, initial analysis of the DIRSI corpus – designed to study directionality (see Table 2.1) – found no language-specific effect on ST-to-TT expansion or contraction (measured in words) or ST-versus-TT linguistic diversity, contradicting results for EPIC. Another promising departure in DIRSI research is an initial study of the use of discourse markers: pragmatic 'so' was used twice as much by the interpreter working into his/her native language (Bendazzoli 2010a). In the semi-experimental CIAIR corpus, Tohyama and Matsubara (2006c) have studied pausing and its effect on listeners. Meyer (2008, 'K6' or CoSi Corpus) has looked at proper names in consecutive and SI, while Setton and Motta (2007) correlated two ways of assessing quality – user reception and transcript scoring – with objective linguistic features of the TT in an experimental corpus of 24 expert and novice renditions of two speeches.

2.4.5.3 Strategies, SI Technique and Process Modelling

Time-aligned authentic corpora are an indispensable source of evidence in the quest to understand and model the cognitive processes and strategies in SI, a traditional fascination of interpreting studies. Abundant new data and technology have brought the prospect of persuasive evidence much closer than in the 'manually-challenged' days of Lederer's (1981) and Setton's (1999) studies. Pending the time-alignment of the EPIC data, ironically the most intensive corpus-fed research on interpreting strategies and processes has come from the machine-interpreting research group at Nagoya, who have published a credible classification of Japanese–English SI strategies (Tohyama et al. 2006a), fine-grained measurements of lag (Ohno et al. 2008), and analyses of human versus machine segmentation for SI – the former being most difficult to implement when discourse markers are involved (Ding et al. 2005).

2.4.5.4 Social and Ideological Aspects

The recent 'socio-cultural turn' in translation studies has reached CIS, with studies of intertextuality, norms or interpreter–participant interaction (Diriker 2001), or using linguistic and discourse indicators such as lexical density, repetition, metaphor and modality to probe issues like politeness and face (Monacelli 2005, 2009), the expression of ideology or cultural identity, or the weakening or strengthening effect of interpretation on ideological discourse (Beaton 2007).

2.4.5.5 Pedagogical Applications

The EPIC corpus is an organized, indexed sample of professional interpretation in an institution to which many European graduates will one day apply for a job, so is not surprisingly already being exploited for training purposes. As live input, EP speeches are extremely fast, and usually read out from text, thus limiting their use to advanced students (Bendazzoli 2010b), but Sandrelli (2010) recommends the use of corpus ST-TT concordances as a reference for students working into a B language. In Japan, the Nagoya group are developing a pedagogical tool (with a feedback function planned) to use the synchronized CIAIR corpus to demonstrate professional strategies and patterns of lag, chunking or self-correction to interpreter trainees (Tohyama and Matsubara 2006b).

2.5 Taking Stock

A doctoral student embarking on a corpus-based research in the early 1990s found little guidance in the interpreting studies literature (Dam 2001), while her (linguist) supervisors were sceptical even of the 'data' of corpus linguistics. Her epistemological and methodological stance, like that of other CIS researchers in the 1990s, is that interpretation of corpus data could be left neither to the individual researcher nor to a machine, but must be validated by 'intersubjective consensus' (see also Pöchhacker 1994; Kalina 1998).

As can be readily imagined, each of the few published corpus analyses described above demanded an enormous investment in time and painstaking manual transcription work, perhaps only possible for a doctoral thesis. Pöchhacker (1994) and Setton (1999) can be considered as exploratory starting points; Pöchhacker erected a comprehensive typological framework but found it ultimately hard to apply to utterance-level data, while Setton no doubt overloaded and over-annotated his corpus with features which did not all lead to interesting findings.

In the heroic phase of CIS, limited sample size and laborious data preparation probably played as great a part as over-enthusiasm for personal intuitions in leading corpus researchers to highlight extracts or examples that best illustrated their theses. Dialogue interpreting studies might present one or two case studies to illustrate points related to power and distance relations, and SI studies might identify examples of anticipation, or rephrasing, and present them as strategies or tactics, or as the natural result of deverbalization, depending on theoretical preference.

Taking the trouble to collect authentic performance data and identify examples in support of a thesis was already a significant step forward from an era in which readers were expected to accept ideal, prescriptive but unsubstantiated descriptions of the art of interpreting. The phenomena revealed were at least shown to exist, allowing for a theoretical debate about how to interpret them. Their value, not to be underestimated, is in forcing us to review assumptions about how psycholinguistic findings apply to interpreting. For example, if a synchronized corpus shows lags varying between 2 and 11 seconds with no corresponding drop in accuracy, one can no longer make general statements, often based on poorly digested psycholinguistic theory, to the effect that ‘items’ *cannot* be held in ‘memory’ for more than n seconds.

From the 1990s, however, a new generation of researchers began to question the representativity of these phenomena. Exemplification alone invited the charge of cherry-picking examples to illustrate theories of interpreting postulated a priori. This perception, added to the difficulty of disentangling all the potential factors in interpreting performance, as well as the pressure of rules of normal science as interpreted by some members of the interpreting research community, led them to disregard the study of natural corpora and seek greater scientific credibility by shifting the emphasis to experimental research, constructing modest controlled experiments in laboratory conditions. The methodological weaknesses of the corpus-analysis paradigm have naturally contributed to this trend.

However, experimental research has not performed visibly better in overcoming two of the main obstacles in the pursuit of scientific credibility for research on interpreting. First, the difficulty of quantifying the most relevant and interesting *variables* continues to beset all approaches: we have not yet seen more convincing replication of procedures and results from controlled experiments than corpus studies. Second, in all interpreting research, *samples* are still too small by the standards of ‘normal science’, whether measured in numbers of subjects in experiments, respondents to questionnaires, or the length of interpreting corpora. Finally, artificial and controlled experiments in the classroom or laboratory pose problems of ecological validity. The remainder of this chapter will examine how corpus-based research can clear these obstacles.

As already stated, the additional dimensions we have to capture in CIS are *orality, multilingualism* and *synchronicity*. This additional complexity will itself impose a choice of variables to study, and a further choice of those features we can represent and display in a transcript. What we can study will remain constrained by what we can reliably (and replicably) capture. This in itself divides potential research topics in two categories:

- those which can be pursued by capturing *measurable features in large corpora*, which can be averaged statistically as indicators over texts or the performances of interpreters;
- investigations into *local microprocesses*. These require synchronized input and output data, and more dauntingly, a framework for capturing pragmatically charged features, not the least of which is prosody.

After a brief discussion of options for transcription and some quantitative methodologies, we will address the prospects for progress on this more distant frontier.

2.6 Corpus Selection, Compilation and Design

Data collection is the first major hurdle in CIS, but is not insurmountable and may be exaggerated. Interpreters and institutions have often been shy to allow ‘raw’ recorded interpretation to be released, and in some cases have explicitly cited the fear that the tapes would compare unfavourably with official translated transcripts. This gives researchers all the more reason to find ways of showing the oral dimension in the transcription or commentary.

All the corpus analysts listed in Table 2.1 (see Section 2.4 above) agree that an initial exploratory corpus should be recorded at a real conference event. This does not mean the loss of all control. Some subject variables, like training and experience, are given by the interpreters’ profiles, although others, like the interpreters’ preparation – the basis for their initial projected mental model of the discourse – are difficult, though not impossible, to pinpoint. Discourse variables like delivery speed, technicality, register, genre (narrative, descriptive, discursive) or last but certainly not least, spontaneity (recited, semi-rehearsed, impromptu), can all be controlled by selecting particular speeches or extracts.

In the past, researchers have often selected a corpus which is broadly representative of professional practice ‘as it should be’, that is exemplifying semi-rehearsed, discursive speech[2] and moderate to average technicality of subject-matter. However, conditions change. In particular, SI of fast recited speeches without access to the prepared text is now a more than legitimate object of research. On the other hand, not all interpreting conditions, even everyday ones, are suitable for exploratory study. Completely improvised speech may be so disconnected and allusive (see, for example, the Watergate transcripts reproduced in part by Pinker (1994: 222–3) that any cohesion apparent at the event disappears in the transcription. This type of input requires complex inferencing, context construction (not to say guesswork) and special production strategies to an extent that it is difficult to correlate output with the incoming utterance, the remoter discourse record or background knowledge. Finally, wordplay and metalinguistic or culture-specific discourse may also call for explanation and paraphrasing strategies not specific to SI.

Finally, however impressive the material, researchers should remember that any corpus only illuminates (and may amplify) one setting in the variety of situations that constitute interpreting as a profession worldwide. In the EP corpora, as researchers recognize, speeches are almost all very short (2–8 minutes) and read out at very high speed (often over 160 wpm), switching rapidly between a wide variety of topics; and all interpreters are working only into their A languages (Bendazzoli 2010b: 62). More widely based generalizations about interpreting will have to await analysis of material from the UN system, for example, and a range of private market events on different continents.

2.7 Data Presentation

2.7.1 Transcription and Choice of Features

Parallel corpora of speeches with their interpretations can be rich resources even without synchronization, but they raise problems of cross-language comparison not found in ordinary corpus linguistics. To capture a full range of lexical, structural and prosodic patterns, both sound recordings and transcripts are necessary. Corpus-based research into the interpreting process usually requires synchronized (dual-track) recordings, or even video, opening up a range of technical problems. In addressing these, the researcher must also decide which features to search for, measure, analyse and/or display.

Researchers publishing *dialogue interpreting corpora* have given priority to readability over detail, usually sticking to ordinary manual transcription following orthographic conventions with a gloss added where necessary. Overlapping speech or interruptions may be marked, as well as signs of emotional involvement like laughter, shouting or sobbing, or significant hesitations, pauses and self-repairs. Additional information felt to be necessary is mentioned in the commentary with its interpretation by the researcher (e.g. Pym 1998 refers to 'increasing rhythm and mounting intonation' in a dialogue with a witness at the O. J. Simpson trial). In the case of dialogue interpreting, this preference for simple manual transcripts over systematic machine-readable text encoding is easily understood for several reasons, since analysis or commentary is usually based on:

(i) unmarkable features like word choice, which no one has as yet claimed to quantify and thus remain fully in the domain of conscious, explicit human analysis and interpretation;
(ii) multiple features taken together, which would obscure a transcript to the human reader; and
(iii) global features such as rhythm which are applicable to the whole text (Wadensjö 1998: 259).

Simplified representation of discourse is, therefore, justified by the type of analysis in this paradigm as well as for reasons of presentation. But authors like Brown and Yule (1983) go further, fearing, for example, that in representing intonation and prosody by punctuation (question marks, commas, dots), 'too much critical attention may be focussed on details of spoken language that were only ever intended by the speaker as ephemeral parts, relatively unimportant, of the working out of what he wanted to say', so 'DA [discourse analysis] is often prone to overanalysis'. They propose to 'generally ignore paralinguistic features in spoken language [. . .] since the data [. . .] is spoken by cooperative adults who are not exploiting paralinguistic resources against the verbal meaning of their utterances but are, rather, using them to reinforce the meaning' (1983: 9–12). Dialect and accent, for example, are often airbrushed out, for almost 'politically-correct' reasons, except where specifically relevant to a communication problem (see, for example, Krouglov 1999).

For researchers looking for clues to cognitive processes, such simplification goes too far and assumes too much, particularly about what a speaker 'intended'. Not all discourse researchers are satisfied with such minimal analysis. Cognitive linguists form another research community that prefers to work on natural corpora, but probes further into psychological and linguistic dimensions than the conversation analysts. For Dubois and Schuetze-Coburn (1993: 228), for example, discourse research needs to identify and address three layers or dimensions of hierarchical structure in language: turn structure, prosodic structure and grammatical structure. Nearly 30 years after Brown and Yule (1983), corpus linguists of virtually all persuasions are now agreed on the need to find ways to mark texts automatically, so as to be able to extract information from large corpora and thus enormously enhance the value of their generalizations.

The potential range of features to be marked and encoded embraces syntactic, semantic, prosodic, paralinguistic and extralinguistic features. Each researcher will have to make common sense decisions about text mark-up at three levels: which features to identify as relevant to their research, which to display in a publication for a human readership, and which to mark for encoding and automatic processing, with a view to possible quantification and statistical analysis.

Several complex transcription schemes have been developed, on the basis of the famous Gail Jefferson system, to show relevant paralinguistic features within or alongside a transcript; but they are unfortunately still largely competing and mutually incompatible. Three or four such systems are described in Edwards and Lampert (1993), for example, Dubois et al. (1993), Ehlich (1993), Gumperz and Berenz (1993). The main difficulties lie at the suprasegmental level, from intonation and rhythm to paralinguistic features like facial expression, posture, gesture and tempo, for which there are no agreed conventions and which are therefore usually ignored. These features pose problems at multiple levels: (i) recording, (ii) classification (iii) measurement of certain variables relative to others, for example, intonation against a baseline, or tempo over a specific passage relative to speaker's overall pace and pattern; (iv) interpretation, particularly of combinatorial meanings: in speech, prosody and paralinguistic features like gesture work in conjunction with segmental features like syntax and lexical choice to produce the basis for meaning construal. The challenges of CIS applied to signed interpreting can readily be imagined (Metzger and Roy forthcoming).

Manual transcriptions can be designed to be read comfortably by the researchers themselves, the readers of the published research report or by a computer program. In the first and last case, the transcript should show those features which correspond in some way to the variables used in the study, for instance as indicators or components of them. For SI, there are at least three possible presentations:

(i) *synchronized interlinear* transcription – with selected prosodic features (pauses, pitch or intensity stress etc.) and optional word-for-word gloss[3];

(ii) a *parallel* tabular presentation by aligned *segments* (up to three interpreted versions side by side), either roughly time-aligned or matched by content (and thus not limited to SI);

(iii) a *'fluent'* or 'clean' transcript, punctuated and with speech errors and hesitations eliminated, which can give a feel for the speakers' and interpreters' discourses as perceived by charitable listeners. In interlinear or tabular presentations, each line can be made to correspond to a fixed time interval (say 2s) provided this does not cramp the transcription (Setton 1997, 1999).

Until recently, SI transcription – and alignment, annotation, etc. – was only possible manually and was very laborious, so transcripts were presented mostly as plain fluent text, with only very short extracts in synchronized interlinear transcription.

New web-based tools are revolutionizing the presentation of corpus data, not least through gradual (at least semi-) automation of the most laborious operations: sound-to-text alignment, ST-to-TT alignment, segmentation and time-coding, as well as the inclusion of various other layers, from the environmental or circumstantial data in file headers to in-text tagging and annotations. With modern tools, it is a trivial matter to include multiple layers and dimensions of annotation that the user can display or hide on demand – as in the TEI (Text Encoding Initiative) standard, for example.

Universal standardization of corpus presentation is not a realistic goal, as institutions will continue to develop and use proprietary software for economic or other reasons. But current technology is already good enough to support targeted and detailed web-based querying of each database by multiple users, opening the way to data-sharing, replication and the whole apparatus of collaborative and competitive science.

This is not to say that corpus recording, transcription and presentation can or should ever be 'fully automated'. Many significant features in the infinitely rich spectrum of authentic, contextualized discourse will resist automatic capture. For this reason, perhaps, some seasoned transcribers and analysts believe that a single ideal transcription system is both unrealistic and counter-productive, and prefer to forego the ideal of multi-user 'flexibility':

> transcription [should] be limited to features to be subsequently analyzed, [which] in turn are determined [. . .] by the specific purpose of the research project [. . .]: only what is to contribute systematically to data analysis should be transcribed, and only what makes the data analysis intelligible should be presented (O'Connell and Kowal 1994).

Certainly, a printed transcript, like a traditional paper map, becomes unmanageable if it contains too much information. But technology should make it a simple matter to generate one transcription system for a scientific analysis, and another, simplified notational system for the reader, with the option to retrieve more features.

2.7.2 Contextualizing the Corpus: File Headers and Commentary

As authors like Pöchhacker (1994) in the *skopos* tradition have shown, no interpretation activity can be satisfactorily understood and studied outside the framework of the communication event (e.g. a conference) or without taking into account the type of discourse. Texts for interpretation cannot be classified in simple categories but must be characterized by a sheaf of parameters – an indispensable framework to ensure comparability between corpus studies. Pöchhacker meticulously described the event in which his corpus was embedded in terms of its place, date, title, aims and participants, then applied the taxonomic approach at a more local level by characterizing each speech along several scalar parameters (1994: 112–14; 158–60).

Table 2.2 shows some of the auxiliary information that different researchers have chosen or may choose to show for some large or multi-text corpora:

Table 2.2 Corpus metadata: classifiers and file header information

Header item/ Parameter	Pöchhacker 1994 (single event)	EPIC (single institution)	Nagoya CIAIR (heterogeneous)
Conference/event			
Date, speech no.	Date (no numbering)	date, speech no.	date, location
Source language	✓	✓	
Combination/direction	✓	✓ 'type', e.g. orig.>IT (sub-corpus)	[uniform JA><EN]
Duration/ timing	✓	✓ e.g. 'short'/85s	recording time,
Equipment	–	[uniform]	AV equipment
Interpreter	(anon. A/B/C)	[all pros]	yrs experience [all pros]
Speech duration/length in words	(duration only)	✓ 'short'/153 words	
Speaker information	(available via conf. program at 'hypertext' level)	Name, gender, nationality, native of SL, political group/ function	Speaker information; + speaker's role
Addressee orientation	dialogue/ monologue, 1–1, 1–many, broadcast etc.		'conversational task'; 'speech type'
Delivery mode	Read, (semi-) rehearsed, improvised ; material given interpreter	Read, impromptu or mixed	–
Topic	[general topic defined by event]	'Politics'/and specific topic	['daily topics']/ 'speech type'?
Media/A-V support	any slides, videos or graphics	yes/no	–
Tempo (speed)	in syllables per min.	wpm /low, med, high	–
Melody	intonation . . . (and?)	–	–
Dynamics	intensity or loudness	–	–
Rhythm	significant deviations from mean	–	–
Voice quality	significant deviations from mean	–	–
Articulation	significant anomalies	– [in separate studies]	–
Comments		– [in separate studies]	–

Pöchhacker also attempts parameters for a typology of interpreted events, such as: *Structuredness, Cultural/group homogeneity, Information density, Visual accompaniment* and *Information Flow.* Pöchhacker's strategy of working 'down' from the hypertext or encompassing event entails a preference for plain orthographic transcription, to allow readers an easy overview of the connections between different levels, placing 'traditional linguistic features' such as lexicon and syntax provisionally in the background and highlighting instead 'temporal and verbal-paraverbal' features: transcripts show pauses, hesitations, ellipses, unusual intonation, throat noises, laughter, applause, heckling and time segmentation. At the other extreme of the 'human-automatic' spectrum, machine-interpretation researchers in contrast show less interest in paralinguistic and extralinguistic factors.

In web-based sharable databases, this 'metadata' can be entered in file headers. Certain parameters common to all speeches in the corpus may be taken for granted [shown in square brackets in the table], but the information can be provided elsewhere, in commentary. In future, agreement on a standard header format is desirable, since parameters that are constant within one corpus and thus easily taken for granted or omitted (e.g. 'event' in EPIC) may become significant variables when comparing a broader range of samples. For example, headers in DIRSI, a corpus of medical conference interpreting corpus with a declared research focus on performance and quality factors (specifically directionality), provide for additional details on:

- the conference: title, topic, date, venue, type of session (open, presentation, discussion, closing), and type of speech event (opening/closing remarks, paper or lecture, floor allocation, procedure or housekeeping announcements, Q or A, comment),
- the participant speaking: speaker, organizer, sponsor, chair, discussant, presenter or lecturer, audience, interpreter; name [except interpreters], gender, country, language, native speaker or not.
- materials supplied to the interpreter: not only *if* supplied but also *when* (in advance, on the spot, or not at all).
- directionality: whether the interpreter is working into an A or a B language.

2.8 Automated Corpus-Analytic Tools

2.8.1 The Soundtrack

Audio display and acoustic analysis tools are both now readily available. Digital recording and speech synthesis software (e.g. Cool Edit Pro 1.2, or Praat 3.9) display the main prosodic features of a sound sample like pitch and intensity (and indeed, for experimental work, usually allow individual manipulation of these features: Seeber (2001), for example, artificially flattened intonation contours to a speaker's mean base frequency and compared interpreting performance from these 'flat' versions and the original 'lively' versions). A transcript can also be aligned with the frequency spectrogram as it is typed, or later, using programmes like Winpitch which allow an operator to click on the text segments as (s)he hears them; the programme automatically aligns the text to the sound, which can then be displayed with the F0 spectrogram. Simultaneous acoustic analysis of both channels of stereo recordings is also possible in principle. Timecoding programmes used for subtitling, like *Aegisub*, may also be adaptable.

Video is an important dimension, not only for specific research purposes – for example, the possible impact on performance of a more or less restricted view of the room, or of speakers' body language – but also to gain a better overall impression of the event. The future researcher will no doubt have access to a complex display of speakers and interpreters in audio and video, with scrolling annotated multitrack transcription, prosodic contour, affect level indicators and other exciting effects; another challenge will be in gradually learning to read it.

2.8.2 Transcription

Reliable fully-automatic sound-to-text transcription is still a long way off. Dictation programmes (such as *DragonNaturallySpeaking*), even if fed with high-quality voice recordings, still need several hours of training for each different voice to show a real benefit in residual editing time over manual transcription, while transcription support programmes (Transcriber, WinPitch Corpus) basically offer little more than user interface comfort for what remains a manual exercise. However, the EPIC group has thought of a simple but clever expedient to streamline the transcription of large corpora: researchers train the dictation programme to recognize their voices then shadow each speech (repeat it aloud), generating a first draft which is then manually corrected and completed against the original. In EPIC's case, the dictation step is only needed for the interpreted versions. For the original speeches, the Parliament's 'cleaned' verbatim reports can be used as drafts to restore them to their original raw oral state.

This still leaves prosodic, para- and extralinguistic features to be inserted if needed for the research purpose.

2.8.3 Alignment and Segmentation

Alignment of source texts with their translations is now possible, and for some texts almost perfect, with programs using sentence-length probabilities, or building constantly refined databases of probable word correspondences, with word and sentence information being polled alternately to achieve near-perfect alignment of sentence to translated sentence in a few passes (Kay and Röscheisen 1994, Gale and Church 1994).

However, performances fall sharply when sentences in the ST and TT are in more complex correspondence (1–2, 2–1, 3–1 or even 3–2). This occurs often in interpretation, where compression, chunking and paraphrase are routine SI processes, making such complex correspondences the norm. More basically, 'sentences' are harder to pin down in speech than in writing, without some artificial segmentation at the transcription stage which might be distorting. In capturing the pragmatic dimension of a discourse, or two parallel discourses, for assessing fidelity, for example, the 'equivalent' of a segmental feature may be suprasegmental, and vice versa. All this suggests that some form of segmentation more suited to the nature of oral speech will ultimately be less trouble, even if we lose the apparent spectacular help of the alignment programmes described in the literature.

The main candidates for segmentation are *syntactic*, *prosodic* and *researcher-defined*. For reasons already given, compounded by cross-linguistic differences (e.g. languages like Chinese which have a Topic-Comment structure), the clause is not necessarily a useful unit, and the proposition is also too fuzzy. Dam (2001) found that propositional analysis *à la* van Dijk and Kintsch (1983) reflected the surface forms of expression too closely, and finally settled on a workable, verb-based unit by trial and error, occasionally merging segments to handle residual problems due to the literate-oral shift (Shlesinger 1994), when interpreters used more verbs than the original (often verbalizing noun phrases). Similarly, Setton (1999) found that a spontaneous Chinese discourse could quite easily be segmented heuristically into 'idea units' (which in consecutive interpretation notes might be marked off by horizontal lines), usually comprising some rhetorical prefacing, a Topic-Comment (or Subject-Predicate) nucleus or complex phrases, and an 'afterthought', as in the example below:

> *'I think – talking about this joining the WTO – in terms of intentions – I think government and opposition– I think in fact already reached unanimous consensus – this is an undeniable fact . . .* '

However, it is not clear how well such a procedure, or indeed any of these expedients, could be standardly formulated to allow reliable replication.

What about *prosodic* segmentation? According to Dubois and Schuetze-Coburn (1993: 229), 'the intonation group (or 'tone group') is coming to be seen as the primary prosodic unit of spoken discourse, revealing both information flow and mechanisms for structuring the interaction', and Svartvik et al. (1982) suggest it should replace syntactic units as the primary unit of linguistic analysis for spoken language. Discourse analysts generally agree that intonation and turn-taking should have the first claim on layout; it is easy then to overlay a representation of syntactic structure (ibid. 247).

Dubois and his colleagues claim that starting from intonation unit structure helps rather than hinders representation of syntactic structure, noting in addition that intonation units are more straightforwardly discrete and sequential than clauses and less variable in size. They propose showing the 'concurrent hierarchies' of syntax and intonation in tree format in either dimension (converted here into nested brackets): [Text [Turn [Intonation Unit [Group [Word]] or [Sentence/Clause [Phrase ('Group') [Word]]. However, their definition of a 'group' again seems rather too loose for reliable replication:

> roughly, [a group] may consist of a NP (minus any phrasal/clausal modifiers), a verb complex (including auxiliaries and most adjacent adverbs), or other nonclausal elements such as predicate adjectives, prepositions, complementisers and so forth. (Ibid. 237–8)

But as they also point out – crucial for our purposes – the system needs to be adapted to different languages. For example, 'in English PPs are counted as a prepositional group plus a noun phrase group (two groups), whereas in German P- and N-headed dependent phrases are all basic-level arguments [. . .] and PPs are to VP as oblique case-marked NPs, so Ps are grouped with NPs . . . ', while on the other hand, 'in German, adverbs are best separated from adjectives, since adverbs like *natürlich, noch, eigentlich* will each form separate units'.

According to these authors it is 'too much trouble for the analyst to sort adverbs into different types by scope and function; such fine distinctions are best left for later, at coding level' (ibid. 250–1). However, in a (cognitive-) pragmatic analysis the function of such items and their scope – whether over a phrase or the whole utterance – may be critical. Recent work has shown how some items in the broad class of adverbs or particles, such as attitudinal and evidential expressions (*frankly, I think, probably, apparently*) contribute to propositional content, while others, like *well, anyway, so, doch, ja, d'ailleurs, pourtant,* have a procedural function, helping hearers to inferences (Ifantidou 2001, Carston 2002). The latter are discourse markers with key signposting functions and cannot be dismissed as 'secondary information' in any study interested in meaning.

For the foreseeable future, the prosodic dimension is likely to remain both elusive in its interaction with syntax and semantics in a specific language, and resistant to cross-linguistic generalizations. Discussing syntax–intonation correlations, Cruttenden (1986: 79) avers that 'the strongest statement that can be made is that syntactic cohesion is generally stronger within intonation groups than across them.' He clings to the idea that there is a link between the traditionally defined clause and the intonation group – they coincide on an estimated 40 per cent of occasions – but he gives a list of common 'exceptions' so long[4] that it suggests a different analysis should be sought.

2.8.4 Time-Coding and Synchronization

Unlike CL and CTS, a corpus suitable for studying SI processes has to be time-coded and synchronized to at least as fine a grain as the proposed segmentation. Among the recent large corpus projects, the CIAIR corpus is automatically segmented prosodically using pauses of at least 200 ms (Tohyama et al. 2005). However, the Nagoya group also reports developing a (proprietary) internet-based tool comprising a text base in which the user can align segments by clicking on the bilingual text display, and a parallel 'time-chart', enabling alignment and synchronization fine-grained enough to study lag as well as interpreters' restructuring of fairly short strings. But again, in this simplified system, 'fillers, and utterances with no obvious counterpart, can have no correspondence' (ibid.).

EPIC began by exploring mixed prosodic/syntactic criteria with some assumed semantic correspondence, segmenting 'into meaning units on the basis of [. . .] intonation and syntactic information' (Monti et al. 2005), and are now exploring automatic pause-based segmentation of audiofiles with *Speechindexer*, and manual alignment of audio, video and text with time-codes, using *Transana*, or on indexes in text and audio files, using *SpeechInterpreter*.

Segmentation by pauses of a certain length may seem easier to automate, although EPIC reports problems with pause detection due to booth noise in the interpreted versions. But as the only criterion, it may not be helpful for analysis, given the multiple reasons for pausing in both original and interpreted speech. In process research, an instructive experiment might be to segment the interpreted speech first, syntactically or prosodically, and see what ST segments the bursts correspond to.

Ultimately it seems that, as with other aspects of data analysis and presentation, a universal principle for segmenting interpreting corpora will remain elusive, and perhaps counter-productive. Segmentation proposed by corpus providers, if motivated and consistent, should not be an obstacle to evaluation and some comparability, failing perfect replication. Ideally, researchers should be able to overlay their own segmentation. An automatic and very fine-grained time-coding system, independent of the choice of segments by different researchers, may be the key to flexibility for different users.

2.8.5 Tagging, Annotating and Indexing

Once again, in deciding on tagging and annotation we must be aware that – as the discussion in the previous section showed – the schemas we inherit from

CL are based on *descriptive linguistics*, which was developed for a different purpose: to describe *langue,* not to illustrate the functional impact of *parole*, for which no agreed or even consensus system has yet emerged.

The type and amount of feature tagging or annotation that is necessary or appropriate will again depend on the research goals pursued. But the prospect of web-based corpora that can be shared by multiple users seems to lead inevitably either to a maximalist approach, in which every kind of annotation is performed and retrievable (though obviously not necessarily simultaneously visible), or to a minimalist provision of the corpus in a form that can be downloaded and annotated by each researcher for his/her own purposes. However, indexing must almost certainly be done by the corpus compilers and depositors.

For researching *lexical* choices and correspondences in parallel corpora, and for pedagogical applications, parallel concordancers (e.g. *ParaConc* – Barlow 2002; http://paraconc.com/ – Barlow 2009) have already made a seminal contribution in CL and CTS. For performance and process studies, concordancers will obviously not yield judgements about the appropriateness of word choices, but might be used to provide an indication of the *amount* of processing (use of context and constructive inference) done by a translator or interpreter, as reflected in the amount of deviation from standard or 'dictionary' lexical equivalents.

As to *syntactic structure*, POS taggers, lemmatizers and automatic parsers are now probably available for a score of languages.[5] Lemmatizers are necessary to count lexical types and tokens and are therefore indispensable for lexical frequency studies, but will also be significant labour savers in the future if they can be calibrated to measure syntactic complexity, making it easier to study controversial issues such as the impact of SL-TL word-order differences on SI performance, and to verify hypotheses or earlier findings about shifts on the oral/literate continuum, simplification, etc. which have syntactic as well as lexical aspects (e.g. hypotaxis). Parser output (bracketing or tree structures) is obviously not displayed in published transcripts, but these should contain references to sites where fully encoded versions can be found in a standard like TEI in which multiple features can be encoded and selectively displayed.

Some *text analysers* compute syntactic data and even semantic and stylistic profiles data for any text over 600 words or so in length. *CordialAnalyseur,* for example, delivers parse trees of French text, POS percentages, TTR and function/content word ratios, as well as an ontology-based semantic profile and an evaluation of style, readability, technicality, abstractness, etc. Of course, language and culture specificity confine the use of such analysers to inter-subject comparisons; and these features alone, common to written and spoken text, are not valid indicators of input difficulty or hearer-friendliness ('readability') of input or output in *spoken discourse for interpretation* (even without taking into account the pattern of information delivery and prosody).[6] Also, CIS researchers will want to construct their own task-specific indicators, for example from data on pausing, lexical repetition, sentence complexity or the use of connectors.

2.8.6 Standards for Data Sharing

Since more than one centre may be compiling and analysing large corpora in the future, sharing makes sense, and some consensus on what to encode and display and how would save a lot of unnecessary work. Given the rarity of usable authentic corpora, compilers should aim to extract, encode and store as much information as possible, with the possibility of searching and displaying it selectively, as provided for example in a standard like TEI, which allows selective display of different layers of information in a corpus database, allowing researchers to generate representations at different levels.

Compilers should aim to provide several layers of encoding :

1. *Digitized audio and or video tracks,* synchronized (for SI) across the various languages provided.
2. *Plain (orthographic) transcription*: a 'fluent', readable text, with optional retrievable display of features like disfluencies, speech errors and prosodic contours, aligned and displayable with the synchronized audio files. Russo (p.c.) suggests transcribers use simple machine-readable processors like TextEdit (.txt) rather than programmes like Word.
3. *Fine-grained time-coding*, for the study of temporal patterns (like EVS), and so that each researcher can measure speeds, for example, as preferred (wpm versus spm).
4. *Prosodic profile*: the plain text transcription can be aligned with frequency and intensity traces (spectrograph) and chunked into intonation groups, both of which could be number-coded in the database.
5. *Syntactic profile*: the output of syntactic analysis (part-of-speech tagging and parsing) should be encoded in to allow selective display, if possible, of individual or multiple parts of speech, and specific sentence structures.

Segmentation grids, proposed by a corpus compiler or researcher (for instance into his/her chosen idea units) and coded for concordancing and alignment with the other profiles, could also be proposed for *optional* use (not hard-wired into the database). One example, from comparative discourse analysis of parallel texts, is the use of *lexical cohesion*-based grids that chunk texts into a sequence of subtopics (van der Eijk 1999).[7]

The peculiarities of SI suggest that interesting findings might be accelerated by capturing oral features alongside the syntactic tagging and lexical analysis we inherit from traditional corpus (text) linguistics. It seems clear that even a rough representation of any interesting aspect of natural speech depends on getting some handle on the prosodic dimension, and on language- and culture-specific information about the import of certain prosodic patterns in speech. Some progress is being made in semi-automatic tagging of intonation (Campione and Veronis 2001), and affect and emotion can also be detected in voice contours by programmes developed for psychotherapeutic applications (Stassen 1988, 2002).

For analysis, research on intonation in different languages (Hirst and Di Cristo 1998) has tentatively correlated some basic contours with implied speaker meaning – for example, falling/closing versus rising/continuing intonation – in different languages (Hirst and Di Cristo 1998), and some language-independent patterns (see Vaissière 1983), which it might be possible to overlay locally on text, or to assess at least whether prosody is helpful or misleading over a particular passage (as a complement to artificially flattening speech as in experimental studies).

Apart from the features that are encoded, accessible or displayable, a key factor in facilitating data sharing is the user-friendliness of the query tools. Alongside XML-based tools like Sara or Xaira (for TEI), new standards being developed include the CorpusWorkBench (CWB-CQP) standard adopted notably by EPIC, DIRSI, via the C-ORAL-ROM project at the Universidad Autónoma de Madrid.

'Practisearchers' have already realized that this will entail working closely with computational linguists and engineers – often with a steep learning curve – just as interdisciplinary collaboration has been necessary, and fruitful, with statisticians, neurolinguists, cognitive psychologists and sociologists.

2.9 Analysis

In dialogue interpreting, in particular, corpus studies have been heavy on text interpretation, but so light on the volume and range of the corpus taken into account in each study that they often fail to qualify as 'scientific' by most accounts. Such studies are not replicable as such, and must therefore either be evaluated on a par with literary criticism or as contributing to an exploratory, hypothesis-forming phase in investigation. Subjectivity can be reduced by using panels of judges – 'intersubjective consensus' – but this is cumbersome and also encounters volume limits, even when latin squares or similar statistical devices are resorted to to reduce the load on each judge. Moreover, qualified judges are as hard if not harder to recruit for research than interpreters.

If interpreting studies are to aspire to contributing not only to philology but also to cognitive science, we must meet contemporary research norms that demand not only a minimum sample size – in terms of numbers of experimental participants, questionnaire respondents or corpus size – but also credible methodologies that can be checked and replicated, transparent processes of inference from data, and discussion that engages and responds to existing theories. As we have seen, there are already bright prospects for the 'spadework' of CIS to be dramatically reduced, leaving the challenge of conceptualizing research questions, and designing analyses, that exploit this rich vein of data while approaching the rigour traditionally associated with experimental studies.

In analysing a corpus to clarify a research question, there are at least two ways of avoiding the charge of 'cherry-picking': one can either justify treating certain features as indicators of a phenomenon, count them globally in a speech or sub-corpus and make statistically based inferences; or one can define more complex items (a particular sentence structure, for example), find all instances in ST, and study their correspondences in TT, refining the hypothesis by iteration, extension to new samples, replication etc., in the light of different theories.

2.9.1 Measurable versus Interesting Variables

There is often a temptation to regard 'quantitative' analysis as intrinsically more rigorous and reliable than 'qualitative' analysis. But unfortunately, the most countable features in a body of speech that was originally used for communication may not be the most interesting ones.

While the dominant research orientation has been somewhat different in community interpreting (mostly sociolinguistic) and conference interpreting (mainly psycholinguistic), CIS in both modes faces similar methodological and epistemological challenges in moving from the identification of features to their interpretation. Figure 2.2 shows some features of discourse, of concern (albeit unequal) to the two research communities, starting from a core of 'objective' features and fanning outward to qualities which may correlate with composites of these core features, but must necessarily be evaluated by human judges.

Both research groups will be concerned with *fidelity*, and thus potentially, with the whole range of its potential determinants, linguistic and paralinguistic (central column); from these, both will make judgements about fuzzier values like *style, register,* and *lexical connotations*. From there, the research focus diverges: dialogue interpreting research typically seeks to infer social values like involvement, distance or face, while conference interpreting research is traditionally – though not exclusively – more interested in cognitive factors like difficulty or effort. In both branches, the more measurable 'inner' values will necessarily be taken as indicators of 'outer' variables of interest which cannot be directly measured.

Moving outwards from the core to the periphery we go from observable and quantifiable facts to inferences and interpretations. The 'observables', or the

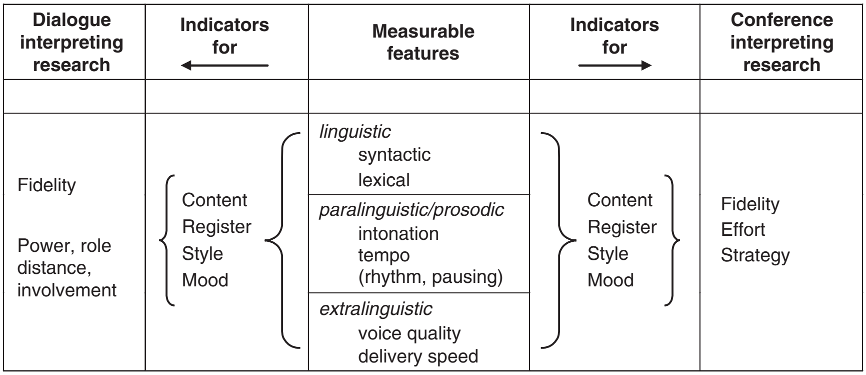


Figure 2.2 Measurable versus interesting variables in interpreting research

nearest we have to them in linguistics – that is, linguistic features of text – range from phonetics through syntax to prosody (measurable changes in pitch, rhythm and intensity). These must then be 'semantically' interpreted, for example, in postulating that a particular word order is marked, denotes new information, contrasts or emphasizes; or that a particular word or phrase means so-and-so, or carries these overtones; finally, we make pragmatic, sociological and psychological inferences: intonation contour N conveys surprise, appeal, concession, irony; lexical choice A connotes low social class, cooperativeness; particle or function word X denotes greater or lesser commitment or certainty.

2.9.2 Composite Indicators for a Complex Phenomenon

Given the range of potential factors in interpreting performance, there might be some promise in measuring a wide range of variables at different levels and experimenting with correlations, including correlating selected clusters of variables treated as composite indicators. This approach has often been fruitful in the exploratory early stages of studies of complex phenomena. In terms of Figure 2.2, to improve 'resolution' across the spectrum from the uninteresting measurables in the middle outward to the more interesting end-variables at the poles, measurable entities must be taken as indicators or proxies for the target qualities we are interested in, betting on the probability that specific clusters of such microvariables, duly tested and developed, will be workable indicators of the more general features.

Lamberger-Felber (1998) took a step in this direction, measuring several low-level features of renditions of SI with text, like word length and frequency, as well as omissions and errors, in a sample of 12 interpreters' versions of the same text; but the range and type of variables remains too modest to reveal more than the most predictable correlations (e.g. long omissions correlated with time lag).

Setton and Motta (2007) tested this methodology by correlating three sets of potential indicators of the quality of 24 French interpreted versions of two English speeches: (a) users' judgements, (b) scored features of the written transcripts (paraphrasing, elaboration, accuracy, style, fluency) and (c) linguistic features generated by a text analyser (TTR, sentence length etc.) – corresponding to Shlesinger et al.'s (1997) 'instrumental (user reception), intertextual and intratextual' dimensions. The study explores the possibility of constructing composite indicators: for example, clustering 'elaboration' and 'paraphrase' as a macrovariable for 'autonomy', and correlating this with errors-omissions-weaknesses on the one hand, and external quality assessments on the other.

Visualizing multiple variables and their interactions ('eyeballing the data') can also reveal patterns that might otherwise be missed and generate hypotheses. The ECIS group is using a pilot application, Trendalyzer, to display groups of variables over time or specific combinations of two or more variables at fixed points (see Barranco-Droege et al., forthcoming).

2.9.3 Accountability and Theoretical Refinement

One recognized method of exploring a hypothesis experimentally is to create instances of an independent variable in a ST (from scratch or by manipulating an authentic sample) and analyse the TT responses. A well-known drawback of natural corpus-based research is the rarity of occurrences of some phenomena, creating the temptation either to use a corpus that may not be representative in other ways, or to generalize from a small number of instances. From this point of view, the advent of large interpreting corpora is a boon, although some research questions will still need preliminary preparation and analysis that will limit the size of the usable corpus. For example, Setton (1999) compared the information contained in the interpreters' production at a given instant to what could be deemed to be encoded in the preceding chunk of input within the span of working memory, as a window on the alleged extra difficulty of SI from SOV into SVO languages, applying cognitive (memory, knowledge schemas) and pragmatic theories (relevance theory on the use of available context in speech communication). However, this required, among other things, a synchronized transcript, with some representation of both syntactic structure and information structure (showing the incremental delivery of information as the discourse unfolds), and a model of likely available context. This was hardly feasible in a large enough corpus (even had one been available), making a merely illustrative methodology inevitable.

To uncover some phenomena in 'parallel' and process analysis, it might be necessary to start from the TT, or even to trawl the corpus 'manually'. As concordancers (and translators and interpreters) know, ST and TT items often cannot be matched one-to-one. A TL utterance with little or no formal lexical or syntactic correspondence to a SL utterance may function perfectly as its equivalent in a specific context, for example, *J'aiprisquelqu'un* for *I hired a worker* (an example from Seleskovitch; the only relevant implicature here is 'instead of doing it myself'), or even in any context, as in *Don't get mad, get even* for *la vengeance est un plat qui se mange froid.* Alternatively, pragmatic procedures, like foregrounding for emphasis or signalling topic change, are often effected by syntactic means in SL, but prosodic devices in TL, or vice versa. Finally, elements 'added' in TT and contributing to effective and faithful interpretation may not be traceable to any element expressed on the surface of the source text (Setton 1999; Bendazzoli 2010b).

In probing this complex material, the challenge will be to tease out the most elusive aspects of interpreting, including stimulus-response patterns in different contexts that reveal something about processes, in addition to general statistical trends, while at the same time ensuring credibility by making the effort to define the variables and hypothesized processes as carefully and explicitly as possible to allow testing and replication by others. Refutation is not failure, as graduate students sometimes think, but an opportunity to refine the theory. We now have the data and technology to assemble a valid database for such studies, but as the above examples show, some fairly sophisticated theory is needed to formulate hypotheses and interpret findings.

2.9.4 Combining Methodologies

Finally, CIS need not compete with other methodologies, but may usefully be combined with them. For example, inferences can be drawn from regular differences between the interpretation of a speech in real-life, and later in controlled laboratory conditions, which is itself a form of manipulation of the situational context variable (Setton 1999: 176). Again, in research on SI strategies, instead of choosing the IV based on a priori assumptions about what may pose a problem – word order, for example – we can first identify real-life 'problem triggers' and truly 'strategic' (i.e. conscious) behaviour, at least, with methods like Ivanova's controlled retrospection technique of playing SI recordings back to interpreters to identify areas of conscious difficulty in an experimental sample, then using this information to choose similar items to focus on in a large natural corpus (Ivanova 2000). Finally, we have already mentioned triangulation, to 'close in' from different angles on factors in quality, for example, user reception, ST–TT equivalence and product features (Setton and Motta 2007). Other fruitful combinations of the standard methodologies – introspection (interviews, questionnaires, surveys), corpus analysis and controlled experiment – will certainly be found, with a little imagination – a key ingredient of successful science.

2.10 Discussion and Conclusion: Theoretical Maturity for Better Data

Automatic text analysis has already yielded valuable findings in corpora that are now large enough (as in EPIC) to be representative at least of a particular setting, and even a particular but mainstream genre of SI, defined by a sheaf of constants (professionals with similar qualifications working in standardized conditions, all into A, on fast, read-out speeches, etc.). This work has already largely 'caught up' with CTS, or has the potential to do so, for both comparable or parallel corpus analysis, and is going beyond CTS to address orality-specific features (e.g. disfluencies).

After collecting raw natural data, then ordering it, the next step in scientific enquiry is to form and test descriptive, then explanatory hypotheses, using the best theories available, while actively imagining and patiently eliminating alternative explanations and adjusting the model accordingly. However, as CIS enters an age of abundance and precision of data, it may soon find itself constrained by a relatively simplistic theoretical apparatus that still bears the traces of earlier mechanistic and literal conceptions of language and cognitive processes. This underdevelopment can be exemplified in the communicative and cognitive dimensions respectively, and is at least partly curable.

(1) Some delay in applying the latest models of *speech communication* being developed in modern pragmatics is understandable. On the one hand, the old tropes dressed up in discourse analysis still do plausible, though impressionistic service, to illustrate differences in register or status between interlocutors, or some emotional colouring of a discourse. Also, the features of speech that contribute to incremental meaning assembly on line, which extend into para- and extralinguistic dimensions, are extremely hard to capture and represent in a corpus, and interact in a highly complex way. And theoretical work in this area (in post-Gricean cognitive pragmatics) is still a work in progress, not to mention ready-made applications to interpreting. Fully grasping orality, so to speak, will take time and patient effort.

(2) Slowness in updating our models of *memory* is less excusable. While some researchers are clearly aware of sophisticated recent models that reflect interaction between different kinds of memory, or the role of knowledge activation and organization, they remain to be applied to the analysis of authentic interpreting corpora, which are rich in examples of varying lag without consequent loss or omission, demonstrating the concurrent operation of some form of memory not purely verbal or phonological, and thus *not* subject to articulatory suppression – a result that in no way contradicts Baddeley, except on the most superficial reading.

Visualizing time, prosody and environmental information around a speech will certainly give us ideas to explain the phenomena we will see in our newly displayed data – recasting, anticipation, added cohesive devices and so on. But it will be impossible to explain them without some model – be it from schema theory, relevance theory, long-term working memory or mental modelling – of how contextual and conceptual representations of meaning are derived and mobilized both from the speech itself and from elsewhere, and contribute to the interpreting process without being subject to the tight temporal constraints of manipulation of verbal forms in the phonological loop.

This theoretical upgrade will entail hard work by a new generation of researchers willing to go beyond the first graduate thesis aimed at demonstrating basic research ability (and where we often find, inevitably, hasty and superficial use of the most convenient heuristic theory). The only sufficient motivation for this, in a small and underfunded discipline, will be solid findings and generalizations that can be applied to all interpreting – and perhaps contribute beyond this speciality – by taking into account the most significant variables of speech type (or event genre), speed and mode of delivery, wide enough multilingual coverage to check for language-pair specificities, and enough comparative novice-expert studies to sort out what can and can't be done with more or less technique and experience – a valuable guide to designing 'last mile' pedagogy for advanced trainees.

For determined researchers, obtaining more and varied corpora to which the new techniques can be applied (from the ECIS and EPIC laboratories – and perhaps from Japan, though currently proprietary) should not be insuperable. UN material, audio and video, has been provided in the past (Chernov 1978). All institutions organize some public interpreted events, some of which are web-streamed with their interpretation, and some large private-market convention organizers already publish full audio proceedings and/or transcripts, or are prepared to let researchers use all or part of them (Pöchhacker 1994; Kalina 1998; DIRSI medical conference corpus, Bendazzoli 2010a). Corpora of television interpreting are being collected, notably in Italy (e.g. FOOTIE (Sandrelli forthcoming) and CoRiT (Straniero and Falbo forthcoming)); see Table 2.1, in Section 2.4 above).

In summary, CIS has the potential to provide reliable and much-needed new evidence about interpreting: nothing can replace the study of production in natural (socially determined) conditions. CIS has learned from CL and CTS, and now has the data and tools to come into its own. However, the added dimensions of interpreting – multilingualism, orality, situatedness and immediacy – pose special challenges for capture, multivariate analysis and theoretical imagination, synthesis and adaptation. Consensus on some basic standards for corpus compilation, encoding and display will allow data-sharing and replication, and the freedom to try new angles and methods of analysis, setting in motion the multiplier effects characteristic of the emergence of any fruitful new paradigm.

Acknowledgements

This chapter would not have been possible without the help of colleagues too numerous to name, but in particular Claudio Bendazzoli and the whole EPIC team, and Angela Collados-Ais and the whole ECIS team, as well as Franz Pöchhacker, Alessandra Riccardi and Miriam Shlesinger, for their valuable advice and help with references.

Notes

[1] The full text and 19 tapes are available to order from Copenhagen. The languages are Da-En-De-Fr-Nl-It (see Table 2.1).

[2] Semi-rehearsed speech is perhaps the most common register in international conference practice, lying between spontaneous dialogue, as in the negotiations in Lederer's (1981) corpus, and speech recited from text, which is also quite common. Furthermore, conference discourse typically belongs to the discursive or argumentative genre; hardly ever is it primarily descriptive or narrative.

[3] Chinese discourse was transcribed in standard Hanyu Pinyin romanization, with a *hanzi* (Chinese character) transcript provided separately.

[4] The list includes including sentence or clause-modifying adverbials (attitudinal, evidential), time and place PPs, Subject NPs (especially when long, postmodified, or topicalized contrastively), Topic/Subject when enlarged or recapitulated at clause end, topicalized object, tags like 'isn't it', *AREn't they?*, agentive by-clause following a passive verb, parentheticals, nouns in apposition (tonal harmony with tone on original noun), and rhetorical parallel phrases.

[5] ConnexorMachinese Syntax (12 languages); Cordial for French; Cosmas, Brill-Tagger or Morphy for German, Freeling for Spanish, French and English, etc.

[6] Significant work is being done in France both in the development of practical systems for annotating prosody in spoken corpora (e.g. Campione and Véronis 2001) and on the analysis of the multilayered functional structure of prosody in speech (Caelen-Haumont and Keller 1997).

[7] Van der Eijk suggests that, while there are no known computational mechanisms to compute coherence (based on relations such as elaboration, exemplification and cause), it is possible to automatically detect a less complex relation, cohesion, which arises from back-references, conjunction or lexical cohesion (reiteration of word forms either directly, or by hypernyms (peach: fruit) or semantically related words (garden, digging)). These relations are used to compute the similarity of adjacent text segments (Van der Eijk 1999: 3–4).

References

Baker, Mona (1995) 'Corpus in Translation Studies: An Overview and Some Suggestions for Future Research', *Target* 7(2): 223–43.

Barik, Henri (1973) 'Simultaneous Interpretation: Temporal and Quantitative Data', *Language and Speech,* 16: 237–71.

— (1975) ‘Simultaneous Interpretation: Qualitative and Linguistic Data’, *Language and Speech*, 18: 272–97.

Barlow, Michael (2002) ‘ParaConc: Concordance Software for Multilingual Parallel Corpora’, in *Language Resources for Translation Work and Research,* Las Palmas, Canary Islands, 20–4. Available online at: http://www.mt-archive.info/LREC-2002-Barlow.pdf (accessed 20 October 2010).

— (2009) *ParaConc and Parallel Corpora in Contrastive and Translation Studies,* Houston: Athelstan. Available online at: http://paraconc.com/ (accessed 20 October 2010).

Barranco-Droege, Rafael; Emilia Iglesias Fernández, José Manuel Pazos Bretaña and Jessica Pérez-Luzardo Díaz (forthcoming) *Maximizadores y minimizadores de la calidad en interpretación simultánea*. Granada: Comares.

Beaton, Morven (2007) ‘Interpreted Ideologies in Institutional Discourse: The Case of the European Parliament’, *The Translator* 13(2): 271–96.

Bendazzoli, Claudio (2010a) ‘Il corpus DIRSI: Creazione e sviluppo di un corpus elettronico per lo studio delladirezionalità in interpretazionesimultanea’. Unpublished PhD Thesis, Alma Mater Studiorum, Università di Bologna.

— (2010b) ‘The European Parliament as a Source of Material for Research into Simultaneous Interpreting: Advantages and Limitations’, in N. L. Zybatow (ed.) *Translationswissenschaft – Stand und Perspektiven. Innsbrucker Ringvorlesungenzur Translationswissenschaft VI (Forum Translationswissenschaft, Band 12).* Frankfurt: Peter Lang, 51–68.

Bendazzoli, Claudio and Annalisa Sandrelli (2009) ‘Corpus-Based Interpreting Studies: Early Work and Future Prospects’, *Tradumatica 7: L’aplicaciódels corpus linguistics a la traducció*. Available online at: http://webs2002.uab.es/tradumatica/revista/num7/articles/08/08art.htm (accessed 20 October 2010).

Bendazzoli, Claudio, Annalisa Sandrelli and Mariachiara Russo (this volume, Chapter 12) ‘Disfluencies in Simultaneous Interpreting: A Corpus-Based Analysis’.

Biber, Douglas, Susan Conrad and Randi Reppen (1998) *Corpus Linguistics: Investigating Language and Use*, Cambridge: Cambridge: CUP.

Brown, Gillian and George Yule (1983) *Discourse Analysis,* Cambridge: CUP.

Caelen-Haumont, Geneviève and Eric Keller (1997) ‘La prosodie, de la parole à la synthèse: l’apport de la sémantique et la pragmatique’, in Eric Keller and Brigitte Zellner (eds) *Les défis actuels de la synthèse de la parole,* Université de Lausanne: Etudes de Lettres, 103–30.

Campione, Estelle and Jean Véronis (2001) ‘Semi-Automatic Tagging of Intonation in French Spoken Corpora’, in P. Rayson, A. Wilson, T. McEnery, A. Hardie, S. Khoja (eds) *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, U.K.: Lancaster University, UCREL, 90–9.

Carston, Robyn (2002) *Thoughts and Utterances,* Oxford: Blackwell.

Cencini, M. (2000) ‘Il Television Interpreting Corpus (TIC). Proposta di codificaconformeallenorme TEI per trascrizioni di eventi di interpretazione in televisione’. Unpublished dissertation. Forlì: SSLMIT.

Chernov, Ghelly V. (1978) *Teoriya i praktikasinkronnogoperevoda* [Theory and practice of simultaneous interpretation], Moscow: Mezhdunarodniyeotnosheniya.

— (1994) ‘Message Redundancy and Message Anticipation in Simultaneous Interpreting’, in Sylvie Lambert and Barbara Moser-Mercer (eds) *Bridging the Gap. Empirical Research in Simultaneous Interpretation,* Amsterdam and Philadelphia: Benjamins, 139–53.

— (2004) *Inference and Anticipation in Simultaneous Interpreting*, Amsterdam and Philadelphia: John Benjamins.

CIAIR Simultaneous Interpretation Database (Nagoya University). Available online at: http://sidb.el.itc.nagoya-u.ac.jp/en.html (accessed 25 July 2010).

Collados Aís, Ángela (2009) 'Marco evaluador de la Interpretación Simultánea', in *Estudios de Traducción: Perspectivas.* Zinaida Lvóskaya in memoriam, Frankfurt: Peter Lang, 145–69.

Cruttenden, Alan (1986) *Intonation,* Cambridge: CUP.

Dam, Helle V. (2001) 'The Manipulation of Data. Reflections on Data Descriptions Based on a Product-Oriented PhD on Interpreting', in Daniel Gile, Helle V. Dam, Friedel Dubslaff, Bodil Martinsen and Anne Schjoldager (eds) *Getting Started in Interpreting Research: Methodological Reflections, Personal Accounts and Advice for Beginners*, Amsterdam and Philadelphia: John Benjamins, 163–79.

Dejean le Feal, Karla (1978) 'Lectures et improvisations'. Unpublished doctoral dissertation, Université de Paris III.

— (1982) 'Why Impromptu Speech Is Easy to Understand', in N. E. Enkvist (ed.) *Impromptu Speech*, Åbo Research Institute, Åbo Akademi Foundation, 221–39.

Ding Zhe, Koichiro Ryu, Shigeki Matsubara and Masatoshi Yoshikawa (2005) 'Interpreting Unit Segmentation of Conversational Speech in Simultaneous Interpretation Corpus', in *Proceedings, COCOSDA 2005 (International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques),* 2–3 December, Bali, Indonesia, 148–52.

Diriker, Ebru (2001) *Contextualising Simultaneous Interpreting: Interpreters in the Ivory Tower?* Doctoral dissertation, Bogazici University, Turkey.

Dollerup, Cay and Leo Ceelen (1996) *A Corpus of Consecutive Interpreting* (Copenhagen Studies in Translation 5), Centre for Translation Studies and Lexicography: University of Copenhagen.

Donovan, Clare (1990) 'La fidélité en interprétation'. Unpublished doctoral dissertation, Université de la Sorbonne Nouvelle, Paris III.

Dubois, John W. and Stephan Schuetze-Coburn (1993) 'Representing Hierarchy: Constituent Structure for Discourse Databases', in J. A. Edwards and M. D. Lampert (eds) *Talking Data: Transcription and Coding in Discourse Research,* Hillsdale: Lawrence Erlbaum, 221–62.

Dubois, John W., Stephan Schuetze-Coburn, Susana Cumming and Danae Paolino (1993) 'Outline of Discourse Transcription', in J. A. Edwards and M. D. Lampert (eds) *Talking Data: Transcription and Coding in Discourse Research*, Hillsdale: Lawrence Erlbaum, 33–44.

Edwards, Jane A. (1993) 'Principles and Contrasting Systems of Discourse Transcription', in J. A. Edwards and M. D. Lampert (eds) *Talking Data: Transcription and Coding in Discourse Research,* Hillsdale: Lawrence Erlbaum, 3–32.

Ehlich, Konrad (1993) 'HIAT: A Transcription System for Discourse Data', in J. A. Edwards and M. D. Lampert (eds) *Talking Data: Transcription and Coding in Discourse Research,* Hillsdale, NJ: Lawrence Erlbaum, 123–48.

Fumagalli, D. (2000) 'Alla ricercadell'interpretese. Uno studio sull'interpretazio neconsecutivaattraverso la corpus linguistics'. Unpublished MA dissertation, SSLMIT, University of Trieste.

Gale, William and Kenneth Church (1994) 'A Program for Aligning Sentences in Bilingual Corpora', in Susan Armstrong-Warwick (ed.), *Using Large Corpora,* Cambridge, MA: MIT Press, 75–102.

García Becerra, O. (In press). 'First Impressions in Interpreting Quality Assessment' in *Interpreting . . . Naturally. Updating Brian Harris' Pioneer Insights.* Selected Papers (Vol. I), Proceedings of the Tenth Conference on Translation and Interpreting, International Symposium on Interpreting Studies in Honour of Brian Harris, Castellón de la Plana, November 2009.

Gile, Daniel (2001) 'Interpreting Research: What You Never Wanted to Ask But May Like to Know'. AIIC website. Available online at: http://www.aiic.net/ViewPage.cfm/article229.htm (accessed May 2010).

Goldman-Eisler, Frieda (1972) 'Segmentation of Input in Simultaneous Translation', *Journal of Psycholinguistic Research* 1(2): 127–40.

Gumperz, John J. and Norine Berenz (1993) 'Transcribing Conversational Exchanges', in J. A. Edwards and M. D. Lampert (eds) *Talking Data: Transcription and Coding in Discourse Research,* Hillsdale, NJ: Lawrence Erlbaum, 91–121.

Hirst, Daniel and Albert Di Cristo (eds) (1998) *Intonation Systems. A Survey of Twenty Languages,* Cambridge: Cambridge University Press.

Ifantidou, Elly (2001) *Evidentials and Relevance,* Amsterdam: John Benjamins.

Iglesias Fernández, Emilia (2010) 'Speaker Fast Tempo and Its Effect on Interpreter Performance: A Pilot Study of a Multilingual Interpreting Corpus', *International Journal of Translation,* 145–70.

Kalina, Sylvia (1998) *Strategische Prozessebeim Dolmetschen,* Tübingen: Gunter Narr.

Kay, Martin and Martin Röscheisen (1994) 'Text-Translation Alignment', in Susan Armstrong-Warwick (ed.), *Using Large Corpora,* Cambridge, MA: MIT Press, 121–42.

Krouglov, Alex (1999) 'Police Interpreting: Politeness and Sociocultural Context', *The Translator,* 5(2): 285–302.

Lamberger-Felber, Heike (1998) 'Der Einflußkontextueller Faktoren auf das Simultandolmetschen: Eine Fallstudie am Beispielgelesener Reden'. PhD dissertation, University of Graz.

— (2001) 'Text-Oriented Research into Interpreting: Examples from a Case-Study', *Hermes* 26: 39–63.

Laplace, Colette (1994) *Théorie du langage et théorie de la traduction. Les concepts-clés de trois auteurs: Kade (Leipzig), Coseriu (Tübingen), Seleskovitch (Paris),* Paris: Didier-Erudition.

Laviosa, Sara (1996) 'The Corpus-Based Approach: A New Paradigm in Translation Studies', *Meta* 43(4): 474–9.

— (2000) 'Simplification in the Language of Translation: Before and after the Advent of Corpora', *Athanor* XI(3).

Lederer, Marianne (1981) *La traduction simultanée,* Paris: Minard Lettres Modernes.

Mason, Ian (1999) 'Community or Dialogue Interpreting,' in Ian Mason (ed.), *The Translator* 5(2), *Special Issue on Dialogue Interpreting.*

Metzger, Melanie and Cynthia Roy (forthcoming) 'The First Three Years of a Three-Year Grant' [provisional title], to appear in Brenda Nicodemus and

Laurie Swabey, *Moving Forward in Interpreting Studies: Methodology and Practice Revisited*, Amsterdam and Philadelphia: Benjamins.

Meyer, Bernd (2008) 'Interpreting Proper Names: Different Interventions in Simultaneous and Consecutive Interpreting', *Trans-kom*, 1(1). Available online at: http://www.trans-kom.eu/ihv_01_01_2008.html (accessed August 2010).

Monacelli, Claudia (2005) 'Surviving the Role: A Corpus-Based Study of Self-regulation in Simultaneous Interpreting as Perceived through Participation Framework and Interactional Politeness'. Unpublished dissertation, Heriot-Watt University.

— (2009) *Self-preservation in Interpreting*, Amsterdam and Philadelphia: Benjamins.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli and Mariachiara Russo (2005) 'Studying Directionality in Simultaneous Interpreting through an Electronic Corpus', *Meta,* 50 (4), 1079–1147. Available online at: http://id.erudit.org/iderudit/019850ar (accessed 10 October 2010).

O'Connell, Daniel C. and Sabine Kowal (1994) 'Some Current Transcription Systems for Spoken Discourse: A Critical Analysis', *Pragmatics* 4(1), 81–107.

Ohno [Ono] Takahiro, Hitomi Tohyama, Shigeki Matsubara (2008) 'Construction and Analysis of Word-level Time-aligned Simultaneous Interpretation Corpus', in *Proceedings of European Language Resources Association (ELRA) Conference, Marrakesh, May 28–30 2008* (LREC 2008).

Oléron, Pierre and Nanpon, Hubert (1965) 'Recherches sur la traduction simultanée', *Journal de Psychologie Normale et Pathologique* 62, 73–94.

Pinker, Steven (1994) *The Language Instinct*, London: Penguin Books.

Pöchhacker, Franz (1994) *Simultandolmetschenalskomplexes Handeln,* Tübingen: Günter Narr.

— (2004) *Introducing Interpreting Studies*, London: Routledge.

Pradas Macías, E. Macarena (2009) 'Identificación del patrón pausístico para la medición de la calidad en interpretación simultánea,' in Wotjak, G., Ivanova, V. and E. Tabares Plasencia, *Translatione via facienda, Festschrift für Christiane Nord zum 65. Geburtstag* [Festschrift for Christiane Nord on her 65th birthday], Frankfurt: Peter Lang, 235–52.

Pym, Anthony (1998) '"Nicole Slapped Michele": Interpreters and Theories of Interpreting at the O. J. Simpson Trial', *The Translator* 5 (1999): 265–83.

Russo, Mariachiara, Claudio Bendazzoli and Annalisa Sandrelli (2006) 'Looking for Lexical Patterns in a Trilingual Corpus of Source and Interpreted Speeches: Extended Analysis of EPIC (European Parliament Interpreting Corpus)', *Forum* 4(1): 221–54.

Sampson, Geoffrey (2001) *Empirical Linguistics*, London/New York: Continuum.

Sandrelli, Annalisa (2010) 'Corpus-Based Interpreting Studies and Interpreter Training: A Modest Proposal', in Lew N. Zybatow (ed.) *Translationswissenschaft – Stand und Perspektiven.* Innsbrucker Ringvorlesungenzur Translationswissenschaft VI (= Forum Translations-wissenschaft, Band 12). Frankfurt am Main: Peter Lang, 69–90.

— (forthcoming) 'Introducing FOOTIE (Football in Europe): Simultaneous Interpreting at Football Press Conferences', paper presented at *Emerging Topics in Interpreting/Nuovipercorsi in traduzione e interpretazione*, Trieste, June 2010.

Sandrelli, Annalisa, Claudio Bendazzoli and Mariachiara Russo (2010) 'European Parliament Interpreting Corpus (EPIC): Methodological Issues and Preliminary Results on Lexical Patterns in Simultaneous Interpreting', *International Journal of Translation* 22(1–2).

Sandrelli, Annalisa, Mariachiara Russo and Claudio Bendazzoli (2007) 'The Impact of Topic, Mode and Speed of Delivery on the Interpreter's Performance: A Corpus-Based Quality Evaluation', unpublished poster presented at *CRITICAL LINK 5, Quality in Interpreting: A Shared Responsibility,* Parramatta, Sydney, Australia, 11–15 April 2007.

Seeber, Kilian (2001) 'Intonation and Anticipation in Simultaneous Interpreting', *Cahiers de Linguistique Française*, University de Genève, 61–97.

Seleskovitch, Danica (1975) *Langage, langues et mémoire: études de la prise de notes en interprétation consécutive*, Paris: Minard Lettres Modernes.

Setton, Robin (1997). *A Pragmatic Model of Simultaneous Interpretation*, Ph.D. thesis, Chinese University of Hong Kong, UMI Dissertations.

— (1999) *Simultaneous Interpretation: A Cognitive-Pragmatic Analysis,* Amsterdam and Philadelphia: John Benjamins.

— (2002a) 'A Methodology for the Analysis of Interpretation Corpora', in Giuliana Garzone and Maurizio Viezzi (eds) *Interpreting in the 21st Century: Challenges and Opportunities. Selected Papers from the 1st Forlì Conference on Interpreting Studies, 9–11 November 2000,* Amsterdam and Philadelphia: John Benjamins, 29–45.

— (2002b) 'Traductologie et théorie de la pertinence', in Fortunato Israël (ed.) *Identité, altérité, équivalence? La traduction comme relation. Actes du Colloque International tenu à l'ESIT les 24, 25 et 26 mai 2000, en hommage à Marianne Lederer,* Paris/Caen: Lettres Modernes Minard, 97–112.

— (2003) 'Words *and* Sense: Revisiting Lexical Processes in Interpreting', *Forum* 1: 139–68.

Setton, Robin and Motta, Manuela (2007) 'Syntacrobatics: Quality and Reformulation in Simultaneous-with-Text', *Interpreting* 9(2): 199–230.

Shlesinger, Miriam (1989) 'Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-Literate Continuum'. Unpublished MA thesis, Faculty of Humanities, Tel Aviv University.

— (1998) 'Corpus-Based Interpreting Studies as An Offshoot of Corpus-Based Translation Studies', *Meta* 43(4): 486–93.

— (2008) 'Towards a Definition of Interpretese. An Intermodal, Corpus-Based Study', in Hansen Gyde, Andrew Chesterman and Heidi Gerzymisch-Arbogast (eds) *Efforts and Models in Interpreting and Translation Research. A Tribute to Daniel Gile,* Amsterdam and Philadelphia: John Benjamins.

Shlesinger, Miriam and Noam Ordan (forthcoming) 'More *Spoken* Than *Translated*: Simultaneous Interpreting Revisited'.

Stassen, Hans H. (1988) 'Modelling Affect in Terms of Speech Parameters', *Psychopathology* 21: 83–8.

Straneiro Sergio, F. (2003) 'Norms and Quality in Media Interpreting: The Case of Formula 1 Press Conferences', *The Interpreters' Newsletter* no. 12, 135–74.

— (2007) *Talkshow Interpreting. La mediazionelinguisticanellaconversazionespettacolo.* Trieste: Edizioni Universitarie Trieste.

Svartvik, J., M. Eeg-Olofsson, O. Forsheden, B. Oreström and C. Thavenius (1982) *Survey of Spoken English*. Lund: Lund University Press.

Tohyama, Hitomi and Shigeki Matsubara (2006a) 'Collection of Simultaneous Interpreting Patterns by Using Bilingual Spoken Monologue Corpus,' in *Proceedings of 5th International Conference on Language Resources and Evaluation, May 2006 (LREC-2006),* 2564–9.

— (2006b) 'Development of Web-Based Teaching Material for Simultaneous Interpreting Learners using Bilingual Speech Corpus', in *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, June 2006 (ED-MEDIA-2006),* 2906–11.

— (2006c) 'Influence of Pause Length on Listeners' Impressions in Simultaneous Interpretation', in *Proceedings of 9th International Conference on Spoken Language Processing, September 2006 (Interspeech-2006),* 893–6.

Tohyama, Hitomi, Shigeki Matsubara, Nobuo Kawaguchi and Yasuyoshi Inagaki (2005) 'Construction and Utilization of Bilingual Speech Corpus for Simultaneous Machine Interpretation Research', in *Proceedings of 9th European Conference on Speech Communication and Technology* (*Interspeech-ICSLP 2005*), 1585–8. Available online at: http://hdl.handle.net/2237/93 (accessed 10 October 2010).

Vaissière, Jacqueline (1983) 'Language-Independent Prosodic Features', in Anne Cutler and D. Robert Ladd (eds) *Prosody: Models and Measurements*, Berlin: Springer, 53–66.

Van der Eijk, Pim (1999) 'Comparative Discourse Analysis of Parallel Texts', in Armstrong S., K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D. Yarowsky (eds) *Natural Language Processing Using Very Large Corpora*, Dordrecht: Kluwer Academic Publishers.

Van Dijk, Teun and Walter Kintsch (1983) *Strategies of Discourse Comprehension,* New York: Academic Press.

Vuorikoski, Anna-Riitta (2004) 'A Voice of Its Citizens or a Modern Tower of Babel? Interpreting Quality as a Function of Political Rhetoric in the European Parliament', PhD thesis, University of Tampere. Tampere: Acta Universitatis Tamperensis 985. Available online at: http://acta.uta.fi/pdf/951–44-5878–8.pdf (accessed 10 October 2010).

Wadensjö, Cecilia (1998) *Interpreting as Interaction,* London and New York: Addison Wesley Longman.

Wallmach, Kim (2000) 'Examining Simultaneous Interpreting Norms and Strategies in a South African Legislative Context: A Pilot Corpus Analysis', *Language Matters* 31: 198–221.

Chapter 3

Translation Units and Corpora

Dorothy Kenny

3.1 Introduction

The unit of translation is a basic concept in translation studies. To date it has received only scant attention in corpus-based translation studies, despite the fact that parallel corpora are likely to contain a wealth of data on translation units, and that corpus linguistics may offer new, as yet barely tapped, ways of looking at the issue. It is precisely this potential that the current chapter seeks to explore.[1] Before this can be done, however, the theoretical ground has to be prepared. In particular, we must take a stand on how 'translation unit' is to be understood and made operational in the context of corpus-based translation studies. The paper thus starts by reviewing existing approaches to translation units, focusing on: comparative stylistics; process and product-oriented descriptive translation studies; and natural language processing. Tentative methods for identifying translation units in parallel corpora are then proffered and illustrated using examples from the German–English Parallel Corpus of Literary Texts (GEPCOLT).[2]

3.2 The Translation Unit in Comparative Stylistics

In their seminal work, Vinay and Darbelnet (1958/1995) define the translation unit as 'the smallest segment of the utterance whose signs are linked in such a way that they should not be translated individually' (1995: 21). Translation units are thus segments of the source text, but they are viewed through the prism of the target language (ibid.: 17). Vinay and Darbelnet further consider translation units to have some kind of cognitive status, and to be conventionalized to the extent that they can be recorded in dictionaries, although they do not subscribe to the idea of absolute, unvarying equivalence between source-language translation units and their translations (ibid.: 21).[3] They do not expect translation units to be co-terminous with syntactic categories, although they are not completely consistent on this matter (ibid.: 21, 27). Some later theorists maintain flexibility on this issue (e.g. Bennett 1994; and the process-oriented studies discussed below), although others argue on theoretical grounds that the clause (Bell 1991; Malmkjær 1998), clause complex (Huang and Wu 2009) or sentence (Zhu 1999, 2005) make the most appropriate translation units. Still others give primacy to the full text (Bassnett-McGuire 1980).

The linguistic rank of translation units is an issue that arises frequently in the literature, and one to which we return again below. Suffice it to say for the moment that authors such as Bennett are sceptical about the ability of translators to work with 'any text other than the very shortest as an undivided UT [unit of translation], for reasons of memory limitations, if nothing else' (1994: 13). For Bennett, the text represents the translation 'macro-unit', that is, 'the largest linguistic unit which the translator needs to consider' (ibid.), which is quite different, for all but the shortest texts, from the smallest segment of text that should be translated as a whole.

Vinay and Darbelnet approach translation units from multiple directions, and pack rather a lot into their treatment of the concept. The tension between source language and comparative perspectives, the elusiveness of 'units of thought', and the (morpho-syntactic) flexibility of their translation units all make Vinay and Darbelnet's notion of translation unit difficult to operationalize in corpus-based research. The purported conventionalized nature of translation units seen as lexicological units, however, offers some potential for their identification in corpora. Bennett (1994: 13) has also helpfully pointed out that the translation unit as conceived by Vinay and Darbelnet is really a 'translation atom', that is, the smallest segment that must be translated as a whole (if one wishes to avoid overtranslation or mistranslation), although such atoms form part of larger units that are somehow operational in translation and may themselves form part of larger units, and so on until we reach the level of the full text. Bennett (ibid.) likens translation atoms to listemes in linguistics, that is, 'listed or memorised linguistic objects' that can range in size from morpheme to sentence. Units of translation are often larger, but should never be smaller than listemes or translation atoms. Bennett (ibid.) continues by noting that 'as one ascends the grammatical hierarchy, the fewer the proportion of items at each level which are listemes: all morphemes are listemes, but only some phrases and a handful of sentences.' As we shall see below, contemporary corpus linguists point out that listeme-like 'prefabricated' units are more common at higher linguistic ranks than previously thought, an observation that may have implications for studies of translation 'atoms'. The picture is complicated, however, by the fact that such 'pre-fabs' exhibit high levels of variation, and capturing them in patterns can mean abstracting away from the surface facts that confront us so directly in corpora.

Bennett's observations on translation atoms are consistent with Vinay and Darbelnet's (ibid.: 23–7) own discussion of the degrees of cohesion between the elements in (multi-word) units of translation: translation units can be highly 'unified' or else represent 'affinity groups' 'whose elements are more difficult to detect and in which the cohesion between the words is less evident' (ibid.: 24). Affinity groups can combine to form 'complex units', which can, where appropriate, be translated as single units.

For our analysis, we propose to distinguish between (1) source text segments, which may correspond to monolingual listemes (after Bennett) or, less commonly, unconventional, creative forms, and (2) 'translation units', which we will understand as empirically validated, mutually defining source and target text segments, after Toury (see below), but also drawing, albeit less directly, on Vinay and Darbelnet's comparative approach. We also find Vinay and Darbelnet's emphasis on multi-word units instructive and take their lead (and that of several other authors reviewed here) in positing translation units that do not correspond to unvarying linguistic units.

Vinay and Darbelnet's approach to the translation unit (or 'translation atom') is thus useful in some regards, but it remains overly oriented towards the source text and idealized translations, factors that limit its ability to account for much of what happens in real translation (see Ballard 1997; Santos 2000; Krings 2001; and below). In empirical translation studies, on the other hand, observations are made on the basis of actually occurring translation processes and products, allowing different insights into the nature of translation units.

3.3 The Translation Unit in Empirical Translation Studies

Empirical translation studies can be divided into two main categories: those that focus on the translation *process*, and in particular on the translator's cognitive activity; and those that study translation *products* – target texts which can be related, amongst other things, to their host cultures, their users, and their respective source texts. In the following sections we consider how translation units have been treated from each of these points of view.

3.3.1 The Translation Unit in Process-Oriented Translation Studies

Alves and Gonçalves (2003) define translation units in terms of attentional focus on the source text and stress the dynamic nature of such units:

> translation units are seen here as segments of the source text, independent of specific size or form, to which, at a given moment, the translator's focus of attention is directed. It is a segment in constant transformation that changes according to the translator's cognitive and processing needs. Alves and Gonçalves (2003: 10–11)

According to this view, it is not possible to identify translation units *a priori* on the basis of source language structures, or stretches of source text of a specified length. Rather the identification of translation units can happen only in real time, as translators translate. In order to isolate such units, researchers have thus traditionally relied on concurrent think-aloud protocols (TAPs), that is, translators' own attempts to verbalize their thought processes as they translate. But the use of TAPs is far from straightforward. There are concerns, for example, about the ability of TAPs to capture information about automatized translation processes. Another problem is that verbalization may affect the very thought processes on which it is supposed to be reporting; more specifically, it has been found to have an effect on the length of source text segments processed by subjects in translation scenarios (Jakobsen 2003: 91). Thus, used as a data elicitation technique, TAP can become a confounding variable in a study of TUs. A further difficulty for researchers interested in translation units defined in terms of attentional focus on the source text, is that only a relatively small proportion of any particular TAP is likely to give information about the source text. Krings (2001: 314), for example, found in one of a suite of experiments that only 6.3 per cent of all verbalizations made by translators were about the source text. Similarly, it is unlikely that the proportion of a TAP that does relate to the source text will account for all of that source text. It can be assumed that completely unproblematic stretches of source text will not result in verbalizations; while subjects' problems, for example, in understanding the source text, will be explicitly commented upon. It is for this reason perhaps that some research within the process-oriented framework links the notion of translation unit/focus to that of translation *problem* (see, for example, Livbjerg and Mees 2003). For Barbosa and Neiva (2003) on the other hand, translation units are not so much defined as problems but rather are demarcated by problems, which cause breaks in the translation 'flow'.

One aspect of the literature on translation units where there appears to be consensus relates to the linguistic ranks at which novice and expert translators operate. Although, as indicated above, translation units cannot be defined in terms of ST structures or pre-specified lengths of ST, some studies (e.g. Kiraly 1990; Lörscher 1996; Krings 2001) have found that professional translators tend to focus on ST structures of higher rank (e.g. phrases, clauses or sentences) than semi- or non-professionals, who tend to translate at the level of syntagma or individual words. It is worth mentioning however, that although Kiraly's study (ibid.) found that professionals (translating albeit into L2) operate on a supra-sentential level more frequently than novice translators, the bulk of their translation was still located below sentence level (Krings 2001: 159).[4]

Thus far we have focused on research based on subjects' introspection. Given the problems inherent in TAP methodology however, process-oriented researchers have begun to use other data collection methods instead of, or in conjunction with, TAPs. In one such study Jakobsen (2003) uses the keyboard monitoring software Translog to investigate the effects of think aloud on translation speed, revision and segmentation. He defines a segment as ‘any length of keystrokes between two pauses of 5 seconds (or more)’ (ibid.: 90). Given that Translog monitors target text production, such segments are, by definition, target segments, and these are related to source texts in quantitative terms only.[5] Jakobsen does not comment on whether the differences in target segmentation he observes reflect differences in source text segmentation, so it is not clear whether we can relate his segments to attentional focus on parts of the source text. Translog does, however, suggest an interesting way of recording what parts of the target text belong together in terms of their genesis in time (as opposed to any structural bonds they may have). Alves and Couto Vale (2009), who use keystroke-logging and eye-tracking software to capture data, offer a related account of the ‘unfolding of translation units in time’ (ibid.: 251) where such units are delimited by pauses registered by the software.

In summary then, translation units as described by those working within the process-oriented paradigm tend to be:

- source-oriented, in that they are defined by attentional focus on the source text, although newer technologies are also directing researchers’ attention to target texts.
- dynamic, in that they emerge during real-time translational processing of source texts and are not restricted to any particular length or structural boundaries, although their extent may vary with varying levels of translator expertise.
- problem-oriented, either because researchers focus on those aspects of translation that prove to be problematic, or because the data elicitation techniques used by researchers favour the discovery of problems in the translation process.
- difficult to identify, given that they are dynamic; but also because access to cognitive processes is always indirect, and different ways of eliciting data about cognitive processes tend to yield different results; and because there is no agreement in the literature on the best way to identify them.

3.3.2 The Translation Unit in Product-Oriented Translation Studies

While process-oriented approaches to translation units give priority to source-text segments, product-oriented approaches start with target texts and view the unit of translation as ‘the target-text unit that can be mapped onto a source-text unit’ (Malmkjær 1998: 286). The process whereby translation units are identified in paired source and target texts has not been discussed widely, although Toury (1980, 1995) gives the problem extensive treatment: Toury deals with 'coupled pairs' of target and source segments. Like some researchers working in process-oriented studies, Toury is interested only in those parts of the source text that have actually posed a problem in translation, a fact that can be established only through 'concurrent identification of the respective solution' (1995: 78). A major issue with coupled pairs, according to Toury, is that it is not clear how their boundaries should be determined, given their dynamic nature and high context dependency. Toury's answer to this difficulty is to propose a 'no leftovers' principle:

> Thus, the analyst will go about establishing a segment of the target text, for which it would be possible to claim that – beyond its boundaries – there are no leftovers of the solution to a translation problem which is represented by one of the source text's segments, whether similar or different in rank and scope. (Toury 1995: 78–9)

Zabalbeascoa (2000) offers a very similar treatment. For him, the identification of translation units prior to the act of translation is largely a matter of ST parsing and is oriented towards minimal units; from a retrospective, descriptive point of view:

> it is more a question of first finding meaningful bitextual pairs, which means that the length and nature of each segment is determined by the type of solution, which provides evidence of the problem as the translator presumably saw it. Zabalbeascoa (2000: 121)

For Toury and Zabalbeascoa, 'problem/solution pairs' and 'translation units' are thus mutually defining, dynamic (in common with much process-oriented research), and specific to individual pairs of texts. While Toury does not make much of the subjective judgements that must be involved in their identification, he does stress that 'whatever units one chooses to work with should be *relevant to the operation which would then be performed on them*' (1995: 88, Toury's emphasis). Toury's coupled pairs serve for the most part in comparative analyses of translations that focus on 'reconstructing rather than implementing translation decisions' (ibid.: 88), but he suggests, drawing on Harris (1988), that the pairing of source and target language segments might have some psychological validity (Toury ibid.: 99).

Again in summary, translation units in product-oriented translation studies tend:

- to be mutually defining source and target text segments taken together, and whose boundaries cannot be predicted on structural grounds;

- to be identified presumably through subjective analysis by matching TT solutions with their ST problems;
- to have the status of analytical categories in descriptive studies but may also reflect 'coupled pairs' stored in translators' long-term memory.

3.4 The Translation Unit in Natural Language Processing

In Natural Language Processing 'translation unit' tends to be used in a way that gives equal emphasis to source and target text, and that reflects the granularity of the automatic or semi-automatic processes that these texts are subjected to. Bowker (2002: 155), for example, defines 'translation unit' as: 'A source text segment and its corresponding translation as stored in a translation memory'. In most commercial translation memory systems translation units are usually stored at *sentence* level, as it is relatively easy to identify sentence boundaries automatically and to 'align' sentences and their translations across two languages.

Alignment involves the explicit pairing of segments that are translations of each other in a parallel corpus (the directionality of translation is not usually considered an issue in automatic alignment).[6] Depending on the ultimate application, the linguistic rank at which alignment is carried out (i.e. the alignment 'granularity') will differ. As already indicated, in translation memory applications, automatic alignment tends to happen at sentence level. Alignment at group or phrase level is useful in corpus-based Machine Translation (see Carl and Way 2003) and alignment at word level is crucial in applications like bilingual lexicon extraction (see Kraif 2003, and the papers in Véronis 2000).

Sentence alignment, although by no means trivial, is generally considered to be 'a mastered technology for most parallel corpora' (Kraif 2003: 2). It often relies on a high degree of one-to-one matching of source and target sentences in parallel corpora, and on the order of sentences in the source text and their translations in the target text not changing (a property known as 'monotonicity'). While alignments do not always have to be one-to-one, the general requirement of monotonicity is only rarely relaxed, but this seems largely unproblematic as radical divergences between source and target texts appear to be rare in the text genres studied to date (Véronis and Langlais 2000; Huang and Wu 2009).

Word alignment is a far more difficult task (and accordingly far less developed), as one-to-one matching and monotonicity at word level are highly unlikely for all but the most closely related languages. Indeed commentators such as Martin Kay remind us that given texts translated by humans (as opposed to machines, which might, on occasion, translate word for word), 'the very notion of alignment falls apart at finer levels of granularity' (Kay 2000: xvii–xviii).

Kraif's (2003) illuminating discussion of word-level alignment includes a number of examples where it is not obvious where the boundaries of units to be aligned should be drawn, the dynamic nature of translation units leading to what he terms 'segmentation inconsistency'. Kraif also introduces the term 'semantic discrepancy' to label those cases where words that can be said to translate each other in a particular context cannot be generally held to be 'synonymous' (my term) across languages, for example because one might be more specific than the other, as is the case with *aux dépens du* (at the expense of) and *involving* in example (1) below (from Kraif 2003: 4), or because of more radical divergences between source and target texts.

(1) Illegal transactions involving the heritage . . .
Transactions illégales aux dépens du patrimoine . . .

Kraif goes on to distinguish between 'lexical correspondences' and 'translational equivalence'. Lexical correspondences are those that would also be found in a bilingual dictionary (ibid.: 4) – or perhaps more usefully, that one would want to include in a bilingual dictionary – as they are not bound to individual contexts and are thus generally reusable. They represent fairly stable semantic equivalents and their pairing in a corpus is a fairly regular occurrence. Kraif's translation equivalents, on the other hand, are highly context-bound and result from choices made by translators on particular occasions, and depending on a variety of factors such as the purpose of the communication, the text type, differing cultural or conceptual backgrounds, and so on (ibid.: 5). Translation equivalents as understood here represent one-off solutions and are likely to involve semantic discrepancies. Although they provide interesting data for descriptive translation studies, they are of less interest to those attempting to extract generally reusable bilingual lexicons from parallel corpora. Kraif's own work relies on the fact that while one-off translation equivalents may complicate the picture, regularly occurring lexical correspondences can still be extracted from parallel corpora using a combination of manual analysis and statistical processing.

Like Kay (2000), Kraif (2003: 13) recognizes different types of lexical alignment, depending on one's aims. The units to be aligned could, for example, be limited to terms, content words, noun phrases or phraseology, and the required relationship between source and target units could be specified as 'semantic identity or similarity', or pairs might have to meet the criterion of being reusable in different contexts, and so on. In Kraif's scheme, source and target units would not be mutually defining: units to be paired with their translations would first be identified monolingually, and if no satisfactory match was found in the target text, no attempt would be made to redefine the source unit as part of a larger source unit, as the one-to-one matching assumption would no longer hold.

Kraif's conceptualization reflects what actually happens in contemporary terminology extraction and word alignment where translations may be sought only for ST units that meet predefined linguistic criteria (e.g. they are units composed of two nouns, or an adjective + noun); or the only units that are of interest are those for which there is enough frequency and distribution information in a parallel corpus to make their proposed alignment reliable (see Bowker 2002: 82–6). It is thus clear that word alignment differs from sentence alignment in that systems attempting to align at lexical level are not normally called upon to account for *all* of the source or target text: they are 'fragmentary', in Kraif's (2002: 285) terminology. Given this difference, and the fact that assumptions of monotonicity and compositionality made in sentence alignment do not hold at the lexical level, it might be better not to see current approaches to pairing source and target lexical units as a case of alignment at all (Kraif 2003: 3, 13).

To sum up, in NLP applications:

- The 'translation unit' is, in common with the other approaches addressed here, a dynamic entity in that it can refer to coupled ST and TT segments of varying rank, depending on the application, although it is normally defined in structural terms.
- At sentence level, equal attention is given to TT and ST, and an attempt is made to assign *all* TT and ST segments to a translation unit.
- At the lexical level, one language can have priority over the other, and no attempt is made to account for all the words in both texts. Source units are often identified before any attempt is made to spot their translations, but researchers often rely on similarities in the distribution of given source units and their supposed translations to identify translation units.
- No claims are made about the psychological validity of the translation units extracted from parallel corpora; NLP work on translation units is motivated by the need to produce useful data for translation-oriented applications rather than the desire to model mental processing in translators.
- Regular 'correspondences' can be identified in parallel corpora and distinguished from highly context-bound 'equivalences'.

We conclude then, that in NLP, translation units below the rank of sentence cannot normally be easily or exhaustively identified by automatic means, and that some level of manual intervention is thus justified. We note also that Kraif's (2003) distinction between conventional (lexical) 'correspondences' and one-off translation 'equivalences' reflects a preoccupation with regularities and departures from those same regularities that have also received attention in related areas of corpus linguistics, discussed below.

3.5 Units of Meaning in Corpus Linguistics

There is an expanding literature in corpus linguistics on (extended) units of meaning, but we will focus here on John Sinclair's seminal work, also summarized in Kenny (2001: 99–104).

For some time, and based on his study of huge quantities of data in electronic corpora, Sinclair argued that much of language use relies on 'semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments' (Sinclair 1987: 320). Speakers and writers draw on such semi-preconstructed units when producing text, and hearers and readers also rely on this 'idiom principle' (ibid.) or 'phraseological tendency' (Sinclair 1996) when interpreting texts. Only when readers are forced to abandon this default interpretation do they fall back on the 'open-choice principle', according to which text is built up and interpreted using a kind of slot-and-filler approach, whereby 'At each point where a unit is completed (a word or phrase or a clause), a large range of choice opens up, and the only restraint is grammaticalness' (ibid.: 319–20). Sinclair exemplified phraseological tendencies in English in a number of publications (1991, 1996, 2004), incorporating into his analysis ideas about the co-selection of lexis and grammar, and the importance of speakers' and writers' attitudes in conditioning how meaning units are supposed to function. He concluded that 'so strong are the co-occurrence tendencies of words, word classes, meanings and attitudes that we must widen our horizons and expect the units of meaning to be much more extensive and varied than is seen in a single word' (Sinclair 1996: 94).[7]

Among the linguists who have lent empirical support to Sinclair's position is Michael Stubbs. Stubbs (2001, 2002) has sought to quantify the extent to which the most common lexemes in English display phraseological tendencies. Like Sinclair (1992), Stubbs has also pursued the idea of delexicalization. Words become delexicalized when their independent semantic content is weakened over time, precisely as a result of their frequent use in phrases. According to Sinclair (1992), adjectives that occur in frequent collocations, for example 'general' in 'general trend', are at least partly delexicalized. They serve not to add semantic features to the head noun, but rather to intensify the content of that noun. Semantically, they thus become somewhat redundant, although they retain important pragmatic functions. Such cases of delexicalization are good illustrations of the phraseological tendency in language, as rather than seek to interpret, even over-interpret, each word in isolation, it makes more sense to look at the extended unit, as the meaning of the phrase is 'dispersed across the phrase as a whole' (Stubbs 2002: 230).

3.6 Translation Units in Corpus-Based Translation Studies

Within corpus-based translation studies (CTS), little attention has been paid so far to the notion of translation units, although some researchers, including Baker (2004) and Dayrell (2004), have sought to describe a particular type of lexical patterning in *target* texts, namely the use of recurring lexical phrases, and in doing so have been concerned with the kind of extended units of meaning that have interested corpus linguists. Relevant work based on parallel corpora includes Danielsson (2003) and Santos (2000), although neither is explicitly concerned with translation units, and Kenny (2001 and 2004), Teubert (2002, 2004) and Kondo (2007).

Danielsson's main interest in translation is that it provides a way of validating findings based on monolingual analysis. Like Sinclair, she argues that the unit of analysis in language should be the 'unit of meaning' rather than the orthographic word, and sets out to develop a methodology for automatically extracting such units of meaning from a monolingual corpus. She then tracks the translation of these units of meaning in an English–Swedish parallel corpus, whereby her main priority is to find 'a single unit of meaning that also occurs in the target language' (ibid.: 123).

The core of Santos's (2000) work involves the development of an abstract model of translation based on translation networks and focusing on temporal aspect in Portuguese and English. Like Kay (2000) and Kraif (2003), she distinguishes between the common and the exceptional in translation, but unlike Kraif, she does not prioritize the common over the exceptional.

Kenny (2001) contains some useful data on translation units in a parallel corpus, GEPCOLT, although their significance is not drawn out. In particular, the anaphoric compounds (2001: 163–7) and repeated clusters (2001: 138–40, 204–6) treated in this source exemplify the text as macro-translation unit (after Bennett 1994). In Kenny (ibid.), corpus evidence is used to show how the interpretation of a creative compound currently in focus is facilitated by access to (sometimes much) earlier instances of lexically related syntagmas, whether or not translators actually turn to such earlier instances in arriving at a translation solution. Likewise, the use of repeated clusters (recurring strings of orthographic words with no internal variation) is seen as creating highly cohesive stretches of source text; and the partial undoing of this cohesion in the target text as questionable.

Kenny (2004: 342–3) draws on corpus linguistic research into units of meaning in an attempt to show how source and target text segments can be mutually defining, in the sense suggested by Toury (1995) and described above. Data published in this source, as well as newer data extracted from GEPCOLT, will be used below to exemplify how a parallel corpus can be used in investigating units of translation.

Teubert (2002, 2004) and Kondo (2007) look at translation units from the combined perspectives of corpus linguistics, lexicography and translation studies. According to Teubert, translation units are source-text segments, usually compounds, multi-word units, collocations or set phrases (2002: 193), 'that are large enough to be monosemous' so that for each translation unit 'there is only one equivalent in the target language, or, if there are more, then these equivalents will be synonymous' (2004: 184–5).[8] Teubert (ibid.: 185) argues that translation units are not necessarily co-terminous with (monolingual) units of meaning: as each language construes a different reality, what might be a translation unit from the point of view of one target language may not be from the point of view of another target language. So, translation units, although segments of source texts, are always seen through the prism of the target language. Teubert also stresses the importance of recurrence (which is so easily observed in large parallel corpora) in the automatic extraction of (established) translation units and their equivalents (2002: 209, 211).

3.7 Searching for Translation Units in a Parallel Corpus

Before attempting to show how parallel corpora can be used to simultaneously isolate units of meaning and units of translation, it is perhaps advisable to recap on some of the basic assumptions made here: in line with the product-oriented studies discussed above, translation units are understood here as mutually defining source/target text segments. In the absence of a reliable, non-fragmentary, method of isolating such units bilingually, however, the manual search for units has to start monolingually – consistent with many of the NLP approaches mentioned above, and with Teubert (2002, 2004) and Kondo (2007) – but will shuttle backwards and forwards between source and target text, with the analyst refining judgements in the process. Given the product-orientedness of the current study, units of translation will be approached primarily as units of analysis, although this does not mean that the translation units we deal with do not have some kind of cognitive status. Taking cognizance of the limitations of memory, theoretical claims made by scholars such as Malmkjær (1998) and Zhu (1999), and the empirical findings of Huang and Wu (2009), we will assume that translation units, while not fixed structurally, will not normally exceed clause or sentence rank. As has already been acknowledged, however, there are cases where the current translation unit is analysed (and, we might presume, processed) taking into account other points in the source and target texts, in which case the text is considered to be the macro (translation) unit, following Bennett (1994).

In the following, we adopt Toury's 'no leftovers' method for identifying problem/solution pairs, although with some differences. First, as indicated above, Toury is interested in more 'problematic' cases, but we are interested in identifying translation units whether or not problems can be unequivocally pinpointed. Secondly, Toury's method applies to individual instances in individual texts, while we are concerned to harness the potential of corpora to uncover groups of related instances across texts and translators.

For this purpose we have elected to investigate, in the first instance, a number of related phraseological units involving the common German preposition *mit* 'with'. This choice is not accidental: as Danielsson (2003: 117) has pointed out, relations between mid-frequency words and words of higher frequency – and the highest-frequency words are nearly all grammatical words, including determiners, pronouns, prepositions, and so on[9] – 'tend to bring out phraseological features', which are of particular interest here. As Stubbs (2002: 235) puts it, 'Grammatical words do not occur on their own: their function is to form larger units.'

We are interested in whether or not the extended units of meaning whose identification is facilitated using corpus techniques coincide with translation units, although we note Teubert's (2002) above-mentioned scepticism on this point. A caveat is called for here, however: it is commonly held that corpus-based studies of lexis should be conducted on the basis of very large corpora; given enough evidence about lexical items, new, previously unrecorded patterns become visible to linguists, and the value of the corpus seems indisputable. With a corpus as limited in size by today's standards as GEPCOLT, which contains a mere one million words in each of German and English, we are less likely to uncover previously obscured patterns, to break new lexico-syntactic ground, as it were. Nevertheless, given that corpus-based translation studies is still at a stage where it is edging towards appropriate methodologies (see Baker 2004), we may have to live with the fact that, in the course of proposing a new technique, we may not generate sufficient data to be able to produce dramatic new evidence about translation in general. We should expect our techniques to suggest fruitful avenues for exploration of larger corpora, however. It is in this spirit that the analysis below is proffered.

3.7.1 The *Mit Aller Kraft* Pattern

A detailed profile of *mit* is beyond the scope of this chapter. We will rather be interested in instances where *mit* shows strong phraseological tendencies, and an inspection of the concordance lines for *mit* suggests that this is the case where it co-occurs with forms of ALL 'all'.[10] There are 80 such co-occurrences in the corpus, but we will focus on a subset of those in Table 3.1,[11] where *mit all* occurs in a circumstantial adjunct, most commonly *mit aller Kraft* 'with all (one's) might' (11 instances) or its synonyms *mit aller Gewalt* (3), *mit aller Macht* (2) and *mit aller Wucht* (1).[12]

The phrases in Table 3.1, along with the extended *mit aller Kraft und Empörung* 'with all her might and indignation' and the variant *mit allen Kräften* (plural form), as well as having a common semantics, share a common syntax, in which the fully inflected determiner ALL is indicative of set phrases (Durrell 1991: 93). Two other variants are also attested in the corpus. In these cases there is extra pre-modification of the head noun: an explicit possessive determiner and a classifier of *Kraft* appear in the phase, for example, *mit aller ihrer Seelenkraft* 'with all her

Table 3.1 *mit all** selected concordance lines from GEPCOLT

1.	saßen in der Klemme. Es war, als ob sie	mit aller	Gewalt versuchten, sich durch	debier.txt
2.	zu ersticken. Sie stößt mit dem Stuhl	mit aller	Gewalt gegen die Türe. Da ersc	dezuern1.txt
3.	worden ist und die Krankenschwester sie	mit aller	Gewalt dazu gebracht hat, aufz	dezuern2.txt
4.	ff angefroren sein mußte. Sie versuchte	mit aller	Kraft, bei klarem Verstand zu	debier.txt
5.	n aus den Gedanken, in denen er steckt,	mit aller	Kraft herausziehen, doch selbs	dehofman.txt
6.	e Stopfnadel aus meinem Handarbeitszeug	mit aller	Kraft, und wenn ich sage Kraft	dejelin2.txt
7.	ahrtendolch in Pension ist, und posiert	mit aller	Kraft wie auf Sophies Bruderfo	dejelin2.txt
8.	hließend ersticht er die tote Schwester	mit aller	Kraft. Dann ist er endlich dam	dejelin2.txt
9.	der Kurt Lukas war, schien sein Messer	mit aller	Kraft in die Hündin zu treiben	dekirch.txt
10.	hle Bäume vor dem Feld, der Mann rannte	mit aller	Kraft, das sah ich an den gewö	deloest.txt
11.	tarre der Meereisdecke ziehen die Hunde	mit aller	Kraft nach der nächstgelegenen	derans.txt
12.	ultern gegen die Matratze, stemmte mich	mit aller	Kraft gegen seinen Amoklauf, e	dewodin.txt
13.	hte lebenden Embryo-Gesichter, scheinen	mit aller	Kraft von der Sonne angezogen	dezuern1.txt
14.	eren Glas-Röhrchen auf ihr Blut warten,	mit aller	Kraft und Empörung auf den Fuß	dezuern1.txt
15.	ich kam.« McEllis keuchte. Er kämpfte	mit allen	Kräften, um diese Bergprüfung	dekirch.txt
16.	it Feigheit zu tun. Er wehrte sich zwar	mit aller	Macht dagegen, verurteilt zu w	deroth2.txt
17.	t, er will nicht sterben, er wehrt sich	mit aller	Macht, mit der brachialen, ele	dewodin.txt
18.	rt; man hieb auf ihn von der Seite ein,	mit aller	Wucht, der Schmerz war so rase	degold.txt

mental power' and *mit allen meinen Geisteskräften* 'with all my mental power'. Such variations appear less commonly than do examples of the *mit* + inflected ALL + *Kraft/Gewalt/Macht/Wucht* pattern, and will not be elaborated upon here.

Unsurprisingly, *mit aller Kraft* and its synonyms co-occur predominantly with verbs that denote a general expenditure of effort to achieve or resist something (e.g. *versuchen* to try, *sich wehren gegen etw.* to resist, *vermeiden* to avoid) or, more specifically, physical exertion or the use of extreme physical force (*rennen* to run, *kämpfen* to fight, *stoßen* to shove, *in etw. treiben* to drive into something, *einhauen* to lash into, *sich stemmen gegen* to brace oneself against something, *werfen* to throw, *erstechen* to stab to death). Despite the small sample available to us here, a fairly clear picture emerges of the semantic 'preferences' (Sinclair 1996) of *mit aller Kraft/Gewalt/Macht/Wucht.*[13] The predictability of co-occurrence of the adjunct with this class of verbs suggests that there is an element of prefabrication at work here, or at least that there is such compatibility between the adjunct and verbs in question that they are predisposed to co-occur.

If we analyse these co-occurrences as prefabricated units, or extended units of meaning, to use Sinclair's term, the question arises as to whether these units of meaning also constitute units of translation. An inspection of the bilingual concordances in Table 3.2 shows that where the verb denotes the use of extreme physical force, the tendency is for the translator to use an adjunct based on 'with all one's might'. In instances where the more general verb *versuchen* 'try' is used, the translations include 'as hard as they could' and 'with all her strength'.

One could thus surmise that the nature of the verb that co-occurs with the adjunct in question has a bearing on how that adjunct is translated, that is that the two are taken as a unit in translation, but the size of the sample investigated here (as well as the distribution of individual examples across texts) militates against any sort of conclusion. It also be argued, that in all of the examples cited, translation could just as easily be compositional, that is, each constituent (verb, adjunct) is translated in isolation from the others, or, indeed, that what happens is something between these extremes: the adjunct is translated in a separate move to the verb, but with the translation of the verb in mind. In the absence of the kind of detailed information about the translation/target text production process that the process-oriented studies addressed above can yield, translation products cannot reveal directly whether they were created in one or a series of 'moves'.

What is clear, however, is that *mit aller Kraft/Gewalt/Macht/Wucht* cannot normally be translated without knowledge of the agent of the action concerned. Interpreted as a discrete entity, the German *mit aller Kraft* does not make explicit whose might is being brought to bear in a particular action. This information is, however, retrievable from the co-text: the agent whose force is being brought to bear is the same as that of the main verb. In English, this information must be made explicit through the use of a co-indexical determiner in front of the noun that translates *Kraft* or its synonyms, as illustrated in examples (2) and (3) below (where the superscripts indicate co-indexation/co-referentiality and

Table 3.2 *mit all** selected concordances and translations in GEPCOLT

1.	bier.de P1260 S8 Es war, als ob sie **mit aller Gewalt** versuchten, sich durch den Mund zu befreien.	bier.en P1260 It was as if they were trying **as hard as they could** to escape through his mouth.
2.	zuern1.de P240 S1 Sie stößt mit dem Stuhl **mit aller Gewalt** gegen die Türe. Da erscheint die Wärterin – mit dem gleichen mitleidigen Gesicht – und führt sie in eine große, andere Zelle, worin sich nichts befindet.	zuern1.en P240 She beats the chair against the door **with all her might**, whereupon the warder appears with the same compassionate face and takes her to another cell, a large one which is completely empty.
3.	zuern2.de P99 S5 Einmal hat sie in einer solchen unerträglichen Depression so lange Zeit im Bett gelegen, bis ihre Haut wund geworden ist und die Krankenschwester sie **mit aller Gewalt** dazu gebracht hat, aufzustehen und sich normal zu bewegen und zu beschäftigen.	zuern2.en P99 Once, during such an unbearable depression, she spent so long in bed that her skin became sore and the nurse had to **force her physically** to get up, walk about and occupy herself in a normal way.
4.	bier.de P951 S4 Sie versuchte **mit aller Kraft**, bei klarem Verstand zu bleiben.	bier.en P951 She tried **with all her strength** to keep a clear head.
5.	hofmann.de P16 S4 Meist müssen wir ihn, ehe er uns hört, lange am Ärmel zupfen und ihn aus den Gedanken, in denen er steckt, **mit aller Kraft** herausziehen, doch selbst dann sagt er nicht viel.	hofmann.en P16 Mostly, before he'll listen to us, we have to tug for a long time at his sleeves and drag him **with all our strength away** from his thoughts, yet even then he doesn't say much.
6.	jelinek2.de P159 S2 Wer sich diese Stopfnadel aus meinem Handarbeitszeug **mit aller Kraft**, und wenn ich sage Kraft, dann meine ich das auch, mitten in der Stunde unter den Fingernagel stößt, ohne dabei laut aufzuschrein, mit dem geh ich ins Bubenklosett, linke Kabine.	jelinek2.en P159 If anyone will **force** this darning needle from my needlework kit under his fingernail during class without shouting out, and when I say **full force** I mean full, I'll go to the boys' toilet with him, the cubicle on the left.
7.	jelinek2.de P543 S3 Heute zieht Rainer ein Pfadfinderfahrtenmesser, das eigentlich, seiner ursprünglichen Bestimmung nach, ein HJ-Fahrtendolch in Pension ist, und posiert **mit aller Kraft** wie auf Sophies Bruderfoto.	jelinek2.en P543 Today, Rainer draws a boy scout's knife (which was originally a Hitler Youth dagger and is now in retirement) and poses like the photo of Sophie's brother **as well as he's able**.

Continued

Table 3.2 Continued

8.	jelinek2.de P756 S3 anschließend ersticht er die tote Schwester **mit aller Kraft**.	jelinek2.en P756 Next he stabs his dead sister **with all his might**.
9.	kirchhof.de P1727 S2 Der Mann, der Kurt Lukas war, schien sein Messer **mit aller Kraft** in die Hündin zu treiben.	kirchhof.en P1727 He seemed to be plunging his knife into the dog **with all his might**.
10.	loest.de P252 S7 Ein Mann rannte einen Weg entlang auf mich zu, rechts und links standen kahle Bäume vor dem Feld, der Mann rannte **mit aller Kraft**, das sah ich an den gewölbten Schultern.	loest.en P252 A man was running along a path towards me, to right and left stood skeletal trees at the edge of the fields, the man was running for **all he was worth –** I could tell that from the hunch of his shoulders.
11.	ransmayr.de P704 S3 Wie immer und wie aus Angst vor der trügerischen Starre der Meereisdecke ziehen die Hunde **mit aller Kraft** nach der nächstgelegenen Küste, nach dem Gletscherabbruch im Norden;	ransmayr.en P704 As always- as if they fear that the solid sheet of frozen sea is a trick – the dogs **strain every muscle** to pull toward the nearest coastline, the glacial wall to the north.
12.	wodin.de P624 S5 ich stürzte zurück ins Zimmer, riß die Decke vom zweiten Bett und legte sie über ihn, ich preßte seine Schultern gegen die Matratze, stemmte mich **mit aller Kraft** gegen seinen Amoklauf,	wodin.en P624 <s>Dashing back into the room, I tore a blanket off the other bed and draped it over him, forced his shoulders down on the mattress, braced myself against his heaving body **with all my might**.
13.	zuern1.de P621 S3 Diese ‘Froschkönige’, diese im Lichte lebenden Embryo-Gesichter, scheinen **mit aller Kraft** von der Sonne angezogen zu werden, als könnte die Sonne etwas an dieser wie unterbrochenen Mensch-Werdung retten.	zuern1.en P621 These “frog kings,” these embryo-faces which live in the light, appear to be drawn **by every fibre of their being** to the sun, as if the sun could rescue something of this seemingly interrupted process of growing up.

14.	zuern1.de P499 S6 Sie wirft das hübsche, hölzerne Gestell, in dem die leeren Glas-Röhrchen auf ihr Blut warten, **mit aller Kraft und Empörung** auf den Fußboden und sieht mit Freude zu, wie das Glas zerbricht.	zuern1.en **P499 With all her might and indignation** she knocks over the pretty wooden rack containing the empty glass tubes which are waiting for her blood, and looks on with joy as the glass shatters on the floor.
15.	kirchhof.de P75 S2 McEllis keuchte. <s>Er kämpfte **mit allen Kräften**, um diese Bergprüfung zu bestehen.	kirchhof.en P75 McEllis was panting, striving **with all his might** to emerge victorious from this mountain trial.
16.	roth2.de P237 S32 Er wehrte sich zwar **mit aller Macht** dagegen, verurteilt zu werden, aber tief in seinem Inneren empfand er seine Verurteilung wie eine nachträgliche Rechtfertigung seines Lebens und die drohende Strafe als Buße dafür, was aus ihm geworden war.	roth2.en P237 Although he resisted being convicted **with all his might**, deep inside he felt the conviction to be a belated justification for his life and the imminent sentence as penance for what he had become.
17.	wodin.de P638 S6 <s>Und eines weiß ich ja doch mit letzter Sicherheit: um dieses Licht kämpft er, er kämpft um die Rückkehr in die Welt, er will nicht sterben, er wehrt sich **mit aller Macht**, mit der brachialen, elementaren Kraft seines Körpers,	wodin.en P638 <s>Of one thing I was sure: he was fighting for the light, struggling to return to the world, reluctant to die, defending himself **with all his might**, with all the brute strength in his body.
18.	gold.de P108 S6 man hieb auf ihn von der Seite ein, **mit aller Wucht**, der Schmerz war so rasend, daß er größer schien als er selbst; <s>das Denken wurde aber trotzdem in ihm immer schneller: wenn man wüßte, wie weh das tut, würde man ihn nicht prügeln.	gold.en P108 he was being lashed by someone at his side **with the utmost weight and force**, the pain was so scorching, it seemed bigger than himself, yet his thought worked all the more swiftly: if they only knew how much it hurt, they would stop whipping him.

the source of the example is indicated in parentheses), and in Table 3.3, which presents the translations of *mit aller Kraft* and so on in summary form.

2a. Er^{i} wehrte sich zwar mit aller Macht dagegen, verurteilt zu werden (roth2.de)
2b. He^{i} resisted being convicted with all his^{i} might (roth2.en)
3a. Sie^{j} stößt mit dem Stuhl mit aller Gewalt gegen die Türe (zuern1.doc)
3b. She^{j} beats the chair against the door with all her^{j} might (zuern1.en)

Although four of the translations of the *mit aller Kraft* pattern, 'as hard as they could', 'as well as he's able', 'for all he was worth', and 'by every fibre of their being', depart from the 'with all one's might/strength' pattern, they do share with the above examples the fact that they make explicit whose effort is being expended.

Applying Toury's product-oriented 'no leftovers' principle, we can say that expressions such as 'with all his might' all contain the solution (with no leftovers) to the problem posed by *mit aller Kraft,* but the source of part of the solution (i.e. knowledge of the agent) lies outside the problem itself, and here we have a case of the clause operating as the 'macro-unit'.[14]

Note, however, that the German adverbial allows knowledge of the agent to remain vague as in example (4a), where the agent is the impersonal pronoun *man*:

4a. man hieb auf ihn von der Seite ein, mit aller Wucht (gold.de)
4b. he was being lashed by someone at his side with the utmost weight and force (gold.en)

Table 3.3 Summary of translations of *mit aller Gewalt/Kraft/Macht/Wucht* in GEPCOLT

mit aller Gewalt	1. as hard as they could
	2. with all her might
	3. *force
mit aller Kraft	4. with all her strength
	5. with all our strength
	6. *force/full force
	7. as well as he's able
	8. with all his might
	9. with all his might
	10. for all he was worth
	11. *strain every muscle
	12. with all my might
	13. by every fibre of their being
	14. with all her might and indignation
mit allen Kräften	15. with all his might
mit aller Macht	16. with all his might
	17. with all his might
mit aller Wucht	18. with the utmost weight and force

As is typical in the translation of such impersonal constructions from German into English, the translator uses a passive voice in English (as well as indicating a very vague human agent in 'someone'). This syntactic fact helps to explain why the translator in this case did not use a conventional 'with all his/her etc. might' translation to translate *mit aller Wucht*. Given that a precise referent cannot be picked out as agent, the translator has to use an (unconventional) impersonal adjunct 'with the utmost weight and force', to obviate the need for a possessive determiner which would be co-referential with an explicit agent. Although this translation is less context dependent than more conventional translations, it remains a one-off solution in GEPCOLT. We cannot assume that context independence correlates with frequency of occurrence. Sometimes context dependence is the norm.

Phrases like *mit aller Kraft/Macht/Wucht* make interesting test cases for Kraif's (2003, and above) distinction between lexical correspondences and translational equivalents. The latter, we recall, are translation solutions that are highly context-dependent, one-off solutions that may involve semantic discrepancies. Solutions like 'with all his might' and 'with all her might' are, as we have seen, highly dependent on context (or co-text), and semantically, they are more specific than the German phrase. They do, however, vary *systematically* with the co-text/context, and thus cannot be seen as idiosyncratic, one-off solutions. Their conventional nature, despite internal variation, is supported by the fact that they find their way into standard German–English bilingual dictionaries, where the variable *one's* (*mit aller Kraft* = with all one's might) acts as a placeholder for the appropriate possessive determiner which is, in turn, generated by the TL speaker/writer on the basis of his/her interpretation of the German ST. If a corpus-based approach to identifying units of translation on the basis of repeatedly observed correspondences between units in source and target texts is to be successful, it must also be able to cope with such paradigmatic variation (in the same way that Danielsson's (2003) monolingual analysis of units of meaning can). Although repeated 'clusters' or 'chains' (see Kenny 2001: 43; Stubbs 2002; Baker 2004) of unvarying strings of tokens are also revealing, patterns, which enable some abstraction from the data, ultimately can account for more data.

Another example from GEPCOLT demonstrates the usefulness of corpora in highlighting the recurrent patterns of the source language, and in identifying extended units of meaning.

The concordances in Table 3.4 show 6 (out of 18) instances where the lemmas Auge 'eye' and aufreißen 'to open wide' collocate in GEPCOLT, namely those instances where *aufgerissene/aufgerissenen* appears as a pre-modifier to the plural head noun *Augen*.

Given the monolingual evidence supplied in Table 3.4, we might surmise that *mit weit aufgerissenen Augen* is a kind of prefabricated phrase, where the expectancy of co-occurrence of *weit* with *aufgerissenen Augen* is so great that its

Table 3.4 Concordance for *aufgerissen* Augen* in GEPCOLT

1. bscheulich aus. Sie hatten aufgerissene Augen und rote, emporstehende Haar
2. en Punkt aus. Sie kam mit aufgerissenen Augen von einem Schrecken zurück,
3. nden dann herum, mit weit aufgerissenen Augen, in besonders gepflegten
4. nblick erstarrt, mit weit aufgerissenen Augen; und dann ein Schrei aus der
5. auens, das in seinen weit aufgerissenen Augen steht, ich rufe ihn, rufe
6. Es war ein Mann mit weitaufgerissenen Augen, vermutlich ein abgestürzter

presence adds little extra meaning to the phrase. Two other contrastive facts support this analysis. First, as outlined in Kenny (2004: 342–3) ‘to open wide’ is the most common translation (i.e. used in five out of ten cases) of AUFREIßEN as a full verb (and without modification by the adverb *weit*) in GEPCOLT. Secondly, and as suggested by the gloss given above, the ‘prototypical’ translation found for AUFREIßEN in bilingual dictionaries is also ‘to open wide’. It appears then that AUFREIßEN already contains the meaning of *weit*, and that the presence of *weit* in the concordance lines above serves more to intensify the meaning of *aufgerissen* than to add any extra information. In other words, *weit* appears here to be delexicalized; its pragmatic function is more important than its semantic one. While we may accept that *mit weit aufgerissenen Augen* is a single unit of meaning, this does necessarily mean that it has a stable, unchanging translation in the target texts, akin to Kraif’s (2003) ‘lexical correspondences’. Rather, as the examples in Table 3.5 demonstrate, translations are anything but stable on a formal level, and include: gawping eyes; gazing wide-eyed at all that was going on; with wide open eyes; open-eyed; in his wide open eyes; eyes popping out of her head.

Despite the variety in the translations of (*mit/in seinen*) *weit aufgerissenen Augen*, there is no evidence that translators have interpreted *weit* as adding extra semantic features if AUFREIßEN is already considered to imply the semantic feature [+wide]. One other example in the corpus, however, points to a possible over-interpretation of the co-occurrence of *weit* and AUFREIßEN:

5a. Die Hühner . . . reißen ihre Augen weit auf und rühren sich nicht, weil Hühner sich beim Eierlegen nicht rühren können. (wodin.de)
5b. The hens . . . open their eyes very wide and don’t move because hens can’t move while they’re laying. (wodin.en)

Here the translators have added the intensifier ‘very’ to correspond, perhaps, to a perceived discrete element of meaning in the adverb *weit*. This example could thus be seen as an instance of over-translation, in Vinay and Darbelnet’s (1995: 16) understanding of the term as an instance where the translator sees ‘two units when there is only one’. In Kenny (2004: 343), I argue that prototypical translations, evidence for which is found in bilingual dictionaries

Table 3.5 *aufgerissen* Augen* selected concordances and translations in GEPCOLT

1.	zuern1.de P268 S2 Sie hatten **aufgerissene Augen** und rote, emporstehende Haare.	zuern1.en P268 They had **gawping eyes** and red hair which stood on end.
2.	gold.de P193 S2 Einige von ihnen setzten einfach ihre Kinder im Heim ab, für ein paar Tage, die standen dann herum, **mit weit aufgerissenen Augen**, in besonders gepflegten Kleidern.	gold.en P193 Some simply left their children at the home for a few days, they would stand around, **gazing wide-eyed at all that was going on**, wearing particularly smart clothes.
3.	roth2.de P21 S3 Es war ein Mann **mit weitaufgerissenen Augen**, vermutlich ein abgestürzter Bergsteiger, der einen Pickel in der Rechten hielt.	roth2.en P21 A man **with wide open eyes**, presumably a mountaineer who had fallen to his death, holding an ice pick in his right hand.
4.	weller.de P2083 S3 Sie kam **mit aufgerissenen Augen** von einem Schrecken zurück,	weller.en P2083 She was returning **open-eyed** from a horror . . .
5.	wodin.de P160 S3 . . . und ich sehe, was ich sehen will, den Ausdruck des tödlichen, über alles hinausgehenden Grauens in Ljudas Gesicht, einen Augenblick erstarrt, **mit weit aufgerissenen Augen**;	wodin.en P160 . . . and I see what I wanted to see: a look of deathly, all-transcending terror on Lyuda's face, frozen for an instant, **eyes popping out of her head**
6.	wodin.de P635 S6 die Mauer des Grauens, das **in seinen weit aufgerissenen Augen** steht,	wodin.en P635 the wall of dread that loomed **in his wide open eyes**.

and parallel corpora, can point up instances of delexicalization in the source language. This in turn allows for extended units of meaning to be identified in source texts, and the translations (prototypical and non-prototypical) of individual instances of such units to be investigated. In this way the analysis of units of translation and units of meaning proceeds in a reciprocal fashion.

3.8 Conclusions

In this chapter we have attempted to show how parallel corpora can be used to investigate units of translation. We started out by reviewing a good deal of the existing literature on translation units, in a bid to clarify the concepts relevant to our discussion. Having now approached the issue from the particular viewpoint of corpus-based translation studies it is time to ask whether our discussion has any implications for those basic concepts. A first conclusion emerges from our discussion of the *mit aller Kraft* examples, namely that the solution to a problem posed by a discrete stretch of source text may be contained within the boundaries of a discrete stretch of target text, but some of the source of that solution may lie outside the problem part of the pair. In other words, the ST part of the coupled pair (often a translation 'atom') and the parts of the ST that must be kept in focus to arrive at a TT solution do not necessarily coincide. It is this fact that motivates the terminological distinction between 'parts of coupled pairs', 'translation foci', and 'translation atoms' in the first place, and our discussion can be seen as lending empirical support to Bennett's (1994) position in particular. A second conclusion has already been suggested above, namely that analyses based on repeated observations of syntagmatic units in source and target texts can be made more powerful if they allow for paradigmatic variation. Thirdly, it is hoped that even the meagre evidence provided in this chapter has shown that monolingual and comparative analyses can reinforce each other. But there lies the rub: given the relatively small size of the parallel corpus used in this study, and in translation studies in general, it is difficult to back up any claims with an abundance of empirical evidence. If this direction is to be pursued, researchers will have to either extend their corpora, or take their leads from monolingual research based on far larger sets of data.

Notes

[1] The research reported on here was funded by an Albert College Fellowship awarded by Dublin City University.

[2] In keeping with accepted practice in corpus linguistics, natural language processing and corpus-based translation studies, 'parallel corpus' is used here to refer to a collection of source texts alongside their translations. GEPCOLT is described in detail in Kenny (2001).

[3] Cf. Zhu (1999: 433–4), who links discourse on translation units to the idea of 'absolute' equivalence.

[4] Cf. Huang and Wu's (2009) study, in which 56 out of 65 professional translators surveyed reported that they translated 'sentence by sentence'.

[5] Thus Jakobsen relates the number of segments in target text production under various conditions (e.g. with or without concurrent TAP) to the number of source text characters, to show, for example, that when translating with think aloud, subjects had an average of 7.49 segments per 100 ST characters, compared to 5.15 without (2003: 90).

[6] A translation memory can be understood as a particular kind of parallel corpus.

[7] We will not elaborate here on the issue of co-selection of grammar and lexis, except to say that this is not a position unique to Sinclair. Hunston and Francis (2000) provide an overview of other relevant research. Cognitive approaches to grammar can also be based on the notion that grammatical constructions are symbolic, that is they carry some meaning, and this meaning is normally in concert with the meanings of lexical items that are found in such constructions (see, for example, Stefanowitsch and Gries 2003).

[8] Kondo (2007) relaxes this requirement somewhat by allowing repeated source text segments that have one *predominant* equivalent (i.e. in at least 85 per cent of cases the same translation is used) to be recognized as translation units.

[9] *Mit* is indeed one of the most common words in German. It occurs 8,911 times in the nearly one million words of German in GEPCOLT. It is thus the 15th most common type in the German side of the corpus, and the most common preposition. Although *auf* occurs 9,028 times and is the 14th most common type in the German corpus, an estimated 1,000 instances of this type are accounted for by separable prefixes rather than prepositions. In fact, both *mit* and *auf* can occur as separable verbal prefixes, and instances where they have become separated from their respective verbs are homographic with prepositional *mit* and *auf*. Such instances of homography are not taken into account in the above statistics for *mit*, but their impact is minimal (a cursory inspection of concordance data suggests there are around 125 instances of separated prefix *mit*, compared with 1,000 for *auf*).

[10] We follow here the convention of using small capitals to denote lemmas, understood to include all relevant inflected forms.

[11] The right-hand column in Table 3.1 indicates the file in GEPCOLT from which the example is taken. In subsequent bilingual Tables related filenames are used, so that debier.txt in Table 3.1 refers to the same file as bier.de in Table 3.2.

[12] Table 3.1 excludes instances where *mit* is, for example, a strongly bound preposition, that is one required by its accompanying verb, or where *all* acts as a pronoun. It also excludes instances of *mit allen Mitteln* 'with all means' (four instances) and the related *mit allen gesetzlichen Mitteln* 'with all legal means', as well as the two fairly fixed phrases *mit aller Sorgfalt* (one instance) and its synonym *mit aller Vorsicht* (one instance) 'with all (due) care'.

[13] One of the main exceptions to this trend, the sentence fragment '[Rainer] posiert mit aller Kraft wie auf Sophies Bruderfoto' translated as '[Rainer] poses like the photo of Sophie's brother as well as he's able', may be interpreted ironically, the incongruity caused by the juxtaposition of *mit aller Kraft* and the usually anodyne verb *posiert* reflecting the mismatch between Rainer's goal and the effort he expends in achieving it.

[14] Although Bennett (1994) describes the text, rather than local structures such as clauses, as the translation 'macro-unit'.

References

Alves, Fabio and Daniel Couto Vale (2009) 'Probing the Unit of Translation in Time: Aspects of the Design and Development of a Web Application for Storing, Annotating and Querying Translation Process Data', *Across Languages and Cultures* 10(2): 251–73.

Alves, Fabio and José Luiz V. R. Gonçalves (2003) 'A Relevance Theory Approach to Inferential Processes in Translation', in Fabio Alves (ed.) *Triangulating Translation. Perspectives in Process-Oriented Research*, Amsterdam and Philadelphia: John Benjamins, 3–24.

Baker, Mona (2004) 'A Corpus-Based View of Similarity and Difference in Translation', *International Journal of Corpus Linguistics* 9(2): 167–93.

Ballard, Michel (1997) 'Créativité et traduction', *Target* 9(1): 85–110.

Barbosa, Heloisa G. and Aurora M.S. Neiva (2003) 'Using Think-aloud Protocols to Investigate the Translation Process of Foreign Language Learners and Experienced Translators', in Fabio Alves (ed.) *Triangulating Translation. Perspectives in Process-Oriented Research*, Amsterdam and Philadelphia: John Benjamins, 137–55.

Bassnett-McGuire, Susan (1980) *Translation Studies*, London and New York: Methuen.

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

Bennett, Paul (1994) 'The Translation Unit in Human and Machine', *Babel* 40(1): 12–20.

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

Carl, Michael and Andy Way (eds) (2003) *Recent Advances in Example-Based Machine Translation*, Dordrecht, Boston and London: Kluwer Academic Publishers.

Danielsson, Pernilla (2003) 'Automatic Extraction of Meaningful Units from Corpora', *International Journal of Corpus Linguistics* 8(1): 109–27.

Dayrell, Carmen (2004) 'Towards a Corpus-Based Research Methodology for Investigating Lexical Patterning in Translated Texts', *Language Matters* 35(1): 70–101.

Durrell, Martin (1991) *Hammer's German Grammar and Usage Revised Second Edition*, London, Sydney and Auckland: Edward Arnold.

Harris, Brian (1988) 'Bi-text: A New Concept in Translation Theory', *Language Monthly* 54: 8–10.

Huang, Harry and Canzhong Wu (2009) 'The Unit of Translation: Statistics Speak', *Meta* 54(1): 110–30.

Hunston, Susan and Gill Francis (2000) *Pattern Grammar*, Amsterdam and Philadelphia: John Benjamins.

Jakobsen, Arnt Lykke (2003) 'Effects of Think Aloud on Translation Speed, Revision, and Segmentation', in Fabio Alves (ed.) *Triangulating Translation. Perspectives in Process-Oriented Research*, Amsterdam and Philadelphia: John Benjamins, 69–95.

Kay, Martin (2000) 'Preface', in Jean Véronis (ed.) *Parallel Text Processing. Alignment and Use of Translation Corpora*, Dordrecht, Boston and London: Kluwer Academic Publishers, xv–xx.

Kenny, Dorothy (2001) *Lexis and Creativity in Translation: A Corpus-Based Study*, Manchester: St. Jerome.

— (2004) 'Die Übersetzung von usuellen und nicht unusuellen Wortverbindungen vom Deutschen ins Englische', in Kathrin Steyer (ed.) *Wortverbindungen – Mehr Oder Weniger fest. Institut für Deutsche Sprache Jahrbuch 2003*, Berlin and New York: Walter de Gruyter, 335–47.

Kiraly, Don (1990) 'Toward a Systematic Approach to Translation Skills Instruction'. Unpublished PhD dissertation, Ann Arbor, UMI.

Kondo, Fumiko (2007) 'Translation Units in Japanese-English Corpora. The Case of Frequent Nouns'. Paper presented at *Corpus Linguistics 2007*, University of Birmingham, 27–30 July 2007.

Kraif, Olivier (2002) 'Translation Alignment and Lexical Correspondences', in Bengt Altenberg and Sylviane Granger (eds) *Lexis in Contrast: Corpus-Based Approaches*, Amsterdam and Philadelphia: John Benjamins, 271–89.

— (2003) 'From Translational Data to Contrastive Knowledge', *International Journal of Corpus Linguistics* 8(1): 1–29.

Krings, Hans P. (2001) *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing*, London and Kent, OH: Kent State University Press.

Livbjerg, Inge and Inger M. Mees (2003) 'Patterns of Dictionary Use in Non-domain-specific Translation', in Fabio Alves (ed.) *Triangulating Translation. Perspectives in Process-Oriented Research*, Amsterdam and Philadelphia: John Benjamins, 123–36.

Lörscher, Wolfgang (1996) 'A Psycholinguistic Analysis of Translation Processes', *Meta* 41(1): 26–32.

Malmkjær, Kirsten (1998) 'Unit of Translation', in Mona Baker (ed.) *The Routledge Encyclopedia of Translation Studies*, London and New York: Routledge, 286–8.

Santos, Diana (2000) 'The Translation Network. A Model for a Fine-Grained Description of Translations', in Jean Véronis (ed.) *Parallel Text Processing. Alignment and Use of Translation Corpora*, Dordrecht, Boston and London: Kluwer Academic Publishers, 169–86.

Sinclair, John (1987) 'Collocation: A Progress Report', in Ross Steele and Terry Treadgold (eds) *Language Topics: Essays in Honour of Michael Halliday*, Amsterdam and Philadelphia: John Benjamins, 219–331.

— (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

— (1992) 'Trust the Text', in Martin Davies and Louise Ravelli (eds) *Advances in Systemic Linguistics*, London: Palmer, 5–19.

— (1996) 'The Search for Units of Meaning', *Textus* IX: 75–106.

— (2004) 'The Lexical Item', in *Trust the Text: Language, Corpus and Discourse*, London and New York: Routledge, 131–48.

Stefanowitsch, Anatol and Stefan Th. Gries (2003) 'Collostructions: Investigating the Interaction of Words and Constructions', *International Journal of Corpus Linguistics* 8(2): 209–43.

Stubbs, Michael (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*, Oxford: Blackwell.

— (2002) 'Two Quantitative Methods of Studying Phraseology in English', *International Journal of Corpus Linguistics* 7(2): 215–44.

Teubert, Wolfgang (2002) 'The Role of Parallel Corpora in Translation and Multilingual Lexicography', in Bengt Altenberg and Sylviane Granger (eds) *Lexis in Contrast: Corpus-Based Approaches*, Amsterdam and Philadelphia: John Benjamins, 189–214.

— (2004) 'Units of Meaning, Parallel Corpora, and Their Implications for Language Teaching', in Ulla Connor and Thomas A. Upton (eds) *Applied Corpus Linguistics. A Multidimensional Perspective*, Amsterdam and New York: Rodopi, 171–89.

Toury, Gideon (1980) *In Search of a Theory of Translation*, Tel Aviv: The Porter Institute for Poetics and Semiotics.

— (1995) *Descriptive Translation Studies and Beyond*. Amsterdam and Philadelphia: John Benjamins.

Véronis, Jean (ed.) (2000) *Parallel Text Processing. Alignment and Use of Translation Corpora*, Dordrecht, Boston and London: Kluwer Academic Publishers.

Véronis, Jean and Philippe Langlais (2000) 'Evaluation of Parallel Text Alignment Systems', in Jean Véronis (ed.) *Parallel Text Processing. Alignment and Use of Translation Corpora*, Dordrecht, Boston and London: Kluwer Academic Publishers, 369–88.

Vinay, Jean-Paul and Jean Darbelnet (1958) *Stylistique Comparée du français et de l'anglais*, Paris: Didier and Montréal: Beauchemin.

— (1995) *Comparative Stylistics of French and English. A Methodology for Translation* [translated and edited by Juan C. Sager and M.-J. Hamel], Amsterdam and Philadelphia: John Benjamins.

Zabalbeascoa, Patrick (2000) 'From Techniques to Types of Solutions', in Alison Beeby, Doris Ensinger and Maria Presas (eds) *Investigating Translation: Selected Papers from the 4th International Congress on Translation, Barcelona, 1998*, Amsterdam and Philadelphia: John Benjamins, 117–27.

Zhu, Chunshen (1999) 'UT Once More: The Sentence as the Key Functional Unit of Translation', *Meta* 44(3): 429–47.

— (2005) 'Accountability in Translation Within and Beyond the Sentence as the Key Functional UT: Three Case Studies', *Meta* 50(1): 312–35.

Chapter 4

Hardwiring Corpus-Based Translation Studies: Corpus Encoding

Federico Zanettin

In this chapter, I try to answer three main questions. The first question is, 'Do corpora for descriptive translation studies need encoding?' I think they do, and I try to explain why. The second question is 'What needs to be encoded?' In order to answer this question, I propose a model consisting of multiple layers of annotation. Finally, in response to a third question, 'How to encode?', I suggest the adoption of standards for electronic texts. In particular, I examine XML/TEI. I begin by providing some terminological background as to what is meant by corpora and encoding, and examine which types of corpora have been used in translation research in relation to corpus encoding. I focus on corpus-based research in descriptive translation studies, and consider the different stages involved in the creation of a corpus as a resource. I then in turn consider the why, what and how questions. Finally, I show examples of texts encoded using the XML/TEI standards. These examples are taken from the *CEXI* project, carried out between 2000 and 2004 at the School for Translators and Interpreters of the University of Bologna and aiming at the creation of a bidirectional parallel English–Italian corpus.

4.1 Corpora in Translation Research

First of all, corpus-based translation studies are not necessarily concerned with computers and electronic texts. One of the best known pieces of corpus-based research, that of Vanderauwera (1985), which relates to Dutch fiction translated into English, is based on a carefully investigated corpus of printed volumes. Translation teachers, as well as translation scholars, have compiled and studied collections of printed texts of various types, for instance collections of documents such as advertisements, technical and institutional texts produced in two different cultures under similar circumstances, known as 'parallel texts'. A corpus can mean a heap of books or a pile of photocopies and cut-outs. More often than not, however, a corpus is taken to mean 'a helluva lotta text stored in a computer in machine readable format' (Leech 1992: 106).

This distinction may perhaps appear obvious but it is nonetheless important to stress that, ultimately, it is not the object of enquiry which is different but the methodology used for the investigation. What defines a corpus is first of all the 'principled way' (Johansson 1995) in which a collection of texts was assembled. The electronic format makes it possible to study a corpus with different methodological tools.

A second distinction I would like to draw is that between native and non-native electronic texts. By 'native' I mean electronic texts which were originally created to be read on a screen rather than printed on paper. On the other hand, many electronic corpora are digitized versions of printed texts, such as books and newspapers. In fact, the very first corpora were thus created (e.g. the *Brown Corpus*, the Cobuild corpus, the *BNC Corpus*).[1] There can also be mixed corpora, such as the *ANC* (*American National Corpus*).

I believe it is important to maintain the distinction between the two. If the medium is the message, native electronic texts are different in kind from non-native electronic texts. For instance, the results of a study conducted within the *ANC project*, which compared written web material with other types of planned written discourse, suggest that 'web texts' can be thought of as a specific genre, more 'cryptic and terse' and consistently containing shorter paragraphs than comparable printed texts (Ide et al. 2002: 843–4). Indeed, it would be worth studying whether translations of native electronic texts differ (and if so on which counts) from translations for the printed medium.

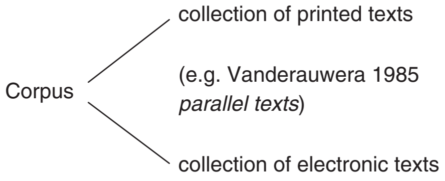


FIGURE 4.1 Printed vs. electronic corpus

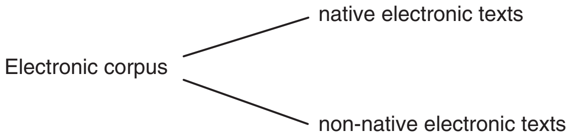


FIGURE 4.2 Electronic text corpora

Still another type of corpora of interest for scholars in translation studies is corpora of interpreting data, that is, spoken corpora consisting of transcriptions of audio or audiovisual data recorded during an interpreted event (see Setton 2003 on criteria for collection and analysis of simultaneous interpreting corpora). A number of additional problems are raised in the case of interpreting corpora by the fact that we are dealing with oral rather than written data. Transcriptions, whether linked to audio(visual) recordings or not, are not only non-native texts, but also secondary data in themselves (see Leech et al. 1995 on spoken corpora in general and Cencini and Aston 2003 on problems related to the encoding of interpreting data). In this chapter, I am primarily concerned with non-native electronic texts, that is, texts for which there exists a printed version (books, newspapers, etc.) which continues to exist alongside the electronic version.[2]

One more terminological clarification is necessary. At this point, text encoding should be understood as 'a process whereby documents are transferred to an electronically searchable format for scholarly research', adopting the broad definition of the Electronic Text Center of an American University Library (University of Nebraska-Lincoln).

4.2 Why Do Corpora Need Encoding?

In translation research, electronic corpora have been used by people working in three main areas:

1. Machine (assisted) translation
2. Translator training/education
3. Descriptive translation studies.

Issues related to encoding play a different role in each of these areas. For the Machine (assisted) Translation and NLP (Natural Language Processing) communities, corpora are essentially parallel corpora, and much research has been concerned either with alignment techniques (i.e. ways to retrieve paired instances of translated + original segments) or with annotation schemes (i.e. ways to add computer-processable information to electronic text corpora). Alignment is essential for the purpose of mapping translation equivalent units, so as, for example, to automatically extract terminology or translation memory units from parallel corpora. Much effort has also gone into devising ways to attach meta-textual and metalinguistic information to texts, in order to increase automatic processability.

Applications include aligned parallel corpora, such as multilingual parliament proceedings (from the European and Canadian Parliament) or computer manuals (cf. for example: http://opus.lingfil.uu.se/).[3] Examples of the second type of research efforts include TMX, which stands for Translation Memory eXchange and is a standard developed by the Localization Industry Standard Association (LISA) for storing and exchanging translation memories created by computer-aided translation (CAT) tools and localization tools – and XCES, which stands for XML Corpus Encoding Standard and has been created in the provision of corpora to be used in language engineering and NLP (Ide et al. 2000). XCES includes specifications for aligned data in parallel corpora.

Translation scholars and teachers may have little interest in delving into the intricacies of seemingly abstruse algorithms or programming languages, which in fact are the domain of language engineers. However, those who wish to use electronic resources in their research or teaching may wish to devote a little time to understanding the principles behind the technology, and the methodological tools which can be used to create and analyse corpus resources, especially since language engineers seem to be more interested in devising schemes for processing the data, rather than in their analysis, and the task of creating a corpus for practical or descriptive purposes is usually left to the translator or the translation scholar.

In the area of translator education, practical experiments carried out with translation students usually involve first-hand access to corpora, with two main purposes, terminology mining and target text writing.

As part of the wider picture of translator education, corpora have been used in the language learning classroom, taking advantage of research in second language pedagogy such as that exemplified by data-driven learning (see, for example, Johns 1986; Aston 2001; Kettemann and Marko 2002). Corpora used in this context range from large linguistic corpora such as the BNC (see, for example, Stewart 2000) through rough and ready collections of internet documents, to the internet itself (see, for example, Zanettin 2002a).

The main concern is with the availability of data rather than with their format, and with the translation task at hand. The definition of *corpus* is rather loose, and researchers and teachers in this field are usually not much interested in encoding standards and corpus building in general. Corpora are usually monolingual or bilingual comparable, much in the guise of time-honoured **parallel texts**. The texts included in these corpora are more often than not *native* electronic texts, for example, web pages and texts taken from CD-ROMs. They may differ extensively in terms of format: some are in plain-text, while others carry with them some form of meta-textual information, like web pages which are written in HTML, the HyperText Markup Language. Small corpora used in translator education tend to be a one-off resource, and they can be as easily discarded as they are created. The focus of users is usually on word and pattern searches, and meta-textual information, if present, is generally ignored.

The tools used to analyse these corpora can either be general purpose concordancing programmes such as Wordsmith Tools (Scott 1996), Monoconc (Barlow 2000), Paraconc (Barlow 2002) and Multiconcord (Wools 1995), the latter for parallel corpora, or custom-built software such as XAIRA for the *BNC* or the in-house software used for the *COBUILD* corpus or the Italian *Coris/Codis Corpus* (Rossini Favretti et al. 2002). The web can also be used as a very large corpus with the aid of search engines such as *Google* and web concordancers such as KwicFinder (Fletcher 2004), Webcorp (Kilgarriff 2001) or Wordsmith Tools' Webgetter utility, or as a source of corpora (Gatto 2009, Kilgarriff 2010).

Scholars in the third main area, the use of computerized corpora for theoretical and descriptive research, have been using a wide variety of types of corpora, monolingual and bilingual, comparable and parallel. Parallel corpora can be uni- or multidirectional, following a one-to-one model as in bilingual parallel corpora or a one-to-many model, as in multilingual or multi-target combinations, the so called stars and diamond models (see Zanettin et al. 2003). The type and degree of encoding of these corpora differ as much as their design, ranging from plain text documents to heavily annotated corpora.

Very often corpora for descriptive research are carefully constructed over a period of time. Three main stages are involved in the creation of a corpus from printed source texts. The first stage is that of corpus design, the definition of a target population which the corpus aims at representing and the definition of criteria for the inclusion of and the description of texts. Once a decision has been reached on the design criteria, the acquisition (second) stage begins; text copies have to be located and copyright permission has to be obtained. The third stage, encoding, involves the transfer of source documents from paper to the electronic medium by means of typing or scanning. Texts acquired through OCR (Optical Character Recognition) software are then proofread and annotated. These three stages can be performed in cycles. Corpus design criteria may be modified because of acquisition restrictions and opportunities, and different levels of encoding may be applied at different times.

In my view, creating electronic corpora from printed texts may be well worth the effort, provided that the resources created meet three main criteria:

1. Stability: Creating an electronic corpus for research purposes is usually a medium- or long-term project, involving teamwork and many difficulties encountered along the way. The resources created may then well be a permanent asset and should be constructed according to sound criteria and principles. Ideally, these resources should be available to the wider research community. To this end it is essential to provide adequate documentation for the corpus and its contents. Stability also means that corpora should not depend on specific people or software.
2. Flexibility: Since the final users of a corpus may be different from the people who created it, a corpus should provide for the possibility of further elaborations and unexpected types of investigations. New users may want to build on existing data, for instance, adding linguistic annotation to plain

texts. Given the comparative nature of most corpus-based studies, researchers may want to use only selected parts of a corpus by creating sub-corpora, or use it in conjunction with other corpora. In order to re-use and exchange corpus resources, the adoption of common encoding standards would seem an advisable choice.

3. Accessibility: The advantages of electronic over paper corpora are that the former allow for larger-scale investigations, with more consistent results. However, results may vary depending on the nature of the data and the tools used to investigate them. Ideally, the data themselves should be openly available and accessible to researchers. I am aware that there may be copyright and other related problems, nonetheless being able to locate resources and knowing they share common encoding standards is a first step to accessibility. Translation-driven corpora could be made accessible to the research community through a cataloguing system such as the Open Language Archives Community (OLAC). The OLAC is an international project aimed at creating 'a worldwide virtual library of language resources' (see OLAC website). The OLAC catalogues are accessible through the Linguist List website.

In any case, the *why*-question is a deceitful one, since all electronic texts are by definition encoded. A first level of encoding is that of characters, and obvious as it may be, it is not something which should be taken lightly by translation scholars. Only recently have computers become capable of dealing with so-called minority languages, as ASCII is being replaced by Unicode.[4] To make sense of the words, it is also necessary to know where these words come from, that is, to be able to recover documental information about the texts in a corpus. Minimally, all texts in a corpus are identified, more or less explicitly, by file names. Often, some kind of description about the text (a 'header') is associated with a text file.

Secondly, encoding is analysing. It is not a matter of deciding whether or not to encode, but rather a matter of deciding *what* to encode and *how*. Corpus encoding may be regarded as making explicit one's analysis of a text (at least a preliminary one). For instance, the extralinguistic information associated with a text may be crucial to the interpretation of the results.

Typically, a corpus created in relation to descriptive research will contain at least one translation component, as in monolingual comparable or bilingual parallel corpora. Creating a corpus of translations is one way of defining what translation is. Whereas a descriptive theory has it that a translation is 'any target language text which is presented or regarded as such within the target system itself, on whatever grounds' (Toury 1982: 27) when it is a question of pragmatic decisions about what to include in a corpus, the concept of translation seems to become one with fuzzy edges, maybe better explained in terms of prototype theory (Halverson 1998). The concept of translation is not stable through time. For some texts the translational status may be debatable. As much information as possible about a translation and why it was regarded as such by the corpus compilers should be recorded and be retrievable.

4.3 What May Be Encoded?

The type of encoding I am referring to is the explicit annotation of a text by the corpus compiler or translation scholar. The issue of textual annotation is a hot one, and there are at least as many advocates of a 'clean text' policy as there are of annotated text. According to Tognini-Bonelli (2001), annotation imposes a pre-existing theory on corpus data. The results of investigations done using a theory-specific annotation system are bound to confirm the theoretical framework in which they were conducted. Only by looking at clean texts, she argues, may we be able to discover new features of language. I would tend in principle to agree with this approach, but in my view, it is important not to rule out the possibility of sharing resources with those who hold a different view. The two different approaches may be compatible within a shared annotation framework. I should also stress that the controversy is about linguistic annotation, not about annotation of any kind. I do not believe that anybody would object to the need to preserve extralinguistic and documental information in some structured way.

We may think of encoding as adding multiple layers of annotation. If a corpus of translations is meant to be a stable resource the optimal solution would be to be able to add different layers of text encoding as the need arises.

At a first level each electronic text may contain information about the text itself (meta-textual and extralinguistic information) and the text content, as plain text. A lot of fruitful investigations can be pursued using so called 'plain text corpora', by using information derived from word lists and keywords, by using collocational information and by conducting pattern searches for words or phrases with the help of a concordancer. However, other types of investigation may require further levels of encoding. This may regard both the language, as in complex pattern searches using Part-of-Speech (POS) tagging, or other features of texts not directly recoverable from lexical content.

A second layer of annotation relates to paratextual and structural information. Information about text structure concerns its subdivision into shorter units such as parts, chapters, paragraphs and sentences. As far as parallel corpora are concerned, aligning text pairs means, in fact, encoding segmentation units and creating bitextual correspondences between them. Alignment maps can be derived from parallel corpora with the aid of aligning software. Other elements which may be encoded are lists, headings and figures, as well as graphical features of the source printed text such as italics, bold or quotes. Finally, all paratextual apparatus, which may be especially important in the case of translations, could be identified as such. This would include features of translated text such as translator notes, glossaries and introductory remarks.

A third layer of annotation could be implemented through automatic or semi-automatic routines and taking advantage of research and tools in MAT and NLP. Some users may want to add linguistic information and cut up a text in shorter segments of language, up to every single word. They may want to attribute a label for part of speech to each word or analyse the syntactic structure of a text. In this case, a number of software tools are available for many languages, from POS taggers to syntactic parsers. Other features which could be encoded in a partly manual and partly automatic way include referring expressions (in order to study cohesion), names and direct discourse.

4.4 How to Encode?

In case anyone is attracted by the idea of using annotation to encode information into a text, the real question becomes *how* to do so. The next few sections are rather more technical than the preceding sections, which is necessary in order to explain how a translation corpus could be usefully encoded. The first answer to the *how* question is to adopt a standard. A viable solution for constructing stable, flexible and accessible corpus resources for translation research and for corpus-based descriptive translation studies in particular may be the adoption of an XML TEI encoding framework.

4.4.1 Standards

While motivations for adopting standards are very general in scope (the following arguments are adapted from XML manuals), they apply very nicely to the purposes of translation-driven corpora. The adoption of a standard makes the exchange of information more efficient, for instance, by eliminating the need to re-key information. It also allows networking, that is, the sharing of resources within a community which adopts the same standard. If each corpus is constructed with its own rules and software, there is a risk that it becomes dependent on a specific software configuration. Proprietary solutions tend to become obsolete quickly, while the adoption of a standard for encoding would help to prevent or at least delay corpus obsolescence. By adopting a common standard, corpus users may also end up with better software, and software writers would be allowed to compete by adding value on top of interoperable core technologies rather than on incompatible ones.

4.4.2 XML

XML means eXtensible Markup Language, and is a world-wide standard for distributed electronic texts. What XML does is to provide a common metalanguage by which electronic texts of all kinds can be stored, transmitted and displayed by different users, for example, on the internet. It does so by explicitly stating a set of instructions, which are enclosed in angle brackets in a format familiar to anybody who has seen the source of a web page. In what follows, I provide a very brief introduction to XML: for more detailed coverage of XML and related technologies, readers are referred to the chapter by Saturnino Luz (Chapter 5, this volume).

XML is a text-based annotation system, that is, data are stored in plain-text format and can be displayed in any text editor. As a consequence, it is freely available and accessible. It is extensible, that is, flexible enough to change as needs change. Because of its modularity it allows for the reduction or expansion of its building blocks. It is a cross-platform independent tool, meaning that it is designed to work on any kind of machine and operating system. And finally, it is becoming the standard mark-up language for internet applications. This aspect is crucial to the future of corpus-based translation studies: the possibility of working with online corpus resources could well increase the types and number of corpus-based studies.

According to a number of XML online tutorials, XML syntax is a piece of cake: the message to those who want to learn how XML works is 'The syntax rules of XML are very simple and very strict. The rules are very easy to learn, and very easy to use.' The message to software writers and designers is: 'Because of this, creating software that can read and manipulate XML is very easy to do' (http://www.w3schools.com/xml/xml_syntax.asp). A simple XML document may look like the one in Figure 4.3:

Meta-textual information is enclosed between angle brackets and is thus separated from textual content, which is here highlighted for better display. Each textual element is enclosed between a start and an end tag (with the same name as the start tag, but preceded by a slash sign), and all elements are arranged into a hierarchical structure nested inside a root element. In the first line we can see an *xml declaration* which identifies the xml document as such. Then we have two textual elements (title and paragraph) which are nested inside the <doc> element.

```
<?xml version="1.0"?>
<doc>
        <title> This is the title</title>
        <p>Text</p>
</doc>
```

Figure 4.3 Simple XML document

There are two possible types of XML documents: so-called *well-formed* XML documents are those which are conformant to XML syntax, as in the example above. XML *per se*, however, does not specify the content of the tags and the relationships between them. A *valid* XML document is instead defined as a well-formed XML document which conforms to a Document Type Definition (DTD), which is a set of instructions about how XML documents which belong to the same class (like texts in a corpus) should be interpreted by an XML application such as a web browser or a concordancer.

A DTD is common to a set of XML documents and makes them portable from one application to another. Often stored in a file external to the main XML document, the DTD states how a specific type of XML document is to be dealt with by a text processing software (this is the part where a language engineer is required).

XML applications can be quite complex, and a large number of DTDs have been written for such diverse applications as business-to-business and financial transactions, publishing and mathematics.

4.4.3 Text Encoding Initiative

The assumption is that any attempt at encoding a corpus of translations as a stable resource should be carried out in the XML framework. This would allow for general compatibility with text processing software, including concordancers and web applications. The question then becomes what type of XML document a text in a translation corpus should be. The obvious answer would be to look at existing standard document types and see if there is anything available to suit a minimalist approach as well as to support deeper encoding. The guidelines for text encoding suggested by the Text Encoding Initiative (TEI), an international consortium which has produced a *de facto* standard for scholarly work with electronic texts, are the most apparent candidate. TEI documents are XML conformant (*valid* XML documents), and the TEI guidelines are designed to allow for encoding a wide range of texts. The TEI website lists 115 projects which have adopted the TEI guidelines. These include literary studies, manuscript studies, dictionaries, language corpora and others. For example, within the corpus linguistics community, TEI guidelines have been implemented in the *British National Corpus* and in XCES, the XML Corpus Encoding Standard already mentioned.[5]

TEI offers the corpus encoder a set of predefined tags for document elements and structural relations among them, providing a framework for the annotation of structured information in a header containing meta-textual information and in the text itself. A simple TEI document could be represented as in Figure 4.4:

```
●  XML declaration
●  DTD declaration
         o  Header
         o  Text
```

Figure 4.4 A simple XML TEI document (simplified version)

```
<?xml version=”1.0”?>
<!DOCTYPE TEI.2 PUBLIC “-//TEI//DTD TEI Lite XML” “teixlilte.dtd”>
<TEI.2 >
         <teiHeader>
Here goes the header
         </teiHeader>
         <text>
Here goes the text
         </text>
</TEI.2>
```

Figure 4.5 Simple XML TEI document (full version)

After the XML declaration there follows a DTD declaration. A minimum of two elements are then required, a header (the electronic title page) and a text, each of which can contain further nested elements. A full version of such a document may look like Figure 4.5:

In this case the DTD is TEI Lite, that is a basic DTD which follows TEI guidelines giving specifications for most commonly used tags.

TEI has been opposed on the ground that it is too complicated and cumbersome to deal with by people interested in studying language or translation rather than in the workings of computers. In other words, the question is whether the possible advantages to be gained from annotating a text are outweighed by the time spent doing it.

Do researchers in translation studies need to be able to understand and write XML/TEI? To some extent, the answer is yes. That is, if one accepts that a basic understanding of XML may be a price to pay for the better processing and hence analysis of electronic texts, a price affordable by a corpus project wishing to create a stable resource for translation studies. Creating a corpus from printed texts involves a lengthy process of digitizing and proof-reading, as well as a laborious process of contacting authors and publishers to obtain copyright permissions. Annotation requires an additional effort, which can, however, be kept to a minimum but still allow for the adoption of a common standard. Only by making corpus resources public and agreeing on standards for their encoding will it be possible for more research to be replicated, findings to be cross-checked, and evidence to be accumulated.

While corpora designed for different purposes may require different degrees of encoding, all the texts composing any given corpus are likely to share a common layer of structural and meta-textual information. Adopting XML/TEI does not mean that all corpora have to be encoded to the same detail or using the same theoretical approach, and TEI seems to be flexible enough to ensure that minimal encoding can be accomplished without too much effort during corpus creation.

Rather than learning a new language in full, it might be enough for translation scholars to learn the ropes. Simple instructions may be followed during text preparation, for example to add structural annotation to a text while proofreading (see, for example, the guidelines for text preparation for the Portuguese–English parallel corpus *Compara*, Frankenberg-Garcia and Santos 2003).

A second way of relating to a computer language is resorting to digital translators, that is user-friendly software which mediates between the corpus compiler and the encoding system. For instance, a simple template with obligatory and optional fields may be devised to put structured information into a text header.

In principle, an XML document can be created with any word processor, but it is true that software specifically designed to create TEI documents is scarce. Indeed, it has been argued that the shortage of TEI conformant tools may be the major hindrance to its becoming a more widely adopted standard. A number of XML authoring tools are also available, as are corpus tools for linguistic annotation for those interested in this layer of annotation. XML is a relatively recent development, so it is hoped that more XML/TEI text processing software will be developed in the future.

4.5 Encoding *CEXI*

I would now like to discuss examples of XML/TEI conformant texts based on the encoding scheme devised for the *CEXI* project. The project envisaged the creation of a parallel bilingual and bidirectional Italian–English corpus containing a selection of text (sample) pairs from books published mostly between 1980 and 2000 (Bernardini 2002; Zanettin 2002a). The annotation scheme, which was tested on a few sample texts, involved the creation of a header and guidelines for basic text annotation.

4.5.1 The Header

The *CEXI* header was created following simple instructions from the TEI Lite manual (Burnard and Sperberg-McQueen 2002). A standard TEI Lite header contains four main parts or elements: **file description**, that is, a full bibliographic description of the electronic file, which includes a title statement, a publication statement and a source description element. The second

```
<fileDesc>
        <titleStmt>
                <title>My Son's Story: in machine-readable form</title>
                <author>Nadine Gordimer</author>
        </titleStmt>
        <sourceDesc>
                <bibl>
                        <author>Nadine Gordimer</author>
                        <title>My Son's Story</title>
                        <imprint>
                                <pubPlace>London</pubPlace>
                                <publisher>Penguin Books</publisher>
                                <date>1991</date>
                        </imprint>
                </bibl>
        </sourceDesc>
</fileDesc>
```

Figure 4.6 *CEXI* header: file description

part is called encoding description and documents the relationship between the electronic file and its source, as well as taxonomies derived from design specifications. The third part, profile description, contains information about the languages used in the text and a text class element which groups information about the nature of the text in relation to the taxonomies adopted. The last part, revision description, is used to document any editorial changes made to the file at different times by different people; it is the history of the electronic text.

The example in Figure 4.6 is taken from the file description part of the TEI Lite header for a CEXI text. The text, which was the first to be annotated, is the electronic version of Nadine Gordimer's *My Son's Story*. The structure of the file description should be self evident.

Figure 4.7 illustrates the encoding description part, and is taken from the same file. Descriptive labels for the CEXI classification schemes are shown in bold.[6] Texts can be either originals or translations, in English or in Italian, and belong to one of many categories within a broad distinction between imaginative and informative texts. Here only categories for the imaginative domain are shown.

Translated texts can then also be classified according to a taxonomy similar to that used for selecting sub-corpora from the Translational English Corpus (Baker 1999). The translation corpus taxonomy may include information about the translator (gender, status), the translated text, the source language(s), bibliographical information about the source text (title, author, date, etc.) and about the source text author.

The profile description contains information about the languages used and the status of the text according to the classification schemes adopted. As can be seen in Figure 4.8, the Italian translation of Nadine Gordimer's novel has words in four languages, is a translation into Italian of an imaginative work, and so on.

```
<encodingDesc>
<projectDesc>
<p>See the project description in the corpus header for information about
the CEXI project.</p>
</projectDesc>
<classDecl>
taxonomy id=“CEXIcodes”>
<category id=“source1”><catDesc>original</catDesc></category>
<category id=“source2”><catDesc>translation</catDesc></category>
<category id=“lang1”><catDesc>English</catDesc></category>
<category id=“lang2”><catDesc>Italian</catDesc></category>
<category id=“mdom1”><catDesc>imaginative</catDesc></category>
<category id=“mdom2”><catDesc>informative</catDesc></category>
<category id=“dom1”><catDesc>quality fiction</catDesc></category>
<category id=“dom2”><catDesc>detective fiction</catDesc></category>
<category id=“dom3”><catDesc>science fiction</catDesc></category>
<category id=“dom4”><catDesc>romantic fiction</catDesc></category>
<!-- other categories here -->
</taxonomy>
</classDecl>
</encodingDesc>
```

FIGURE 4.7 *CEXI* header: encoding description

```
•  language usage
         o  English
         o  French
         o  Afrikaans
         o  Italian

•  text class
         o  class code (Italian/imaginative/translation/quality fiction)
         o  class code (information on translator (Franca Cavagnoli),
            translation, source text and source text author)
```

FIGURE 4.8 Profile description information for Nadine Gordimer’s *Storia di mio figlio*

The last section of the header, the revision description part (not shown here), records editorial changes to the file.

4.5.2 The Text

The text itself can be annotated as follows (optional elements are shown in italics): within the main TEXT element there is an obligatory BODY element and optional front and end matter (introduction, glossaries, etc.). The body of a text can be further subdivided into one or more hierarchically lower levels, as in parts and chapters. Each major subdivision is marked by a DIV tag. Paragraphs and sentences can be marked-up semi-automatically, while part-of-speech tagging can be carried out automatically using appropriate tools.

```
<TEXT>
        <frontMatter>
        here goes the front matter
        </frontMatter>
        <BODY>
                <DIV1>
                        <DIV2>
                                <P>
                                        <S>
                                                <W>
                                                here goes a word
                                                </W>
                                        </S>
                                </P>
                        </DIV2>
                </DIV1>
                <DIV1>
                . . .  . . . .
                </DIV1>
        </BODY>
        <endMatter>
        here goes the end matter
        </endMatter>
</TEXT>
```

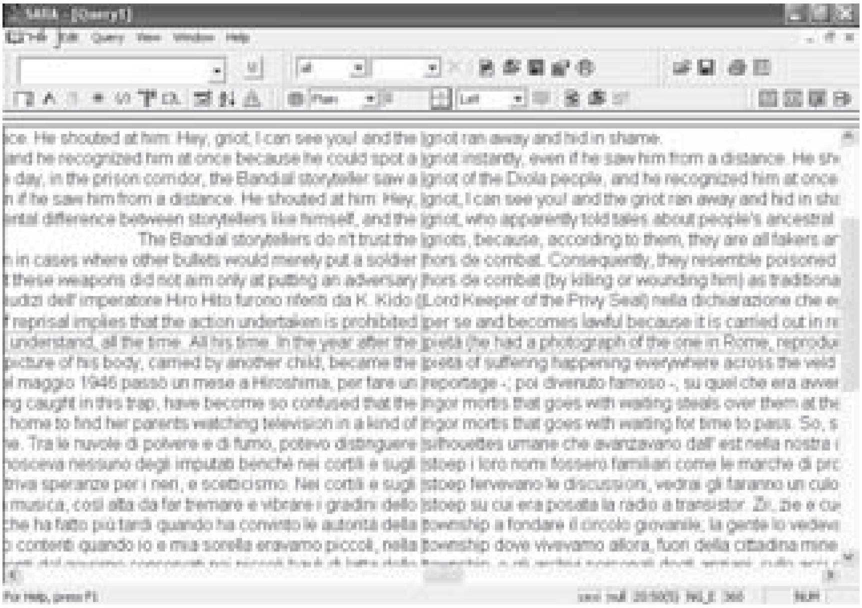
Figure 4.9 Text structure

Other types of textual annotation which can be routinely carried out as new texts added to a corpus are those relating to the annotation of direct speech (<q>quotation</q>), names (person, place, institution, other, e.g. <name type="pers"> Benoni </name>), and distancing and emphasizing devices such as quotation marks and italics. Such graphical features can be interpreted and assigned to one of various categories such as <foreign> (when highlighting is taken to indicate that the highlighted text is a foreign word or expression), or <soCalled> (when highlighting indicates the author or narrator disclaims responsibility for the word or expression used).

A number of text analysis tools were tested during the lifetime of the project. Wordsmith Tools was extremely useful for basic wordlists and statistics, and concordancing. A few text pairs were aligned (using different software) and ParaConc proved to be a very versatile tool for parallel concordancing. Some experimenting was also done with a preliminary version of Xaira, the XML version of SARA, the software originally developed for use with the BNC. Xaira is a generic purpose tool for searching XML (TEI) corpora such as CEXI, and can function both as a standalone programme accessing local corpus files or as a client accessing a remote corpus over the internet (Burnard and Dodds 2003).

4.5.3 Sample Searches

Xaira allows for the selection of sub-corpora using the information contained in the profile description of the header, so it is possible to create a sub-corpus



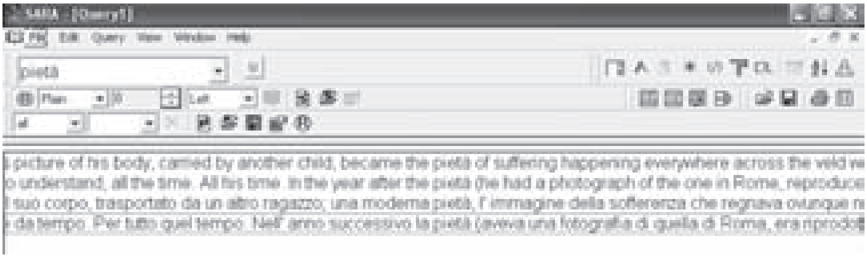
```
ce. He shouted at him: Hey, griot, I can see you! and the griot ran away and hid in shame.
and he recognized him at once because he could spot a griot instantly, even if he saw him from a distance. He sh
 day, in the prison corridor, the Bandial storyteller saw a griot of the Diola people, and he recognized him at once
n if he saw him from a distance. He shouted at him: Hey, griot, I can see you! and the griot ran away and hid in sha
ental difference between storytellers like himself, and the griot, who apparently told tales about people's ancestral
                  The Bandial storytellers do n't trust the griots, because, according to them, they are all fakers ar
n in cases where other bullets would merely put a soldier hors de combat. Consequently, they resemble poisoned
t these weapons did not aim only at putting an adversary hors de combat (by killing or wounding him) as traditiona
udizi dell' imperatore Hiro Hito furono riferiti da K. Kido (Lord Keeper of the Privy Seal) nella dichiarazione che e
f reprisal implies that the action undertaken is prohibited per se and becomes lawful because it is carried out in re
 understand, all the time. All his time. In the year after the pietà (he had a photograph of the one in Rome, reprodu
picture of his body, carried by another child, became the pietà of suffering happening everywhere across the veld
l maggio 1946 passò un mese a Hiroshima, per fare un reportage -; poi divenuto famoso -, su quel che era avver
ng caught in this trap, have become so confused that the rigor mortis that goes with waiting steals over them at the
 home to find her parents watching television in a kind of rigor mortis that goes with waiting for time to pass. So, s
e. Tra le nuvole di polvere e di fumo, potevo distinguere silhouettes umane che avanzavano dall' est nella nostra i
noscева nessuno degli imputati benchè nei cortili e sugli stoep i loro nomi fossero familiari come le marche di pro
triva speranze per i neri, e scetticismo. Nei cortili e sugli stoep fervevano le discussioni, vedrai gli faranno un culo
 musica, così alta da far tremare e vibrare i gradini dello stoep su cui era posata la radio a transistor. Zii, zie e cu
che ha fatto più tardi quando ha convinto le autorità della township a fondare il circolo giovanile; la gente lo vedeva
 contenti quando io e mia sorella eravamo piccoli, nella township dove vivevamo allora, fuori della cittadina mine
```

FIGURE 4.10 A search for 'foreign' words

containing only texts translated into Italian and belonging to the quality fiction category, or only texts translated by the same translator. A search can be performed using the information contained in the tags, for instance, by searching for names, foreign words or expressions, translator's notes, and so on. Information about the source of each concordance line is immediately available. The following examples (based on a sample of three full text pairs from the *CEXI* corpus) illustrate the types of investigations which can be performed using Xaira.

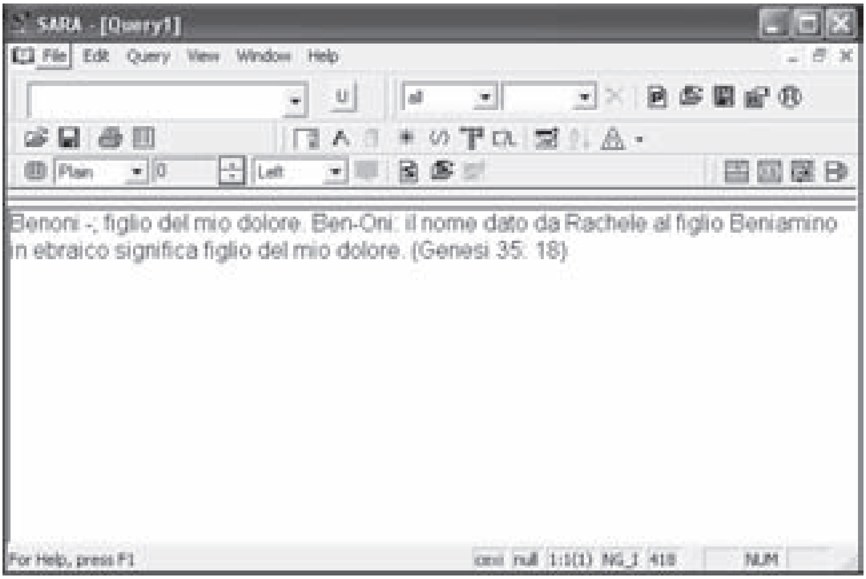
Figure 4.10 shows part of the output of a concordance for words which were in italics in the source texts and were labelled as foreign words during annotation. Most examples are from *My Son's Story*, both from the English original and the Italian translation. It can be noticed, for instance, that the words *griot* and *pietà* were rendered in italics (and tagged as foreign) in the original text but not in the Italian translation. In contrast, the South African English words *stoep* and *township* were foreignized in the Italian translation.

Looking further at a concordance of *pietà* (Figure 4.11), we can see how the translator resorted to an explicitation strategy to compensate the loss of information inherent in translating a word like *pietà*, which is also a general word meaning *pity*, into Italian.



picture of his body, carried by another child, became the pietà of suffering happening everywhere across the veld w
o understand, all the time. All his time. In the year after the pietà (he had a photograph of the one in Rome, reproduce
l suo corpo, trasportato da un altro ragazzo, una moderna pietà, l' immagine della sofferenza che regnava ovunque n
da tempo. Per tutto quel tempo. Nell' anno successivo la pietà (aveva una fotografia di quella di Roma, era riprodo

Figure 4.11 A search for pietà in My Son's Story/Storia di mio figlio



Benoni -; figlio del mio dolore. Ben-Oni: il nome dato da Rachele al figlio Beniamino in ebraico significa figlio del mio dolore. (Genesi 35: 18)

Figure 4.12 A search for footnotes in *Storia di mio figlio*

By translating *una moderna pietà* (a modern pietà) the reference to the statue by the Italian sculptor Michelangelo is re-established. Finally, Figure 4.12 shows the results of a search for translator's footnotes, which in this text produces only one hit, a note explaining the meaning of the name Benoni.

It is interesting to note that in resorting to this translation strategy the translator has explained something which was already explained in the text, both in the original and in the translation, that is, that Benoni means 'son of sorrow'/ *figlio del mio dolore* (literally of 'my sorrow').

4.6 Conclusion

To sum up, it can be said that all texts in electronic format are encoded, and that annotation is but one part of corpus construction. Annotation helps creating stable, flexible and accessible corpus resources for translation studies, and there seems to be very few choices between TEI/XML and not caring about encoding. The advantages of adopting the XML/TEI annotation framework can be recapitulated as follows:

- It is a standard
- It provides a sound scheme for annotating extralinguistic information
- It can be partially implemented during text preparation
- It can be partially automated
- It follows a modular approach, and it is possible to build on existing annotation, or ignore existing tags
- It allows for a wider variety of investigations than plain text
- It is the optimal solution for internet access to corpus resources.

Findings from corpus-based studies need to be based on cumulative evidence, and hence resources and procedures of analysis need to share some degree of comparability, notwithstanding differences in their original research purposes and design. Such resources also need to be made accessible to the wider research community, allowing for the replication of findings. In my view, corpus-based translation studies may benefit from implementing a consistent policy of corpus encoding and by adopting encoding standards, and this can be done with relatively little effort.

Notes

[1] See Francis and Kucera (1979) on the Brown corpus, Sinclair (1987) on the COBUILD corpus and Burnard (1995) on the British National Corpus.

[2] When a printed text is converted into electronic format it begins a life of its own, and the distinguishing characteristics of source material in a corpus may blur in the sea of bites, that is, unless they are explicitly encoded into texts. The danger for the researcher is to loose sight of the original context of production and reception of a text.

[3] Interest in the first area is evidenced by publications and conferences such as Véronis (2000) or the HLT-NAACL 2003 Workshop *Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 31 May 2003.

[4] ASCII stands for American Standard Code for Information Interchange, a character set finalized in 1968 which accounted for 128 characters in Latin script.

Unicode is the new standard which provides in principle for millions of characters and scripts.

[5] Language engineering standards such as XCES allow for applications in machine-assisted translation.

[6] Classification schemes and other information common to all texts in a corpus actually only need to be contained in a general corpus header, and are reproduced here only for explanatory purposes.

References

Aston, Guy (ed.) (2001) *Learning with Corpora*, Houston, TX: Athelstan.

Baker, Mona (1999) 'The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators', *International Journal of Corpus Linguistics*, 4(2): 1–18.

Barlow, Michael (2000) *MonoConc Pro 2.0. Software: Text analysis Program,* Houston, TX: Athelstan Publications.

— (2002) 'ParaConc: Concordance Software for Multilingual Parallel Corpora', in Elia Yuste Rodrigo (ed.) *Language Resources for Translation Work and Research* (CDROM), LREC 2002. Third International Conference on Language Resources and Evaluation. Proceedings, ws8.pdf, 20–4.

Bernardini, Silvia (2002) 'Educating Translators for the Challenge of the New Millennium: The Potential of Parallel Bi-Directional Corpora', in Belinda Maia, Johann Haller and Margherita Ulrych (eds) *Training the Language Services Provider for the New Millennium*, Porto: Faculdade de Letras da Universidade do Porto, 173–86.

Burnard, Lou (1995) *British National Corpus. Users Reference Guide. Version 1.0.* Available online at: http://www.cbs.dk/departments/vista/download/BNCreference.pdf

Burnard, Lou and C. M. Sperberg-McQueen (2002) *TEI Lite Manual.* Available online at: http://www.tei-c.org/Lite/

Burnard, Lou and Tony Dodd (2003) 'Xara: An XML-Aware Tool for Corpus Searching', in Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds), Proceedings of the Corpus Linguistics 2003 conference, *UCREL Technical Papers* 16, Special issue, 142–4.

Cencini, Marco and Guy Aston (2003) 'Resurrecting the Corp(us|se): Towards an Encoding Standard for Interpreting Data', in Giuliana Garzone and Maurizio Viezzi (eds) *Interpreting in the 21st Century*, Amsterdam and Philadelphia: John Benjamins, 47–62.

Fletcher, William (2004) 'Facilitating the Compilation and Dissemination of Ad-hoc Web Corpora', in Guy Aston, Silvia Bernardini and Dominic Stewart (eds) *Corpora and Language Learners*, Amsterdam: John Benjamins, 273–300.

Francis, W. N. and H. Kucera (1979) *Brown Corpus Manual.* Available online at: http://helmer.aksis.uib.no/icame/brown/bcm.html

Frankenberg-Garcia, Ana and Diana Santos (2003) 'Introducing COMPARA, the Portuguese-English Parallel Translation Corpus', in Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds) *Corpora in Translation Education*, Manchester : St. Jerome Publishing, 71–87.

Gatto, Maristella (2009) *From 'body' to 'web'. An Introduction to the Web as Corpus*, Bari: Laterza – University Press Online <http://www.universitypressonline.it>

Halverson, Sandra (1998) 'Translation Studies and Representative Corpora: Establishing Links between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study', *Meta* 43(4): 494–514.

Ide, Nancy, Paul Bonhomme and Laurent Romary (2000) 'XCES: An XML-Based Standard for Linguistic Corpora', in *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Athens, Greece, 825–30.

Ide, Nancy, Randi Reppen and Keith Suderman (2002) 'The American National Corpus: More Than the Web Can Provide', in *Proceedings of the Third Language Resources and Evaluation Conference* (LREC), Las Palmas, Canary Islands, Spain, 839–44.

Johansson, Stig (1995) 'Mens sana in corpore sano: On the Role of Corpora in Linguistic Research', *The European English Messenger* IV(2): 19–25.

Johns, Tim (1986) 'Microconcord: A Language-learner's Research Tool', *System* 14(2): 151–62.

Kettemann, Bernhard and Georg Marko (eds) (2002) *Teaching and Learning by Doing Corpus Analysis*, Amsterdam and New York: Rodopi.

Kilgarriff, Adam (2001) 'Web as Corpus', in Paul Rayson, Andrew Wilson, Tony McEnery, Andrew Hardie and Shereen Khoja (eds) *Proceedings of the Corpus Linguistics 2001 conference, UCREL Technical Papers: 13,* Lancaster University, 342–4.

— (2010) 'Corpora by Web Services', in *Workshop on Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation (LREC 2010)*, 45–51.

Leech, Geoffrey (1992) 'Corpora and Theories of linguistic Performance', in Jan Svartvik (ed.) *Directions in Corpus Linguistics*, Berlin: Mouton, 105–22.

Leech, Geoffrey, Greg Myers and Jenny Thomas (1995) *Spoken English on Computer: Transcription, Mark-up and Application*, London: Longman.

OLAC: Open Language Archives Community. Available online at: http://www.language-archives.org/

Rossini Favretti, Rema, Fabio Tamburini and Cristiana De Santis (2002). 'A Corpus of Written Italian: A Defined and a Dynamic Model', in Andrew Wilson, Paul Rayson and Tony McEnery (eds) *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Munich: Lincom-Europa.

Scott, Mike (1996) *WordSmith Tools*, Oxford: Oxford University Press.

Setton, Robert (2003) 'A Methodology for the Analysis of interpretation Corpora', in Giuliana Garzone and Maurizio Viezzi (eds) *Interpreting in the 21st Century*, Amsterdam and Philadelphia: John Benjamins, 29–45.

Sinclair, John M. (ed.) (1987) *Looking Up. An account of the COBUILD Project in Lexical Computing*, London: Collins.

Stewart, Dominic (2000) 'Conventionality, Creativity and Translated Text: The Implications of Electronic Corpora in Translation', in Maeve Olohan (ed.) *Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects*, Manchester: St. Jerome, 73–91.

TMX: Translation Memory Exchange. Available online at: http://www.lisa.org/tmx/

Tognini-Bonelli, Elena (2001) *Corpus Linguistics at Work,* Amsterdam and Philadelphia: John Benjamins.

Toury, Gideon (1982) 'A Rationale for Descriptive Translation Studies', in André Lefevere and Kenneth David Jackson (eds) *The Art and Science of Translation, Dispositio* 7: 22–39.

Vanderauwera, Rita (1985) *Dutch Novels Translated into English: The Transformation of a Minority Literature*, Amsterdam: Rodopi.

Véronis, Jean (ed.) (2000) *Parallel Text Processing: Alignment and Use of Translation Corpora,* Amsterdam: Kluwer Academic Publishers.

Wools, David (1995) *Multiconcord: The Lingua Multilingual Parallel Concordancer for Windows*, CFL Software Development.

Zanettin, Federico (2002a) 'CEXI. Designing an English Italian Translational Corpus', in Ketteman, Bernhard and Georg Marko (eds) *Teaching and Learning by Doing Corpus Analysis*, Amsterdam: Rodopi, 329–43.

— (2002b) '*DIY Corpora: The WWW and the Translator*', in Belinda Maia, Jonathan Haller and Margherita Urlrych (eds) *Training the Language Services Provider for the New Millennium,* Porto: Facultade de Letras, Universidade do Porto, 239–48.

Zanettin, Federico, Silvia Bernardini and Dominic Stewart (eds) (2003) *Corpora in Translator Education*, Manchester: St. Jerome Publishing.

Chapter 5

Web-Based Corpus Software

Saturnino Luz

5.1 Introduction

What is a web-based corpus and what is web-based corpus software? The answer is, strictly speaking, that there is no such thing as *web-based corpus software.* However, one should not be discouraged by this rather negative assessment. In fact, if one examines the title closely, different bracketings of the phrase might suggest interesting possibilities. For example, if one chooses to write it as '(web-based corpus) software', the emphasis falls on the idea of the World Wide Web as a large corpus. It is, however, a very chaotic one. It is chaotic in the sense that it is difficult for its users to account for and control the sort of phenomena such a large and dynamic repository might reflect when explored, say, through an ordinary search engine. This makes the task of formulating and testing hypotheses extremely difficult. All sorts of 'noise' might creep in: there are texts written by native and non-native writers, computer-generated text (e.g. text resulting from the ubiquitous web-page translation services currently on offer), duplication, and other forms of text which do not conform to standard norms. Little, if anything, can be done to guarantee the quality or integrity of the data being used. Still, this chaotic, noisy environment can be of some use to the statistically minded (computational) linguist. To borrow an example from Manning and Schütze (1999), one could use the web to decide which of the following word sequences to treat as a language unit: 'strong coffee' or 'powerful coffee'. A quick search reveals over 30,000 occurrences of 'strong coffee' against just over 400 occurrences of 'powerful coffee', thus indicating that the former forms a collocation pattern while the latter apparently does not.

In contrast, should one wish to write 'web-based corpus software' as 'web-based (corpus software)', the emphasis clearly falls on 'corpus software', of which web-based corpus software would simply be one type. In other words, one could simply regard the Web as the medium through which better constructed, human-designed corpora can be searched and studied by a large community of geographically dispersed researchers. Many tools have undoubtedly been designed which do just that, including the Translational English Corpus (TEC) system (Luz and Baker 2000). The main shortcoming of this approach stems from the very fact that it is better defined. One of the attractive aspects of the World Wide Web is that all stages of information exchange are distributed. That is, in principle, anyone can provide and access any data. Information consumers benefit from a simple and intuitive access model (hypertext) and the widespread availability of web-browsing software on virtually all platforms. More importantly, all an information consumer who wishes to become an information provider needs to do is learn a few idioms of a simple mark-up language (HTML). By using more specialized tools, such as corpus servers and clients, this flexibility is lost. And with flexibility goes the dream of a massive (and therefore statistically attractive), self-maintained corpus.

This chapter presents and discusses recent advances in tools, technologies and standards which might enable the corpus research community to bring about the basics of an infrastructure for creating and sharing distributed, dynamic and widely accessible corpora. The end result may still be a far cry from the dream of using the entire web as an evolving corpus, but it will certainly advance the idea that implementing a world-wide web of corpora is feasible, thus rendering moot the ambiguity in the title of this chapter. Since the above-mentioned infrastructure does not exist at present, we will necessarily have to focus on corpus software and tools that use the Web (or, more appropriately, the internet) as a communication medium. We start by presenting an overview of the technology involved, describe the model adopted by the TEC system to support remote access to a corpus of translated English, and finally discuss perspectives for future research in this area.

This overview of web technologies covers the main tools for data storage and mark-up, text indexing and retrieval, and the issue of distributed storage of corpora. It also addresses, though superficially, the issue of creating and maintaining metadata, including storage and database management. The aspects of text indexing and retrieval covered include tokenization, data structures for indices and search. Ways of moving from storage of an individual corpus to distributed storage of collections of corpora, and the non-technical issue this entails, namely, copyright, are also discussed. These issues are illustrated through a case study: the Translational English Corpus (TEC) system. We end the chapter by presenting a vision for web-based corpus software: its overall architecture and development philosophy.

5.2 The Internet and Corpus Software

The internet technologies presented and discussed below have not been developed specifically to deal with corpora, or even text, though text comprises the vast majority of data available on the internet. In what follows, we describe a selection of those technologies which have been found to be particularly useful for text storage and retrieval in the Translational English Corpus (TEC) project, namely: mark-up languages, indexing techniques and the client-server architecture. This represents a large array of technologies which are covered only partially and superficially, as the contribution aims to give the reader a perspective of the bigger picture and the possibilities it suggests for corpus research, rather than its details. References to the relevant literature are included in each section.

The TEC corpus is a collection of English texts translated from a wide variety of languages, both European and non-European, and is held at the University of Manchester (Baker 1999). It consists of four sub-corpora: fiction, biography, news and in-flight magazines. The corpus is an expanding collection which contains more than ten million indexed words, and 'headers' which store a variety of extralinguistic data about each text in the collection (known as *metadata*).

Although the individual texts that make up TEC are not directly available in full, researchers can access them over the internet, and run different types of analyses on the corpus or selected subsets of it through a set of software tools available from the TEC website. These tools support standard corpus analysis functionality, such as concordancing, as well as functionality specifically designed to allow translation researchers to explore different hypotheses by constraining search parameters derived from metadata. Possible search modes include selection by source language, by translator, by author, gender, nationality, and so on. For examples of how this functionality can be used and a discussion of its significance for translation studies see Baker (1999).

As a stochastic entity, the usefulness of a corpus is invariably dependent on its size and composition. Large amounts of text are needed if a corpus is to significantly reflect general language phenomena. For the sort of investigation carried out in corpus-based translation studies, where the corpus user is interested in discovering patterns inherent to a particular type of language, the constraints and mechanisms these patterns might reflect and so on, careful attention must be paid to text selection and documentation of extralinguistic features that might have a bearing on the sort of phenomena of which the corpus is supposed to provide samples. Together with legal constraints such as the need to protect copyright, corpus size and composition concerns dictate the main requirements for the software infrastructure: it must provide means for efficient physical storage and retrieval of data, and a loosely coupled, but secure and fine-grained database layout. The first step towards meeting these requirements is to supply the corpus designer with a simple and flexible way to annotate the data and encode metadata for further processing by other software modules.

5.3 A Brief Introduction to XML

A mark-up language is a collection of mechanisms that allow a corpus designer to annotate data with various forms of meta-information in a standardized and portable way. Probably the best known example of this kind of device is the hypertext mark-up language (HTML), which allows its users to specify text formatting and layout information. However, formatting mark-up languages such as HTML do not suffice for corpus annotation. In fact, different corpora designed by different people are likely to require different sets of tags, depending on the sort of metadata their user communities require. What is needed in addition to a standard syntax for annotation is a way of specifying the vocabulary (i.e. tags), syntax and constraints of the annotation scheme itself. The standard generalized mark-up language (SGML) was the product of the first comprehensive attempt to meet this need. SGML defines a set of simple, basic conventions (syntax) that should apply universally to all SGML-annotated documents,[1] and mechanisms for constraining this basic syntax so that it can be used in different contexts. Particular ways of constraining SGML syntax are known as SGML *applications*. HTML itself is an SGML application. Although the primary goal of SGML was standardization of document storage, its development was also greatly influenced by the goals of SGML application writers, namely: flexibility and user-friendliness. The result was a complex formalism described in over 150 pages of an ISO standard (ISO8879, 1986). That complexity ended up working against the primary goal of user-friendliness. The success of a data storage standard depends on it being well supported by software libraries and tools. In the case of SGML, library and application developers often chose to support only those parts of the standard they found essential or useful for certain applications, thus effectively undermining document portability.

The extensible mark-up language, XML, originated from an attempt to remedy this situation. Its designers started from SGML and simplified it by removing a number of exotic, rarely used features. XML is simple and flexible in that it has no pre-defined set of tags. It provides a uniform means of encoding metadata which is easily interpretable by humans as well as machines. It has been widely adopted by industry and academia, and its development has been coordinated by the World-wide Web Consortium (Bray et al. 2006). A large number of software libraries and tools exist which support all aspects of XML parsing and validation in most programming languages.

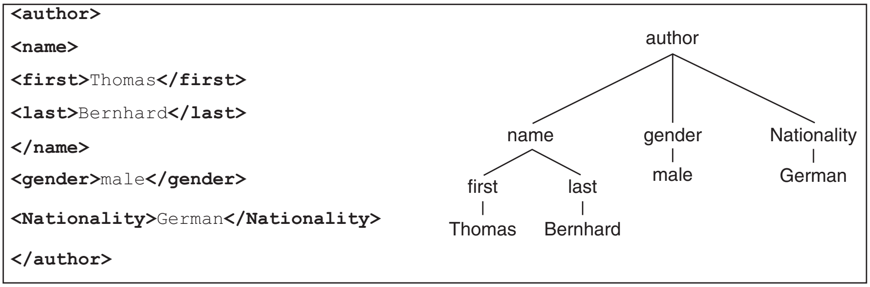
XML documents can be encoded in plain text, or in a variety of (language) encoding standards. The default encoding is UTF-8, a Unicode standard that supports virtually every character and ideograph from the world's languages. As with SGML, there is a basic (built-in) syntax which defines which documents (or text fragments) are *well formed*. This basic syntax can be constrained in a variety of ways, defining which documents are *valid*. Analogously to SGML, a specific way of constraining basic XML syntax yields what is called an *XML application*. From a parsing perspective, *well-formedness* and *validity* are the most important concepts in XML. XML also provides mechanisms for incorporating a form of semantics for XML documents with respect to XML applications. In the following sections, we examine each of these aspects in some detail.

5.3.1 Well-Formed XML

The basic syntax to which all XML documents must conform consists of a small set of simple rules governing essentially: the placement of tags, which strings are allowed as element names, and how attributes should be attached to elements. XML tags are text strings enclosed by angle brackets. Figure 5.1 shows two XML tags for author data: a begin tag and an end tag. This is a typical pattern in XML documents: annotated text appears between a begin tag and an end tag. Begin tags start with <. End tags start with </. In general, XML syntax is similar to HTML syntax. However, the following exceptions should be noted: in XML one is allowed to “invent” one’s own tag names, start tags must always be matched by end tags unless they are *empty elements* (as in *<br/>*), tags are case-sensitive (i.e. *<author>* is not the same as *<Author>*), and tag overlapping is not allowed.

XML documents can be represented as tree structures. Each XML document must have (and each XML application must define) a unique root element of which all other elements are descendants. Figure 5.1 shows a well-formed XML document and its corresponding syntax tree.

The kind of XML encoding shown in Figure 5.1 is called data oriented. It resembles, in a way, the sort of structure one would find in a database management system. XML is obviously also appropriate for annotation of ordinary text, in which the structure is implicit, or more loosely encoded. This encoding style is often referred to as narrative-organized. The distinction, however,



```
<author>
<name>
<first>Thomas</first>
<last>Bernhard</last>
</name>
<gender>male</gender>
<Nationality>German</Nationality>
</author>
```

Figure 5.1 A simple XML document represented as a tree

```
<author>
The <Nationality>Irish</Nationality> writer <name> <first>Samuel</first>
<last>Beckett</last> </name> was born in <city>Dublin</city> in
<date><month>April</month> <year>1906</year></date>.
</author>
```

Figure 5.2 Narrative-organized XML fragment

is purely pragmatic. Note the presence, in Figure 5.2, of data-oriented elements (data and name) within the narrative-organized author element. In TEC, for instance, translated texts are encoded in a narrative-organized form, while the metadata encoding of its header files is data oriented.

5.3.2 Basic Syntax Rules

Angle bracket delimited tags such as the ones shown above form the basic units of XML markup. Pairs of tags and their contents are known as XML elements. XML elements can also have attributes. Attributes are name-value pairs attached to a begin tag, as shown in Example (1). An attribute name is separated from its values by an equals sign. Values must always be enclosed in single or double quotes.

(1) <author nationality=‘Irish’ sex=‘male’> Samuel Beckett </author>

The lexicon of an annotation scheme is the set of tokens which name its elements and attributes. In XML, the form these names can take is precisely defined by the following set of rules:

- Names may contain any alphanumeric characters (and also non-Latin characters and ideograms) as well as the following punctuation characters: ‘_’, ‘-‘ and ‘.’.
- White spaces and other punctuation characters are not allowed
- Names may start only with letters (or ideograms), or the underscore character.

Because certain characters play special roles in XML documents, their print format needs to be escaped in certain contexts to avoid ambiguity. The ‘less-than’ symbol (<), for instance, is always interpreted as a marker that an XML tag is about to appear. It must always be matched by a ‘greater-than’ character (>). Therefore, a fragment such as *<maths> 2 < x </maths>* is not well-formed XML. In such contexts, one needs to replace the < character by a special symbol sequence. The symbol sequence &lt; serves this purpose. This syntax (& . . . ;) allows XML users and document designers to specify any symbol. Symbols so defined are called entity references. The logical consequence of using the ampersand character to signal the beginning of an entity reference is that the ampersand symbol itself will need to be escaped. Therefore we have *&amp;* as the entity reference corresponding to the ampersand character.

In addition to tags, angle brackets are used in XML to encode comments and processing instructions. Comments can be added to improve the legibility of the document, and are simply ignored by applications. Comments start with a *<!–* and end with a –> sequence. Example (2) shows an XML comment. Comments are not allowed inside other comments or inside tags.

(2) <!– this is a comment ––>
(3) <?robots index=‘yes’ follow=‘no’?>

XML documents can pass processing instructions to other applications by enclosing them in tags of the following format: *<? . . . ?>*. An example of processing instruction is shown in Example (3), which is actually used in websites to suggest appropriate behaviour to indexing ‘software robots’ like the ones used by most search engines.

Although comments and processing instructions are markup, they are not XML elements. Therefore they can appear anywhere in the document. An exception to this rule is the XML declaration, a processing instruction which declares a number of facts about the document for use by parsers and other applications. An XML declaration is not a compulsory part of an XML document. However, if it is added, it must be the first string to appear in the document. Common attributes of XML declarations include: version, for backward and forward compatibility purposes, encoding (if none specified, UTF-8 is assumed), and standalone, which tells the parser whether to look for an external document definition. A typical XML declaration is shown below:

(4) <?xml version=“1.0” encoding=“ISO-8859–1” standalone=“yes”?>

5.3.3 An Online XML Parser

Readers willing to follow a tutorial introduction to XML will find a set of resources at http://ronaldo.cs.tcd.ie/tec/CTS_SouthAfrica03/. One of the resources available at the site is a simple, online XML parser. In order to use it, one needs a web browser and Java Web Start,[2] which is available with all recent versions of the Java Run-time Environment. The data needed for all XML-related exercises can be accessed through the website or downloaded directly in an archive file.[3] Exercises covering well-formedness of XML documents can be found in data/01/ (a directory created by decompressing the archive file). Descriptions of the exercises can be found in data/01/README.txt.

5.3.4 Document Type Definitions and XML Validation

In addition to the basic conventions described above, XML provides mechanisms to enable the syntax of an annotation scheme to be fully specified. In TEC, for instance, two schemes have been defined: *techeader*, for encoding data about the corpus, including information about translations, translators, authors and so on, and *tectext*, which supports light annotation of the corpus itself in narrative-oriented style. Figure 5.3 shows a fragment of an actual TEC header file.

An XML application such as *tectext* or *techeader* can be defined through document type definitions (DTD). DTDs provide a formalism that allows document designers to specify precisely which elements and entities may to appear in a document, their hierarchical relationships, and their admissible contents and attributes. The following are examples of constraints one can specify through DTDs: ‘a book must have an author’, ‘an author must have a first and a last name, and optionally a middle name’, and so on.

Consider the DTD shown in Figure 5.4, for instance. Each line defines a valid XML element. The string *!ELEMENT* is a reserved word. It is followed by an element name (e.g. *book*) and a specification of the element’s admissible children in the XML tree. Children elements can simply be parsed non-annotated data (including entity references) or other XML elements. The former is specified through the reserved word *#PCDATA*. In Figure 5.4, title has book as the parent element, and parsed data as children. Children can be specified as sequences or choices, or combinations of these. Sequences are comma-separated lists of element names (and possibly *#PCDATA*), as in (*title,author+,publisher?,*

```
<?xml version=“1.0” encoding=“ISO-8859-1” standalone=“no”?>
<!DOCTYPE techeader SYSTEM “techeader.dtd”>
<techeader>
<title subcorpusid=“fiction” filename=“fn000001.txt”>
<subcorpus>fiction</subcorpus>
<collection>Restless Nights.</collection>
</title>
<section id=“fn000001.000”>
<translator sexualOrientation=“heterosexual” gender=“male”>
<name>Lawrence Venuti</name>
<nationality description=“American”></nationality>
<employment>Lecturer</employment>
</translator>
. . .
</section>
</techeader>
```

Figure 5.3 TEC header file fragment using techeader.dtd

```
<!ELEMENT book (title,author+,publisher?,note*)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT author (fname,mname?,lname)>
<!ATTLIST author sex (male|female) #REQUIRED>
<!ELEMENT publisher ANY>
<!ELEMENT fname (#PCDATA)>
<!ELEMENT mname (#PCDATA)>
<!ELEMENT lname (#PCDATA)>
<!ELEMENT note (#PCDATA|p)>
     <!ELEMENT p EMPTY>
```

Figure 5.4 A simple DTD

note), where title, author, publisher and note must appear in the order specified by the declaration. Choices are represented by lists of elements separated by vertical bars, such as *(#PCDATA|p)*, which specifies that parsed data or <p/> tags (and possibly both, in any order) can appear as children. The number of children allowed is determined by special suffixes appended to element names in the DTD: ‘?’ for zero or one occurrence, ‘*’ for zero or more occurrences, and ‘+’ for one or more occurrences. In the example below title must appear first, followed by one or more *author* elements, possibly followed by a *publisher* and so on. One can also combine constraints by using parentheses. An element name followed by XML reserved word ANY indicates that any (possibly mixed) content may appear as text tagged by this element.

Among the declarations in Figure 5.4 is one which specifies the allowable values for the *sex* attribute of element *author*. This declaration is identified by an *ATTLIST* token. The element author requires an attribute *sex*, which may take value ‘*male*’ or ‘*female*’, but no other value. Supplying another value would cause a validating parser to signal an error. Attributes appear inside element tags. The following is a sample *tectext* tag which tells the user why a certain text fragment should be ignored by the indexer: *<omit desc=‘picture’/>*. Figure 5.5 shows how this tag can be declared in a DTD.

An advantage of using attributes is that they allow contents to be more tightly constrained. The type of data acceptable as attribute values can be specified through certain reserved words. DTDs also allow document designers to specify default values for attribute slots or restrictions on how such slots should be handled (e.g. whether values are required or optional, fixed, defined as literals, etc.).

The user tells an XML parser how to validate a document by declaring the document to be of a certain type, as defined by a DTD. This is done through the *DOCTYPE* declaration, which provides a link between a document and its syntax specification. TEC headers, for instance, declare their DTD as follows: *<! DOCTYPE techeader SYSTEM ‘techeader.dtd’>*. The word *techeader* indicates the document’s root element, and the string following reserved word *SYSTEM* indicates the location of the DTD. This could also be a fully specified URL or,

<!ATTLIST omit	desc	CDATA	#REQUIRED>
↓	↓	↓	↓
declaration tag and element name	attribute name	data type	default

Figure 5.5 DTD declaration of a desc attribute for a tectext omit element

in fact, any URI. An alternative to *SYSTEM* is *PUBLIC*, which is used for publicly recognized DTDs, such as the DTD for the XML version of HTML.[4]

Despite the fact that they provide a simple and flexible mechanism for defining XML document syntax, DTDs have a number of limitations. First of all, there is a question of consistency. Although similar, DTD syntax isn't really XML. Some consider this to be a serious drawback. In addition, many applications require more expressive ways of constraining document syntax than DTDs can provide. This prompted the development of competing XML document specification schemes. The most widely used alternative to DTDs is XML Schema, a formalism developed by the World Wide Web Consortium whose syntax is itself XML-compliant.[5]

5.4 Adding Some *Semantics* to an Annotation Scheme

An attractive aspect of using XML for corpus annotation is that, once annotated, documents can be viewed from a variety of perspectives ranging from application needs to user tasks to corpus maintainer goals. Checking a text for well-formedness helps prevent typos and other simple annotation errors. Validating helps ensure consistency across the collection. Once these basic requirements are met, the user or maintainer can interpret the annotated text and benefit from the corpus designer's work in many ways. Although simple visual inspection of raw XML might provide the user with greater understanding of the data, the main benefit of a consistently annotated XML document derives from its amenability to processing by computers. Post-processing of annotated text may facilitate different levels of analysis. Annotation could be used, for instance, to control what should be indexed[6] or displayed, how the text should be laid out for presentation on a web browser or formatted for printing and so on. Markup adds structure to text and ultimately enables applications to assign meaning to that structure. A general processing architecture for annotated data is shown in Figure 5.6. First the original is annotated, then the resulting document is checked for well-formedness and validity, and finally the valid document is transformed for presentation, storage in a database,

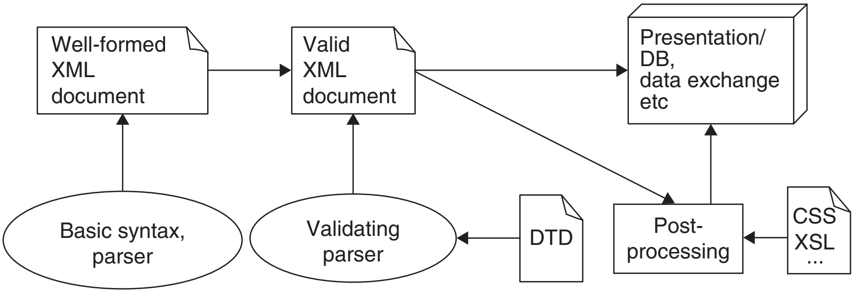


Figure 5.6 Overall XML processing architecture

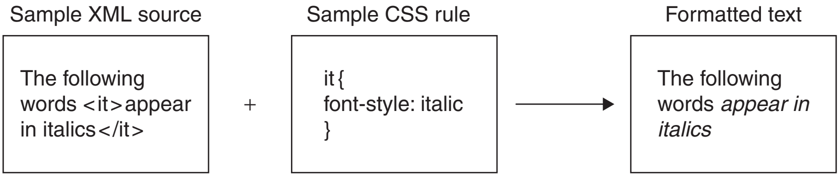
data visualization, or whatever other application need the annotation scheme happens to support.

The semantics of a document structure can be defined in numerous ways. Some of them are quite ad hoc, such as the way TEC handles storage of metadata into an efficient database system in order to speed up retrieval within subcorpora. Others, however, have been standardized, as they range over a large set of applications. These include formatting instructions, which prompted the development of CSS, a language for defining Cascading Style Sheets (Bos et al. 2009), as well as more powerful formalisms for specifying general document transformations such as the Extensible Stylesheet Language (XSL) and its formatting counterparts (Clark 1999).

CSS is a simple and widely used language for formatting of XML-annotated text. CSS can also be used in conjunction with HTML documents, being supported by nearly all modern web browsers. Figure 5.7 illustrates how a CSS can be used in conjunction with an XML document to produce a formatting effect.

A CSS file consists of a list of elements followed by a list of style specifications to be applied to these elements. The most commonly defined properties include the display attribute, which specifies the way an element should be laid out on the main document (as block, inline, list-item, table, or none), how elements are formatted inside a table (inline-table, table-row, table-column, table-cell etc.), and general lengths, such as font-size, border-width, width and height. There are many other properties (e.g. font style and colours) whose enumeration is outside the scope of this chapter.

In order to link an XML file with a given CSS one uses a specific processing instruction of the form *<?xml-stylesheet type='text/css' href='mystyle.css'?>*, where the value of the *href* could have been any valid URI. If viewed on a CSS-compliant web browser, an XML text containing the above processing instruction will be formatted according to the rules specified in *mystyle.css*.



```
The following
words <it>appear
in italics</it>
```

```
it{
font-style: italic
}
```

FIGURE 5.7 XML fragment formatted as specified by CSS rules

In corpus applications, CSS can be used for formatting concordance lines (i.e. with keywords aligned at the centre of the screen and contexts flushed left and right), for example. The tutorial website contains an exercise on creating a CSS file for formatting XML-annotated concordance lines.

XML semantics does not necessarily need to be implemented through style sheets or any standard language, for that matter. A common form of post-processing of XML documents is storage and retrieval of elements in a database system. Since XML itself is not meant to be a database system, application-specific code is often used to load selected elements into a database management system (DBMS) for efficient access. The TEC system uses a native XML DBMS (Meier 2002) to store its metadata.

Should one wish to learn more about XML, there are many books on XML, ranging from short introductions to XML itself to comprehensive guides to XML-related technologies. *XML in a Nutshell* (Harold and Means 2004) is a good introductory book which manages to contain a reference guide to the main XML-related technologies. Less readable but more detailed are the official W3C technical reports (Clark 1999; Clark and DeRose 1999; Bray et al. 2006; Bos et al. 2009) which can be used as concise reference texts once one has mastered the basics of XML.

5.5 Text Indexing and Retrieval

In addition to markup standards, corpus software in general needs to be able to perform basic text indexing and retrieval functions. This section provides an introduction to the main issues in indexing and retrieval. As before, the contribution is set against a TEC backdrop, so the focus will be on the choices made in the implementation of the TEC system.

Text retrieval involves four of basic operations: tokenization, indexing, compression and search. Tokenization, or lexical analysis consists of deciding what counts as a token (i.e. selecting strings for indexing). Indexing consists of storing information about where those tokens can be found. Compression aims at keeping indices within acceptable space bounds. Search is the basic operation on which corpus analysis tools such as concordancing and collocation are built. The first three operations are often performed offline, in a single step, while search is usually performed online.

5.5.1 Tokenization

As mentioned above, tokenization is the process of converting a stream of characters into a stream of tokens (or words). At a higher level of abstraction one finds what is known as lexical analysis. The key question in tokenization is: what counts as a word delimiter? In English, for instance, the best candidates are blanks (including tabulation spaces, line breaks, etc.), hyphens and punctuation characters. One might, in principle, define a token as a string of characters delimited by those characters. In certain cases, however, this definition is inadequate. The hyphens in *state-of-the-art* separate out legitimate tokens, whereas the hyphen in *B-52* does not. Punctuation characters can be equally deceptive, as in *360 B.C.* A related issue is how to deal with capital letters. Should *White* as in *The White House* be indexed separately from *white* as in *white cells*? Sometimes capitalization provides valuable hints about collocation. The TEC system adopts the approach of preserving as much information as possible: hyphenated words are indexed both as individual words and as compounds, superficial lexical analysis rules out tokens such as *B* in the examples above, and case is preserved through the use of two separate indices (case-sensitive search is assumed by default). Figure 5.8 shows the overall layout of the inverted index used in TEC.

Tokenization is an essential phase of corpus processing and involves reading through the entire string of text and testing for pattern matches. Fortunately, the computational cost of performing tokenization is relatively low. Finite-state automata (FSA) can tokenize in linear time. Regular expressions provide convenient syntax for token matching which can be converted into suitable FSA. Regular expressions are well supported in modern programming languages. Interpreted languages such as Perl and Python provide built-in support for regular expression matching. Other languages such as Java, C, and C++ provide a variety of libraries for dealing with pattern matching.

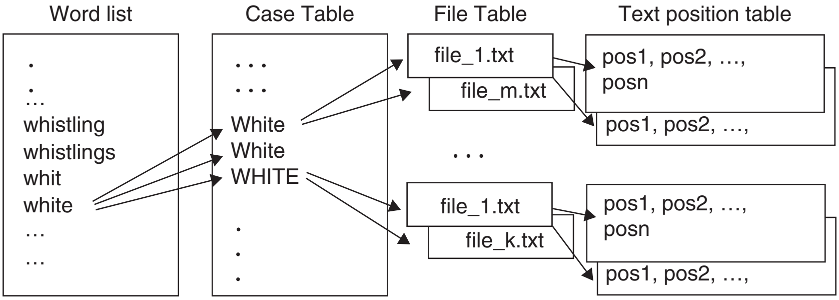


FIGURE 5.8 Structure of case-preserving inverted indices in TEC

A regular expression is a pattern that denotes a (possibly infinite) class of strings to match. Basic regular expression syntax is very simple. Its atomic symbols are: e, which matches any character, and ε, which matches the empty string. In addition to ordinary atomic symbols, regular expressions can contain suffix operators, binary operators and parentheses. Suffix operators quantify over atomic symbols much in the same way as suffix operators quantify over elements in DTDs (see Section 5.3.4). The suffix operator * in e* denotes a sequence of zero or more characters e. All the other suffix operators can be defined in terms of * (plus atomic symbols and binary operators). Binary operators | and, are also analogous to their counterparts in DTDs. Regular expression $e1|e2$ matches character $e1$ or character $e2$, while the expression $e1,e2$ matches character $e1$ immediately followed by character $e2$ (and is usually abbreviated as $e1\ e2$). Parentheses are used for grouping sub-expressions.

The following simplification might help illustrate the use of regular expressions. Imagine a language whose alphabet consists entirely of two letters: a and b, and in which white spaces are word separators. A regular expression for tokenizing texts written in this language could look like this: (a|b)(a|b)*_.

5.5.2 Indexing

In small corpora, it might be practical to use simple string or even regular expression matching to determine the exact occurrences of a query, typically a keyword. Although matching is reasonably fast, it is still not fast enough for use in real-time with larger corpora. Even for medium-sized corpora such as TEC one needs to use pre-compiled indices in order to attain an acceptable search speed. Pre-compiling an index simply means matching all possible keywords in advance (offline) and storing the results into a data structure that can be searched more efficiently online.

Text indexing techniques have been extensively used in the area of Information Retrieval (IR). IR techniques can be straightforwardly adapted for use in corpus processing, and there are many indexing techniques on offer. Each has advantages and disadvantages. As with most applications, one's choice of indexing technique for corpus processing will depend on the size of the database, search speed constraints, and availability of storage space. In what follows, we focus on our choice for the TEC corpus, inverted indices, and describe it in detail. However, we also say a few words about other common indexing strategies, with a view to contextualize this choice, and refer the interested reader to Baeza-Yates and Ribeiro-Neto (1999) for further details.

Generally speaking, indexing strategies differ according to the data structures they use. Inverted indices (also known as inverted files) are tables whose keys are words and whose values are lists of pointers to all positions in which the key occurs in the corpus. In suffix tree and suffix array indexing, each position in the text is considered a suffix. This facilitates search for phrases and longer text fragments as well as word prefixes. Suffix arrays and trees allow fast search but have heavy storage space requirements. Signature files, on the other hand, have much lower space requirements at the price of search speed. Signature files divide the text into blocks each of which is assigned a bit mask. Words are mapped to these bit masks by means of a hashing function. The bit map of a block is obtained by combining the signatures of all individual words in the block. Therefore, searching for a word consists in finding all blocks whose bit maps match the word's signature and performing sequential search on them. Search time in signature files is high compared to inverted indices and suffix trees. A comparison of the main indexing strategies is shown in Figure 5.9. The graph indicates that although inverted files do not exhibit the fastest performance, they represent a good compromise between space requirements and search speed.

In TEC, index construction is very simple: as the text is tokenized, the exact position of each token is recorded into a file (called the inverted index, or inverted file). Figure 5.10 shows a possible index file consisting of a table mapping words to lists of word positions.

Inverted indices are generally space-efficient. Space required to store the vocabulary is rather small in comparison to corpus size. Heaps' Law states that the size of the vocabulary grows sub-linearly on the size of the text.[7] Storing the position of each word in the text, however, takes by far the most space and might cause space requirements to increase greatly if one is not careful. The

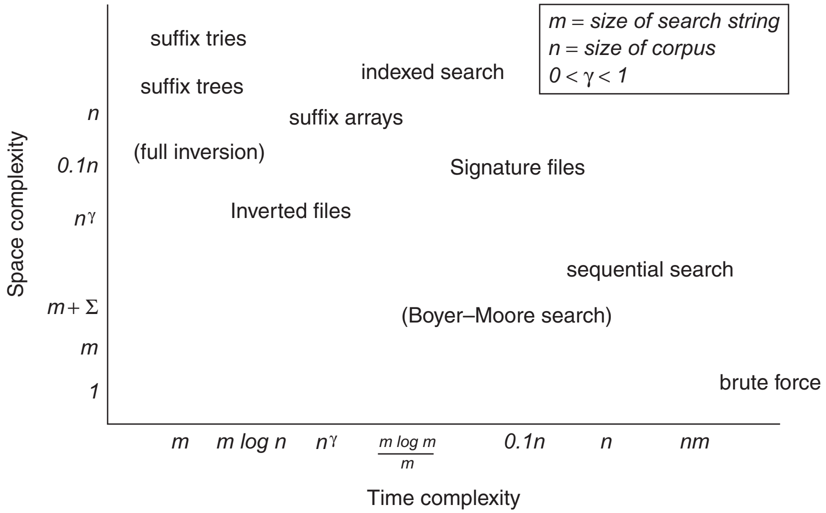


FIGURE 5.9 Space and time complexity of most common indexing techniques, adapted from Baeza-Yates and Ribeiro-Neto (1999)

Each	Word	is	analysed.	The	position	of	each	word . . .
1	6	11	14	25	29	38	41	46

vocabulary		**positions**
each	→	1, 44
word	→	6, 46
is	→	11
. . .	→	. . .

FIGURE 5.10 Sample text fragment (top) and corresponding inverted file (bottom). Numbers indicate word positions

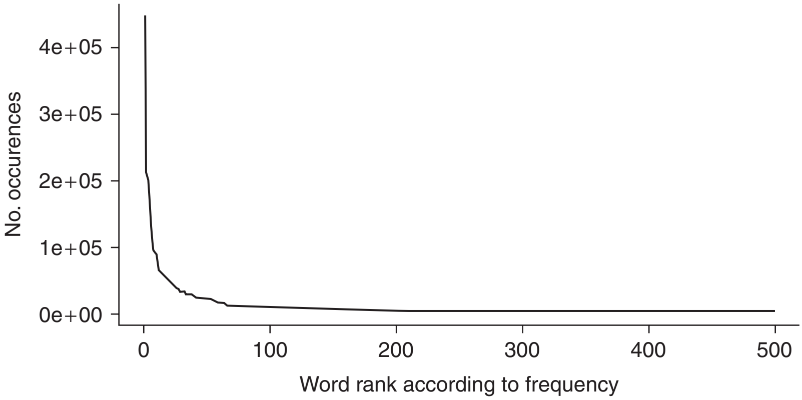


FIGURE 5.11 Word distribution in TEC

number of positions to be stored is distributed according to Zipf's Law. In its simplest formulation, Zipf's Law states that the most frequent word occurs *x* times as frequently as the most frequent word. This pattern has been observed in TEC, whose word distribution curve is shown in Figure 5.11. As inverted indices consist of lists of integers each of which points to a position in the text, keeping space requirements within acceptable bounds implies finding an economical way of storing such lists.

A number of index compression techniques exist. In TEC, we have opted for a simple but effective approach. Instead of storing absolute word positions, we simply store the position of the first occurrence followed by offsets pointing to subsequent occurrences, as shown in Figure 5.12. This technique of storing relative rather than absolute values achieves a 57 per cent index size reduction for a 33,398-word file.

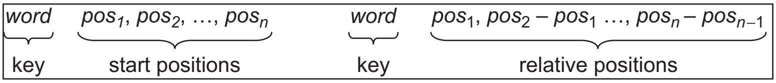


Figure 5.12 Schematic representation of two indexing strategies: uncompressed index versus index compression used in TEC

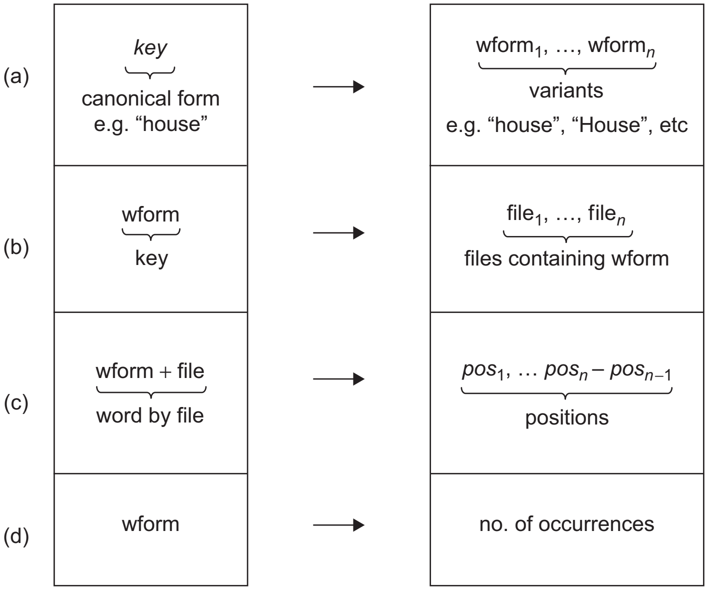


Figure 5.13 TEC inverted index structure

Because most of the search task (tokenization and indexing) is really performed offline, searching for a keyword via inverted indices is usually fast. Actual performance varies depending on how the index is stored. If the index is simply stored as a list sorted in lexicographical order, it can be searched in binary mode which yields search time proportional to the size of the vocabulary (i.e. $O(\log n)$), where n is the number of words in the vocabulary. However, if indices are stored as *hash tables* or as *tries* very fast retrieval can be achieved. Search times drop to linear in the size of the word searched for (i.e. $O(m)$, where m is the size of the word).

The TEC system stores its inverted index in hash tables. In fact, in order to optimize various kinds of queries the system builds a set of interrelated tables. Its index structure is shown in Figure 5.13. The system allows case-sensitive as well as case-insensitive search, so it creates separate entries for different forms of a word with respect to the capital letters that occur in it (Figure 5.13 (a)) but also stores a canonical word form for case-insensitive searches (Figure 5.13 (b)). Tokens are primarily associated to the files in which they occur, so start positions are recorded for each file (Figure 5.13 (c)). Thus files can be easily removed or re-indexed if needed. Finally, the system stores a pre-compiled frequency list (Figure 5.13 (d)). This last table is redundant, as its content can be retrieved from the other three tables. However, being able to access frequency information instantly enables the system to provide immediate feedback to the user on the progress of the search. This is an essential feature in web-based interfaces.

5.6 Data Storage and the Internet

The overall architecture of the internet is based on a client-server model. In this kind of model, data are stored in a central location (the server) and accessed through specialized software running on remote machines (the clients). This model exhibits two interesting features: it allows for resources to be distributed while preserving data integrity (as data are centrally maintained), and it allows for heterogeneity of access, as long as different clients agree to abide by the protocol implemented by the server.

Why are these features interesting for corpus users? First of all, the client-server model helps overcome copyright issues. A great deal of material of interest to translation studies scholars is subjected to copyright restrictions.[8] Distributing large volumes of copyrighted material as an integral corpus would be very costly, if at all feasible. However, not all data need to be available in order for corpus research to be carried out. Corpus researchers are mostly interested in uncovering patterns, often aided by metadata, rather than sequential access to an entire text. The client-server model enables corpus maintainers to restrict access to indirect retrieval of text fragments, thus protecting copyright. By removing this barrier, the client-server model facilitates sharing of resources among researchers, and might help provide the research community with larger volumes and variety of data than has been available so far. In addition, using the Web as a distributed storage medium has other advantages, such as the existence and ubiquity of standards and ease of access.

On the other hand, the model also has potential disadvantages. The most conspicuous disadvantage is the fact that the performance of a web-based corpus tool tends to be poor if compared to single-user tools. The speed of execution of a web-based client is directly affected by the bandwidth and latency of its communication channel to the server (the network bottleneck problem). Although the general trend is for network connections to become faster and more reliable, there is little corpus providers and developers of corpus clients can do to alleviate the network bottleneck problem. Another problem is the instability introduced by relying on a centralized server. If a server crashes, or has its network connection interrupted, all clients are affected.

5.7 The TEC Browser: A Tutorial Introduction

The TEC browser can be started directly via the Web.[9] If accessed through a web browser, the corpus browser will automatically detect and use the network settings of the web browser. Because the TEC system uses the same communication protocol as a standard web server, the browser can operate seamlessly across network security devices such as firewalls.

The main functionality of the TEC browser is the retrieval of concordances. Concordances are retrieved by entering a query on a search box. The general format of a query expression accepted by the browser is the following: *word_1(+([no_of_intervening_words])word_2 . . .)*, where brackets denote optional items. The expression *the+end*, for example, will match all occurrences of the immediately followed by end. Note that if *[no_of_intervening_words]* is omitted, the system assumes it to be 0. So *the+end* is equivalent to *the+[0]end*. One can retrieve all concordances for the phrases *the very end, the end* and *the wrong end* simply by specifying *the+[1]end* as query expression.

In addition to word sequences and intervening words, the server also accepts 'wildcards'. Wildcard syntax allows the user to select word prefixes by appending an asterisk to the query. For example, typing in *test** will retrieve all words which start with test, including test, testament testimony etc. Any *word_n* token in the query syntax expression described above can be replaced by a word or a wildcard. An expression such as *a*+test** is a perfectly valid query which will return *a test, acid test, about testing, a testament* among other phrases. The same result could obviously be achieved by searching for, say, *test** and sorting by the left context. This last strategy, however, would be less efficient since all concordances would need to be transmitted to the client, and transmission delay is the main factor affecting the performance of the browser, as discussed above. Another factor that affects performance in searching for combined keywords is the choice of the primary keyword. The following example illustrates this point. Suppose one is interested in retrieving the expression *the lonely heart*. Although TEC contains a single instance of this phrase, the corpus contains over half a million occurrences of the article *the*, over 3,500 occurrences of the word *heart*, and about 30 occurrences of the word *lonely*. If one chooses *the* as the primary keyword, the system will have to read though over half a million word sequences in order to find that single instance. Searching for *heart* would certainly improve things, but the best choice would be to take *lonely* as the primary keyword, as this would reduce the computation to at most 30 comparisons. If a word sequence query such as *the+lonely+heart* is submitted, the system will automatically choose lonely as the primary key, thus minimizing search time. It is nearly always a good idea to submit a sequence query if a sequence of words is what one is after.

There are situations, however, in which one needs to be able to retrieve all concordances for a given word and then explore possible collocations. A sorting function can be used to support this kind of exploration. TEC allows sorting by left and right contexts of various sizes. ‘Sort context’ pull-down menus located next to the sort buttons allow the user to specify how many words to the left or right of the keyword the list will be sorted by. Sort keywords appear highlighted by colour on the concordance list. One could, for instance, search for test, and then use sorting to group together occurrences of a test, acid test, the test etc. Once a concordance list has been downloaded, sort can be activated by the sort buttons. If one attempts to start sorting before downloading is complete, the system will ask whether the user wants to interrupt the download operation and truncate the concordance list. Although sorting is quite efficient, approaching linear performance in certain cases, it illustrates one of the advantages of a distributed architecture: since sorting is done directly on concordances, it can be entirely performed by the client (i.e. the corpus browser), thus freeing the server to perform searches.

Search can also be constrained by selection of sub-corpora. We have seen above that the information in TEC header classifies the various text files according to number of attributes of its authors, translators and so on. These attributes can also be used to define sub-corpora. Since the original XML-encoded metadata contained in the header files has been parsed and stored in a native XML database (Meier 2002), sub-corpus selection can be done through standard XPATH syntax. For example, the following query will select a sub-corpus containing texts written by Belgian or Brazilian men, and translated by either British or Canadian women[10]:

(5) ($s/author/@gender=‘male’) and ($s/translator/@gender=‘female’) and
($s/author/nationality/@description=‘Belgian’ or
$s/author/nationality/@description=‘Brazilian’) and
($s/translator/nationality/@description=‘British’ or
$s/translator/nationality/@description=‘Canadian’)

The sub-corpora selection tool is activated via the ‘Options’ menu on the TEC browser menu bar. Choosing ‘Select sub-corpus . . . ’ brings up the selection tool. The sub-corpus selection tool is a window which contains a number of selection boxes representing metadata attributes and their possible range of values. This allows a form of sub-corpus selection by direct manipulation. For both authors and translators, the user can specify the author’s (or translator’s) name directly, or a combination of the following attributes: gender, nationality and sexual orientation. Selections are activated by highlighting the items on the selection boxes. Alternatively, the user can enter XPATH queries such as the one in Example (5) directly.

In addition to its core functionality of concordancing and sub-corpus selection, the TEC browser also supports *Plug-ins*. Plug-ins are tools that perform specific functions, building on the main functionality provided by the core browser (GUI, network connection, concordancing, etc.). TEC has a few external plug-ins which serves to illustrate the concept: a frequency list browser, a corpus descriptor and dynamic concordance visualization tool. Other TEC plug-ins currently under development include: selection by part-of-speech tags and collocation analysis.

5.8 A Vision of Web-Based Corpus Software

Before discussing the future of the TEC system and perspectives on the future of web-based corpus software in general we must make a few remarks on the current status of the software. The current TEC architecture employs the standard client-server model depicted in Figure 5.14. Whereas the basic functionality for corpus indexing and access is handled at the server side, the client handles not integral texts directly but concordances generated by the server. The client itself is implemented in a modular architecture that enables new functionality to be easily incorporated. The TEC client may also communicate with a standard web (HTTP) server for requests that involve direct retrieval of metadata. The reasons for splitting the server functionality between a specialized concordance retrieval module and a generic content server are related to performance and security issues. The concordance server can perform more efficiently if it is dedicated to retrieving concordance data. Furthermore, having a dedicated concordancer makes it easier to

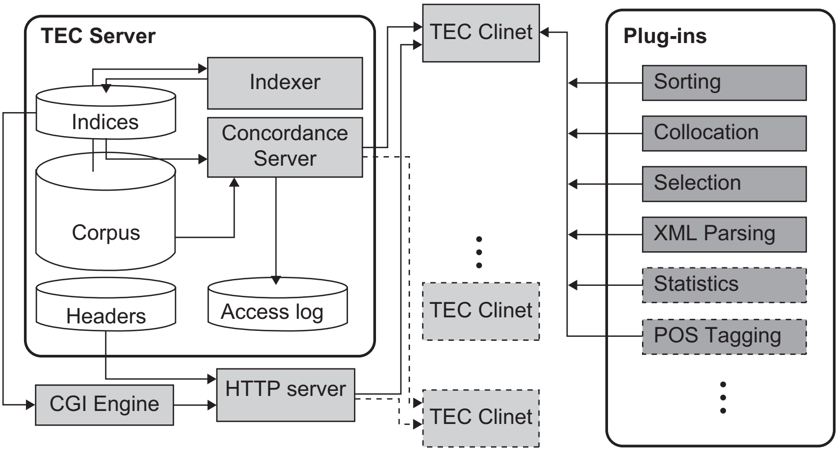


Figure 5.14 The current architecture of the TEC system

guarantee that copyrighted material will not be accidentally made available on the internet.

The TEC tools have been implemented in Java, as part of a suite of language processing tools called *modnlp*. The client can run as a stand-alone program, or over the Web using Java Web Start[11] technology. The entire system has been developed as free software and is distributed under the GNU Public License (GPL[12]). Code, licence and documentation are available at the modnlp website.[13]

5.8.1 Processing Models and Web-Based Corpora

In computer science, the areas of parallel and distributed computing study ways in which processors can be combined in order to improve the efficiency and flexibility of a system. In computational terms, distribution can be achieved in two forms: distribution of data and distribution of processing power. One uses the expression *data stream* to refer to the flow of data from a processor to another, and analogously, *processing stream* when referring to the flow of instructions from a processor to another. There are four classes of (logically possible) architectures with respect to the distribution of computational resources: SISD (single instruction, single data streams), MIMD (multiple instruction, multiple data streams), SIMD (single instruction, multiple data streams) and MISD (multiple instruction, single data streams). Although these classes are primarily used as a taxonomy of computer hardware, at least the first two of them are relevant to the way corpus software works.

While the model adopted by most single-user tools for corpus indexing is analogous to SISD, the TEC architecture depicted in Figure 5.14 can be considered as a MIMD architecture. A few differences must be noted, however. In MIMD architectures, subtasks are usually allocated different processors on the same (multiprocessor) machine. Coordination, data exchange and control therefore often relies on shared memory-based inter-process communication mechanisms. TEC processing tasks, on the other hand, will typically be allocated to remote processors and communication will therefore use network-based communication mechanisms and protocols (e.g. TCP/IP). For example, in the current system, while the server processor is occupied in retrieving concordances and sending them across the internet to the various clients, each client might be displaying the concordances it has received, and placing other requests (e.g. metadata) to the server. This type of loosely coupled interaction is characteristic of distributed architectures.

In the context of our corpus-software application, we can perhaps establish a further distinction within the MIMD class: single-server versus multiple-server architectures. The current TEC system operates as a single-server architecture. The entire corpus is stored at a central location and manipulated by a single processor. Although this kind of architecture helps improve access by remote users, it does little, if anything, to improve information providing. Different user groups that share an interest in corpus research have similar needs, some of which might be met by in-house resources (e.g. a small-scale corpus of translated text). Corpus research often involves analysing data from different source corpora. TEC users, for instance, use corpora like the BNC in order to compare language usage in translated and non-translated English. Ideally, it should be possible for diverse corpus sources to be pooled into a common framework for corpus processing. One might argue that such framework could be implemented as a single-server model simply by using a large, centralized corpus. However, copyright issues and other practical constraints present problems for centralized models. An alternative approach would be to develop a uniform interface that would mask the complexities and physical locations of various, heterogeneous services, as suggested by Sharoff (2006). However, such an approach would still require centralized management and manual updating.

Arguably, the essence of a multiple-server approach to corpus processing is to enable geographically distributed research groups to build smaller-scale corpora and share them through their own servers, on their own terms. Clients would then be able to discover and query selected servers, and autonomously combine the resulting responses into a coherent presentation. Some progress has been made towards automating this process of service description and discovery in the area of service-oriented computing (Papazoglou et al. 2008) but tools to enable easy sharing of language resources in this manner are still scarce.

A typical usage scenario for this improved architecture would involve the client selecting a set of corpora to be searched and broadcasting queries to selected corpus servers, each server evaluating the query (independently and in parallel), and the client receiving and combining the results from each server into a final result to be displayed to the user. An extension of the current architecture of the TEC system to implement this scenario is shown in Figure 5.15.

5.8.2 Challenges

The key challenges to truly distributed web-based corpus software are corpus selection and server capability description.

Corpus discovery and selection can be framed in the context of over a decade of efforts aimed at encoding metadata. These efforts range from language-specific standards such as the ones promoted by the Text Encoding Initiative (Ide and Veronis 1995) to more ambitious proposals (e.g. Zanettin, this volume, Chapter 4). The issues involved are essentially how to encode metadata and what information to encode. Although some progress has been made regarding the former, the latter is still very much an open problem. The development of credible standards such as XML helps solve the problem of how to

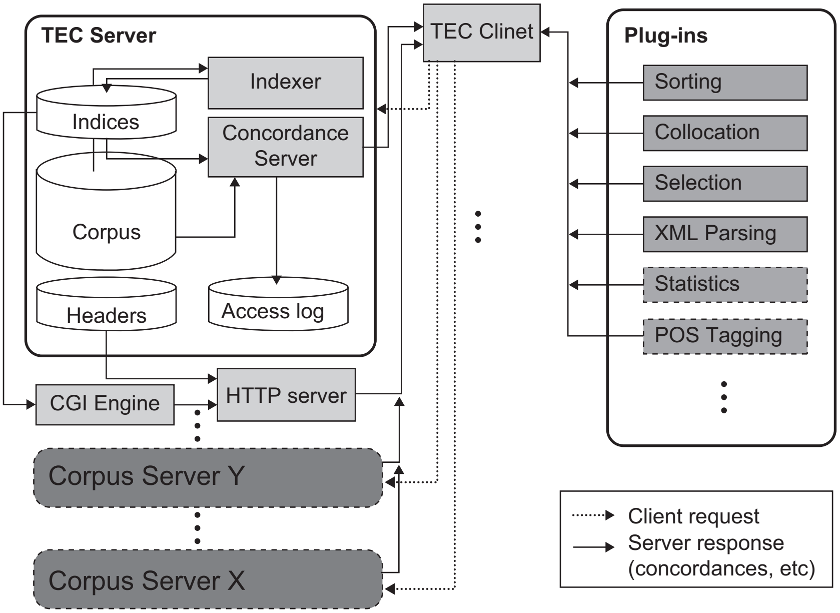


FIGURE 5.15 Multiple-server TEC system architecture

encode metadata so as to allow interoperability of different applications. New XML-based standards are being developed which aim to do the same at the semantic level. What is less clear is how this semantically structured information can be used. A typical dilemma of metadata encoding concerns determining how strict its semantics should be. If the semantics is too restrictive, it will yield over-specific metadata. If it's too lax, it will make it difficult for software clients to handle the potentially large variety of metadata created by the various servers.

Solutions to these problems might well have to be domain-specific. In the domain of translation studies, for instance, one could aim to define an extensible but minimal set of metadata items relevant to the community of corpus users and build basic software functionality around it.

Describing operational capabilities of servers is an issue that appears to have received considerably less attention from web researchers and language resources organizations. This is perhaps due to the fact that it is generally assumed that the sole function of a server of language resources is to make such resources available to its users, via the web or by other means of distribution. However, a quick look at the TEC server's processing capabilities described above suffices to reveal that this assumption is inadequate. In order to function properly, clients might need access to details such as whether the server supports case-sensitive search, what exactly it considers as a token, whether it supports search by part-of-speech tags, and numerous other features. Flexible, loosely coupled distribution of corpora and corpus software cannot be achieved unless it is supported by reliable capability description mechanisms. Emerging technologies such as multi-agent systems and peer-to-peer computing might play an important role in bringing about these mechanisms.

5.9 Conclusion

Although there are still many obstacles to the implementation of a model for the processing of distributed corpora as a complement to existing systems for distributed processing of corpora, recent developments in the areas of annotation standards and internet technologies suggest that this goal is achievable. This chapter has described several technologies currently used in web-based applications which might help bring this to fruition.

Notes

[1] The term *document* has in this chapter, as in the SGML/XML literature, the connotation of a unit of data which is often, though not necessarily, of a textual nature.

[2] http://java.sun.com/products/javawebstart/

[3] http://ronaldo.cs.tcd.ie/tec/CTS_SouthAfrica03/data.tgz

[4] <!DOCTYPE html PUBLIC '-//W3C//DTD XHTML 1.0 Strict//EN' 'http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd'>

[5] See http://www.w3.org/XML/Schema for details.

[6] As in TEC, where omit tags are used to inhibit indexing of non-translated material which would otherwise contaminate the corpus.

[7] In other words, vocabulary size $v = O(n\beta)$, where $0<\beta<1$ and n = text size.

[8] TEC, for instance, consists largely of translated fiction and biographies, all of which is copyrighted material.

[9] http://ronaldo.cs.tcd.ie/tec2/jnlp/

[10] The XQUERY selection statement is automatically added by the browser.

[11] http://java.sun.com/products/javawebstart

[12] http://www.gnu.org/

[13] http://modnlp.berlios.de/

References

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto (1999) *Modern Information Retrieval*, London: Addison-Wesley-Longman.

Baker, Mona (1999) 'The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators', *International Journal of Corpus Linguistics* 4(2): 281–98.

Bos, Bert, Tantek Çelik, Ian Hickson and Håkon Wium Lie (eds) (2009) *Cascading Style Sheets Level 2 Specification*. Available online at: http://www.w3.org/TR/CSS2/ (accessed 31 May 2010).

Bray, Tim, Jean Paoli and C. M. Sperberg-McQueen (eds) (2008) *Extensible Markup Language (XML),* Version 1.0. W3C recommendation REC-xml-19980210. Available online at: http://www.w3.org/TR/REC-xml

Clark, Jim (ed.) (1999) *XSL Transformations (XSLT) Version 1.0*. W3C recommendation REC-xslt-19991116. Available online at: http://www.w3.org/TR/xslt/

Clark, Jim and Steve DeRose (eds) (1999) *XML Path Language (XPath) V. 1.0*. W3C recommendation REC-xpath-19991116. Available online at: http://www.w3.org/TR/1999/xpath/

Harold, Elliotte Rusty and W. Scott Means (2004) *XML in a Nutshell,* third edition. Cambridge: O'Reilly & Associates, Inc.

Ide, Nancy M. and Jean Veronis (1995) *Text Encoding Initiative: Background and Contexts*, Dordrecht, The Netherlands: Kluwer Academic Publishers.

ISO8879 (1986) *Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML),* International Organization for Standardization, Geneva, Switzerland, first edition.

Luz, Saturnino and Mona Baker (2000) 'TEC: A Toolkit and API for Distributed Corpus Processing', in Steven Bird and Gary Simmons (eds) *Proceedings of Exploration-2000: Workshop on Web-Based Language Documentation and Description*, Philadelphia: University of Pennsylvania, 108–12.

Manning, Christopher D. and Heinrich Schütze (1999) *Foundations of Statistical Natural Language Processing*, Cambridge and Massachusetts: The MIT Press.

Meier, Wolfgang (2002), 'eXist: An Open Source Native XML Database', *Web-Services, and Database Systems*, Berlin: Springer-Verlag, 169–83.

Papazoglou, M. P., Paulo Traverso, Schachram Dustdar, Frank Leymann and B. J. Krämer (2008) 'Service-Oriented Computing: a Research Roadmap', *International Journal of Cooperative Information Systems* 17(2): 223–55.

Sharoff, Serge (2006) 'A Uniform Interface to Large-Scale Linguistic Resources', in *Proceedings of the Fifth Language Resources and Evaluation Conference, LREC 2006*, Genoa, Italy, 538–42.

Zanettin, Federico (this volume, Chapter 4) 'Hardwiring Corpus-Based Translation Studies: Corpus Encoding'.

Part II

Methods for the Qualitative Analysis of Contrastive Patterns in Large Corpora

Chapter 6

Lexical Priming and Translation

Michael Hoey

In this chapter I outline some suggestions for a new theory of lexis and language, and to consider how this might impact on translation and translation theory. The theory has been described in more detail in Hoey (2003, 2004a, 2004b, 2006, and most comprehensively in 2005a) and is here presented as briefly as is compatible with clarity. My suggestions in the area of translation will be tentative rather than confident, and will take two forms. First, I will look at a tiny piece of actual translation by amateurs, and then I will generalize to the role of translation in promoting or inhibiting linguistic imperialism.

6.1 The Notion of Lexical Priming

Traditional theories of language have worked on the assumption that every language has a grammar and a lexicon which are described as operating on each other in complex (or sometimes alarmingly simple) ways. The lexicon is represented in a variety of ways but usually amounts to a discrete set of words which can be added to in a range of ways and can be organized as a list (e.g. as a dictionary) or in terms of perceived sense relations (e.g. as a thesaurus). This view of lexis is supported by the way in which lexical facts are usually reported. We have dictionaries, lexicons and thesauri, compilations of words in lists and columns. Words are allowed history (etymology), they are assigned meaning(s), their pronunciations (idealized from many accents and without account being taken of context) are provided, they are assigned to one or more grammatical classes (though their behaviour in syntactic strings is not usually commented on) and selected idiomatic expressions are singled out, though these are rarely comprehensive. Words are considered so separate from grammar that they are played with in grammar-independent ways, as in crosswords, anagrams and word-searches (see, for example, Eckler 1996).

Grammars, on the other hand, are usually represented as complex, abstract and largely coherent (as opposed to lexis, which is represented as simple, specific and largely unorganized). Grammars provide statements about syntactic distribution, legislate on the borders between grammatical and ungrammatical, and talk about the conditions of operation and uses of grammatical words. They use ordinary lexis with no concern for its selection – lexis is only there to illustrate the operation of some grammatical rule or pattern – and most lexical items are neither used in illustration nor mentioned as having an effect, apart from when awkward exceptions are being handled.

Recent years have seen an increasing number of models of language for which the generalizations just given would be inadequate (e.g. Hunston and Francis 2000; Goldberg 1995, and Sinclair 2004 to name only a few). Some advanced learners' dictionaries (e.g. *Macmillan's English Dictionary* 2002) have detailed sections on the grammatical behaviour of certain lexical items, and some practical grammars (e.g. Biber et al. 1999; Francis et al. 1996, 1998) take care to link their grammatical claims to specific lexical items rather than to any old lexis. Overall though, my sweeping account does not travesty, I think, the majority of current theories.

Translation research has concerned itself more with lexis than has linguistics in general and has often looked at grammatical implications in conjunction with lexical choice, but even in such research the division between lexis and grammar has in general been implicitly maintained.

I want to argue that a new perspective is demanded by the evidence we now have of the nature of language. In particular, the problem with these theories is that they do not take sufficiently seriously the challenge posed by 'collocation', the phenomenon whereby words, with varying degrees of arbitrariness, co-occur more often than would be explicable in terms of their distribution across the language. As we shall see, collocation taken seriously proves very subversive of traditional linguistic positions.

In several papers (Hoey 2003, 2006) and also in Hoey (2005a), I have considered in some detail the following sentence by Bill Bryson: *In winter Hammerfest is a thirty-hour ride by bus from Oslo, though why anyone would want to go there in winter is a question worth considering.*

In this sentence, which begins Bill Bryson's book *Neither Here Nor There*, the words *in* and *winter*, *hour* and *ride, ride* and *by, ride* and *bus*, and *ride* and *from, though* and *why*, *why* and *anyone*, *anyone* and *would*, *would* and *want*, *want* and *to, want* and *go*, *go* and *there*, and *worth* and *considering* are all instances of pairs of words that collocate. Furthermore, of course, the collocations group – *hour ride from, why anyone would* and *want to go* are obvious instances. The case made by Sinclair (1991) and others, that when we construct our sentences we do not begin from scratch – choosing grammatical structures for the words we have chosen or words for the grammatical structures we have chosen – but select strings of inter-collocating words as a largely single choice, seems compelling.

The problem that collocation poses is not merely that the lexical choice is syntagmatic as well as paradigmatic. It is that the very existence of collocation needs to be explained. If traditional theories of language and the mind are to be believed, there is no sensible explanation for its existence. If words are the last choices to be made, or are the bits of language that get mapped onto the syntactic structure, or belong to a different mode or dimension, then collocations should not exist. The fact, though, is that without the least hesitation we routinely make use of largely arbitrary collocations, from which it follows that the collocation in question must be stored in a manner that makes it ready for immediate use. In other words, the brain must be storing language in a manner analogous to (though obviously not identical to) the way a concordance represents language. If it was the case that the language we heard was almost immediately decomposed into semantic abstractions, it would be a matter of truly incredible coincidence that collocations ever occurred, except where they transparently reflected some real world relation.

In accordance with the need to account for the existence of collocations, I claim that when we encounter language we store it much as we receive it, at least some of the time, and that repeated encounters with a word (or syllable or group of words) in a particular textual and social context, and in association with a particular genre or domain, prime us to associate that word (or syllable or group of words) with that context and that genre or domain. Each use we make of the word (etc.) and each new encounter with it either has the effect of reinforcing the priming or, if the new encounter does not conform to our previous experiences of the piece of language in question, weakens it. So when we repeatedly read the word *winter* in travel writing in the immediate context of *in* (as opposed to *over, through* or *within*), the experiences prime us to expect it in such a context and ultimately to reproduce the combination, especially if we write or talk about travel.

It follows that collocations are not a permanent feature of the word (etc.). They may well drift in the course of an individual's lifetime. If they do, and to the extent that they do, the word (etc.) will drift slightly in meaning and/or function or in terms of the social context, genre and/or domain in which it typically occurs. Drifts in the primings of a community of speakers are the engine of language change.

In any case, collocational priming is sensitive to the social context, the interpersonal context, the domain, the genre and other kinds of context. Thus *shut* may be primed for a particular language user to occur with *up* in a family context and to be uttered by friends or family members in casual conversation. It is unlikely to be primed in such a way as a result of the language user's encounters with *shut* in committee meetings. If it is primed at all in such a context, it is likely to be in contexts such as *shut the door, an open and shut case* and *shut down.*

If we accept these positions, and my challenge to the linguistics and psychology communities is to come up with alternative explanations if mine do not seem satisfactory, we can see just how subversive collocation actually is.

The priming does not stop at collocation. As a result of our repeated encounters with a word or phrase, we can also become primed to associate it with a particular semantic set. In the Bill Bryson sentence, the word *hour* happens to occur with *thirty* but it is quite possible that our encounters with *hour* will not have primed us to expect *thirty* in conjunction with *hour* – such encounters may well, however, have primed us to expect *hour* to occur with a number, and *thirty* would therefore be understood as a member of the expected semantic set. Likewise the two words *ride by* may well for many readers not be primed to co-occur with *bus*, but previous encounters with words or phrases such as *train* or *four wheel drive vehicle* may well have primed us to expect a vehicle of some kind. (Notice that priming is a personal experience, being the result of a unique set of encounters with the word (etc.); it can therefore never be automatically assumed that every speaker will be primed in the same way – it is necessary always to say, in advance of the evidence, that someone *may well be primed* rather than s/he *will be primed*.)

The examples given in the previous paragraph are all instances of priming for semantic association. A slightly more abstract phrase such as *in winter* in sentence-initial position can be shown to associate with timeless truths (unlike *in the winter, during the winter* or *that winter*) (Hoey 2006). This is another instance of a typical semantic association that we make, as a result of our probable priming.

Not unconnected, *in winter* also has a bias for occurring with the present tense (such a tense being conventionally associated with timeless truths). This is an example of still another type of priming, that of colligational priming. A 'colligation' is an association that a word makes with a particular grammatical condition, or more accurately that we make as a result of our priming (for discussion of colligation and its origin in Firth 1957, see Sinclair 2004; Hoey 2005a). So, to take further examples from the Bill Bryson sentence, the word *ride* when used as a noun avoids subject-function and when preceded by NUMBER-MEASURE OF TIME/DISTANCE tends to occur in complement-function ('complement' here being defined as in Sinclair 1972, as the function following equative verbs, typically the verb *be*). Avoidance of **subject** is an instance of a negative colligation and association with **complement** is an instance of a positive colligation of *ride*. Colligation may also take the form of a preference for particular positioning in the sentence (or indeed the text). Thus the phrases *in consequence* and *as a consequence* have in English a strong tendency for most speakers to occur in sentence or clause-initial position.

The above examples also illustrate another feature of priming – that of nesting. When we are primed to use (or recognize) a word, syllable or word combination in particular linguistic or non-linguistic contexts, these contexts in turn become conditions for further primings. So, for instance, *ride* becomes primed for us as being available for noun use. When so used, it becomes primed for us as frequently associating with NUMBER-MEASURE

OF TIME/DISTANCE. When it so associates, the whole phrase is typically colligationally primed for us to occur as **complement** in the clause in which it appears.

6.2 A Challenge for Translation

What I have been hypothesizing is that when we acquire a lexical item, it is itself the result of priming and in turn becomes primed for collocation, grammatical category, semantic association, colligation and textual colligation (among other things, not discussed here). All of this raises a number of significant questions for translation and translation theory. Translation has concerned itself with problems of lexical equivalence and textual equivalence (e.g. Beekman and Callow 1974; Baker 1999) and of course with syntactic equivalence. But if anything of what I have been saying about lexical priming is true, then a whole range of new issues emerge.

Let us look again for a moment at the characteristic primings of a specific word. The word *consequence* was discussed briefly above in connection with the phrases *in consequence* and *as a consequence.* These expressions indicate two of the word's typical primings – collocation with *as a* and *in.* I also noted that when these collocations are favoured, the combinations *as a consequence* and *in consequence* characteristically are primed for sentence-initial position. To these primings we can add others. To begin with, the word *consequence* occurs, when accompanied by an evaluative adjective, with predominantly negative evaluations. Another priming that the word has for many people is that it avoids object-function (unlike its plural *consequences,* which occurs quite frequently as the head of a nominal group functioning as **object**).

The word *consequence* has an apparently exact equivalent in a number of languages – *conséquence* in French, *consecuencia* in Spanish, *Konsequenz* in German, *conseguenza* in Italian and *consequencia* in Portuguese, for example. The issue that lexical priming raises is whether it follows that these apparent exact equivalents will be equivalently primed. Will the apparent exact equivalent of *consequence* also characteristically collocate with prepositions to form conjunct-style expressions? If they do, will they favour sentence-initial position? Will these languages' equivalents of *consequence* occur with negative adjectives more than positive? Will they also be primed to avoid object-function?

Let me ground this discussion by consideration of a single equivalent – Portuguese *consequencia* – from the point of view of just two of the possible primings listed above for *consequence.* We saw that *consequence* collocates with *as a* and *in.* The Portuguese *consequencia* proves to have similar collocational primings; it collocates with *como* (roughly *as* in English) and with *em* (roughly *in* in English). In this respect, then, the primings of the near-equivalents are the same.

However, primings may have different weightings. One priming may be very strong, another much weaker. The collocation of *consequence* and *as a* is stronger in English for many speakers than that between *consequence* and *in*. Indeed informal inquiry amongst my British students suggests that for many of them *consequence* is actively primed for occurrence with *as a*, but only passively primed for collocation with *in* (i.e. they recognize the latter but would not themselves reproduce it). Examination, however, of a one million word corpus of Portuguese, constructed for me by Tony Berber-Sardinha, shows that *como consequencia* (which roughly translates as *as a consequence*) is less common in Portuguese and that *em consequencia* (which roughly translates as *in consequence*) is more common. Put in priming terms, many English speakers are strongly primed to associate *consequence* with *as a* and more weakly primed to associate it with *in*. Portuguese speakers in contrast are typically primed to associate *consequencia* with *em* more than with *como*. Straight translations of the two expressions will result in a correct denotational translation but not necessarily in a translation of equal naturalness or (in)formality.

This is even more the case if we examine the textual colligations of the two pairs of phrases. *As a consequence* and *in consequence* share, as we have seen, a preference for sentence-initial or clause-initial position. Indeed, 20 per cent of all instances of *consequence* in my corpus occur as part of one of these phrases in such a position. In contrast, virtually no instances of *em consequencia* and *como consequencia* appear at the beginning of a sentence or a clause. Taking these two sets of facts together, it would appear that we have a new class of false friends – words or phrases whose denotations in the two languages are largely similar but which are characteristically differently primed for the speakers of these languages.

I have tried to couch my questions regarding *consequencia* in a way that does not give primacy to English. One must first find the characteristic primings of the Portuguese word (an emic description [Hoey and Houghton 1998]) and then see whether they map onto those of English, and it does not automatically follow that the categories that worked for the one set of primings will necessarily be appropriate for the other.

I have also tried to avoid referring to the colligations and semantic associations of *consequencia* (though practically such a wording is inevitable and indeed acceptable). The reason is that such a wording assumes that the question is answerable for the language as a whole. But in fact colligations, semantic associations, etc. are domain, genre and context specific. So a word does not have a colligation; it has a colligation in newspaper writing, in academic discourse, in detective fiction. Furthermore, what may be a colligation in the primings of one user of the language may not exist or be much weaker for another user. My primings are the product of my experiences and use of language; yours are the product of a quite different set of experiences or uses.

So what enables us to communicate with each other? There are, I suggest, a number of harmonizing factors that ensure that users of language do not spin out of contact with each other. One of these is education; I would go so far as to say that one of the important reasons for education is to ensure a degree of harmony between speakers. As well as testing factual content, examinations serve to test whether language users have been primed appropriately for the discipline. A second is that of the media, since these represent shared experience for a considerable number of language users in a community. A third factor is that of the literature of the language, perhaps a declining force in ensuring the harmonizing of the primings of users (though Harry Potter must be having a vast harmonizing effect) but one of particular relevance to translators.

Perhaps the most important harmonizing factor is the *six degrees of separation effect*. Our utterances form a vast cohesive interconnecting web and because our encounters with words are typically in utterances that conform to – rather than defy – previous primings, our primings harmonize themselves. It is this factor, and the fact that education impinges on most of the kinds of writing and talk that translators and interpreters concern themselves with, that makes it still sensible to compare, in connection with the same genre, the collocations, colligations and semantic associations of apparently equivalent words in two languages. Any claims made, though, about the possible harmonization of primings within and across two languages must be made cautiously and their scope carefully limited.

What I want to do in the remainder of this chapter is to consider translations into Portuguese of the first clause of the Bill Bryson sentence discussed above, provided for me by three Portuguese-speaking students.[1] For the first half of the original sentence, they offered the following alternatives:

No inverno, Hammarfest fica a trinta horas de ônibus de Oslo.
In winter Hammarfest is thirty hours by bus from Oslo.

Hammarfest fica a trinta horas de ônibus de Oslo, no inverno.
Hammarfest is thirty hours by bus from Oslo, in winter

One favoured the first version, two the second. As will be immediately apparent, the only difference in their translations lies in the positioning of the phrase *no inverno*, a matter to which I shall return shortly. We have in effect, then, a single translation.

I examined the choices made in this translation (these translations), using a corpus kindly constructed for me by my colleague, Mike Scott. The corpus is as close to comparable with my 100 million word corpus of English as could be achieved quickly, in that both corpora were built out of news stories. However, there are the following important differences:

1. Brazil makes greater use of news journals (*Veja*, *Istoë*, both comparable to the US's *Newsweek*) than does the UK, which has no mass market news weeklies. This means that quite a lot of the data are different in detail and kind.
2. My English corpus comprises all of the *Guardian* newspaper (and no other newspaper) for four years, and those years are drawn from the early 1990s. The Portuguese corpus is taken from *Folha de São Paulo*, *Veja*, *Istöe* and other similar sources and is therefore more heterogeneous.
3. The corpus Mike Scott constructed for me was estimated to have been approximately 10–20 million words, but my system crashed whenever I loaded much more than 6 million words, probably because of the number of files in the corpus. Consequently the analysis that follows is based on this smaller sample.

For all these reasons, and because in most cases the number of instances I was able to inspect was small, the observations I am about to make should be seen as hypothesis-forming, rather than claim-making.

6.3 Some Lexical Comparisons between the Original and the Translation(s)

6.3.1 *In Winter* versus *No Inverno*

The first point of interest in the translations by my Brazilian informants is the choice of two of them to move *no inverno* (in winter) to a position at the end of the clause. In Bill Bryson's original, the phrase *in winter* is of course beginning a sentence, a paragraph and a text, since the sentence is (more or less) the first in the book. None of these positions for *in winter* are especially rare in my *Guardian* corpus, and sentence-initial position is common. In contrast, there is little evidence of any tendency in the Portuguese corpus for *no inverno* to occur sentence-initially and none whatsoever for its occurrence in paragraph-initial or text-initial position. It is possible, therefore, that the two translators who moved *no inverno* to final position in the clause were doing so to avoid conflict with the phrase's characteristic colligational priming for non-initial position in Portuguese and because using it in initial position would not serve the text-colligational function that its English equivalent serves. It is possible, too, that place names have the same text-initial function for Portuguese as they have for English and that therefore the movement of *no inverno* allowed the text-initial potential of *Hammarfest* to come to the fore.

The next point of interest lies in a subtle but apparently unavoidable mistranslation by the Brazilian students. The Portuguese phrase *no inverno* means *in **the** winter*. A search of my Portuguese corpus revealed no instances of *em inverno* (which would accurately translate *in winter*), though the phrase *em pleno inverno* (in full winter) occurs a couple of times. (A search on Google supports this pattern: there were approximately 795,000 hits for *no inverno*, a comparatively small but respectable 29,300 hits for *em pleno inverno*, and a tiny 524 hits for *em inverno*.) So the question must arise: is *no inverno* primed for most Portuguese speakers to have the same colligations and semantic associations as *in winter* has for most English speakers or is it in these respects more like its closest translation *in the winter*? Given that Portuguese would appear, on the corpus evidence, not to have a choice between the indefinite and definite forms, it might be predicted that the semantic associations of *in winter* with 'timeless truths' (Hoey 2005a) and of *in the winter* with specific events would be neutralized. Likewise, we might expect *no inverno* to manifest neither the colligations of *in winter* with the present tense nor of *in the winter* with the past tense.

The size of my Portuguese corpus meant that there were too few instances of *no inverno* to be able to conclude anything with confidence, but I undertook the analysis anyway. The results are to be found in tables 1 and 2. Table 1 compares the two English expressions with the Portuguese expression in respect of their characteristic primings for semantic association with specific events or general statements of long-standing validity (that I have labelled 'timeless truths').

I reiterate that 16 instances of *no inverno* are too few to permit any sensible conclusions to be drawn, but the data here suggest that *no inverno* might align itself with *in the winter* (its closest translation) in associating with specific events. If a larger corpus were to replicate the distribution found here, we would have to conclude that there has been no neutralization of the semantic associations associated with the two English translations of *no inverno*.

Turning now to colligation, Table 2 compares the two English expressions with the Portuguese expression in respect of their tendency to colligate (or

Table 6.1 A comparison of *in winter, in the winter* and *no inverno* with respect to their occurrence in statements of specific events or 'timeless truths'

	in winter	***in the winter***	***no inverno***
specific event	29 13%	179 **54%**	10 **62.5%**
'timeless truth'	197 **87%**	152 46%	6 37.5%
	strong semantic association with timeless truth	weak semantic association with specific event	apparent semantic association with specific event

Table 6.2 A comparison of *in winter, in the winter* and *no inverno* with respect to their occurrence in clauses with present or past tense

	in winter	***in the winter***	***no inverno***
clause with present tense	133	111	8
	59%	34%	50%
clause with past tense	40	165	7
	18%	50%	44%
none or other	53	55	1
	23%	16%	6%
	colligation with present tense	negative colligation with present tense	apparently no colligation with either tense

otherwise) with either the past or present tense. Halliday and James (1993) argue that the past and present tense are in even distribution in English; any marked deviation from such a distribution in connection with a particular expression would indicate that, for many users of English or Portuguese, the expression was primed to occur with one tense more than the other.

As can be seen, whereas *in winter* has a colligation with the present tense for most speakers and *in the winter* has a negative colligation with the present tense (i.e. speakers avoid using it in present tense clauses), *no inverno* appears (on the basis of scant data) to occur with both tenses equally. Therefore, the apparent absence of choice in Portuguese has resulted in the predicted loss of colligational priming, though not, it would appear, in the loss of its priming for semantic association.

6.3.2 *Fica* versus *Is*

The next point of interest relates to a decision required of the Brazilian translators that did not need to be made by the original American writer. Portuguese has a choice as to how to translate *is*, a choice that has no equivalent in English. The choice is that between *fica* and *está*. Again on limited evidence, it would appear that *fica* is primed in distinctive ways, and that the students were correct to choose *fica* in place of *está*. To begin with, *fica* appears to have a semantic association with PLACE in Subject function, whereas *está* has no strong association of this kind. PLACE (subject) + *fica* occurs 59 times in my data, as opposed to 12 instances of PLACE (subject) + *está*. Of the 59 *fica* instances, 46 (78 per cent) also are associated with LOCATION, whereas only two of the *está* instances occur with LOCATION. The 46 instances of PLACE (subject) + *fica* + LOCATION are further categorizable as follows:

11 (19 per cent)		measurement of location by time or distance from another location
12 (20 per cent)	+	*onde* [where]
7 (12 per cent)	+	point of compass

The two instances of PLACE (subject) + *está* + LOCATION comprise one instance of measurement of location and one instance of *onde*.

We can tentatively conclude from these data that for most speakers of Portuguese the combination of PLACE (subject) + *fica* is primed for semantic association with LOCATION, MEASUREMENT and POINT OF COMPASS and for collocation with *onde*. The combination *onde* + PLACE (subject) + *fica* is primed to occur in the sequence *onde* + VERB + SUBJECT. The implications for translation of these details will be drawn out below.

6.3.3 The Absence of *Ride* in the Portuguese Version

One word is missing in the translation(s), and that is the word *ride*. The reasons for this appear to be complex and lie in the fact that Bill Bryson has chosen to override the typical primings of *ride* in connection with *bus*. The fact is that *ride by bus* in the original English is colligationally unusual. *Bus* and *ride* indeed collocate but almost always in the combination *bus ride*. In my data, *ride by* is associated with difficulty – *ride by tractor, ride by four-wheel drive vehicle* are instances – and Bill Bryson's decision to override the expected *bus ride* may have been to imply (as much of the chapter that follows goes on to show) that the journey is not an easy one in mid-winter. In the Portuguese translations of the Bill Bryson sentence, however, the noun *ride* has been omitted, and the replacement *de ônibus* has in consequence lost the challenge implied in the original. Nor is this the only feature of *ride by bus* lost in the translation. Of 107 instances of *de ônibus,* 32 are (in my judgement) translatable as *by bus*. Of these 32 instances, only seven (22 per cent) are associated with LOCATION, only one (3 per cent) is associated with MEASUREMENT OF DISTANCE and again only one is associated with MEASUREMENT OF TIME. So, unlike *ride by* VEHICLE, *de ônibus* is only weakly primed for association with LOCATION and not primed for MEASUREMENT OF DISTANCE OR TIME in Portuguese news writing. We consequently have lost in the translation the sense of *bus* as engaged on a difficult journey and we have lost the connection with *hour* through the semantic association with MEASUREMENT OF TIME; the words are both there in the translation, but they no longer have a special association.

On the other hand, *de* does behave like *by*. In my Portuguese corpus, I found the following instances:

de ônibus, burro, carro,
by bus, donkey, car

de ônibus e barco
by bus and boat
em comboio *de ônibus e jipes*
in a convoy of buses and jeeps
de ônibus ou de automovel
by bus or by car

We can probably infer that *de* is primed to have a semantic association with MODE OF TRANSPORT, as is *by* in English.

6.3.4 *Hora* versus *Hour*

One pair of words where the Portuguese and the English are very much alike in their characteristic primings is that of *hora* and *hour*. Consider the data in Table 6.3 for *hora(s)*. It will be seen that *hora* shares with *hour* a semantic association with number of the kind already noted. It also shares a semantic association with means of journey (or route) and with place of origin (or arrival). So here we have equivalents in the two languages that, in these respects at least, are also equivalent in their characteristic primings for colligation and semantic associations.

6.4 The Implications for Translation

Space and the paucity of data have only permitted the briefest of accounts of the possible implications of lexical priming for translation, but I hope it can be seen that, with a larger corpus of Portuguese, we *might* be able to say that the translation from English to Portuguese has resulted in subtle shifts of colligation, semantic association and textual colligation, with the new sentence reinforcing or adding some of these features and weakening others. It remains now to consider what the implications of these shifts might be for the significance and role of translation in the world.

If we stay with accepted theories of language, we must conclude that translation is not responsible for more than a modicum of language change. The syntax of the target language stays the same and, mostly, appropriate word-equivalences are found. Translation has led to a small amount of **borrowing** (sometimes of course much more than this: see Baker (1998) for a number of languages where translation had a major initial impact) and connotational change has probably been considerable. Overall, though, translation, according to traditional theories of syntax or lexis, has a limited effect on target languages.

What, however, would the effect have been if all three students had left *no inverno* in front position? What would have been the implications for the target

Table 6.3 Lines from the concordance for *hora** classified according to the semantic associations of *hora**

a menos de uma	*hora*		*da ilha*
less than one	hour		from the island
a apenas uma	*hora*	*de avião*	*de Argel*
only one	hour	by plane	from Argel
a pouco mais de uma	*hora*	*de avião*	*da Argélia*
little more than one	hour	by plane	from Argélia
a cinco	*horas*	*de barco e Caminhada*	*da BR-230*
five	hours	by boat or road	from BR-230
a meia-	*hora*	*de carro*	*de Paris*
half an	hour	by car	from Paris
a duas	*horas*	*de carro*	*das praias*
two	hours	by car	from the beaches
a uma	*hora*	*de carro*	*na direção leste*
an [one]	hour	by car	to the east
a meia	*hora*	*de carro*	*de Manhattan*
half an	hour	by car	from Manhattan
a meia	*hora*	*de carro*	*de Nova York*
half an	hour	by car	from New York
a poucas	*horas*		*de Jerusalem*
a few	hours		from Jerusalem
a uma	*hora*		*de São Paulo*
an [one]	hour		from Sao Paulo
a meia	*hora*	*de trem*	*do centro de Buenas Aires*
half an	hour	by train	from the centre of Buenas Aires
a sete	*horas*	*de trem*	*de Nova York*
seven	hours	by train	from New York
por duas	*horas*	*de trilha*	*na mata*
for two	hours	by track	in the jungle
NUMBER	*hora(s)*	VEHICLE OR ROUTE	FROM (OR IN) A LOCATION

language if the primings for *hora* and *hour* for speakers of the two languages were different but the translation remained the same? The answer is that any Portuguese speaker who would have read the translated sentence would have had an encounter with the words *inverno* and *hora* that did not chime in with his/her previous primings for those words.

Now suppose that every travel book that the Portuguese speaker reads is a translation from English. Suppose, further, that every translator is careless about the colligations and semantic associations of the Portuguese equivalents s/he uses – and to some extent the translator cannot do other than reflect at least some of the colligations and semantic associations from the source language. In such a situation, every Portuguese speaker would be exposed to collocations, colligations, semantic associations and textual colligations that were the product of English primings, and they would thereby become, for those speakers, Portuguese primings. The conclusion to be drawn might be that a translation should only display its foreignness as opposed to disguising its translated status (cf. Venuti 1998) when the translation is *into* English rather than *from* English. If every one of the translations in the travel section displayed their translated status, the effect might be to alter the primings of the Portuguese speakers who read them. Only if the translators successfully obscured the translated origins of the text would the primings of Portuguese readers be preserved, and then only in part.

I have not been painting an unreal picture. I have, for example, regularly found in bookshops around the world a substantial proportion of the fiction section taken up by translations from English. Perhaps this is another area where Anglo-American imperialism is (unintentionally) at work – the invasion of priming snatchers! The language looks the same but it has an alien within it and it will behave differently and in accordance with its own alien culture unless it is confronted and controlled. (I am not oblivious to the fact that I am here relying on the reader picking up on an American cinematic reference, and I am of course not American myself).

The issue then arises of what the effects on English might be of translations into English that retain the primings of typical speakers of the source languages. Perhaps surprisingly, I believe that the effects are likely to be benign (and are equally benign for any other language, as long as the translations do not overwhelmingly come from a single source language).

To understand why translations may have a benign effect on the primings of a language as long as they do not overwhelm that language, it is necessary to step back and look at how a theory of lexical priming might account for creativity. I have argued elsewhere (e.g. Hoey 2005a; 2005b) that creativity appears in a language when a user chooses to override one or more primings, though if all primings are overridden, what is uttered or written will be entirely incomprehensible. Even James Joyce's *Finnegan's Wake* retains some collocations. Creativity is therefore a matter of a selective overriding of primings. This is, I take it, what critics are referring to when they have talked of poets renewing the language.

More generally, and outside the context of literature, as we build up our primings of the words we encounter, and in particular as we build up semantic associations and colligations, our primings have the potential to become more abstract. In short we begin to construct a personal semantics and a personal grammar. This semantics and this grammar will always be imperfect, inconsistent and incomplete, and will differ from person to person both in the detail and in the degree to which they are incomplete, but they are together the basis, in part, of our ability to say new things.

The primings guarantee fluency and we can say very much anything we want to with them as long as what we want to say conforms to our experience of the language. There is a degree of abstraction in colligation, textual colligation and semantic association that will ensure that anyone who does not construct for themselves a personal grammar can still function fully as a human communicating being. The personal semantics and personal grammar allow, however, expressions of thoughts (and the comprehension of thoughts) that go beyond our immediate linguistic experience.

It is for this reason that we value creativity where we find it (even sometimes when its *only* value is that it goes beyond our linguistic experience). Thus it is that we should also value translations, especially when they retain the otherness of the source language. Of course, as I began by saying, this positive value would disappear if 75 per cent of what we read took the form of translations from a single linguistic culture. Allowing for this, though, we can say that translation, like poetry, refreshes the language.

6.5 A Brief Conclusion

To sum up, lexical priming suggests there are many more factors involved in translation than one might have thought, and the translator has the choice of either preserving the primings of the target language or importing the primings of the source language (or, of course, a mixture of both). Which of these choices is the better depends on whether it is a solitary migrant or a swarm and on whether the recipient has the option of turning away. Whatever the choices that are made, drifts in a speaker's primings are, as noted earlier, the engine of language change, and translation is a potential source of drifts. I conclude therefore that translation, whether literary or non-literary, considered or impromptu, is in fact one of the most important linguistic activities undertaken by human beings and, despite the recognized importance of the translation of key texts in the development of certain languages (the *King James Bible* for English, Luther's translation of the same for German), it is an activity whose implications for diachronic linguistics may not yet have been fully explored.

Note

[1] I am grateful to Dr Tania Shepherd for her help in arranging for the collection of these data.

References

Baker, Mona (1992/1999) *In Other Words: A Coursebook on Translation*, London: Routledge.

Baker, Mona (ed.) (1998) *Routledge Encyclopaedia of Translation Studies*, London: Routledge.

Beekman, John and John Callow (1974) *Translating the Word of God*, Grand Rapids, MI: Zondervan.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English*, Harlow: Longman.

Eckler, Ross (1996) *Making the Alphabet Dance: Recreational Wordplay,* New York: St. Martin's Press.

Firth, J. R. (1957) *Papers in Linguistics*, London: Oxford University Press.

Francis, G., S. Hunston and E. Manning (1996) *Collins COBUILD Grammar Patterns 1: Verbs*, London: HarperCollins.

— (1998) *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*, London: HarperCollins.

Goldberg, Adèle (1995) *Constructions: A Construction Grammar Approach to Argument Structure*, Chicago: University of Chicago Press.

Halliday, M. A. K. and Z. L. James (1993) 'A Quantitative Study of Polarity and Primary Tense in the English Finite Clause', in J. M. Sinclair, M Hoey and G. Fox (eds) *Techniques of Description: Spoken and Written Discourse*, London: Routledge, 32–66.

Hoey, Michael (2003) 'Why Grammar is Beyond Belief', in J. P. van Noppen, C. Den Tandt and I. Tudor (eds) *Beyond: New Perspectives in Language, Literature & ELT, Special Issue of Belgian Journal of English Language & Literatures, New Series* 1: 183–96.

— (2004a) 'Textual Colligation: A Special Kind of Lexical Priming', in Karin Aijmer and Bengt Altenberg (eds) *Advances in Corpus Linguistics, Papers from the 23rd International Conference on English Language Research on Computerized Corpora* (ICAME 23), Göteborg 22–26 May 2002, 171–94.

— (2004b) 'Lexical Priming and the Properties of Text', in A. Partington, J. Morley and L. Haarman (eds) *Corpora and Discourse,* Bern: Peter Lang, 385–412.

— (2005a) *Lexical Priming: A New Theory of Words and Language*, London: Routledge.

— (2005b) 'Saying Something New: The Hidden Patterns in Corpora', *The Fourth Sinclair Open Lecture (2004)*, Birmingham: University of Birmingham.

— (2006) 'Language as Choice: What is Chosen?' in G. Thompson and S. Hunston (eds) *System and Corpus: Exploring Connections*, London: Equinox.

Hoey, Michael and Diane Houghton (1998) 'Contrastive Analysis and Translation', in Mona Baker (ed.) *Encyclopedia of Translation Studies*, London: Routledge.

Hunston, Susan and Gill Francis (2000) *Pattern Grammar,* Amsterdam: John Benjamins.

Macmillan English Dictionary for Advanced Learners (2002) (edited by Michael Rundell), Oxford: Macmillan Education.

Sinclair, John M. (1972) *A Course in Spoken English: Grammar,* London: Oxford.

— (1991) *Corpus, Concordance, Collocation,* Oxford: Oxford University Press.

— (2004) *Trust the Text: Language, Corpus and Discourse* (ed. Ronald Carter), London: Routledge.

Venuti, Lawrence (1998) *The Scandals of Translation: Towards an Ethics of Difference*, London: Routledge.

Chapter 7

Looming Large: A Cross-Linguistic Analysis of Semantic Prosodies in Comparable Reference Corpora[1]

Jeremy Munday

7.1 Introduction

This chapter is concerned with semantic prosody, an area that has been studied in monolingual (mainly English) corpus-based work but which until recently had received relatively little attention among translation studies theorists. It sets out to define the concept, review some of the key literature, describe methods of analysis with examples from comparable corpora of Spanish and English, and discuss possible applications for translation.

7.2 Definition and Theory of Semantic Prosody[2]

The term 'semantic prosody' has been variously defined but is used to refer to 'a consistent aura of meaning with which a form is imbued by its collocates' (Louw 1993: 157). A complementary perspective is provided by Hunston and Francis (2000: 137), who state that 'a word may be said to have a particular semantic prosody if it can be shown to co-occur typically with other words that belong to a particular semantic set'. As an illustration, we can say that semantic prosody often refers to how what might be expected to be a semantically neutral form, such as the lemma *CAUSE*, in fact tends to be used with words that give it a particular hue – negative in the case of *CAUSE* (see Stubbs 1996). In this way semantic prosody, while being 'strongly collocational' (Xiao and McEnery 2006: 107), may be said to blend collocation and connotation and even have an attitudinal function (Louw 2000: 60, in Stewart 2010: 14). Hunston (2007) also stresses that this may vary depending on other factors such as genre and register.

Louw acknowledges that he borrowed the then unnamed concept from Sinclair (1991), who studied positive and negative semantic sets in his account of early corpus work with the COBUILD project in Birmingham. One example given by Sinclair is the phrasal verb *SET IN*, in the sense of 'become established'. The instances provided by his corpus led Sinclair to conclude that this verb has a very strong negative prosody since it collocates with generally negative subjects, in examples such as:

Before *bad weather* sets in . . .
Desperation can set in . . .

Semantic prosody is thus about the way that sense and connotation spread surreptitiously across collocates or from the typical surrounding co-text. Since it is often not overtly controlled by the text producer, it may reveal a writer or speaker's underlying attitude or evaluation. However, it may also be used for particular effect, including a deliberate violation of the expected semantic prosody to produce irony (Xiao and McEnery 2006).

The study of irony and ironists – 'people who mean the opposite of what they say' – is the focus of the key article by Louw (1993: 169–71) in which he describes a radio interview where the Director-General of the British Council refers to contacts with UK universities being 'symptomatic of the University of Zimbabwe'. Using the Bank of English corpus (37 million words at that time), Louw analyses the phrase *symptomatic of* to show it has a clear negative prosody. Among the examples given are:

symptomatic of *clinical depression*
symptomatic of a *problem*
symptomatic of *other management inadequacies*
symptomatic of *something deeply wrong*.

Louw argues that the use of the phrase in conjunction with *the University of Zimbabwe* conveys a strong hint of irony, since *symptomatic of* is so closely associated with failings. In the instance in question, the use might therefore suggest that the University of Zimbabwe is a minor partner, needing to compensate for its inadequacies by enlisting the help of UK institutions. The issue then arises as to whether this irony, brought about by an unexpected semantic prosody, is deliberate, in which case the encoders or text producers, to use Louw's words (Louw 1993: 171), are 'writing the device' (i.e. consciously manipulating the semantic prosody); or they may be involuntary, where 'the device writes the encoder' (i.e. the markedness of the prosody then impacts on the sense transmitted by the writer). This is an important point to which we shall return at the end of this chapter.

Of course, to be able to gauge the ironic potential of a specific utterance it is necessary to know the probable semantic prosody of the term in question. This can only really be done by analysing the examples found in a large representative corpus, permitting probabilistic conclusions to be drawn based on the comparison between the trends in the corpus and the analysis of the individual utterance – comparison, in Tribble's terms, between 'global' and 'local' prosodies (Tribble 2000). Louw's contention, even back in 1993, was that 'it will be obvious [. . .] that semantic prosodies must furnish one of the most compelling arguments for building and using larger and larger corpora'. Since, he claimed, prosodies are not evident to intuition, their identification requires the analysis of large amounts of language to reveal underlying trends (Louw 1993: 164).

Later and more detailed corpus-based studies by Stubbs (1995; 1996; 2001) include a combination of manual semantic prosody analysis and quantitative collocational work on larger corpora to provide a semantic profile of words. Thus, in his 1995 paper entitled 'Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies', Stubbs manually analyses concordance lines from the Lancaster-Oslo-Bergen and COBUILD corpora to show that the lemma *CAUSE* has an 80 per cent negative prosody, collocating with unpleasant events such as *accident*, *concern*, *damage* and *death*, with only 2 per cent of examples classed as positive and 18 per cent as 'neutral'. Other lemmas analysed by Stubbs are *CREATE*, which is shown to have a more balanced profile, and *HAPPEN*, which is decidedly negative.

The negative/positive poles of analysis – 'polarity', to use Channell's term (Channell 1999) – are extended by Hoey (1997) to include 'events' and 'profession', although later (Hoey 2005: 23) he prefers the term 'semantic association' for this phenomenon.[3] For event words, such as *consequence*, Hoey gives four associations, which are, in order of frequency:

the logic of an outcome (*likely*, *inevitable consequence*, etc.)
a negative outcome (*dire consequence*, etc.)
a serious outcome (*decisive consequence*, etc.)
an (un)expected outcome (*unintended consequence*, etc.)

For professions, Hoey draws on Campanelli and Channell's (1994) analysis of the verbal phrase *train as a* . . . . This is seen to collocate with occupations such as *lawyer*, *nurse* and *teacher*, which generally have a positive connotation, but also unusual roles such as *boxing second* and *kamikaze pilot*, which many might regard with more circumspection. Hoey (1997: 2) makes the important point that semantic prosody, or association, extends beyond collocation. Thus, whereas *train as a lawyer* may be a relatively common collocation, the organization of the collocates is by non-exclusive semantic sets. This allows the absorption of other, including newly coined, collocates, for example *computer programmer* or, in the negative set, the modern-day *suicide bomber*. A basic Google™ search for this last term in conjunction with parts of the lemma *train* shows the frequency of this collocation, the following being typical examples:

And the majority of 36 women known to have been trained as suicide bombers are still out there. (http://news.scotsman.com/topics.cfm?tid=610&id=768132003)

Suicide bomber was trained at saboteur camp. (http://newsfromrussia.com/accidents/2003/06/21/48515.html)

Korkmaz suspects that his daughter was taken to a PKK camp in Germany or the Netherlands to be trained as a suicide bomber. (www.turkishdailynews.com/old_editions/06_26_99/for.htm)

This fits well with Hoey's theory of lexical priming by which he is specifically referring to the way that 'the word is learnt through encounters with it in speech and writing, it is loaded with the cumulative effects of those encounters such that it is part of our knowledge of the word that it co-occurs with other words' (Hoey 2003). Hoey (2005) develops this discussion of how words are thus 'primed' to occur not only in lexical but also in certain grammatical constructions, such as the passive. Whereas the lexical company a word keeps concerns collocation, organized by associated semantic sets into semantic prosodies, a word's grammatical company (the structures in which it habitually occurs) pertains to what Firth (1957) termed 'colligation'. Collocation, colligation and cumulative priming are all critical to a fuller understanding of semantic prosody. Thus, even the small sample above would suggest that *train as a suicide bomber* may most likely occur in a passive construction.

7.3 Semantic Prosody and Contrastive/Translation Studies

Although the literature mentioned in the previous section has been written from a monolingual English perspective, the importance of semantic prosody analysis for translation has also been stressed, notably by Partington (1998: 77):

> Semantic prosody is an important area of research for translation studies [. . .] It seems to be the case that cognate [. . .] words in two related languages can have very different semantic prosodies.

Partington has in mind words such as *impresionante*, which in Italian, and in Spanish too, has a negative prosody, whereas the English cognate *impressive* is generally positive. One might describe this as a subtler variation on the old false cognate (*faux ami*) translation chestnut. Detailed contrastive studies of *non*-cognate terms have in fact revealed much more subtle differences. Thus, Berber-Sardinha (2000) carried out an insightful corpus-based analysis of the Portuguese dictionary equivalents for Sinclair's example *SET IN*. Analysis of concordance lines of *MANIFESTAR-SE* and *ESTABELECER-SE*, among others, shows that: the collocates of *ESTABELECER-SE*, such as *relação* [relation] and *país* [country], do not reflect the negative semantic prosody of *SET IN*; *MANIFESTAR-SE* retains the negative prosody only with the collocate *doença* [disease]; *CAIR* does form patterns with the Portuguese word for 'night' [*noite*] but in positive senses. Berber-Sardinha's conclusion is that there are no full direct equivalents in Portuguese for the English *SET IN*. This may be true from the perspective of the lexicographer, keen to find a one-to-one mapping of the global use, including global semantic prosody, of a dictionary entry (Catford's well-known 'formal correspondent', Catford 1965: 27), but the translator needs a 'textual equivalent' (Catford ibid.) for a specific translation instance that would encompass the local semantic prosody. The study is also limited to analysis of prosodic polarity with no real consideration of the colligational priming of the different alternatives.

Despite this, Berber-Sardinha's (2000) work does have clear translation-related applications: it is of great assistance for the translator or bilingual lexicographer with the identification of typical collocates and semantic sets. The results are then represented either through the use of typical collocates as sense discriminators for a dictionary correspondent or, in a specific target text context, by the choice of one equivalent over another (e.g. *MANIFESTAR-SE* over *ESTABELECER-SE*). Additionally, it is valuable in objective descriptive source-text target-text (ST-TT) analysis to identify shifts in prosody between ST and TT items, a useful tool for the uncovering of translation shifts above and beyond the denotational, syntagmatic and discoursal translation shifts discussed by the likes of van Leuven-Zwart (1989, 1990). The hypothesis would be that in some cases the translator may not intuitively be aware of the prosody or, influenced by ST lexis and structure, might inadvertently choose an equivalent which has a different prosody from the original. Such a semantic prosody shift may be due to interference from ST to TT (*impressive* for *impresionante*) or to a more subtle lack of match between ST and TT prosodies (*MANIFESTAR-SE* for *SET IN*). The result would then be a blurring or distortion of effect on the reader, whose own lexical priming may well be jolted by an unexpected prosody.

7.4 Reference Corpora Used in This Study

Access to a large number of examples of a particular search term is more or less essential for the analysis of semantic prosody, and this is only really possible when using large electronic corpora (see Olohan 2004: 82). The main corpora employed in this comparative study of Spanish and English are representative comparable monolingual corpora[4]:

(i) the British National Corpus (BNC) (www.natcorp.ox.ac.uk/) designed as a representative collection of mainly British English from the 1980s up to 1995 and comprising around 110 million running words;
(ii) the online CREA Spanish Real Academia corpus (www.rae.es), a constantly updated collection of around 400 million words of different varieties of Spanish from 1975 to the present;
(iii) collocation statistics from the Leeds internet corpora of English and Spanish (http://corpus.leeds.ac.uk/internet.html), based on 100 million word samples of web-held texts.

As an example of how comparable corpora can assist analysis, let us look at how the Spanish equivalent of the lemma *CAUSE* [*causar*] is used. As we saw above, *CAUSE* tends to collocate with words such as *accident*, *concern*, *damage* and *death*. Investigation of the 891 instances of the search term *causan* (third person plural present tense) from the Spanish CREA shows similarly negative collocates such as *amnesia* [amnesia]*, considerables daños* [considerable damage]*, destrozos* [damage]*, dificultades* [difficulties]*, dolor* [pain]*, enfermedades* [illnesses]*, estragos* [devastation]*, preocupación* [concern]*, problemas* [problems]. It is also interesting to note that, in cases such as *efectos* [effects], the negativity is even transferred from another part of the sentence:

los truenos ya causan efectos psicológicos muy deprimentes
[thunder already causes very depressing psychological effects]

where *muy deprimentes* [very depressing] gives a strong negative spin to *efectos*.

Analysis of the CREA shows that *efectos psicológicos* does generally have a rather negative prosody, particularly in certain medical contexts: *efectos psicológicos del estrés* [psychological effects of stress], *de los sonidos electrónicos* [of electronic sounds], and so on. However, in other texts there may be a neutral or even positive prosody:

los efectos psicológicos del cannabis: relax, sensación de paz, predisposición a la amigabilidad y a la risa
[the psychological effects of cannabis: relaxation, feeling of peacefulness, predisposition to friendliness and to laughter]
los efectos psicológicos producidos por la actividad físico-deportiva
[psychological effects produced by physical sporting activity].

Such examples indicate the complexity of the semantic prosody of what might even have been thought of as 'neutral' noun objects.

Analysis of the Spanish corpus for the equivalent of the more neutral lemma *CREATE* (*CREAR* in Spanish) also reveals a varied picture. Collocates range from strongly positive to neutral to negative:

- strongly positive: *un ambiente acogedor y cálido* [a welcoming, warm atmosphere]
- positive: *nuevas fuentes de empleo* [new sources of employment], *nuevos cuerpos de profesorado* [new groups of teaching staff]
- neutral: *un comité* [a committee], *un color apagado* [a muted/dull colour]
- negative: *un ambiente de inseguridad* [an atmosphere of insecurity], *dudas* [doubts], *problemas* [problems], *perspectivas ilusorias* [illusory perspectives].

Again, it is noteworthy how the connotation is often transmitted by those words that occur around the noun: *ambiente* [atmosphere] appears in both the negative and positive classifications because it is strongly affected by its accompanying adjectival phrases, while *nuevos/nuevas* [new] provides the positive connotation in the two examples in which it appears. The example of *un color apagado* could also be neutrally descriptive ('muted') or negatively evaluative ('overly dull') depending on context.

7.5 Contrastive Analysis – *Loom (large)* versus *Cernerse*

The brief contrastive analyses of *CREAR/CREATE* and *CAUSAR/CAUSE* suggest that some of the basic positive-negative semantic prosody patterns may be similar in the two languages. However, as we noted above, we are working with probabilistic patterns and sometimes the writer deliberately violates the expected prosody. This is the case of the main example we shall now consider: *LOOM LARGE* and one of its Spanish dictionary correspondents, *CERNERSE*.

This pair of verbs has been chosen because of the large number of varied and metaphorical uses associated with them, and, initially, because of an unusual and strikingly apparent violation of the intuitively negative or 'ominous' prosody observed in the following example (from *The Metro*, London edition, 16 September 2003):

Mosley must be seeing dollar signs looming large.

Here, Mosley is the triumphant boxer in a World Championship fight and seems sure of more lucrative bouts in the future. The article is generally complimentary to him, and yet, intuitively, *looming large* has a threatening and negative connotation, so there may be hints of irony here or it may be an indication of the relative fluidity of the prosody associated with this verb phrase.

For the verb *loom* on its own, the *Concise Oxford English Dictionary* (11th edition, 2004) gives two senses:

(i) 'come into sight dimly, esp. as a vague and often magnified or threatening shape'; and
(ii) '(of an event or prospect) be ominously close'.

Both these senses encompass negativity, associated with threats or some ominous occurrence. It is, therefore, not surprising that in the British National Corpus the 105 examples of the lemma *LOOM LARGE* (comprising 33 instances of *loom*, 25 of *looms*, 31 of *loomed* and 16 of *looming*) are overwhelmingly negative. If we classify these examples according to the semantic prosody of its subjects, we get the following categories, with illustrative example collocates:

(a) Ominous 'landscape' subjects
Dark water, peatstacks, impressively blackened rocks, dark crevasses. The sense is often simply of the first *Concise Oxford English Dictionary* sense of 'come into sight'. The negative and threatening nature of the collocates is shown by the following larger context where the expected inference of the first half of the sentence is that dark crevasses will cause problems:

Dark crevasses loomed large, but didn't present any real problems.

(b) Metaphorical negative of failure, upheaval and prospect
Awful prospect of defeat, blackmail, chaos, failure of British industrial relations, fear, long commuting times, oppression, problems, shadow of de Gaulle, trade war, unrest. Interestingly, none of these collocates occurs with great frequency, so the negativity comes as much from the semantic set as a whole as from any individual collocate. And, in a similar way to the *efectos* example in Section 7.4 above, so the *awful prospect of defeat* instance sees what otherwise might be thought to be the neutral noun *prospect* (another interesting candidate for semantic prosody study) coloured by the preceding adjective *awful* and the subsequent noun *defeat.*

(c) Negative by context
The following examples are more difficult to evaluate, but the context suggests they are negative:

(i) *The prospect of her stay at Balmoral loomed large in Diana's mind.*
(ii) *. . . only when the prospect of African government loomed large . . .*
(iii) *In recent years, the word 'fibre' has loomed large in our lives.*
(iv) *Economic issues had loomed large in the primaries . . .*
(v) *. . . the conviviality of eateries will loom larger in deciding where to locate . . .*
(vi) *The plight of the rural poor does not tend to loom large in the minds of businessmen.*

Interestingly, (i) and (ii) are further examples of *prospect* and are negative because, in (i), the reference is to the then Princess Diana's dread of staying with the Royal family at Balmoral castle and, in (ii), the British government's sudden, hypocritical concern with the impartiality of the media in African countries as independence approached in the 1960s. Examples (iii) and (iv) are in articles that claim that the earlier fad for high-fibre diets and the focus on economic issues in the early stages of the 1980 US elections were misplaced. Examples (v) and (vi) are ironic: those deciding company relocations are more interested in local restaurants than business logic (v), and the rural poor have little importance for businessmen (vi).

Finally, there is one seemingly positive example about an individual's recovery from illness:

As she thought of the good golfing years she had allowed to go to waste, so practice and praying loomed large in the recovery programme she set herself.

The wider context contains no suggestion of irony here.

7.5.1 Cernerse

An equivalent for *LOOM LARGE* given in the *Collins Spanish Dictionary* (2009) is *CERNERSE*, the origin of which is the transitive verb *cerner* [to sieve – flour, soil, etc.]. This equivalent is an interesting choice by the lexicographer, since it is a strongly metaphorical use compared to, for instance, the entry in the *Oxford Spanish Dictionary* (2008), which prefers more explicitatory target language equivalents.[5] Given the inevitable limitations on the scope of this chapter, with its interest in simply testing the phenomenon of semantic prosody in a bilingual context, we shall restrict ourselves to a contrastive comparison based on the semantic prosody profile of *CERNERSE* alone as a potential equivalent to *LOOM LARGE*. To do this, a search was made for the corresponding third-person uses of the lemma *CERNERSE* in the CREA Real Academia corpus, using the search terms *se cierne* and *se ciernen* (present tense, third-person singular and plural) and *se cernía* and *se cernían* (imperfect tense, third-person singular and plural). The resulting concordance lines were examined manually to filter out the 'sieve' sense. The resulting number of occurrences was:

se cierne(n)	282 instances
se cernía(n)	140 instances.

A classification of the semantic sets of the subject collocates produced the following (these are indicative of the most frequent collocates but, for reasons of space, are not fully comprehensive):

(a) Nature subjects

los buitres ya se cernían en espiral sobre sus cuerpos
[the vultures were already looming/circling in a spiral over their bodies]

buitres, gavilanes y águilas se ciernen majestuosamente por los aires
[vultures, sparrowhawks and eagles loom/circle majestically overhead].

Buitres and *CERNERSE* seem to form a strong collocation; the second example is actually positive, thanks to the addition of the evaluative adverb *majestuosamente.*

(b) Meteorological phenomena
borrasca, canícula, cielo, crespúsculo, humareda acre, invierno, noche, nubes oscuras, oscuridad, solazo, tormenta [gale, dog days, sky, dusk, acrid smoke, winter, night, dark clouds, darkness, blazing sun, storm]. Though descriptive, these are also negative, as can be seen by the following examples:

La canícula se cernía implacable sobre Barcelona.
[The dog days/heatwave loomed implacably over Barcelona]
el cielo se cernía sobre el mundo como un ultimátum
[the sky/heavens loomed over the world like an ultimatum].

This last example, where *cielo* could be read either as 'sky' or 'heavens', is indicative of the blurring between the descriptive and the figurative. Not only is the weather bad, it is also metaphorically ominous (see c(i) below).

(c) Metaphorical negative

(i) Ominous meteorology
una atmósfera enrarecida, nubarrones, nubes de la corrupción, nubes de guerra [rarefied atmosphere, large clouds, clouds of corruption, clouds of war]. Here, the negativity of the weather is transferred to become a threat to the future:

Densos nubarrones se ciernen sobre una orquesta . . .
[Dense clouds loom over an orchestra . . .]
la tempestad que se cernía sobre su cabeza . . .
[the storm which loomed over his head . . .].

(ii) Dangers, threats and uncertainty
amenazas, crisis, dudas, fantasma, hostilidades, incertidumbre, males, peligro, riesgo, sombras, sospecha [threats, crisis, doubts, ghost, hostilities, uncertainty, evils, danger, risk, shadows, suspicion]. This is the most frequent prosody of the verb, with the most common individual collocates being *amenaza(s)* [threat(s), 47 instances] and *peligro(s)* [danger(s), 31]. One example suffices to illustrate the similar *cernerse* + *sobre* pattern of (c)(i):

Sombras ominosas se ciernen sobre nuestro héroe . . .
[Ominous shadows loom over our hero . . .].

(d) Negative by context
In the following two examples the negativity results from the context and the apparent stance of the writer:

los diferentes proyectos urbanísticos que se ciernen sobre ésta [ciudad]
[the different town planning projects that loom over it (the city)].

The wider context shows that the writer feels that these building projects are likely to damage the town.

La gran expectativa se cierne sobre el partido de fondo
[Great expectation looms over the crucial match].

One might expect that *expectativa* [expectation] would be positive. Seemingly, its collocation with *se cierne* violates the semantic prosody. However, the surrounding context in fact suggests that the overriding climate is one of fear since the game is a vital relegation match. This would support a hypothesis of the strength of influence of the global semantic prosody of the verb which is so great it can even withstand a collocation with a normally very positive noun subject.

7.5.2 Typicalities

Such comparative analysis demonstrates that both *LOOM LARGE* and *CERNERSE* have generally negative semantic prosodies. However, there are some notable differences particularly as regards the most common collocates and the syntactic structures in which the two verb phrases are used. Some of these can be identified using automatically generated distribution statistics such as are provided by online tools such as the *Sketch Engine* (Kilgarriff, online) and the *Leeds collection of internet corpora* (http://corpus.leeds.ac.uk/internet.html). These allow the analysis of huge amounts of data, and helps to identify major trends. However, since, as we saw above, semantic prosody seems to depend on semantic fields and interpretation of context and stance, it is only by combination with close analysis of specific examples that a more delicate picture can be constructed.

For *LOOM LARGE*, 43 per cent of the examples studied are followed by a prepositional phrase beginning with *in*, the most frequent being *in . . . mind*: *in Diana's mind, in her mind, in the minds of the new English bourgeoisie*, and so forth. This association with mental pressure, problems and preoccupations is generally absent from *CERNERSE*, where there is only one instance of its use with *problema(s)*. The most frequent collocates are *amenaza(s)* [threat(s)] and *peligro* [danger]. Syntactically, 47 per cent (200 out of 422) of the examples of *CERNERSE* are used with the relative pronoun, 67 per cent in the present tense and a massive 88 per cent (373) with the preposition *sobre* [over]. *CERNERSE* and *sobre* form such a strong collocation they may almost be considered a multi-word unit, a Spanish equivalent of a phrasal verb. An extension of this analysis to include the object categories that follow *sobre* reveals that the most common are place, mentioned either by name [*sobre Europa, Granada*, etc.] or with a phrase such as *sobre el país, el planeta* [over the country, the planet], a person or group of people (*sobre nosotros, los ciudadanos, la raza humana* – over us, the citizens, the human race) or the future itself.

The most common collocational, colligational and (negative) prosodic priming pattern of *CERNERSE* is with the noun subject *amenazas* [threats]:

amenazas + relative pronoun + present tense of *CERNERSE* + *sobre* + noun

The following example illustrates this most clearly and typically:

Cuando hablamos de las amenazas que se ciernen sobre el planeta . . .
[When we speak of the threats that loom over the planet . . .].

The noun *threat* does not occur at all as a subject of *LOOM LARGE*. Even if we extend the analysis to consider all 899 BNC instances of the lemma *LOOM* (i.e. including those without *large*), there are only five occurrences with *threat*, four of which occur with the present participle form *looming* (e.g. *now there's a new threat looming*). This may give a clue that this is the typical English structure. In marked difference to the priming pattern of *CERNERSE*, there are only 22 instances of *LOOM* (0.25 per cent) that occur with a relative pronoun (e.g. *This was one of two problems that loomed large at the time*), none with the noun *threat*.

In order to identify common patterns in English, we can approach the contrastive analysis of this last point from the perspective of the grammatical subject, starting with a British National Corpus (BNC) search of collocates of 2,000 examples of the noun *threat*. The results show that only 18 (0.9 per cent) occur as a subject with a relative pronoun. On the other hand, there are 559 examples (28 per cent) of the phrase *threat to*, of which the two below, in environmental texts, are typical:

all potential threats to the stratospheric ozone layer
the threats to our coastline are mounting.

This would suggest the following possible correspondence of colligation between the Spanish and English:

amenazas + relative pronoun + present tense of *CERNERSE* + *sobre* + noun
threats + *to* + noun.

Moving on to consider the other prominent typicality of *CERNERSE*, its use with metaphors from meteorology, a search for the English equivalent of *nubes* [*clouds*] reveals the prominence of a different verb, namely *gather* rather than *loom large*:

War clouds were soon to gather.
When storm clouds gather on the horizon . . .

It is interesting that the second example chimes very closely with two of the small minority of instances of *CERNERSE* which are followed by a preposition other than *sobre*:

un par de nubes oscuras se ciernen en el horizonte
[a pair of dark clouds loom/gather on the horizon].

The metaphorical nature of both the English and Spanish examples (referring to economic problems and to a general strike, respectively) merely confirms the correspondence.

7.5.3 Triangulation

This contrastive study of examples from comparable English and Spanish corpora is useful for identifying the global semantic prosody of a term, and even local prosody since it is supported by close analysis of individual examples, allied to the preferred syntactic structures or colligation. An important further step will be to supplement and extend it by a study of translations of the terms as found in actual translated texts. This may be achieved by studying a corpus comprising the Spanish daily *El País Digital* (www.elpais.es) and the American English version published in conjunction with the *International Herald Tribune* (available in the archives of *El País Digital*). In this way it is hoped that it will be possible to make the following triangulation, using *CERNERSE/LOOM LARGE* as an example (see Figure 7.1):

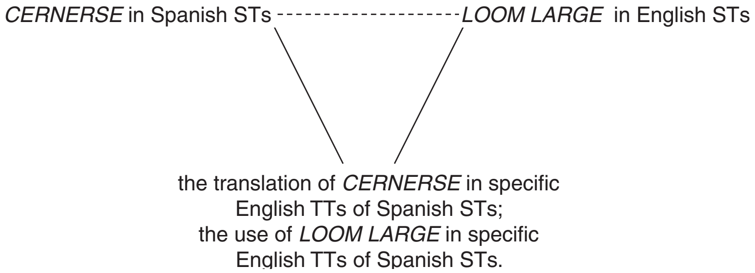

CERNERSE in Spanish STs
LOOM LARGE in English STs
the translation of *CERNERSE* in specific English TTs of Spanish STs; the use of *LOOM LARGE* in specific English TTs of Spanish STs.

Figure 7.1 Triangulation of *CERNERSE/LOOM LARGE*

A note of caution is necessary, however. As Pearson (2003: 18) emphasizes when analysing a corpus of articles from *Scientific American* and their French sister publication *Pour la Science*, the 'translations' in this kind of publication are often subject to cuts and adaptations for the target audience. In our case, the American English version is an eight-page collection of articles from the same or the previous day's Spanish newspaper and some of these articles are a summary version of the ST. Many examples of *CERNERSE* in the Spanish therefore do not figure at all in the TT.

The tight translation deadline for the publication will certainly mean that the translators have little time for revision and may therefore adopt a Mini-max, cognitively less demanding 'literal' translation strategy. Since one hypothesis is that semantic prosody is not intuitive, this may mean that the translator will calque prosodies from ST to TT, for example, consistently translating *una amenaza que se cierne* as *a threat which looms large*, even when this is not the most appropriate for the given collocate. However, initial results, though tentative, do not confirm this. Indeed, the picture is further complicated by some Spanish articles being translations from the English themselves. That is, a comment from an English ST is embedded in the *El País* text even if the fact of translation is concealed. The following example is from one such article on the then President George Bush's policy on terrorism, translated into Spanish and published in *El País* on 25 June 2004:

la amenaza terrorista que se cierne sobre su país
[the terrorist threat that looms over his country]

This derives from the typical, condensed, English ST pattern *the terrorist threat to his country*. The syntactic amplification in the Spanish demonstrates the strength in the mind of the translator of the typical *CERNERSE* collocational and colligational patterns we have identified.

7.6 Awareness of Semantic Prosody

In our discussion of the theoretical writing on semantic prosody, we noted Louw's distinction between 'the encoder writ[ing] the device' and 'the device writ[ing] the encoder'. The potential for conscious manipulation of semantic prosody would seem to be limited, as discussed by Xiao and McEnery (2006: 106) who maintain that it is 'at least as inaccessible to a speaker's conscious introspection as collocation is'. This may be the case if an individual is asked to describe the prosody of a specific word, such as *CAUSE*. But the fact that a violation of prosody (e.g. Louw's *symptomatic of* . . .) can be identified by the reader who then interprets the irony of the choice, shows that as readers we are at least to some extent aware of a word's profile. This, and the *amenaza terrorista* example at the end of the previous section, would suggest that the translator may be aware of the general semantic prosody of target text alternatives (since these are in his/her native language) even if he/she is sometimes less sensitive to subtle prosodic distinctions in the foreign source language. If true, this would be an interesting point of comparison to the findings of an empirical study by Dam-Jensen and Zethsen (2008), who found that Danish MA translation students translating into English did seem to be aware of the prosodies generally associated with *CAUSE* and *LEAD TO*.

One of the subtleties is possible genre variation of semantic prosody, which deserves more attention. Nelson (2000), Tribble (2001) and Hunston (2007) suggest that there seem to be genre-specific values attached, Nelson finding that *CAUSE* is generally neutral in academic texts but not in business texts. Tribble examines project proposals. His findings usefully merge collocation, prosody and colligation in the example of *experience* which has a specifically positive prosody associated with 'a form of professional capital'. Hunston shows that *CAUSE* does not necessarily have a negative prosody in scientific writing. Interestingly, Hoey (2005; see also Chapter 6 this volume) demonstrates similar domain-specific sensitivity for lexical priming as well, giving the example of *in winter*, used in travel literature, and the semantically similar *during the winter months*, in gardening books. This type of phenomenon merits further investigation, both monolingually and in the context of translation.

7.7 Concluding Comments

The corpus-based approach to semantic prosody analysis adopted in this chapter is really just a preliminary study to what is bound to grow as an area of investigation and which should provide very fruitful possibilities for collaborative research. As collaboration grows between translation studies theorists and monolingual corpus linguists and software developers, analysis will become increasingly refined. Indeed, this is already the case, as we noted with the *Sketch Engine* and word sense disambiguation software, which offers a package that will give a visual 'picture' of the most common right- and left-placed collocates of a given search term (Kilgarriff, online); a similar presentation is offered in the Bank of English (http://www.titania.bham.ac.uk/) and the representative WordBanks of French and Spanish corpora (http://www.collinslanguage.com/wordbanks/default.aspx, see Tucker 2004 for a comparative analysis of *war* and *guerre* using these resources). The Leeds collection of internet corpora (http://corpus.leeds.ac.uk/internet.html), for example, provides a wide range of collocational statistics that easily lend themselves to the analysis of semantic prosody. While an overall semantic prosody may be best revealed by corpus-based methods (cf. Louw 1993, see Section 7.2 above), these methods in fact support the intuition, introspection and contextual evaluation of the analyst

(Stewart 2009: 44; 2010: 148). Thus, this combination of computer-assisted tools and close textual analysis allows both a quantitative and qualitative evaluation of semantic prosody and colligation patterns, leading to greater insights into contrastive differences and the translation process.

Notes

[1] The initial work for this chapter was supported by a British Academy Overseas Conference grant, for which grateful acknowledgement is given. My thanks to those who commented on earlier drafts, including Timothy Ennis.

[2] See Stewart (2010) for a full-length discussion of the development of the concept.

[3] There is a terminological argument here: Hoey (2007: 40) describes his own shift from 'semantic prosody' to 'semantic association' and also discusses an alternative term, 'semantic preference' (Sinclair 1998; Partington 2004; Stubbs 2006).

[4] Here, we follow the suggestion of Bernardini et al. (2003: 5) in adopting the terminology of 'comparable' for 'similar originals in two languages' and 'parallel' for 'originals and their translations'. To be more precise, and following Laviosa (2003: 106), our corpora are 'monolingual comparable' and 'bilingual parallel'. There is, of course, considerable controversy over the representativeness of 'representative' corpora, which purport to provide a broad representation of a whole language, itself fraught with questions surrounding genre and dialectal and regional variants, amongst others.

[5] The *Oxford Spanish Dictionary* gives two illustrative examples, with different translations:
the problem loomed large in his mind
el problema dominaba sus pensamientos
[the problem dominated his thoughts]
superstition looms large in these tales
la superstición ocupa un lugar preponderante en estos relatos
[superstition occupies a preponderant place in these tales].

References

Berber-Sardinha, Tony (2000) 'Semantic Prosodies in English and Portuguese: A Contrastive Study', *Cuadernos de Filología Inglesa (*Universidad de Murcia, Spain) 9(1): 93–110. Available online at <http:*//revistas.um.es/cfi/article/view/66701/64341*> or http:*//revistas.um.es/cfi/issue/view/5361* (accessed 8 October 2010).

Bernardini, Silvia, Dominic Stewart and Federico Zanettin (2003) 'Corpora in Translator Education – An introduction', in: Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds) *Corpora in Translator Education*, Manchester: St. Jerome, 1–14.

Campanelli, Pamela and Joanna Channell, with Liz McAuley, Antoinette Renouf and Roger Thomas (1994) *Training: An Exploration of the Word and the Concept with An Analysis of the Implications for Survey Design* (Research Series No. 30), London: Department of Employment.

Catford, J. C. (1965) *A Linguistic Theory of Translation*, Oxford: Oxford University Press.

Channell, Joanna (1999) 'Corpus-Based Analysis of Evaluative Lexis', in Susan Hunston and Geoff Thompson (eds) *Evaluation in Text: Authorial Stance and the Construction of Discourse*, Oxford: Oxford University Press, 38–55.

Dam-Jensen, Helle and Karen Korning Zethsen (2008) 'Translator Awareness of Semantic Prosodies', *Target* 20(2): 203–21.

Firth, J. R. (1957) 'A Synopsis of Linguistic Theory, 1930–1955', in *Studies in Linguistic Analysis*, 1–32, reprinted in F. Palmer (ed.) *Selected Papers of J R Firth 1952–59*, London: Longman, 168–205.

Hoey, Michael (1997) 'From Concordance to Text Structure: New Uses for Computer Corpora', in Barbara Lewandowska-Tomaszczyk and Patrick James Melia (eds) *The Proceedings of PALC 97*, Lodz: Lodz University, 2–22.

— (2003) 'Lexical Priming and the Properties of Text'. Available online at: http://www.monabaker.com/tsresources/LexicalPrimingandthePropertiesofText.htm (accessed 18 March 2011).

— (2005) *Lexical Priming: A New Theory of Words and Language*, London and New York: Routledge.

— (2006) 'Language as Choice: What is chosen?' in Geoff Thompson and Susan Hunston (eds) *System and Corpus: Exploring Connections*, London and Oakville, CT: Equinox, 37–54.

Hunston, Susan (2007) 'Semantic Prosody Revisited', *International Journal of Corpus Linguistics* 12(2): 249–68.

Hunston, Susan and Gill Francis (2000) *Pattern Grammar: A Corpus-Driven Approach to Lexical Grammar*, Amsterdam and Philadelphia: John Benjamins.

Kilgarriff, Adam, *Sketch Engine, Corpus Query System*, Lexcom Ltd. Available online at: http://www.sketchengine.co.uk/ (accessed 8 October 2010).

Laviosa, Sara (2003) 'Corpora and the Translator', in Harold Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam and Philadelphia: John Benjamins, 105–17.

Louw, Bill (1993) 'Irony in the Text or Insincerity in the Writer: The Diagnostic Potential of Semantic Prosodies', in Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair*, Amsterdam and Philadelphia: John Benjamins, 157–76.

— (2000) 'Contextual Prosodic Theory: Bringing Semantic Prosodies to Life', in Chris Heffer and Helen Saunston (eds) *Words in Context: In Honour of John Sinclair*, Birmingham: ELR, 48–94.

Nelson, Mike (2000) 'A Corpus-Based Study of Business English and Business English Teaching Materials'. Unpublished PhD thesis, Manchester: University of Manchester. Available online at: http://users.utu.fi/micnel (accessed 20 January 2008).

Olohan, Maeve (2004) *Introducing Corpora in Translation Studies*, London and New York: Routledge.

Partington, Alan (1998) *Studies in Corpus Linguistics 2. Patterns and Meanings: Using Corpora for English Language Research and Teaching*, Amsterdam and Philadelphia: John Benjamins.

— (2004) ' "Utterly Content with Each Other's Company": Some Thoughts on Semantic Prosody and Semantic Preference', *International Journal of Corpus Linguistics* 9(1): 131–56.

Pearson, Jennifer (2003) 'Using Parallel Texts in the Translator Training Environment', in Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds) *Corpora in Translator Education*, Manchester: St. Jerome, 15–24.

Sinclair, John M. (1991) *Corpus, Concordance, Collocation*, Oxford: OUP.

— (1998) 'The Lexical Item', in Edda Weigand (ed.) *Contrastive Lexical Semantics*, Amsterdam and Philadelphia: John Benjamins, 1–24.

Stewart, Dominic (2009) 'Safeguarding the Lexicogrammatical Environment: Translating Semantic Prosody', in Alison Beeby, Patricia Rodríguez-Inés and Pilar Sánchez Gijón (eds) *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Amsterdam and Philadelphia: John Benjamins, 29–46.

— (2010) *Semantic Prosody: A Critical Evaluation*, New York and Abingdon: Routledge.

Stubbs, Michael (1995) 'Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies', *Functions of Language*, 2(1): 23–36.

— (1996) *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*, Oxford: Blackwell.

— (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*, Oxford: Blackwell.

— (2006) 'Corpus Analysis: The State of the Art and Three Types of Unanswered Questions', in Geoff Thompson and Susan Hunston (eds) *System and Corpus: Exploring Connections*, London and Oakville, CT: Equinox, 15–36.

Thompson, Geoff and Susan Hunston (eds) (2006) *System and Corpus: Exploring Connections*, London and Oakville, CT: Equinox.

Tribble, Chris (2000) 'Genres, Keywords, Teaching: Towards a Pedagogic Account of the Language of Project Proposals', in Lou Burnard and Tony McEnery (eds) *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora* (Lodz Studies in Language), Hamburg: Peter Lang. Available online at: http://www.ctribble.co.uk/text/Genre.htm (accessed 27 June 2007).

Tucker, Gordon (2004) 'A Corpus Linguistic Investigation of the French and English Nouns *war* and *guerre*: The Use of General Reference Corpora in Translation Studies', in Ian Kemble (ed.) *Using Corpora and Databases in Translation: Proceedings of the Conference Held on 14 November 2003 in Portsmouth*, Portsmouth University, 89–106.

Van Leuven-Zwart, Kitty (1989) 'Translation and Original: Similarities and Dissimilarities, I', *Target* 1(2): 151–81.

— (1990) 'Translation and Original: Similarities and Dissimilarities, II', *Target* 2(1): 69–95.

Xiao, Zhonghua and Anthony McEnery (2006) 'Collocation, Semantic Prosody and Near Synonymy: A Cross-Linguistic Perspective', *Applied Linguistics* 27(2): 103–29.

Zanettin, Federico, Silvia Bernardini and Dominic Stewart (eds) (2003) *Corpora in Translator Education*, Manchester: St. Jerome.

Chapter 8

Using Translation and Parallel Text Corpora to Investigate the Influence of Global English on Textual Norms in Other Languages

Juliane House

'Language'– said Edward Sapir some 80 years ago – 'moves down time in a current of its own making. It has a drift.' In this chapter I ask whether a language can today still drift or whether it is being pushed by the one language that is more equal than others: global English. In trying to tackle this question, I deal with translation as the classic meeting point of different languages, and corpus studies as a useful tool. Concretely I wish to show how corpus-based translation studies can be fruitfully used to investigate – both qualitatively and quantitatively – the role that global English as the language with the greatest communicative potential and the most valued symbolic capital, has come to play today in initiating and propelling language change via language contact in translation and multilingual text production. I will first give an overview of a project carried out at the German Science Foundation's Research Centre on Multilingualism and briefly sketch some major results of a series of qualitative case studies based on the multilingual corpus we set up in the framework of this project. Research assistants in this project have been: Nicole Baumgarten, Viktor Becher, Claudia Böttger, Svenja Kranich, Julia Probst, Demet Oezcetin and myself as principal investigator. I then describe some of our quantitative diachronic corpus analyses, discuss some tentative results and develop a set of explanatory hypotheses. In the third and last part of my chapter I address some more general methodological issues arising from such corpus-based work, focusing on the benefits and drawbacks of qualitative and quantitative multilingual translation-related corpus work.

8.1 'Covert Translation' – A Corpus-Based Translation Project Investigating the Impact of Global English on Textual Norms in Other Languages

Globalized and internationalized communication in many areas of contemporary life has led to an increasing demand for texts which are meant for members of different linguistic and cultural communities. Such texts are either parallel texts produced simultaneously in several languages or texts first produced in one language (most frequently English) and later translated 'covertly' into different languages, that is, in such a way as to take account of differences in communicative styles in the source and target language communities. In the literature the term 'parallel text' is used with at least three different meanings: in the first case, text A is an independently produced text in language A, and text B is the translation of A into language B; in the second case, texts A and B are translation equivalent texts that are produced in parallel, that is simultaneously in languages A and B, and in the third case, texts A and B are both original texts in languages A and B, but are comparable in terms of their function, topic, genre and conditions of production. In our project, we use the term 'parallel texts' in the third sense.

A brief remark on the term 'covert translation' is appropriate here. In House (1977, 1981) and House (1997, 2009) I have distinguished between two types of translation: 'overt' and 'covert' translation. While an 'overt translation' is embedded in a new speech event in the target culture, and operates in a new frame, it also simultaneously co-activates the source text alongside the discourse world of the target text. An overt translation is thus a text which serves two masters, the source and the target linguistic-cultural communities. Owing to this dual nature, overt translations are not adapted – or culturally filtered to fit target text conventions. Therefore, an overt translation can never achieve functional equivalence; only a second level functional equivalence is possible.

In contrast, a 'covert translation' is a translation which appears to be, and functions as a second original. The translation is covert because it is not marked pragmatically as a translation, but may, conceivably, have been created in its own right. In a covert translation, the translator creates an equivalent speech event, that is, she reproduces the function the original has within the linguistic-cultural context of the target language. Texts to be translated covertly are of potentially equal concern for members of different cultures, who may have different expectation norms with regard to communicative conventions and textual norms, and it is through the application of a cultural filter that a covertly translated source text is adapted to fit these new norms. Source culture-specific communicative preferences evident in the source text are thus filtered with a view to make them compatible with target textual norms, such that the resulting text gives the reader the impression that this text is *not* a translation but a local text exemplifying a local genre. Cultural filtering requires, of course, reliable information about the relevant language- and culture-specific communicative preferences. This kind of information can be drawn from the findings of contrastive pragmatic discourse analyses for the language pairs involved. With regard to the language pair English/German, a number of dimensions of communicative preferences have been suggested (House 1996, 2002, 2003, 2006). The most important ones of these are the following three:

Dimensions of Communicative Preferences: English–German:

Interpersonal Orientation	<––––––––––→	Content orientation
Implicitness	<––––––––––→	Explicitness
Directness	<––––––––––→	Indirectness

English speakers were found to give preference to an interpersonal orientation, to implicitness and indirectness, whereas German speakers tended to show a tendency towards greater content-orientation, explicitness, directness and the use of ad-hoc formulations. In terms of the Hallidayan meta-functions of language – the ideational and the interpersonal function – German discourse tends to emphasize the ideational function of language whereas in English discourse, equal weight is given to the interpersonal function. These communicative preferences in discourse conventions (supported by other German–English contrastive work by, among others, Heidi Byrnes (1986), Michael Clyne (1987), Monika Doherty (2002) and Catherine Fabricius-Hansen (2004)) are, of course, mere tendencies on a continuum; they do not constitute a clear-cut dichotomy.

Given the dominance of the English language as a global lingua franca in many influential political, economic and cultural contexts, one may reasonably assume that covert translations from English into other European languages, as well as parallel text productions, are influenced by the omnipresence of Anglo-American linguistic-cultural norms. It is this assumption which underlies the project *Covert Translation*, carried out at the German Science Foundation's Research Centre on Multilingualism. The research question of this project is whether target culture-textual conventions in translation and multilingual text production are disregarded so that source and target norms converge. In other words, we are trying to find out whether the obvious, well documented impact that English, as a lingua franca, has on lexical items in other European languages and their local 'shadow meanings' (Chafe 2000), is now also spreading to the more hidden conventions of local text production, such that target text conventions are superimposed by Anglophone ones, with the effect that cultural filtering is discontinued. In the case of English and German texts – which this project initially handled – these adaptations or pragmatic shifts might be located along parameters of communicative preferences. Concretely we have set up the following working hypotheses formulated on the basis of English–German contrastive work:

- a shift from the conventionally strong emphasis in German discourse on the ideational function of language to an Anglophone interpersonal orientation;
- a shift from a conventionally strong emphasis on informational explicitness in German texts to Anglophone inference-inducing implicitness and propositional opaqueness;
- a shift in information structure from packing lexical information densely, integratively and hierarchically to presenting information in a more loosely linearized, sentential way; and
- a shift in word order such that the German *Satzklammer* with its two discontinuous left and right parts give way to more continuous, juxtaposed positions of the two parts.

These hypotheses are tested using a multilingual translation and parallel text corpus, which can be displayed graphically as in Figure 8.1.

This corpus is a dynamic, implicitly diachronic translation and parallel text corpus. It contains texts from three genres: computer instructions, popular science texts and (external) business communication. These genres were selected because they represent areas in which globalization and internationalization processes are arguably most marked. In the first phase of the project work, the texts were prepared for in-depth qualitative analysis, that is they were scanned, transcribed and segmented according to orthographic utterance units. In order to guarantee comparability of the three genres, we have selected on the basis of text-external and text-internal characteristics, a textual stretch functioning as an introduction ('scene setter') for what follows in the body of the text. In the three genres selected, these stretches are: (1) introductory remarks in computer instructions; (2) letters to shareholders, visions and mission statements in economic texts; and (3) opening paragraphs and editorials of popular science texts (e.g. from the journals *Scientific American, UNESCO Courier* and *National Geographic*). For our quantitative analysis we consider the texts in their entirety.

The corpus is made up of three mutually contextualizing parts which will be characterized in what follows. The *Primary Corpus* is a translation corpus and comprises the original English texts and their translations into German. We use this sub-corpus to investigate the translation relation English–German. The *Parallel Corpus* contains English and German authentic texts from the same three genres with comparable topic orientation. We analyse these texts

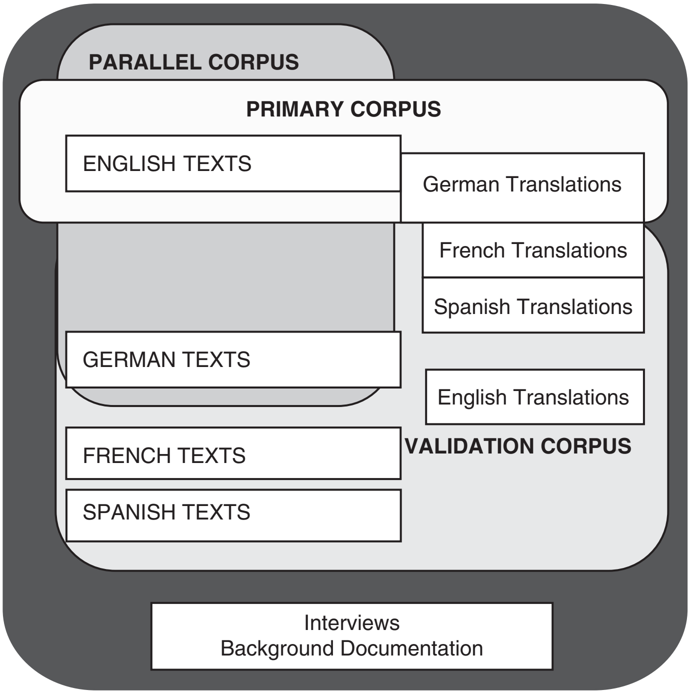


Figure 8.1 The project corpus of the Hamburg 'Covert translation' project: a translation and parallel text corpus

to establish genre-specific and language-specific text norms and conventions, which we then relate to the text norms and conventions that govern the translations in the respective languages. The *Validation Corpus* serves to validate the results of the analyses of texts from the *Primary Corpus* and the *Parallel Corpus*. The *Validation Corpus* holds translations from the same three genres into the opposite direction, that is, from German into English, as well as translations from English into French and Spanish. We analyse this part of the corpus to verify whether the findings of the analysis in the translations and parallel texts are specific for the language pair German and English and the translation relation English–German or not.

At the time of writing this chapter, there were 550 texts in our three subcorpora of about 800,000 words. Over and above these corpora we have also collected a newspaper corpus (English–German parallel editions as well as texts taken from French and Spanish newspapers). Additionally, we have collated background documentation for all three parts of the corpus: documents describing institutional contexts, translational briefs, and so on. To further

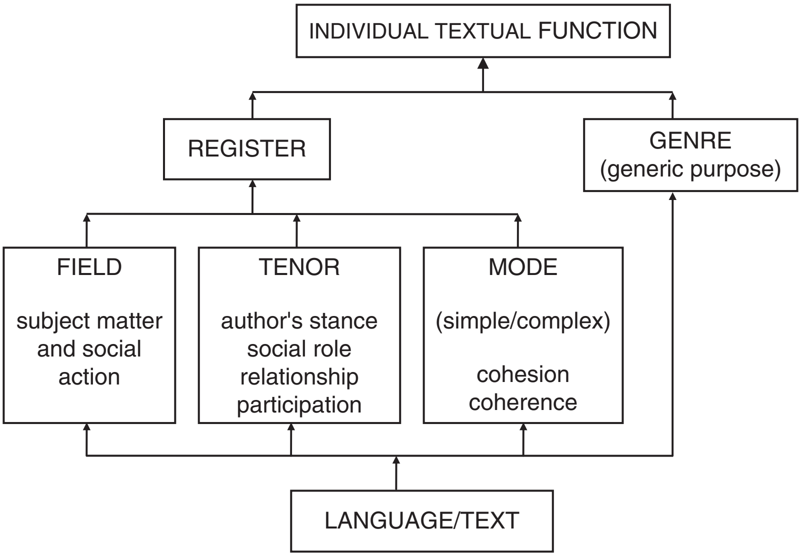


Figure 8.2 The model for analysis: a scheme for analysing and comparing original and translation texts (House 1997, 2009)

validate the results of our analysis, we have conducted in-depth interviews with translators, editors, writers and other persons involved in the text production and reception.

The model for the qualitative analyses in this project is my systemic-functional translation evaluation model (House 1977, 1997 and 2009, see Figure 8.2).

This model is primarily based on Halliday's (1994) systemic-functional theory, but also draws on the Prague School functional stylistics, speech act theory, discourse analysis and contrastive pragmatic analysis. The model provides for a detailed analysis of texts in terms of the three levels of language, register and genre, which are interrelated as follows: The parameters field, tenor and mode are used to open up the frozen textual material such that a particular textual profile, which characterizes the text's function, can be revealed. Once the characteristics of a particular text have been established through detailed textual analysis along the three contextual parameters field, tenor and mode, a particular genre can be described. Genre is a socially established category characterized in terms of the texts' communicative purpose. It links an individual text to a class of texts united by a common communicative purpose and reflects language users' shared knowledge about the nature of texts of the same kind – taking into account both text producer's and recipient's norms of expectation. Analysis and comparison by means of this model allows one to reveal similarities and differences between originals and their translations, and in the case of this project the model can serve as a basis and means for diagnosing changes in German textual norms in selected genres. The claim is that in-depth comparison of linguistic forms in specific, describable contexts and genres can reveal the impact that the English language has on translation and parallel texts.

Analysis and comparison of some 80 English originals and their German translations, as well as 30 German originals in three selected genres (economic texts, popular science texts and computer-related texts) have shown that our hypothesis that cultural filtering is discontinued due to the powerful influence of English textual conventions is essentially *not* confirmed. In other words, we have shown that German translations of English texts of selected genres continue to use a cultural filter. The interpersonal functional component is still conventionally given less prominence in the German texts than is the case in the English originals. Further, the German translations examined are generally more explicit in the description of states of affairs and events giving priority to detailed explication of information content, thus leaving less room for inferencing and showing – in comparison to the English originals – a less strongly marked addressee-orientation. The lexico-syntactic changes in the German texts hypothesized to accompany potential changes in textual conventions (information distribution and word order) were also not noticeable. These results are generally supported by our interview data elicited from authors, translators, editors and translation commissioners.

One explanation for the fact that our working hypotheses were *not* confirmed might be that in translations, language contact takes place in the heads of highly professional, linguistically trained experts. As qualified professionals, such persons tend to be very aware of the type of contrasts and similarities between the two language systems. As a consequence, they will positively avoid interference in the act of translation – a behaviour supported by our interview data. If this assumption is correct, one might hypothesize a potentially stronger influence of the English language over time in the parallel texts, as these 'innocent' monolingual texts are not affected by the translation relation. Analyses of German and English parallel texts, which would test this hypothesis are being undertaken (for a detailed description of the project, see Baumgarten et al. 2004).

However, to make matters more complicated, analyses of several popular science and economic texts *do* point to a shift in the weight given to the interpersonal function in the German texts, in particular with regard to expressing (authorial) 'stance' (Biber et al. 1999), 'subjectivity', 'perspective' and 'point of view' (Carlota Smith 2002, 2003) as well as addressee-orientation and involvement (Nuyts 2001). No major conclusions about these findings can be drawn as yet, and neither can Anglophone influence on German textual norms be diagnosed. Suffice it to say that we may be at a stage of transition, where German textual norms are beginning to be adapted to Anglo-American norms in the process of translation. For instance, in modern letters to shareholders we discovered a tendency towards imitating the Anglo-American originals in the German translations in terms of integrating more narrative elements into the texts instead of simply reporting facts, states of affairs and development, as used to be the case in this particular German genre. We might therefore be faced with a new type of 'genre-mixing' (Böttger 2004). And in modern popular science texts, too, there is a move away from the usual strict 'scientificness' that used to characterize German popular science texts, towards – admittedly in a moderate way – a greater degree of popularization in the sense of 'infotainment' and 'edu-tainment'. The German popular science texts may thus be in a process of becoming more person-oriented, that is more similar to comparable English texts in terms of the linguistic expression of an authorial stance and audience design or addressee involvement. Genre-mixing, too, seems to have begun creeping in to modern German popular science texts, that is, these texts show elements known to be part of journalistic texts and advertisements, for example, heavier use of personal deixis, and the simulation of an interaction between author and reader via mood switches and particular ways of framing the text – phenomena previously found primarily in English popular science texts. While the results of our analyses are also supported by evidence from narrative interviews with translators, editors and other persons relevant to the translation activity, it is essential that we follow up our case studies with quantitative and more consistently diachronic analyses. It is this quantitative diachronic work which is presented in the second part of this chapter.

8.2 Quantitative Analysis of a Special-Domain (Popular Science) Translation and Parallel Text Corpus

On the basis of the results of our qualitative textual analyses, we conducted quantitative and diachronic corpus analyses in order to test our project hypotheses. We started with the popular science sub-corpus and the testing of the first hypothesis mentioned earlier, that is, that there is a different weighting of the interpersonal functional component in German translations over time due to an influence of Anglophone norms on German texts. We decided to first examine the domain of subjectivity. For this, we had to first clarify the concept of subjectivity in order to be able to operationalize it in order to enable quantitative corpus research.

Subjectivity plays an important role in how meaning is construed. Linguistic forms which involve subjectivity can be defined with Lyons (1977: 739) as 'devices whereby the speaker, in making an utterance, simultaneously comments upon and expresses his attitude to what he is saying'. Subjectivity therefore has to do with the expression of a particular attitude, a particular point of view or perspective. It can be characterized as a semantic-pragmatic process whereby meanings become increasingly based on a speaker's affect and attitude towards the proposition. It is also related to Biber's (1988) and Biber et al.'s (1999) notion of stance as well as Chafe's (1994) and Nuyts' (2001) notion of an evidential dimension in discourse. Of the three functional-semantic components postulated in Hallidayan theory, namely the ideational, the interpersonal and the textual, we are here primarily concerned with the interpersonal component. And in looking more closely at the phenomenon of subjectivity, we are at the same time concretizing the above mentioned rather vague project hypothesis (that there is a shift in the weighting of the interpersonal and the ideational function) by pinpointing the interpersonal as subjectivity and addressee orientation (in a later project step). In hypothesizing a shift (over time) towards a more subjective presentation of states of affair in discourse, we are also in line with the assumptions put forward by adherents of grammaticalization theory that language change tends to occur in the direction of categories that are related to the speaker/writer themselves and their relationship to the utterance. But how is subjectivity realized linguistically? We are certainly aware of the fact that – as Carlota Smith (2003) demonstrated – subjectivity encompasses a dazzling multitude of different linguistic phenomena which interact. One can, however, start with an ensemble of at least the following linguistic makers:

1. modal auxiliaries
2. sentence adverbials
3. matrix clauses
4. personal deixis
5. deictic/phoric procedures (pronominal adverbials or composite deictics)
6. modal particles
7. mental state predicates
8. parenthetical commentaries
9. mood switches
10. modal adjectives and nouns.

This list of subjectivity-indexing phenomena (which is far from complete) shows that MODALITY plays a central role in indexing subjectivity, and with it, of course, the classic distinction between root, agentive, non-deictic or deontic modality expressing forms on the one hand and forms expressing epistemic or deictic modality on the other. At first glance, one would assume that it is exclusively epistemic modality which one would need to examine with regard to the phenomenon of subjectivity as it expresses speakers' involvement in the utterance. I do not wish to discuss the complex area of modality here. It is sufficient to say that we follow a rather broad conception of modality as expounded, for instance, by Radden (1999). In other words, we go beyond viewing epistemic meanings as capturing mainly necessity and possibility, and deontic meanings capturing mainly obligation and permission. In a wider sense, deontic modality also comprises functions such as desirability, determination and assumption. It is this wider sense, then, which we follow in our work positing that deontic modality also captures subjectivity, in that it too is related to non-factual events, to speakers' attitudes and decision processes in a wider sense of the evaluation of the propositional content. All modal verbs, as 'grounding predicates', can in this view be interpreted as providing a shift from an objective to a subjective presentation of states of affairs, where the speaker is involved in the utterance offstage, and this is also documented from an evolution of the modal verbs from originally lexical verbs to deontic and epistemic modal verbs. While epistemic modality is centrally related to a speaker's assessment of possibility and probability, and root modality is centrally related to permission and obligation, the latter also encodes speakers' commitment to the necessity or permissibility of an action – and the lines are blurred, especially as the same linguistic forms are used for both. Thus, for the purpose of the project analyses of subjectivity, both root and epistemic modality count as they can both be described as capturing subjective performative functions.

In German, modality as an independent subjective resource tends to be realized to a much higher degree than in English via modal adverbs, and in particular modal particles (*Modalpartikeln*) – a typical feature of the grammar of German. While more frequent in spoken discourse, they are also used in written discourse. German thus provides alternative means of expressing modality, that is, the responsibility for expressing is split up over different competing linguistic means. Modal particles do not exist in English, in oral discourse they are often rendered via intonation (cf. Salkie 2002 for a contrastive analysis of German and English modal verbs).

In order to trace the influence of English on German texts diachronically through translational language contact over time, a quick review of the results of existing diachronic corpus studies in the domain of modality or subjectivity is in order. Thus, for instance, Krug (2000) found a decline of the use of central modal verbs (*can, must, may, might,* etc.) in terms of frequency, and an increase of the semi-modals such as *be to, be going to, have to, have got to* and *want to.* Similarly, Leech (2003) claimed in a diachronic quantitative study that central English modal verbs are declining in both American and British English with semi-modals gradually gaining ground. Leech explains this decline with reference to the current processes of Americanization, colloquialization and democratization leading to a less authoritarian role by the speaker/writer. Leech's findings as well as Mair's (1997) are based on corpus evidence from major written and spoken British and American English corpora (The *LOB*, *FLOB*, *BROWN* and *FROWN* corpora). Nicholas Smith (2003) also confirms this decline of the use of central modal verbs, particularly in the obligation/necessity domain in *LOB* and *FLOB*. He explains the decline of *must* by pointing to a trend towards eliminating overt power markers and an increasing emphasis on equality of power, or at least an appearance of an equality of power. With reference to the use and distribution of *may* in the *International Corpus of English* (ICE), Facchinetti et al. (2003) showed that *may* is the most frequently used epistemic modal verb in scientific and popular scientific texts – *may* being used to tone down, and to withhold complete commitment, thus allowing information to be presented as an opinion, rather than as a fact.

Returning to our project's quantitative diachronic analysis of the phenomenon of subjectivity, we have investigated the use and distribution of the following phenomena: modal verbs, sentence adverbials, speaker–hearer deixis, composite deictics and modal particles. Our research question governing this diachronic quantitative work was to find out whether these different subjectivity phenomena were realized in German over time in such a manner that it reveals English influence. Thus modal verbs would increase at the expense of modal particles. Sentence adverbials, speaker–hearer deixis would increase following the Anglophone norm model, and typically German grammatical phenomena such as modal particles and composite deictics would decrease because comparable devices do not exist in English.

The counts were carried out in five subsets of our popular science translation and parallel text sub-corpus:

1. English original texts from 1978 to 1982 (42,497 words)
2. The German translations of these texts (37,830 words)
3. German original (monolingual) texts from 1978 to 1982 (82,480 words)
4. English original texts from 1999 to 2002 (122,866 words)
5. The German translations of these English texts (113,420 words)
6. German original (monolingual) texts from 1999 to 2002 (100,648 words).

All texts examined in our quantitative analysis are published texts, and all the texts have appeared in the popular scientific journals *Scientific American, New Scientist* and *Spektrum der Wissenschaft.* All counts have been normalized on the basis of 10,000 words. The corpus is fully tagged, but not yet completely aligned. This means that our analyses do not yet include an analysis of translation relations, rather we have looked at the sub-corpora separately, and the results to be described in what follows are at present simple frequencies based on concordancing values.

Next, we present some of our tentative results with regard to modal verbs, sentence adverbials, speaker–hearer deixis, composite deictics and modal particles.

8.2.1 Modal Verbs

As opposed to the findings by Krug, Mair, Leech and others described above that central modal verbs in English are declining, the quantitative analysis of

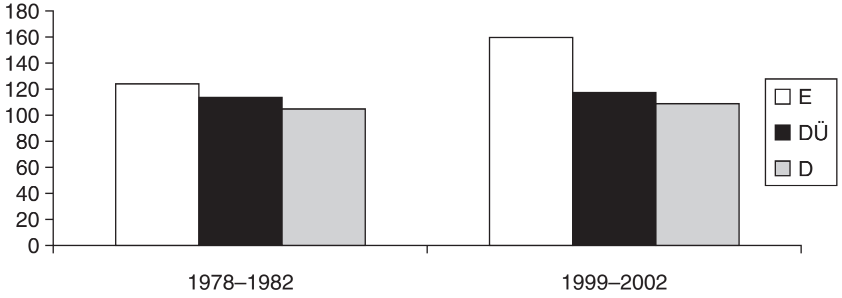


Figure 8.3 Modal verbs (E = English, DÜ = German Translation, D = German).

our small, special domain translation and parallel text corpus of English and German popular science texts, has shown that there is an increase in the use of central modals verbs in the time spans 1978–1982 and 1999–2002 in both languages (see Figure 8.3).

An explanatory hypothesis might be that this particular genre, popular scientific texts, lends itself to subjectivity-marking in that scientific facts and findings are popularized, that is evaluated and set into perspective for the lay reader who is made to see these facts through the eyes of either a science writer or a scientist in the role of commentator on their own research. Interestingly, the exception from this increase in the English sub-corpora is the use of *must* and *should,* both of which decline, in German it is *müssen* which declines. The decline of *must/müssen* and *should* in the English popular science texts might be explained with reference to what Leech (2003) called 'the global processes of democratization and colloquialization', which would leave no room for strong obligation or logical conclusion modals where participants are set up as in authority. In the German translations, *müssen* (the modal with the second highest frequency) has shrunk over time in a similar way to the English originals, with *sollen* used with increasing frequency in its stead. It is impossible to say on the basis of our data whether this is so because of the trend observed in the English texts, or because of a general trend towards democratization.

The German modals *dürfen* and *mögen* (*may/might, will*) – both used predominantly as epistemic and thus strongly subjective devices – occur more frequently in the German originals, and their use in the translations increases over time – an interesting feature given the concomitant decline of the more forceful authoritarian *müssen/must,* and the markedness of the modal *mögen* as formal, written and semi-archaic. We suggest that given the frequency of *may/might/will* – the most likely English equivalents that are used with increasing frequency in the English originals – an influence of English on German should not be excluded. A closer look at the uses of *mögen* in later German translated texts shows, however, that *mögen* is often used in direct quotations or in the formulaic use of routinized structures, collocating with, for example, *was (auch) immer* (whatever). Thus:

Was immer Besonderes uns auszeichnen *mag* . . .
Was auch immer die ursprüngliche Gestalt sein *mag* . . .

This may be interpreted as a sign that the use of *mögen*, a rather rare *old-fashioned* modal as contemporary German usage goes, as translation equivalents of may/might/will is limited to a rather restricted formal environment. In the original German texts, however, *mögen* is not used inside formulaic structures.

Taken together, the findings in the area of modal verbs are ambiguous between Anglophone influence on the German translation and – more likely – the impact of powerful global trends towards colloquialization and democratization not only in the field of modality but in Western discourse generally.

8.2.2 Modal Particles

Modal particles can be interpreted as lexico-grammatical devices that implicitly realize the speaker/writer in the text. They form a rich and variegated system in German, and comprise such little but pragmatically potent words as *ja, doch, denn, eigentlich, wohl, etwa, schon, nur, vielleicht* and so forth. Modal particles have little stable lexical meaning themselves. We found that in our data, modal particles provided a sort of *entourage* for the modal verbs in the German corpora acting like satellites additionally upgrading or downgrading the modal values of the verbs with which they co-occur.

The use of modal particles, a particularly German semantic-pragmatic device, has substantially increased in both the translations and the parallel texts over time (see Figure 8.4). Surprisingly, the older German original texts featured fewer occurrences of modal particles than the older German translations, and there is a marked increase in the newer German translations.

As concerns the tokens of the modal particles used in the different sub-corpora, the more recent German translations also feature a much greater variety of modality markers than the earlier ones. This might be interpreted as a sign of a healthily productive indigenous system. In this domain, then, which is very important for the expression of subjectivity in German – German translations remain *very German* texts indeed. And this result also means that the German translations have become more openly subjective – presumably as a reflection of the type of global communicative trend mentioned before.

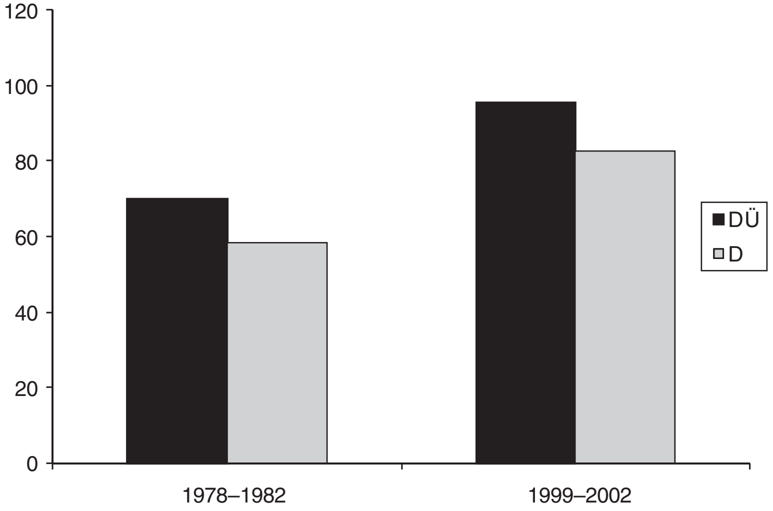


FIGURE 8.4 Modal particles (DÜ = German Translation, D = German).

8.2.3 Speaker–Hearer Deixis

There is a marked increase in the use of speaker–hearer deixis in both English and German (see Figure 8.5).

The increase in the use of speaker–hearer deixis is particularly striking in the case of the first person plural personal pronoun, *we/wir*, with possessive determiners *my/mein* and in the English texts also with the plural possessive determiner *our*. The presence of human participants in the text – as overt marking of subjectivity – has, therefore, substantially increased in both English originals and German translations. In the German originals, nearly half of the cases of personal deixis occur inside direct quotations, co-occurring with mental and verbal processes:

Ein eingefügtes 'Äh' oder 'Soll *ich* mal sagen, dass . . . '
Das wird von dem Meteorologen Wladimir Koeppen vom 6. November 1911 so formuliert: '*Ich* glaube doch, Du hältst meinen Urkontinent für phantastisch'

Outside of such quotations, the personal pronoun *we* is used impersonally, synonymously with *one* in the German originals in 61 cases out of 152:

Heute wissen *wir*, dass die Flugsaurier dem großen Sauriersterben an der Wende . . .
Müssen *wir* aber überhaupt zwischen den beiden entgegengesetzten Erklärungen entscheiden?

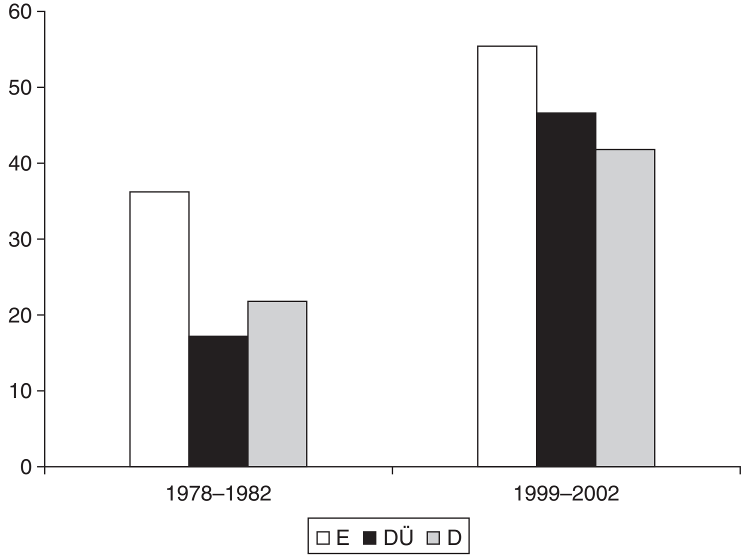


FIGURE 8.5 Speaker–hearer deixis (E = English, DÜ = German Translation, D = German).

In the older German translations, we found 22 such impersonal uses of *wir* out of 50 instances. By contrast there is a comparatively low incidence of such impersonal uses of *we* in the older English originals, that is 13 out of 117.

This personal deictic use of *we* collocates in both the early and – much more marked – the later English originals, with reference to collective entities, that is, *I*, *we*, and particularly *my/our* frequently co-occur with nouns such as group, colleagues, research team, laboratory and doctoral students as for instance in:

More recently *my* colleagues and *I* at the Centre Nationale de la Recherche Scientifique . . .
The system *my* group at university college London uses . . .
My colleagues and *I* (then at Clark University) devised the following experiment . . .
The most interesting findings have come from experiments in which *we* and *our* colleagues produced mouse chimeras.

There is no such embedding of individuals in a collective identity in the older German translations and the older German originals. However, in the newer German translations this has changed. True, here we also find many direct

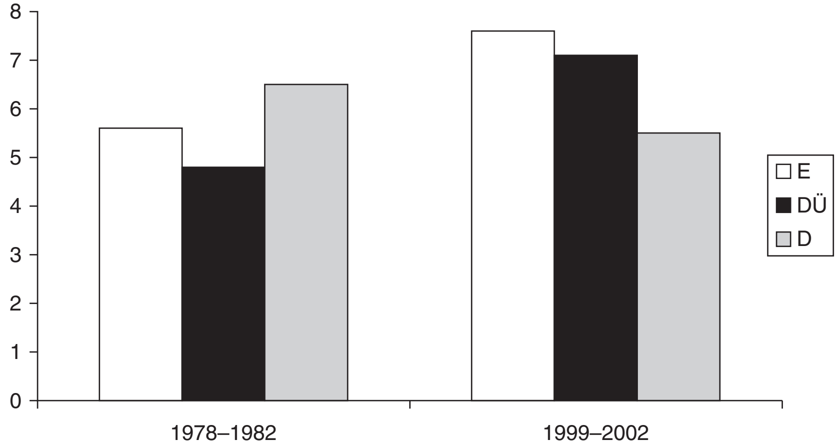


Figure 8.6 Sentence adverbials (E = English, DÜ = German Translation, D = German)

quotations with personal deixis, as well as personal anecdotes and narrative inserts forcing as it were, the use of personal deixis, but we also find clusters of collective nominals: *meine Gruppe, unsere Doktoranden, meine Mitarbeiter* and so on. The use of speaker–hearer deixis thus seems to be a good candidate for Anglophone influence, but it may also be the case that this usage simply reflects global changes in the research culture or it may be that global Anglo-Saxon ways of conducting and reporting on research have influenced research in other countries.

8.2.4 Sentence Adverbials

The use of sentence adverbials increases over time in both the English originals and the German translations, but decreases in the German parallel texts (see Figure 8.6).

Given liberal German word order rules, one might assume that sentence-initial positions as stylistically marked cases in German would increase under the influence of the English language, where such extrapositions are a less marked stylistic option. However, first, sentence adverbials occur more frequently in the older German originals than in the older German translations. Secondly, of the three types of sentence adverbials: modal (*obviously, disturbingly,* etc.), locally structuring (*consequently, accordingly,* etc.) and globally structuring (*finally, firstly,* etc.), it is only modal sentence adverbials which seem to imitate the English trend (but the increase in German is higher than it is in English). Interestingly, German globally text-structuring adverbials resist English influence – a result which confirms our qualitative analyses and also reminds one of

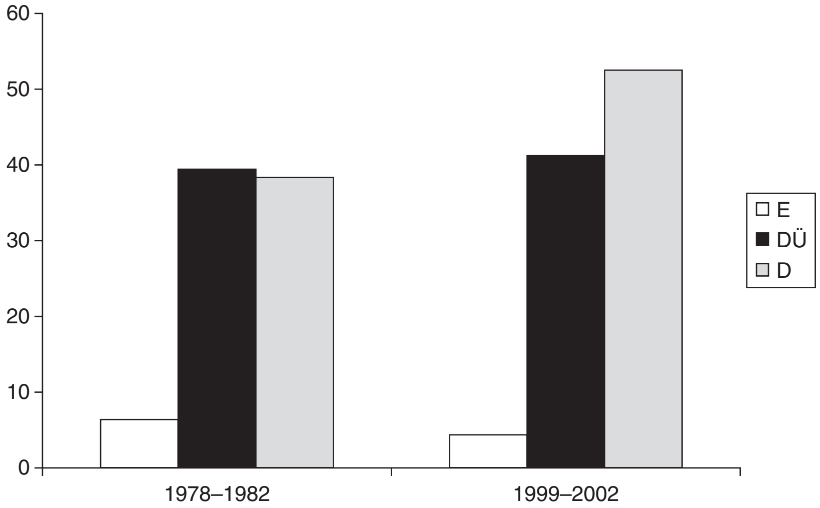


Figure 8.7 Composite deictics (E = English, DÜ = German Translation, D = German).

the results of Michael Clyne's (1987) classic contrastive German–English discourse studies. Overall, the findings with regard to sentence adverbials cannot be taken as supporting the thesis of a direct impact of English textual norms on German norms, but they do point to an overall trend towards an increased expression of speaker's perspective, of subjectivity.

8.2.5 Composite Deictics (i.e. Pronominal Adverbials or Phoric/Deictic Devices)

The occurrence of this particularly German lexico-grammatical textual phenomenon (cf. Rehbein (1995), who calls these devices 'zusammengesetzte Verweiswörter') has slightly increased from the older German originals to the older German translations to the newer translations, and has substantially increased from the older German originals to the newer ones (see Figure 8.7).

German composite deictics are words with tokens like *dazu, dabei, damit, hierzu* and *hieran*. They function as mental signposts guiding readers through the text, summarizing portions of the text for them, accumulating and condensing textual stretches as well as commenting on textual information for them. Since there is no English equivalent for these devices, direct influence from the English original texts to the German translations can be ruled out.

From the older translations to the newer ones we can detect recourse to a greater variety of these devices. This is over and above the 30 different tokens of composite deictics which we found in the older German originals and translations, namely:

dadurch, dagegen, daher, dahinter, damit, danach, darauf, daraus, darin, darüber, darum, darunter, davor, dazu, dementspechend, demnach, deshalb, hieraus, hierbei, hierfür, hierzu, stattdessen, wobei, wodurch, woher, worauf, woraus, worin, wovon, wozu.

The newer translations feature an additional fourteen different tokens, namely:

daneben, dazwischen, demgegenüber, demgemäß, demzufolge, deswegen, hieran, hiermit, hierin, hierüber, hiervon, wofür, womit, wonach.

This added variety in the case of this particular brand of German connectives clearly speaks against an intrusion into German textual norms by English conventions, but rather suggests a certain resistance, or at least a conservatism of local German norms.

8.3 Results and Discussion

What are we to make of these results? Obviously, they do *not* unambiguously confirm or disconfirm our hypotheses, that is, the results of our investigation of the different domains of subjectivity cannot be said to support our hypotheses that there is a direct influence of English on German texts in terms of a greater weight of the interpersonal function through the process of translation.

With respect to the use of modal verbs, we found a greater increase in the occurrence of modal verbs in English than in German. By contrast, the use of modal particles as a particularly German way of expressing modality has considerably increased over the two time frames examined. Also, we found a greater variety in the use of modal particles in German.

In the domain of speaker–hearer deixis, we found a remarkable increase in its occurrence in both English and German, and thus potentially a strong increase in overtly marked subjectivity.

With regard to the other fields of expressing subjectivity, that is sentence adverbials and composite deictics, the situation is unclear. The use of sentence adverbials in the German texts did not unambiguously follow the English model, but there is an increase in a particular type of sentence adverbials, namely modal sentence adverbials, and this may indicate a rise in subjectivity. The occurrence over time of composite deictics can be interpreted as a gain in subjectivity, but it cannot be taken as an indication of direct English influence.

Taken together, while some of our preliminary findings support our hypothesis of direct English influence via translation, others disconfirm it, and still others may be interpreted as supporting a general shift in discourse towards marking subjectivity to a greater degree. In order to come to terms with these fuzzy results, and in order to see more clearly what their relevance is both for the research question of our project and for the study of translation as a locus of contact between hegemonic English and other languages, I suggest the following three explanatory hypotheses or models:

Model 1: Translation as mediator of the English take-over: the translation process effects change. Here translation is in fact the locus of change: translation as a means of language contact aids the original influence of its translations. This would be the case where texts formerly lending themselves to being covertly translated are no longer culturally filtered but anglicized, their own norms and conventions being eclipsed and appropriated.

Model 2: Universal impact of globalization: translation as reflector of change and not instigator thereof. The translational process reflects change, as agents in the service of globalization and international Zeitgeist. Despite its role as locus classicus of language contact, it is not through translation that norms in the L2 change, it is rather through the overpowering and omnipresent presence of hegemonic English that changes in textual conventions arise.

Model 3: Translation as cultural conservation: the translational process resists change. This is the type of hypothesis I suggested earlier with regard to the bulk of our qualitative case studies where there was no change in the rites of covert translation, that is, translators as professionals, as language experts who are highly aware of contrasts and differences between languages and thus preserve the norms of the target text genre.

8.4 Using Corpus Analysis in Translation Studies: Towards a New Paradigm?

In conclusion, I want to make a few general remarks about corpus translation studies and qualitative and quantitative analysis, and suggest a combination of the two. In the case of the work presented above this would mean, for instance, to link our quantitative results to the type of richly contextualized translation analysis and comparison of our case studies.

A corpus in translation studies has a useful function as *one* of many tools of scientific inquiry. Regardless of frequency and representativeness, corpus data are useful because they are often better data than those derived from accidental introspections, and for the study of certain problems such as overall development of the use of modal verbs, corpus data are indeed the only available data. But if the use of corpora is to maximally fulfil its potential, it should be used in conjunction with other tools, that is, introspection, observation, textual and ethnographic analysis. In translation studies, as in other disciplines, we must assess the relative value of the analytical–nomological paradigm on the one hand, where already existing hypotheses (and categories) are to be confirmed or rejected, and where variables are explicated and operationalized, and the explorative–interpretative paradigm on the other hand, where in-depth case studies are conducted to develop categories for capturing newly emerging phenomena. It is important that these two lines of inquiry, the qualitative and the quantitative, are not considered to be mutually exclusive, rather they should be regarded as supplementing each other.

Corpus evidence, and especially impressive statistics, should not be seen as an end in itself, but as a starting point for continuing richly (re)contextualized qualitative work with values one finds interesting – and these must not necessarily be the most frequent phenomena, for the least frequent values can also catch one's attention. In the last analysis, the object of corpus translation studies should not be the explanation of what is present in the corpus, but the understanding of translation. The aim of a corpus is not to limit the data to an allegedly representative sample, but to provide a framework for finding out what sort of questions should be asked about translation and about language used in different ways. The value of corpus translation studies lies in how it is used. Corpus translation studies is not a new branch of translation studies, it is simply a methodological basis for pursuing translation research. In principle, it should be easy to combine corpus translation studies with many other traditional ways of looking at translation. If this is done, corpus translation studies can greatly enrich our vision.

References

Baumgarten, Nicole, Juliane House and Julia Probst (2004) 'English as Lingua Franca in Covert Translation Processes', *The Translator* 10(1): 83–108.

Biber, Douglas (1988) *Variations across Speech and Writing*, Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (1999) *Longman Grammar of Spoken and Written English,* London: Longman.

Böttger, Claudia (2004) 'Genre Mixing in Business Communication', in Juliane House and Jochen Rehbein (eds) *Multilingual Communication*, Amsterdam and Philadelphia: John Benjamins, 115–32.

Byrnes, Heidi (1986) 'Interactional Style in German and English Conversations', *Text* 6: 89–106.

Chafe, Wallace (1994) *Discourse, Consciousness and Time*, Chicago: University of Chicago Press.

— (2000) 'Loci of Diversity and Convergence in Thought and Language', in Martin Pütz and Marjolyn Verspoor (eds) *Explorations in Linguistic Relativity*, Amsterdam and Philadelphia: John Benjamins, 101–24.

Clyne, Michael (1987) 'Cultural Differences in the Organization of Academic Texts: English and German', *Journal of Pragmatics* 11: 211–47.

Doherty, Monika (2002) *Language Processing in Discourse. A Key to Felicitous Translation,* London: Routledge.

Fabricius-Hansen, Catherine (2004) 'Paralleltext und Übersetzung in sprachwissenschaftlicher Sicht', in Harald Kittel, Armin Paul Frank, Norbert, Greiner, Theo Hermans, Werner Koller, José Lambert and Fritz Paul (eds) *Übersetzung. Translation. Traduction. Ein internationales Handbuch zur Übersetzungsforschung. An International Encyclopedia of Translation Studies. Encyclopédie Internationale de la Recherche sur la Taduction,* Berlin: Mouton de Gruyter, 322–9.

Facchinetti, Roberta, Manfred Krug and Frank Palmer (eds) (2003) *Modality in Contemporary English*, Berlin: Mouton de Gruyter.

Halliday, M. A. K. (1994) *An Introduction to Functional Grammar,* London: Arnold.

House, Juliane (1977/1981) *A Model for Translation Quality Assessment,* Tübingen: Narr.

— (1996) 'Contrastive Discourse Analysis and Misunderstanding: The Case of German and English', in Marlis Hellinger and Ulrich Ammon (eds) *Contrastive Sociolinguistics,* Berlin: Mouton, 345–61.

— (1997) *Translation Quality Assessment*: *A Model Revisited,* Tübingen: Narr.

— (2002) 'Maintenance and Convergence in Covert Translation English-German', in Bergljot Behrens, Cathrine Fabricius-Hansen, Hilde Hasselgård and Stig Johansson (eds) *Information Structure in a Cross-Linguistic Perspective,* Amsterdam: Rodopi, 199–213.

— (2003) 'Misunderstanding in Intercultural University Encounters', in Juliane House, Gabriele Kasper and Steven Ross (eds) *Misunderstanding in Social Life. Discourse Approaches to Problematic Talk*, London: Longman, 22–56.

— (2006) 'Communicative Styles in English and German', *European Journal of English Studies* 10: 249–67.

— (2009) *Translation,* Oxford: Oxford University Press.

Krug, Manfred (2000) *Emerging English Modals: A Corpus-Based Study of Grammaticalization,* Berlin: Mouton de Gruyter.

Leech, Geoffrey (2003) 'Modality on the Move: The English Modal Auxiliaries 1961–1992', in Roberta Facchinetti, Manfred Krug and Frank Palmer (eds) *Modality in Contemporary English,* Berlin: Mouton de Gruyter, 223–40.

Lyons, John (1977) *Semantics*, vol. 2, Cambridge: Cambridge University Press.

Mair, Christian (1997) 'Parallel Corpora: A Real-time Approach to Language Change in Progress', in Magnus Ljung (ed.) *Corpus-Based Studies in English,* Amsterdam: Rodopi, 195–209.

Nuyts, Jan (2001) 'Subjectivity as an Evidential Dimension in Epistemic Modal Expressions', *Journal of Pragmatics* 33: 383–400.

Radden, Günter (1999) 'Modalverben in der Kognitiven Linguistik', in Anglika Redder and Jochen Rehbein (eds) *Grammatik und Mentale Prozesse,* Tübingen: Stauffenburg, 261–94.

Rehbein, Jochen (1995) ‘Über zusammengesetzte Verweiswörter und ihre Rolle in argumentierender Rede’, in H. Wohlrapp (ed.) *Wege der Argumentationsforschung,* Stuttgart: Fromman-Holzboog, 166–98.

Salkie, Raphael (2002) ‘Probability and Necessity in English and German’, in Bergljot Behrens, Catherine Fabricius-Hansen, Hilde Hasselgard and Stig Johansson (eds) *Information Structure in a Cross-linguistic Perspective,* Amsterdam: Rodopi, 81–96.

Sapir, Edward (1949 [1921]) *Language,* London: Harvest Books.

Smith, Carlota (2002) ‘Perspective and Point of view: Accounting for Subjectivity’, in Bergljot Behrens, Cathrine Fabricius-Hansen, Hilde Hasselgard and Stig Johansson (eds) *Information Structure in a Cross-linguistic Perspective.* Amsterdam: Rodopi, 63–80.

— (2003) *Modes of Discourse*, Cambridge: Cambridge University Press.

Smith, Nicholas (2003) ‘Changes in the Modals and Semi-modals of Strong Obligation and Epistemic Necessity in Recent British English’, in Roberta Facchinetti, Manfred Krug and Frank Palmer (eds) *Modality in Contemporary English,* Berlin: Mouton de Gruyter, 241–66.

Part III

Studies in Specific Sub-Fields

Chapter 9

Off the Record and On the Fly: Examining the Impact of Corpora on Terminographic Practice in the Context of Translation

Lynne Bowker

> *. . . the use of corpora in terminography work will no doubt raise a host of new issues. One of the most important will be the effects of the new tools on the terminographer's traditional work methods.*
>
> *Meyer and Mackintosh (1996: 278)*

In the words of Anobile (2000: vii), 'the transformations in the language industry following the groundswell of globalized trade are nothing short of revolutionary'. Globalization is a driving force behind the dramatic increase in the demand for translation, and one of the by-products of this has been a renewed interest in the potential of technology for helping translators to work more efficiently.

When new technologies are first introduced into a profession, it is not possible to know immediately what the long-term effects will be. It is only with time that the impact of applying a given technology can be observed. Electronic corpora and associated tools for corpus processing have now been in widespread use in the translation profession for approximately fifteen years. Therefore, it is a good time to look back over the use of these tools during this period, and to reflect on the changes that they have brought about. By assessing and understanding these changes, we will be in a better position to develop strategies for moving forward, such as designing new types of resources for translators or making modifications to the curricula of translator training programmes in order to reflect the new reality of the translation profession.

This chapter will assess the impact that corpus-based resources and tools have had on terminological research and particularly on the terminographic practices of the terminologists and translators who use them. For the purpose of this chapter, corpus-based resources and tools include not only collections of electronic texts that can be interrogated with the help of tools such as concordancers and word listers, but also pairs of texts that have been aligned and stored in specially designed databases for processing by tools such as translation memory systems or automatic term extractors. Moreover, although terminology work can be carried out in monolingual contexts, the focus here will be on terminology research carried out to meet the needs of translators.

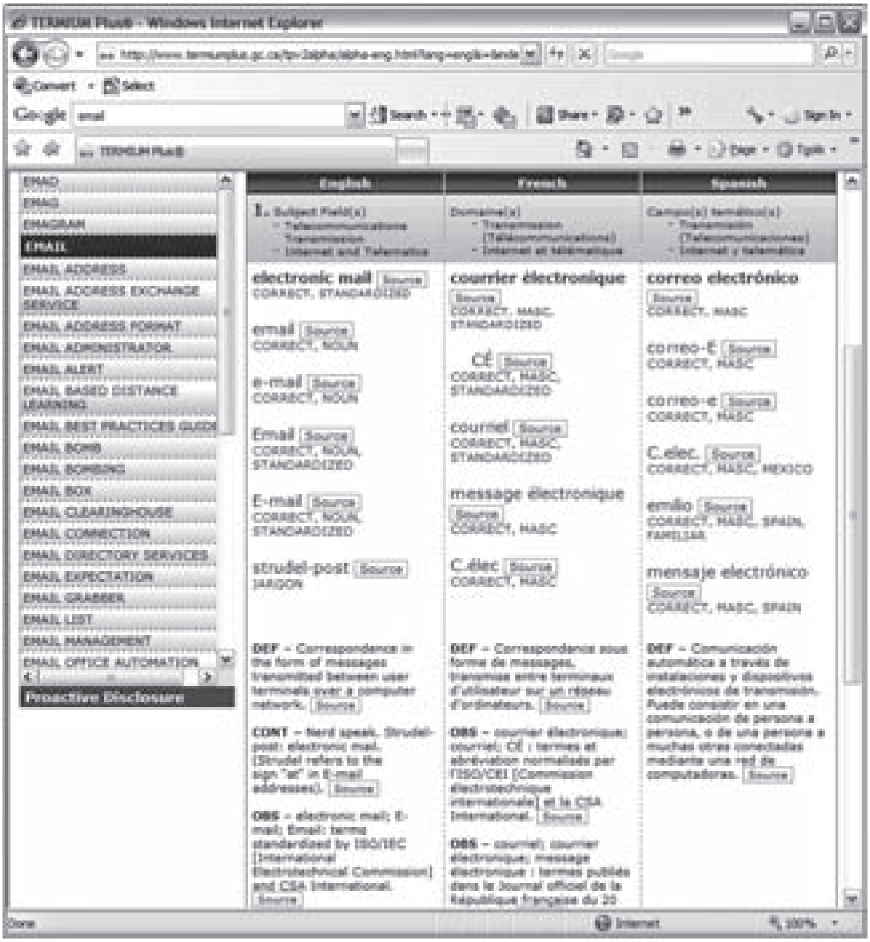
The chapter is divided into four main sections. Section 9.1 briefly presents the discipline of terminology proper and investigates the work carried out by terminologists, looking specifically at how technology has been integrated into this profession. In Section 9.2, the attention shifts to the use of corpus-based technologies by translators, who not only use the terminological resources (e.g. term banks) produced by terminologists, but who also engage directly in terminological research in order to find solutions to translation problems that they encounter. Section 9.3 contains a comparative discussion of the impact of corpora on the process and the product of terminological research as carried out by terminologists and by translators. Finally, Section 9.4 concludes by suggesting that, since the use of corpus-based tools and resources has resulted in a divergence in terminographic practices in these two professions, it may be time to consider modifying the nature of terminology training offered to translators.

9.1 Terminologists and Term Banks

Terminology is generally understood to refer to 'the scientific study of the concepts and terms used in specialized language' (Pavel and Nolet 2001: xvii). Terminologists typically engage in what is known as thematic or subject field research, where they attempt to map out all the concepts in a subject field and to provide detailed information on these concepts, including terms and their equivalents, textual support and basic usage information. Once this data has been gathered, analysed and processed, the end result is usually a terminological resource such as a term bank.

Term banks were among the earliest computer applications in the language fields. The first was *Eurodicautom* (now known as *InterActive Terminology for Europe (IATE)*) in 1963, and others soon followed, including the *Banque de terminologie du Québec* (now known as *Le Grand dictionnaire terminologique (GDT)*) in 1969, and *TERMIUM* in 1974 (Rondeau 1984: 160). Their primary purpose was to act as a repository for storing and disseminating terminological data.

The basic unit of information in a term bank is the **term record**, which contains information about terms and the concepts they represent. As described by Pavel and Nolet (2001: xix), 'all of the collected information is analyzed, filtered, structured, and condensed into a terminological record'. Many templates have been proposed, but most bi- or multilingual term records allow for at least the following basic types of information to be recorded in each language: domain, entry term, grammatical information, synonyms, definition, context, observations, usage information (e.g. register, regional restrictions) and sources. Figure 9.1 shows a sample term record from *TERMIUM*. One important thing to note is that, as stated above by Pavel and Nolet, the information is most definitely condensed.



English	French	Spanish
1. Subject Field(s) - Telecommunications Transmission - Internet and Telematics	Domaine(s) - Transmission (Télécommunications) - Internet et télématique	Campo(s) temático(s) - Transmisión (Telecomunicaciones) - Internet y telemática
electronic mail Source CORRECT, STANDARDIZED	**courrier électronique** Source CORRECT, MASC, STANDARDIZED	**correo electrónico** Source CORRECT, MASC
email Source CORRECT, NOUN	CÉ Source CORRECT, MASC, STANDARDIZED	correo-E Source CORRECT, MASC
e-mail Source CORRECT, NOUN	courriel Source CORRECT, MASC, STANDARDIZED	correo-e Source CORRECT, MASC
Email Source CORRECT, NOUN, STANDARDIZED	message électronique Source CORRECT, MASC	C.elec. Source CORRECT, MASC, MEXICO
E-mail Source CORRECT, NOUN, STANDARDIZED	C.élec Source CORRECT, MASC	emilio Source CORRECT, MASC, SPAIN, FAMILIAR
strudel-post Source JARGON		mensaje electrónico Source CORRECT, MASC, SPAIN
DEF – Correspondence in the form of messages transmitted between user terminals over a computer network. Source	**DEF** – Correspondance sous forme de messages, transmise entre terminaux d'utilisateur sur un réseau d'ordinateurs. Source	**DEF** – Comunicación automática a través de instalaciones y dispositivos electrónicos de transmisión. Puede consistir en una comunicación de persona a persona, o de una persona a muchas otras conectadas mediante una red de computadoras. Source
CONT – Nerd speak. Strudel-post: electronic mail. (Strudel refers to the sign "at" in E-mail addresses). Source	**OBS** – courrier électronique; courriel; CÉ : termes et abréviation normalisés par l'ISO/CEI [Commission électrotechnique internationale] et la CSA International. Source	
OBS – electronic mail; E-mail; Email: terms standardized by ISO/IEC [International Electrotechnical Commission] and CSA International. Source	**OBS** – courriel; courrier électronique; message électronique : termes publiés dans le Journal officiel de la République française du 20	

FIGURE 9.1 Term record from *TERMIUM*

The degree to which the information is condensed can sometimes be problematic given the wide range of users that a term bank is intended to serve. Sager (1990: 197–200) identifies a number of distinct user groups including (a) subject specialists, who often consult such resources for reassurance, but who sometimes need to ascertain the meaning of an unknown term; (b) information scientists, who perform tasks such as indexing of specialized documents; (c) language planners, who are charged with developing or maintaining a language and have a particular interest in standardization; (d) teachers and students, whether studying language or a specialized field; and (e) professional communication mediators, including specialized translators. According to Rondeau (1984: 148) and Sager (1990: 197), the latter make up the largest user group of term banks.

Each group has different needs and the range of information recorded on the term records is not sufficient to meet all of these needs equally well. For instance, Sager (1990: 197), when discussing translators as term bank users, notes 'their use of reference tools is more conditioned by the need to produce specialized texts and less by the need for comprehension [of an unknown term]'. In other words, translators really want usage information – such as contexts or collocations – which will help them to produce a target text that reads well. Instead, what they typically find in term banks are definitions and terms presented out of context, or in only a single context.

There are several possible explanations as to why limited information was included on a term record in the early days. First, early computers had less storage capacity, which may have restricted the amount of information that could be stored on a given record.

Secondly, although computers were being used as information repositories as early as the 1960s, L'Homme (2004: 48) notes that the actual collection and analysis of data was still done manually – by gathering and poring over printed documents – for many more years. Manual analysis is time-consuming and labour-intensive and thus necessarily limits the amount of data that can be gathered and analysed. Although electronic corpora were actually first developed by linguistic researchers in the 1960s (e.g. Brown Corpus, Lancaster-Oslo/Bergen (LOB) Corpus), such resources were not easily accessible outside academic institutions in those days. By the 1980s, general language corpora began to enjoy increasing attention among lexicographers, especially following the 1987 publication of the first corpus-based dictionary – the *COBUILD Dictionary* (Renouf 1987). However, even then, terminologists continued to lag behind lexicographers in turning to electronic corpora as a source of information. As pointed out by Bowker (1996: 30), the reason was most likely that, at that time, it was difficult to access specialized documentation in electronic form.

Finally, because the research process was so labour-intensive when using printed documentation, terminologists were no doubt trying to spare users the pain of having to invest their own time in conducting research. They therefore sought to provide quick answers by presenting their findings in a highly distilled form, even though, as pointed out by Sager (1990: 197), such condensed answers did not fully meet the needs of all users.

Today, nearly fifty years after the first term banks were developed, some of these limitations need no longer apply. Clearly, in this age of the terabyte, storage capacity is no longer a significant issue.

In addition, terminologists are no longer limited to working solely with printed documentation as it has become increasingly easier to access specialized material (e.g. via the internet, databases, electronic journals). Therefore, terminologists now find it easier to create specialized corpora, which can be interrogated with the help of corpus analysis software, including automatic term extractors, concordancers, word listers and collocation generators. In fact, not only is it now possible for terminologists to carry out their work using corpus-based resources and tools, it is virtually unthinkable for them not to do so (Pavel and Nolet 2001: xx; Bowker and Pearson 2002: 20; L'Homme 2004: 119). Because corpora can be consulted more quickly and easily than printed texts, terminologists working today can consult a wider range of documents, and they can use tools to process, sort and display the data in ways that facilitate the identification of a considerable variety of types of terminologically meaningful information, such as terms (Cabré et al. 2001), equivalents (Gaussier 2001), synonyms (Hamon and Nazarenko 2001), definitions (Pearson 1999), collocations (Heid 2001) or semantic relations (Meyer 2001; Marshman 2007).

The fact that the research process has become significantly less labour-intensive means that the quick or condensed answer is becoming less acceptable. On a traditional term record, for example, one term is designated as the preferred term and others are designated as synonyms. However, only the preferred term is shown in context (and in only a single context), while the synonyms are simply listed without any textual support. Faced with such a term record, a translator may well find it difficult to select and correctly use the most appropriate term.

It would be more helpful for translators to have access not simply to term records that provide a single 'best' term with a solitary context, but rather to information that would allow them to see all possible terms in a range of contexts and thus find the solution that works best in the target text at hand. Nevertheless, in spite of the possibilities offered by technological advancements, the term records found in today's term banks remain similar to those dating back to the 1960s. The record that was presented in Figure 9.1, for example, is dated as having been produced in 2011, yet it still contains very little in the way of contextual information or other types of usage information (e.g. collocations) that would be valued by translators, who are still being targeted as the primary user group. Looking at the *TERMIUM* record shown in Figure 9.1, only a single context has been provided for the English term *electronic mail* to supplement the one-line definition.

In summary, it would seem that while the corpus-based approach has had an impact on the terminological research process in the field of terminology, the impact on the product (i.e. term records in a term bank) has been negligible.[1] In other words, terminologists make full use of corpora and corpus analysis tools to help them collect and analyse data more efficiently; however, they do not pass the benefits of their corpus-based research on to the translator in any appreciable form. Rather than providing additional contextual information, collocations or other usage information gleaned from a corpus, they are continuing the traditional practice of condensing information into a formulaic term record.

9.2 Translators and Personal Terminology Management Systems

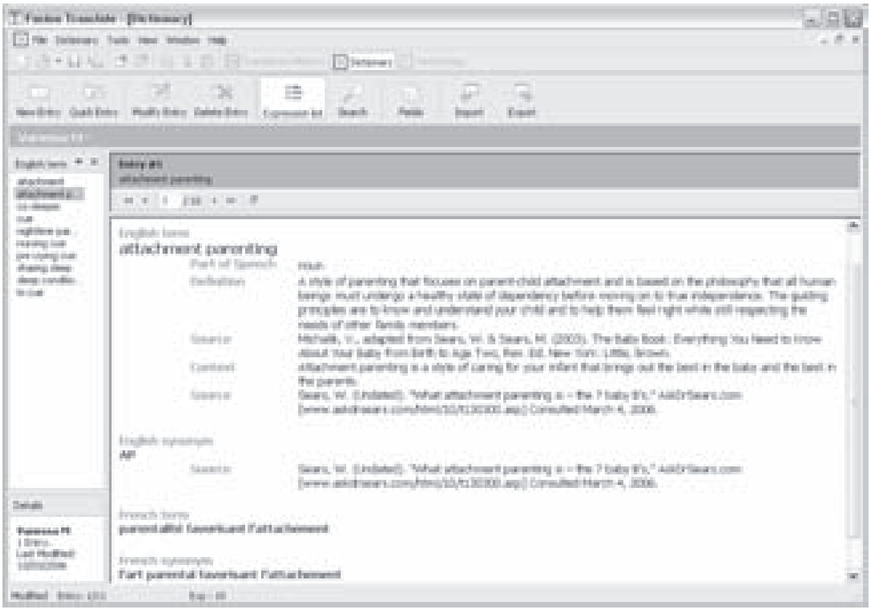
In contrast to the thematic research typically carried out by terminologists, translators are more likely to engage in 'ad-hoc research' (Cabré 1999: 152), which is conducted on individual terms or concepts and it is intended to provide answers to specific terminological problems that translators encounter as they work on a translation project. Although some terminologists engage in this type of investigation, the majority of translators conduct this type of research themselves. A study by Champagne (2004: 30) concludes that experienced translators invest 20–25 per cent of their time in terminology tasks, while inexperienced translators can invest up to 40–60 per cent. Moreover, even when a company employs a terminologist, Cabré (1999: 48) notes that 'the time constraints within which translators often have to work may not allow them to hand the task over to a terminologist'.

Like terminologists, translators traditionally recorded the results of their research on term records. In the days before desktop computers, such collections were typically stored on index cards, but when desktop computers became commonplace in the 1980s, personal terminology management systems (PTMSs) were among the first types of language software to be offered. Not surprisingly, however, the early systems, which operated as standalone tools, took as a model the template structure used in term banks. These early PTMSs were relatively rigid, permitting only a set of predefined fields and sometimes even insisting that each field be filled in, mirroring the completeness of records that terminologists strive to produce.

Gradually, PTMSs became more flexible, allowing users to customize records by defining and filling in the fields relevant to their needs. Translators could, for example, skip the definition, or add an extra context if desired. Interestingly, however, the sample databases that come with even the most current systems still typically reflect the traditional templates found in most term banks, and these sample databases are often used as teaching aids in terminology courses on translator training programmes.[2] Figure 9.2 shows a sample term record created using a PTMS.

Once it became easier to compile specialized corpora (see Section 9.1 above), translators warmed quickly to the advantages offered by these resources for finding solutions to the terminological problems they encounter (cf. Bowker 1998; Zanettin 1998; Lindquist 1999; Kübler 2003).

As corpora became more popular among translators, slight modifications began to appear on the term records created by these translators. For instance, following her detailed investigation and development of a corpus-based approach to terminology, Pearson (1998: 200) devised a model template in which she suggests the addition of some extra fields containing collocational information. For example, as shown in Figure 9.3, on the record for the



Entry #1
attachment parenting

English term
attachment parenting

Part of Speech noun

Definition A style of parenting that focuses on parent-child attachment and is based on the philosophy that all human beings must undergo a healthy state of dependency before moving on to true independence. The guiding principles are to know and understand your child and to help them feel right while still respecting the needs of other family members.

Source Michalik, V., adapted from Sears, W. & Sears, M. (2003). The Baby Book: Everything You Need to Know About Your Baby From Birth to Age Two, Rev. Ed. New York: Little, Brown.

Context Attachment parenting is a style of caring for your infant that brings out the best in the baby and the best in the parents.

Source Sears, W. (Undated). "What attachment parenting is – the 7 baby B's," AskDrSears.com [www.askdrsears.com/html/10/t130300.asp] Consulted March 4, 2006.

English synonym
AP

Source Sears, W. (Undated). "What attachment parenting is – the 7 baby B's," AskDrSears.com [www.askdrsears.com/html/10/t130300.asp] Consulted March 4, 2006.

French term
parentalité favorisant l'attachement

French synonym
l'art parental favorisant l'attachement

FIGURE 9.2 A sample term record created using the personal terminology management system included in the Fusion Translate tool suite

```
Language                  English
Term                      ankyrin 2.2
Grammatical category      noun
 . . .
Corpus attested collocations within 3 words to the left and right of the
node
Verbs:                    activates, insert, lacks, identified
Nouns:                    protein, ankyrin, variant
Adjectives:               erythroid
```

FIGURE 9.3 Term record that includes collocational information gathered from a corpus (Pearson 1998: 202)

term *ankyrin 2.2*, Pearson has recorded information about the verbs, nouns and adjectives that typically appear in the vicinity of this term.

More recently, as illustrated in Figure 9.4, the corpus-based specialized dictionary *DiCoInfo* (Dictionnaire fondamental de l'informatique et de l'Internet) seeks to systematically include in each entry descriptions of the paradigmatic and syntagmatic relationships between terms in an effort to foreground the lexical structure of the domain (L'Homme 2008).

Figure 9.4 Entry for the term *courriel* in the *DiCoInfo* that includes information on lexical relations and actantial structures gathered from a corpus

9.2.1 Integrated Tool Suites

A major development came when PTMSs, which had previously been standalone tools, began to be integrated into computer-aided translation (CAT) tool suites, most notably with tools such as translation memory systems and automatic term extractors.

9.2.1.1 Translation Memory Systems

A translation memory (TM) is essentially a database of aligned source and target texts. It can be considered as a type of parallel corpus, which is interrogated in a largely automated way with the help of specialized TM software. Works such as Bowker (2002) and Somers (2003) contain detailed descriptions of how TM systems operate, so only a brief summary will be provided here.

Essentially, when a translator has a new text to translate, the TM system will automatically compare the segments (which typically correspond to sentences) contained in this new text against those stored in the database, and if a match is found, the previous translation is proposed to the translator. The goal of this corpus-based technology is to allow a translator to 'recycle' relevant portions of previously translated texts.

Segment from new source text	The specified operation was interrupted by the system.
Fuzzy match retrieved from TM database	EN: The operation was interrupted by the application. FR: L'opération a été interrompue par l'application.

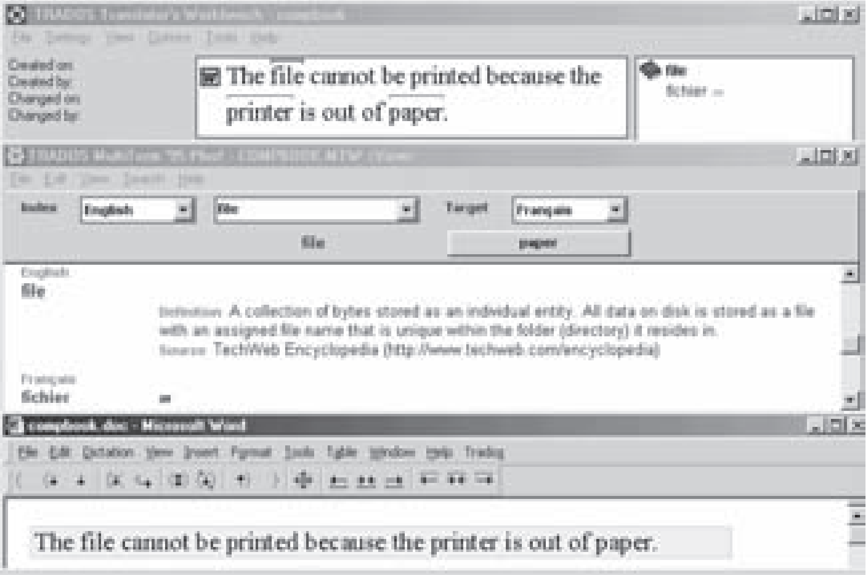
Figure 9.5 A fuzzy match retrieved from the TM database (differences between the new segment and the previously translated segment are underlined)

The first TM systems looked primarily for exact matches; in other words, the TM searched for segments in the new ST that were identical to segments contained in a previous text. However, even minor changes (e.g. from singular to plural, or from present to past tense), would not qualify as an exact match. Therefore, the powerful concept of fuzzy matching was introduced.

A fuzzy matching algorithm looks for segments in the TM database that are similar to the new ST segment, and which may therefore be useful for helping the translator to produce a new target text. Figure 9.5 gives an example of a fuzzy match.

While fuzzy matching increased the amount of text that could be recycled, TMs also sought to integrate PTMSs, which were the personal terminology collections built up in advance by translators. In other words, translators were responsible for conducting the terminological research and then transcribing the results onto term records in the PTMS. Once these records had been created, the TM system could interact with the PTMS and automatically search for relevant term records, as well as searching for matches in the TM database. Any time there was a term in the ST for which there was a matching entry in the PTMS, the term record could be displayed for the translator's perusal, and if desired, with a single click the equivalent could be pasted from the term record directly into the target text in a word processor to save the translator the effort of typing it. Figure 9.6 shows an example of this type of interaction between three different tools:

1. the ST segment that the translator needs to translate in the MS Word word-processing system (in the bottom pane);
2. the term matches identified by the Trados Translator's workbench TM system (in the top pane, with lines over the terms for which there is a corresponding term record); and
3. the actual terminology records for those terms for which the translator previously created records using the MultiTerm PTMS (part of the record for *file* is visible in the scrollable centre pane).


Created on:
Created by:
Changed on:
Changed by:
Index
Target
File Edit Dictation View Insert Format Tools Table Window Help Trados

The file cannot be printed because the printer is out of paper.

file
fichier =

Index: English | file | Target: Français

file | paper

English
file

Definition A collection of bytes stored as an individual entity. All data on disk is stored as a file with an assigned file name that is unique within the folder (directory) it resides in.
Source TechWeb Encyclopedia (http://www.techweb.com/encyclopedia)

Français
fichier =

compbook.doc - Microsoft Word

The file cannot be printed because the printer is out of paper.

FIGURE 9.6 An example of the interaction between a TM, a PTMS and a source text in a word processing system

9.2.1.1.1 Recording Non-Canonical Forms of the Term

This interaction of TM, PTMS and word processor led to an interesting change in the way information was recorded on a term record. Kenny (1999: 70) observes that while traditional terminology textbooks (e.g. Auger and Rousseau 1978: 74; Sager 1990: 145) dictate that terms should be recorded in their canonical form, the actual form of terms (and their translations) can vary considerably in texts. For instance, while the canonical form of a noun is the singular, it may appear in the text in the plural, and while a verb may be recorded on a term record in the infinitive, it can appear in many different conjugated forms in a text. Things become even more complicated when languages have declensions or are marked for gender, since adjectives may need to agree with the nouns they are modifying. Moreover, as noted by Kenny (1999: 70–1), in languages such as Irish, the grammatical or phonetic environment in which a word appears can change its form considerably (e.g. through initial mutations known as eclipsis or through aspiration). In such cases, pasting the canonical form from the term record directly into the target text leaves the translator with some editing work to do.

As Kenny (1999: 74) observes, to really benefit from the possibility of automatically pasting a term from a term record into the target text, some translators began to record their terms not in the canonical form dictated by the textbooks, but in the most commonly used forms, thus optimizing interaction between the TM, the PTMS and the word processor.

9.2.1.1.2 Streamlining the Contents of Term Records

Around the same time, O'Brien (1998: 118) was observing other changes with regard to the nature of term records being created by translators: 'While terminology tools generally allow the user to input detailed information such as gender, plural form, examples of usage, etc. glossaries in the localization industry often contain only the source term and the target term.'

O'Brien suggests several reasons for this streamlining. First, she notes that translators are facing ever-shorter deadlines, so they have less time to create extensive records. Secondly, in fast-paced fields, the concepts and terms date very quickly, so it is not worth creating a detailed record that will never be consulted again because it is out of date. Finally, O'Brien notes that the translator is really only interested in the equivalent and the context in which it can appear. This echoes Sager's (1990: 197) observation that translators have a greater need for information to assist them in producing a target text than they do for information to help them understand a source text.

9.2.1.1.3 Using Translated Material as a Resource for Terminological Research

Another way in which the integration of PTMSs with TMs has affected terminological research carried out by translators is that they are conducting this research using translated material as a resource. Terminology textbooks (cf. Cabré 1999: 134; Pavel and Nolet 2001: 8; Dubuc 2002: 164) emphasize that, wherever possible, documents used for terminology research should be original language documents on the grounds that translations are more likely to contain conceptual or terminological errors, awkward syntactic constructions, or non-idiomatic expressions and are therefore inappropriate reference sources for the production of other documents. Although this reasoning may be sound, translators working under extreme time constraints may feel pressured to use the translated material contained in a TM database, rather than compiling corpora containing original language texts. Or it could be that in cases where translators know the source of the material in the TM, they feel comfortable recycling it. For example, if a translator did the translations in the TM him- or herself and knows that the terminological research conducted was sound, he or she may be happy to rely on it. Moreover, increasingly, clients are providing a TM database and insisting that it be consulted, leaving the translator no choice but to use translated material as a resource.

9.2.1.2 Term Extractors

A term extractor is a tool that attempts to automatically identify all the potential terms in a corpus and to present a list of candidates to the user for verification. Several different underlying approaches can be used, including frequency- and recurrence-based techniques, part-of-speech and pattern-based techniques, corpus comparison techniques and various combinations of these.[3] However, one of the most popular methods is based on statistics, whereby the term extractor essentially looks for repeated sequences of lexical items. The frequency threshold (i.e. the number of times that a sequence must be repeated) is often user-specified.

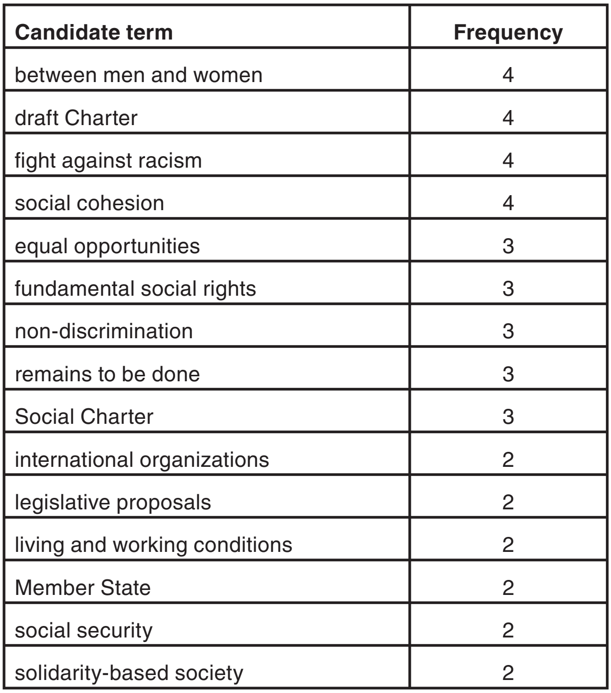
Term extractors have been used for some time by terminologists doing thematic research, in order to help them try to identify all the relevant terms in a subject field (Pavel and Nolet 2001: 71–3). However, since translators typically engage in ad hoc research rather than subject field research, these tools were not initially of much interest to them. This changed, however, once term extractors became routinely integrated with TM systems, starting in around 2003.

With such integration, a translator can use a term extractor to generate a list of all the sequences of lexical items that appear a specified number of times in the TM database, as illustrated in Figure 9.7, which shows a list generated using a term extractor.

Subsequently, the translator can ask the tool to try to automatically identify possible translation equivalents for all the lexical items on the list using the aligned target texts contained in the TM database. Again, this is done using statistical techniques, and the results may not be perfect. However, for each item on the list, the tool attempts to suggest one or more potential equivalents, which the translator can verify by referring to the original paired text segments from which the equivalent was extracted, as shown in Figure 9.8. Once a translator has confirmed or corrected the proposals, this list of source and target equivalents can be automatically uploaded into the PTMS.

9.2.1.2.1 Including 'Non Terms' on Term Records

Looking at the generated lists shown in Figures 9.7 and 9.8, one striking observation that can be made is that the items on the list do not necessarily correspond to 'terms' as understood in the field of terminology. For a terminologist, a term is a lexical item used to designate a specialized concept. However, the items appearing on the list generated by this term extractor are simply strings of characters that appear multiple times in the TM database. Some of these items, such as *social cohesion* or *Member State* may actually correspond to terms in the strict sense, but many others do not. In fact, the items on the list may not even represent coherent semantic units. For example, *fight against racism, between men and women* and *remains to be done* cannot be considered terms per se, nor can they even be considered fixed



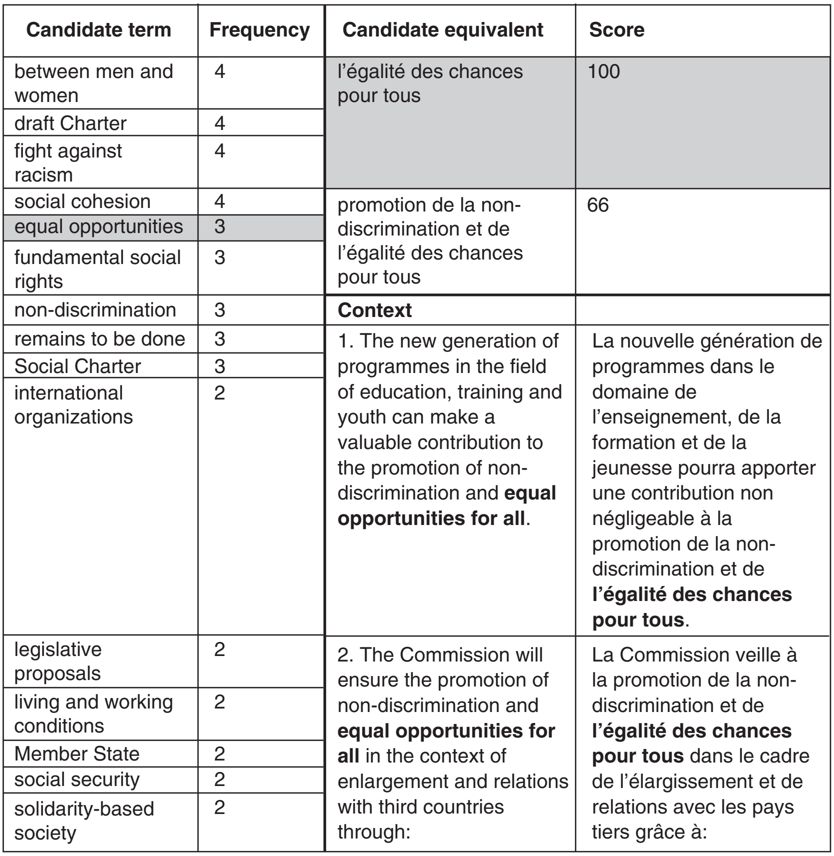
Candidate term	Frequency
between men and women	4
draft Charter	4
fight against racism	4
social cohesion	4
equal opportunities	3
fundamental social rights	3
non-discrimination	3
remains to be done	3
Social Charter	3
international organizations	2
legislative proposals	2
living and working conditions	2
Member State	2
social security	2
solidarity-based society	2

Figure 9.7 A list of candidate terms extracted from a TM database

expressions. Rather, they are simply chunks of language that happened to appear several times in the texts contained in the TM database. If a terminologist encountered such items on an automatically generated list of candidate terms, these would certainly be removed from the list (Pavel and Nolet 2001: 45). However, for a translator whose goal is to produce an acceptable translation within a short timeframe, any type of information that may be recyclable – term or otherwise – could potentially be useful. Therefore, translators who are working with such tools are beginning to fill up their PTMSs with records containing a mixture of terms and non-terms, which contradicts the strict principles put forth in the terminology literature (e.g. Dubuc 2002: 33).

9.2.1.2.2 Adopting a More Semasiological Approach

Another deviation from terminology principles is that this approach is more semasiological, or form-based, rather than onomasiological, or concept-based. Traditionally, terminologists have been encouraged to work in an onomasiological fashion and to create one record per concept as well as to record all the



Candidate term	Frequency
between men and women	4
draft Charter	4
fight against racism	4
social cohesion	4
equal opportunities	3
fundamental social rights	3
non-discrimination	3
remains to be done	3
Social Charter	3
international organizations	2
legislative proposals	2
living and working conditions	2
Member State	2
social security	2
solidarity-based society	2

Candidate equivalent	Score
l'égalité des chances pour tous	100
promotion de la non-discrimination et de l'égalité des chances pour tous	66
Context	
1. The new generation of programmes in the field of education, training and youth can make a valuable contribution to the promotion of non-discrimination and **equal opportunities for all**.	La nouvelle génération de programmes dans le domaine de l'enseignement, de la formation et de la jeunesse pourra apporter une contribution non négligeable à la promotion de la non-discrimination et de **l'égalité des chances pour tous**.
2. The Commission will ensure the promotion of non-discrimination and **equal opportunities for all** in the context of enlargement and relations with third countries through:	La Commission veille à la promotion de la non-discrimination et de **l'égalité des chances pour tous** dans le cadre de l'élargissement et de relations avec les pays tiers grâce à:

Figure 9.8 Proposed equivalents for candidate terms automatically extracted from the TM database

information pertinent to a given concept on that single record (Cabré 1999: 38). Use of an automatic term extractor results in the creation of a new record for every character string that is identified. This means that synonyms each have their own record, even though they refer to a common concept. Even different forms of the same lemma would appear on separate records, which would achieve a result similar to that implemented by the translators observed by Kenny (1999: 70–4), summarized above. To maximize recyclability, those translators were recording the most commonly used form of a term, or even multiple forms of a term, rather than the canonical form. A term extractor would capture all of the frequently used forms and make them available for recycling.

9.2.1.2.3 Adapting to Active Terminology Recognition and Pre-Translation

As described in Section 9.2.1.1, any information stored in the PTMS can be automatically compared by the TM system against a new source text to

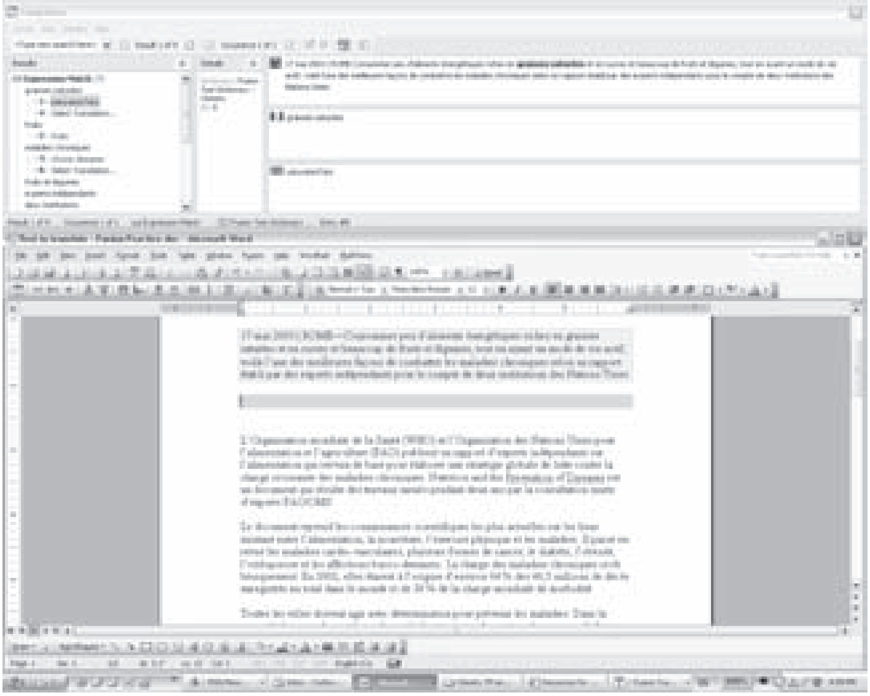
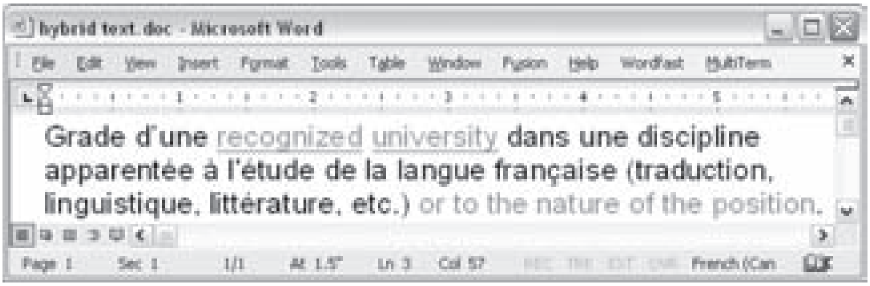


Figure 9.9 Information from the PTMS is available for the translator to consult during translation

be translated. Matches can be displayed for the translator to consult (see Figure 9.9), or, if desired, all matches from the PTMS can be automatically inserted directly into the text, replacing the original source text terms. The result of this process, known as pre-translation, is a partially translated or hybrid text, as shown in Figure 9.10. Sections that remain in the source language are those for which no match could be found in the PTMS, and so the translator must translate these sections from scratch or using other resources.

During pre-translation, the computer is simply doing pattern matching and is not able to understand or make use of information such as definitions, part of speech and so on. Therefore, the streamlining of term records observed by O'Brien (1998: 118), and described above, remains in effect. Moreover, there is no need for translators working interactively with the TM and PTMS to select and copy contexts onto their term records since, if they wish to see a term in context, they can simply call up all the examples contained in the TM database.[4]


hybrid text.doc - Microsoft Word
File Edit View Insert Format Tools Table Window Fusion Help Wordfast MultiTerm
Page 1 Sec 1 1/1 At 1.5" Ln 3 Col 57 REC TRK EXT OVR French (Can

Grade d'une recognized university dans une discipline apparentée à l'étude de la langue française (traduction, linguistique, littérature, etc.) or to the nature of the position.

Figure 9.10 An extract from a hybrid text produced during the pre-translation process, whereby material from the PTMS can be inserted directly into the text

9.3 Discussion

It seems clear that the introduction of electronic corpora and associated software has changed the way that terminology research is conducted and recorded.

For terminologists, the change has been one of gradual evolution. Terminologists have always used corpora, so the change from print-based to electronic corpora was not radical. Corpus-based tools and resources have simply made it easier for terminologists to carry out their tasks, allowing them to consult a wider range of documents and to achieve better coverage of the subject fields. However, some of the significant benefits afforded by corpora (e.g. viewing terms in multiple contexts, accessing frequency information about different terms, providing more detailed information about semantic relations), have not as yet been passed on to the end users of many terminology products. Instead, term records found in term banks such as *TERMIUM*, *GDT* and *IATE* still retain the same basic form that they had before electronic corpus-based research took hold. However, numerous terminology researchers (e.g. Pearson 1998; Condamines and Rebeyrolle 2001; Marshman 2007) have begun to advocate the benefits of providing such information on term records, so perhaps we will see changes in this regard in the future.

In spite of their shortcomings, it seems that translators do still consult term banks, but perhaps not with the same degree of 'blind faith' as in the past, when they had fewer options for conducting independent research. Nowadays, translators seem to use term banks as part of a system of checks and balances rather than as a definitive resource. For instance, translators might begin their research by consulting a term bank, and then use the term bank suggestions as initial search terms in their own corpus-based resources. However, they also take the findings from their corpora and compare them against the contents of term banks. In other words, translators treat term banks and corpora as complementary resources. Translators respect the fact that the terminologists have conducted very detailed and thorough investigations – something they themselves do not always have time to do – but they are also mindful that specialized subject fields and the language used to describe these fields are constantly expanding and changing, so no term bank can provide exhaustive up-to-date coverage. Moreover, clients may have terminological preferences that are not reflected in the term banks maintained by other institutions. Therefore, most translators still find it necessary to conduct their own terminological research to ensure that the appropriate subject fields and client preferences are adequately represented. However, in contrast to the gentle evolution of terminological practices on the part of terminologists, there has been a revolution in the way that translators conduct terminological research and record their findings. Several factors have contributed to this revolution.

O'Brien (1998: 118), for example, cited time pressure as one reason that translators have begun recording minimal information on term records. It is widely recognized that in today's market, translators face tight deadlines as the volume of text to be translated continues to increase, as more organizations begin to make their information available worldwide via the internet, and as more companies look to enter the global market. Practices such as simultaneous shipment or 'simship' (where all language versions of a document or product are released at the same time) have resulted in shorter deadlines for translators. Add to this the fact that baby-boomers are retiring but there are comparatively few professionally qualified translators entering the market (CTISC 1999: 18).

However, to be fair, translation has always suffered from this general type of pressure. In the past, translation was often an afterthought or a last-minute decision on the part of a client, but once made, the client typically wanted the work 'done by yesterday'. At least in today's market, more clients are planning for translation in advance and including it as part of the overall production process. Why then are today's translators recording such comparatively minimal information on their term records?

One reason is related to the amount of time required to conduct the terminology research in the first place. As previously noted, in the past, translators had to conduct their research using only printed documentation, which was hugely time-consuming and labour-intensive. Translators could literally not afford to re-read entire documents to find a good explanation or an illuminating context for a term – they had to record this information on a term record so that they could find it again easily.

Today, thanks to the availability of electronic documents, coupled with specialized interrogation tools, not only does it take less time to gather a corpus, but translators can also quickly zero in on those parts of it that are most relevant. It takes almost no time to generate concordances or frequency lists 'on the fly'. In fact, it would probably take longer to thoroughly evaluate all the material in order to try to select just one 'ideal' context to copy onto a term record than it would to simply re-generate the information from the corpus on a future occasion. As an added bonus, by generating the concordances 'on the fly' instead of looking at information recorded on a term record, a translator gets to see the search term in a range of contexts, rather than in just a single context. This is important because although one particular context could prove helpful for dealing with a given source text, a different context may be more useful when tackling another text. As Lauriston (1997: 180) notes, 'for translators . . . the quantity of information provided is often more important than the quality. They are usually able to separate the wheat from the chaff and even turn the chaff into palatable solutions to a particular communications problem.'

Of course, it may be asked whether it is overly time-consuming for a translator to have to wade through a series of contexts, rather than just consulting a single context on a term record. Although it would take a considerable amount of time to read all the documents in a corpus from beginning to end, corpus analysis tools present information to translators in an easy-to-read format (e.g. KWIC concordance, frequency list) so that a translator can quickly isolate the information that is most pertinent for a given task.

Another reason that translators working with printed corpora found it useful to create detailed term records is that documentation had a longer lifespan in the past. Publishing printed documents is a slow process, so the rate at which new material was made available to translators for incorporation into their printed corpora was slower. Therefore, it was worth recording information from a printed corpus because this information would remain valid for a longer period of time. Of course, the term records needed to be updated periodically, but not at the rate deemed necessary in today's market, where, as observed by O'Brien (1998: 118), the terminology may change with each new text the translator tackles. Similarly, terminologists who work with electronic corpora when conducting thematic research are still required to compile corpora of a substantial size in order to ensure broad enough coverage of a field. The larger a corpus, the more effort is required to compile it, and therefore the longer it must be used in order to have a reasonable return on investment.

In contrast, information is published in electronic form at such a rapid rate that it is not uncommon for translators to compile so-called disposable or do-it-yourself (DIY) corpora on a regular basis – often for each new job they undertake (Zanettin 2002; Varantola 2003). A disposable corpus is a one that has been collected for the purpose of translating a particular source text or a group of related texts. Therefore, its content has been selected specifically with the characteristics of that text or group of texts in mind (e.g. client, readership, style, communicative function, currency of information). Since disposable corpora are being created or updated on such a regular basis, it almost seems redundant to create detailed term records since these would also need to be updated, thus doubling the task. In addition, disposable corpora are not usually very large – at least not by corpus linguistic standards – which is another reason why, as noted above, it is not unrealistic or overly time-consuming for translators to generate queries 'on the fly' and to re-consult the corpus, rather than recording selected information on a term record.

A possible drawback to using disposable corpora is that, even though the texts should be carefully selected, it is still possible that they may contain some errors, such as spelling mistakes or non-standard usage. According to Grefenstette (2002: 206), this should not deter users from adopting a corpus-based approach because 'the correct form is always orders of magnitude more frequent than the erroneous form'. However, Grefenstette was speaking about very large general language corpora used for lexicography. In the case of disposable corpora for specialized translation, the size of the corpus may be small enough that the results of a search could be skewed by erroneous material, so translators should keep this in mind when analysing corpus data.

9.4 Rethinking the Nature of Terminology Training for Translators

This leads us to a discussion of what today's translators-in-training need to learn about conducting terminology research. This discussion will focus on the situation in Canada, since this is the market with which I am most familiar. At present, all translator training programmes in Canada include a course in terminology as part of the curriculum. However, the textbooks used in many of these courses (e.g. Cabré 1999; Pavel and Nolet 2001; Dubuc 2002; L'Homme 2004) are written from the perspective of terminologists rather than translators. Accordingly, students in these classes spend time learning how to do tasks such as designing a term record template, constructing a well-formed terminological definition or selecting a single ideal context from a range of potential contexts. Although such tasks were relevant to translators in the past, they now seem to be more the remit of terminologists since, as demonstrated above, the needs of these two groups appear to be diverging more and more in this corpus-based era. Given this situation, it might be time to put more thought into developing a curriculum aimed at the specific terminological needs of translators, rather than one that tries to provide training in terminology proper, where the focus is on thematic research instead of on the term research skills needed by translators. After all, a single course in terminology offered as part of an undergraduate translation programme is unlikely to provide sufficient training for someone to work as a terminologist. Therefore, the focus of a terminology course that is a core component of a translator training programme should fall squarely on the needs of the translator. Those more theoretical aspects of terminology, such as conceptual classification, which often come up during the course of thematic research, but rarely during ad hoc research, should perhaps be relocated to courses in a graduate programme. It may even be worth exploring the merits of setting up separate training programmes – possibly in the form of post-graduate diplomas or professionally oriented master's programmes – specifically to train terminologists. No such programme currently exists in Canada, but it is not inconceivable to think that this type of programme might attract students from various backgrounds, including translation, linguistics, modern languages, information science, or even those with previous training in specialized subject fields where there is a need for terminological research.

Returning to the discussion of what might usefully be included in a terminology course for translators, it is clear that it can still be helpful for them to have a general understanding of the overall aims of the field of terminology, the methodologies used, and the products developed by terminologists. However, instead of learning to craft intensional definitions or to winnow down a selection of possible contexts to find one ideal context, translators may be better off learning to identify the characteristics of a knowledge-rich context (Meyer 2001) – one that contains useful information – rather than a context that simply attests to the existence of a term. In addition, translators need to learn strategies for isolating such knowledge-rich contexts in corpora, and once they have been identified, translators need to know how to critically evaluate these contexts, so that they can glean the necessary information – whether comprehension- or production-oriented – that will help them solve the problem at hand. These skills will stand translators in better stead than simply being able to select a single context to record on a term record – a context that may or may not be helpful for the translation of a future text.

Moreover, since translators do not engage in thematic research, but rather seek answers to particular translation problems, it does not really seem necessary for them to spend a lot of time learning how to distinguish between a term and a non-term – a topic addressed in detail in standard terminology textbooks. For a translator, any lexical item that poses a challenge, even if it is not a term per se, must be researched, so this distinction is not especially useful in a course for translators. This is particularly the case when working with integrated TM tool suites, where translators are now using term extractors to simply produce lists of lexical items – some of which may be terms, but others which may simply be expressions or even just strings of words that crop up frequently in the database – along with their equivalents.

This raises the point that the use of carefully selected corpus-based resources may eliminate the need for translators to produce term records in the traditional sense. Records produced when using TMs are often simply lists of equivalents, meanwhile, when using disposable corpora, translators may discard the idea of creating any type of record at all, preferring to simply generate searches 'on the fly' as the need arises. With this in mind, it seems that rather than teaching students how to design templates for term records, it might be better to place more emphasis on designing and compiling disposable corpora. Evaluating documentation has always been a crucial component of terminology courses – even in the times of printed corpora – but it is becoming particularly important to make students aware of the potential pitfalls associated with Web-based documentation, since this is the first place that many translators will turn to for research or as a potential source of electronic texts to put in their disposable corpora. As pointed out by Grefenstette (2002: 206), the Web is an unvetted resource, so extreme care must be taken when selecting texts to serve as a source of terminological information, especially since disposable corpora are not usually all that large and so are at risk of being unfavourably skewed by the inclusion of an inappropriate text. As Pearson (2000: 237) notes, choosing texts wisely when surfing the internet is not something that comes naturally to student translators. Too often, they merely enter key words into a search engine and simply retrieve the first ten hits, thinking that any text containing these key words will be a useful source of information for their purposes. As a result of this naïve approach, the texts retrieved may be inappropriate in terms of register, technicality and type. Therefore, terminology classes for translators should really include guidelines to help students learn to take a more refined approach to text selection. Some more recent textbooks, such as L'Homme (2004), have already taken steps towards addressing this issue.

An additional change that needs to be made in terminology courses is that students should be given guidance regarding the potential and drawbacks of using translated material as a resource. As discussed above, translators are beginning to source terminology from resources containing translated material, such as TM databases. Rather than simply forbidding the use of translated texts, which has typically been the approach taken in terminology courses in the past, trainers need to accept that translators may not always have a choice (e.g. if a client mandates the use of a TM). With this in mind, it would be helpful to train students how to design their TM databases in such a way as to minimize potential terminological clashes. This might include advising students to create separate TMs for each client, or for different text types, or maybe even for documentation relating to individual products. In the absence of this type of guidance, students frequently take the naïve *bigger is better* approach, piling all the texts they can find into one large database in the hope of increasing the number of matches retrieved. While this may indeed produce a greater number of matches, the resulting translation may lack terminological coherence since the terms may have been sourced from a wide range of documents having different styles, registers or even the preferred terminology of different clients. According to Topping (2000: 60) and Bowker (2006: 182), mixing terminology from different clients or domains has a 'train wreck' effect on a target text. In addition to learning how to design databases to optimize high-quality recycling, students could also benefit from guidance on how to properly maintain and update a TM database in order to reduce the risk of problems being perpetuated in subsequent translations.

There is also room for improvement in providing opportunities for students to use the tools in a way that better reflects how they are employed by practising translators. At present, students in terminology classes may use term extractors, but this is done from a terminologist's perspective, and so the resulting list of candidates is divested of non-terms. Next, students are typically taught how to use PTMSs in standalone mode. However, as discussed in Section 9.2.1, working translators tend to use PTMSs in conjunction with term extractors, TMs and word processors, creating stripped-down records for terms and other repeated expressions, which are then consulted automatically by a TM system and sometimes even pasted directly into a target text. To understand how these tools really operate, and the impact that they have on terminology and translation practices, students must work with the integrated tool suite. This means giving students realistic exercises that allow them to see all the components of the tool working together. Since students tend to do translation work in one class, terminology in another, and technology in yet another, they may never have the opportunity to bring all these elements together. This means that when they finally enter the workforce, they are forced to learn a new way of working – one that may equal more than the sum of its parts. To offer this kind of integrated experience to students would clearly entail extra effort on the part of the educators, who may need to coordinate with their colleagues in order to ensure that all the elements are brought together in a logical and coherent way, but it would surely be of great benefit to students if this could be achieved (Bowker and Marshman 2009).

In conclusion, it seems clear that corpus-based technology has impacted both the process and product of terminology research. In this chapter, an attempt was made to take stock of the changes that have been introduced, and to explore what motivated these changes. It is only by gaining a deeper understanding of these changes that we can begin to develop a logical plan for moving forward. The suggestions put forward here for modifying the curriculum with regard to terminology courses for translators are tentative at best, but given that corpora and corpus analysis tools are becoming increasingly entrenched in the translation profession, we can no longer put off thinking about the best way to prepare students for working in a corpus-based world.

Notes

[1] I am speaking here of conventional, widely used and easily accessible term banks (e.g. *TERMIUM*, *GDT*, *IATE*). Several researchers (e.g. Meyer et al. 1992; Condamines and Rebeyrolle 2001; Cabré et al. 2004) have designed and produced small-scale highly specialized term banks containing additional information drawn from corpora. However, these prototypes are usually developed as part of a research project, may have a limited lifespan, and are not easily accessible to

average translators. Hopefully, this will change, and it is at least encouraging to note that such researchers see the value of conceiving and developing products that include a wider variety of corpus-based information.

[2] Although users are free to design their own term records, in terminology classes many students adopt the traditional template and make only minor modifications.

[3] For a more detailed description of different approaches see, for example, Cabré et al. (2001) and Lemay et al. (2005).

[4] Note that a valid criticism of TM systems in the past was that because these systems simply stored pairs of segments rather than complete texts, it was not possible to see a context that was larger than a given segment (Macklovitch and Russell 2001). However, some TMs, such as the *MultiTrans* system developed by MultiCorpora R&D (www.multicorpora.com), preserve the complete text in the database.

References

Anobile, Michael (2000) 'Foreword', in R. C. Sprung (ed.) *Translating into Success: Cutting-edge Strategies for Going Multilingual in a Global Age*, Amsterdam and Philadelphia: John Benjamins, vii.

Auger, Pierre and Louis-Jean Rousseau (1978) *Méthodologie de la recherche terminologique,* Québec: Office de la langue française.

Bowker, Lynne (1996) 'Towards a Corpus-Based Approach to Terminography', *Terminology* 3(1): 27–52.

— (1998) 'Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study', *Meta* 43(4): 631–51.

— (2002) *Computer-Aided Translation Technology: A Practical Introduction*, Ottawa: University of Ottawa Press.

— (2006) 'Translation Memory and "Text"', in L. Bowker (ed.) *Lexicography, Terminology and Translation: Text-Based Studies in Honour of Ingrid Meyer,* Ottawa: University of Ottawa Press, 175–87.

Bowker, Lynne and Elizabeth Marshman (2009) 'Better Integration for Better Preparation: Bringing Terminology and Technology More Fully into Translator Training Using the CERTT Approach', *Terminology* 15(1): 60–87.

Bowker, Lynne and Jennifer Pearson (2002) *Working with Specialized Language: A Practical Guide to Using Corpora*, London and New York: Routledge.

Cabré, M. Teresa (1999) *Terminology: Theory, Methods, and Applications,* Amsterdam and Philadelphia: John Benjamins.

Cabré, M. Teresa, R. Estopà and J. Vivaldi (2001) 'Automatic Term Detection: A Review of Current Systems', in Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme (eds) *Recent Advances in Computational Terminology,* Amsterdam and Philadelphia: John Benjamins, 53–87.

Cabré, M. Teresa, C. Bach, R. Estopà, J. Feliu, G. Martínez, and J. Vivaldi (2004) 'The GENOMA-KB Project: Towards the Integration of Concepts, Terms, Textual Corpora and Entities', *LREC 2004 Fourth International Conference on*

Language Resources and Evaluation, Lisbon: European Languages Resources Association, 87–90.

Canadian Translation Industry Sectoral Committee (CTISC) (1999) *Survey of the Canadian Translation Industry: Human Resources and Export Development Strategy*. Final Report of the Canadian Translation Industry Sectoral Committee. Available online at: http://www.uottawa.ca/associations/csict/rap-e.htm#stre (accessed 30 May 2010).

Champagne, Guy (2004) 'The Economic Value of Terminology: An Exploratory Study'. Unpublished Report submitted to the Translation Bureau of Canada, Public Works and Government Services Canada.

Condamines, Anne and Josette Rebeyrolle (2001) 'Searching For and Identifying Conceptual Relationships via a Corpus-Based Approach to a Terminological Knowledge Base (CTKB): Methods and Results', in Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme (eds) *Recent Advances in Computational Terminology*, Amsterdam and Philadelphia: John Benjamins, 127–48.

Dubuc, Robert (2002) *Manuel pratique de terminologie* (4th edition), Brossard, Québec: Linguatech.

Gaussier, Eric (2001) 'General Considerations on Bilingual Terminology Extraction', in Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme (eds) *Recent Advances in Computational Terminology*, Amsterdam and Philadelphia: John Benjamins, 67–183.

Grefenstette, Gregory (2002) 'The WWW as a Resource for Lexicography', in Marie-Hélène Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins*, Grenoble: Euralex, 199–215.

Hamon, Thierry and Adeline Nazarenko (2001) 'Detection of Synonymy Links between Terms: Experiment and Results', in Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme (eds) *Recent Advances in Computational Terminology*, Amsterdam and Philadelphia: John Benjamins, 185–208.

Heid, Ulrich (2001) 'Collocations in Sublanguage Text: Extraction from Corpora', in S. E. Wright and G. Budin (eds) *Handbook of Terminology Management* (vol. 2), Amsterdam and Philadelphia: John Benjamins, 788–808.

Kenny, Dorothy (1999) 'CAT Tools in an Academic Environment: What Are They Good For?', *Target* 11(1): 65–82.

Kübler, Natalie (2003) 'Corpora and LSP Translation', in Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds) *Corpora in Translator Training*, Manchester: St. Jerome Publishing, 25–52.

Lauriston, Andy (1997) 'Terminology and the Computer', in Robert Dubuc *Terminology: A Practical Approach*, Brossard, Quebec: Linguatech, 179–92.

Lemay, Chantal, Marie-Claude L'Homme and Patrick Drouin (2005) 'Two Methods for Extracting "Specific" Single-word Terms from Specialized Corpora: Experimentation and Evaluation', *International Journal of Corpus Linguistics* 10(2): 227–55.

L'Homme, Marie-Claude (2004) *La Terminologie: Principes et Techniques*, Montréal: Les Presses de l'Université de Montréal.

— (2008) 'Le DiCoInfo: Méthodologie pour une nouvelle génération de dictionnaires spécialisés', *Traduire* 217: 78–103.

Lindquist, Hans (1999) 'Electronic Corpora as Tools for Translation', in Gunilla Anderman and Margaret Rogers (eds) *Word, Text, Translation: Liber Amicorum for Peter Newmark*, Clevedon, UK: Multilingual Matters, 179–89.

Macklovitch, Elliott and Graham Russell (2001) 'What's Been Forgotten in Translation Memory?' in J. S. White (ed.) *Envisioning Machine Translation in the Information Future: Proceedings of the 4th Conference of the American Machine Translation Association (AMTA),* Berlin: Springer, 137–46.

Marshman, Elizabeth (2007) 'Towards Strategies for Processing Relationships between Multiple Relation Participants in Knowledge Patterns: An Analysis in English and French', *Terminology* 13(1): 1–34.

Meyer, Ingrid (2001) 'Extracting Knowledge-rich Contexts for Terminography: A Conceptual and Methodological Framework', in Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme (eds) *Recent Advances in Computational Terminology*, Amsterdam and Philadelphia: John Benjamins, 279–302.

Meyer, Ingrid and Kristen Mackintosh (1996) 'The Corpus from a Terminographer's Viewpoint', *International Journal of Corpus Linguistics* 1(2): 257–85.

Meyer, Ingrid, Lynne Bowker and Karen Eck (1992) 'COGNITERM: An Experiment in Building a Knowledge-Based Term Bank', *Proceedings of the Fifth EURALEX International Congress (EURALEX '92),* Tampere, Finland: Tampereen Yliopisto, 159–72.

O'Brien, Sharon (1998) 'Practical Experience of Computer-Aided Translation Tools in the Software Localization Industry', in L. Bowker, M. Cronin, D. Kenny and J. Pearson (eds) *Unity in Diversity? Current Trends in Translation Studies*, Manchester: St. Jerome Publishing, 115–22.

Pavel, Silvia and Diane Nolet (2001) *Handbook of Terminology/Précis de terminologie*, Ottawa: Minister of Public Works and Government Services Canada. Also available online in English at: http://www.btb.gc.ca/publications/documents/termino-eng.pdf and in French at: http://www.btb.gc.ca/publications/documents/termino-fra.pdf (accessed 11 March 2011).

Pearson, Jennifer (1998) *Terms in Context*, Amsterdam and Philadelphia: John Benjamins.

— (1999) 'Comment accéder aux éléments définitoires dans les textes spécialisés?', *Terminologies Nouvelles* 19: 21–8.

— (2000) 'Surfing the Internet: Teaching Students to Choose their Texts Wisely', in L. Burnard and T. McEnery (eds) *Rethinking Language Pedagogy from a Corpus Perspective*, Frankfurt am Main: Peter Lang, 235–9.

Renouf, Antoinette (1987) 'Corpus Development', in J. M. Sinclair (ed.) *Looking Up*, London: Collins, 1–40.

Rondeau, Guy (1984) *Introduction à la Terminologie* (2nd edition), Boucherville, Québec: Gaëtan Morin.

Sager, Juan C. (1990) *A Practical Course in Terminology Processing*, Amsterdam and Philadelphia: John Benjamins.

Somers, Harold (2003) 'Translation Memory Systems', in Harold Somers (ed.) *Computers and Translation: A Translator's Guide,* Amsterdam and Philadelphia: John Benjamins, 31–47.

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.

Varantola, Krista (2003) 'Translators and Disposable Corpora', in F. Zanettin, S. Bernardini and D. Stewart (eds) *Corpora in Translator Training*, Manchester: St. Jerome Publishing, 55–70.

Zanettin, Federico (1998) 'Bilingual Comparable Corpora and the Training of Translators', *Meta* 43(4): 616–30.

— (2002) 'diy Corpora: The www and the Translator', in B. Maia, H. Haller and M. Ulrych (eds) *Training the Language Services Provider for the New Millennium*, Porto: Faculdade de Letras da Universidade do Porto, 239–48.

Chapter 10

Style of Translation: The Use of Foreign Words in Translations by Margaret Jull Costa and Peter Bush

Gabriela Saldanha

Leech and Short (1981: 13) declare that the goal of literary stylistics is 'to gain some insight into the writer's art', and they point out that 'we should scarcely find the style of Henry James worth studying unless we assumed it could tell us something about James as a literary artist'. It follows from this that only if we can demonstrate the existence of stylistic patterns that can only be described as the *translator's* art, rather than to a more or less skilful reproduction of the *writer's* style, will we be able to assert that translations are interesting in their own right, that is, for their own stylistic value. This chapter describes patterns in the use of foreign words in literary translations from Spanish and Portuguese into English by Margaret Jull Costa and Peter Bush and argues that those patterns are the result of personal stylistic choices by the translators. These choices are shown to be revealing of the translators' art, in the sense that they tell us something about how the translators conceive their role as intercultural mediators.

Discussions of style in relation to translation have generally been presented from the perspective of the source texts, and focusing on translators' attempts, and in particular *unsuccessful* attempts (see, for example, Parks 1998/2007) to convey that style in the target text. Three works on literary style in translation have paved the way for an account of style from the perspective of the target texts: Baker (2000), Malmkjær (2003) and Munday (2008). Malmkjær introduced the concept of 'translational stylistics', which she describes as explaining why, given the source text, 'the translation has been shaped in such a way that it comes to mean what it does' (Malmkjær 2003: 39). In her illustration of a translational stylistic analysis, Malmkjær compares several translations of Hans Christian Andersen's stories by Henry William Dulcken. Although Malmkjær describes clear patterns of choice in the translations, these are explained mainly in terms of differences between source and target cultures. Malmkjær takes the personal histories of Andersen and Dulcken into account but her interest lies mainly on the style of the text, the *translation*, rather than the style of the *translator*.

Baker (2000), in contrast, focuses on the *translator's* style, which she describes as 'a kind of thumb-print that is expressed in a range of linguistic – as well as non-linguistic – features', including open interventions, the translators' choice of what to translate (when the choice is available to the translator), their consistent use of specific strategies, and especially their characteristic use of language, their 'individual profile of linguistic habits, compared to other translators' (Baker 2000: 245). Baker is primarily concerned with the latter, the 'subtle, unobtrusive linguistic habits which are largely beyond the conscious control of the writer and which we, as receivers, register mostly subliminally' (Baker 2000: 246).

Baker (2000) outlines a methodological proposal for the study of translator style. She starts from the target texts and suggests using the source texts only as a way of testing the effect of conflicting variables. Baker proposes first establishing stylistic patterns in several translations by the same translators, thereafter proceeding to filter the possible variables that may affect such patterns in order to see if those patterns can be attributed to the style of the translator or if they are simply carried over from the source texts as a feature of the source language in general, the poetics of a particular group, or the style of the authors (Baker 2000: 258). The final stage in Baker's proposed methodology involves exploring potential motivations for the patterns revealed using extralinguistic information on the translation, the translation process and the translators' professional backgrounds.

In order to illustrate the methodology proposed, Baker (2000) carries out a corpus-based analysis of five English translations by Peter Bush (one from Portuguese and four from Spanish) and three Arabic to English translations by Peter Clark. She compares the type/token ratio, average sentence length and the use of reporting structures with the verb SAY by each translator.[1] The corpus available to Baker was not a parallel corpus and it was not possible to examine all the source texts. Still, Baker points out that some of the patterns identified as distinctive of Clark's translations, such as the heavy use of modifiers with the verb SAY, may be largely carried over from his Arabic source texts. The source texts could also be influencing the figures for type/token ratio and sentence length given by Baker. Although the type/token ratios in all texts translated by Bush are higher than in those translated by Clark, they are particularly high in translations of Goytisolo's works (three out of the five texts in the Peter Bush corpus), and the same pattern emerges when looking at the averages for sentence length. In addition, the greater variation among texts translated by Bush noted by Baker could be due to the fact that they are translations of texts by three different authors, and two are autobiographies while the rest is fiction. The three translations by Clark, in contrast, are by two different authors and are fiction. Baker's study does not offer definitive results;

its main strength lies in the fact that it proposes a new concept and opens previously unexplored avenues of research.

Munday (2008), like Baker, looks at style from the point of view of the translator. The two main questions addressed in Munday's work are: 'What are the prominent characteristics of style, or "linguistic fingerprint", of a translator compared with the style of the ST author and of other translators?' and 'what is the relationship of the style of the translation to the environments of the target texts, (. . .) in other words, how far is it possible to determine the impact of external factors on the translators' decision-making?' (Munday 2008: 7).

In order to look at style from both source and target text perspectives, Munday looks at several translations by the same translators (Harriet de Onís and Gregory Rabassa) and at the work of one author (Gabriel García Márquez) translated by different translators (Gregory Rabassa and Edith Grossman). Munday's main concern is the links between stylistic choices at the micro-level and the macro-contexts of ideology and cultural production; this inevitably draws him to pay closer attention to those linguistic features that can more easily be explained as meaningful choices (syntactic calquing, syntactic amplification, compound pre-modifiers, creative or idiomatic collocations) rather than the kind of patterns at the lower syntactic level that form the basis of Baker's study and have proved more relevant in revealing the habitual aspects of composition. The discourse analytical approach allows him to establish a clear link between micro-level choices and contextual and ideological factors. The patterns revealed in the work of specific translators are often compared with reference corpora to see whether the strategies used by the translator respond to normalized/idiomatic preferences of the TL or are unusual/original uses.

Saldanha (2005, 2011) further develops the methodology proposed by Baker (2000), combining elements that are also used in Munday (2008) such as parallel texts and comparable corpora, and offers a more in-depth analysis of the stylistic traits that can be attributed to the translator. That methodology is briefly summarized here before focusing on one particular set of results, concerning the use of foreign words in the work of Peter Bush and Margaret Jull Costa, and discussing what these results can reveal about the two translators' approach to their task.

10.1 Methodology

10.1.1 When Can We Talk about 'Translator Style'?

A defining element of style is 'distinctiveness': in every work of fiction, certain linguistic features stand out because they depart from a norm; that is, they are frequent or infrequent in relation to a relative norm of comparison. In other words, for a stylistic trait to be distinctive or characteristic of an author or translator, it has to appear more or less frequently in the work of that author/translator rather than in that of others. This is the basis of what Leech and Short (1981) call *deviance*, which is perceived by the reader as *prominence*.

From a methodological point of view, to demonstrate that a certain stylistic preference is distinctive of the work of one translator, we would need a range of texts by the same translator and a reference corpus, which would consist of, minimally, works by one other translator, and ideally, a representative sample of translations by other translators, working in the same genre and language combination. This is the type of corpus used here, although the reference corpus has some limitations as pointed out below.

However, distinctiveness is a necessary but not sufficient condition. Halliday (1971) argues that the fact that a linguistic feature is prominent does not necessarily mean that it has literary relevance, since there are idiosyncrasies of style that have no discernible literary function. Literary relevance is related to the Prague School notion of 'foregrounding', understood as artistically motivated deviation (Leech and Short 1981: 48) or, in Halliday's words, as prominence that is motivated (1971: 339). For a prominent feature of style to achieve literary relevance it has to form a coherent pattern of choice, together with other features of style, and impact on the meaning of the text as a whole (Halliday 1971). In Halliday's model, whether a pattern is motivated or not depends on whether it contributes to how the text functions at the ideational, interpersonal or textual levels and does this in a way that is coherent with other patterns.

Thus, in order to talk about 'translator style', we need to identify stylistic traits that: (1) are felt to be recognizable across a range of translations by the same translator, (2) distinguish that translator's work from that of others, (3) are 'motivated', in the sense that it has a discernable function, and (4) constitute a coherent pattern of choice. However, all these elements could point to the style of a translator only in principle, since other variables – such as author style or the specific characteristics of a sub-genre or a particular linguistic variety which the translator is dealing with, to name but a few – could also explain the recurrence of certain stylistic features. Therefore, it is also necessary to demonstrate that the stylistic traits can be attributed to the translator and cannot be explained as directly reproducing the source text's style or as the inevitable result of linguistic constraints. A certain degree of diversity – for example, in terms of authors, genre or sub-genre, date of publication, place of publication, language variety and, where possible, language – in the corpus of translations by the same translator would go some way towards ensuring a minimal effect from each variable. However, comparison with the source texts is still essential in order to identify possible triggers for the choices made and explore explanations for such choices.

A final methodological point to make is that, in view of the ultimate goal of literary stylistics described above, finding stylistic traits should not be an aim in itself, but, as Baker (2000: 258) points out, it is only worthwhile if it tells us something about 'the cultural and ideological positioning of the translator'. In order to find a meaningful explanation for the results these have to be triangulated using extralinguistic information about the translator and the context of translation.

10.1.2 The Corpora (CTMJC and CTPB)

The study presented here focuses on the work of Margaret Jull Costa and Peter Bush. These translators were chosen for a number of reasons, mainly having to do with the number and wide range of authors they have translated, and with the similarities in their cultural and professional backgrounds. This allowed for a suitable degree of diversity across the translations by each translator while retaining comparability between the works of the two translators. Both Margaret Jull Costa and Peter Bush have translated from Spanish and Portuguese into English. The source texts they have worked with have been produced in very different cultural backgrounds, and in the case of Margaret Jull Costa, in different historical periods. Regarding the translators' own cultural and professional backgrounds, both are British and lived in Great Britain for most of their adult lives. Neither has an explicitly endorsed political agenda – such as feminist or minoritizing translation – in relation to their professional work. This meant that differences in their style would be unlikely to be attributable to different translation traditions or to allegiance to different schools of thought.

The question of which translations to include in the two corpora was partly determined by the fact that it was necessary to obtain authorization to scan the texts. Kenny (2001: 115) notes that copyright holders introduce an element of self-selection into the corpus, and this was clearly the case in this instance, therefore no attempt at a stratified sampling was made.

The full contents of the *Corpus of Translations by Peter Bush* (CTPB) and the *Corpus of Translations by Margaret Jull Costa* (CTMJC) are listed in Saldanha (2005, 2011). Each corpus consists of five source texts, by five different authors, and their translations. CTPB contains approximately 436,000 tokens and CTMJC approximately 260,000 tokens. The different sizes of the two corpora respond to constraints in the number of texts that could be used for the reasons explained above concerning copyright. Peter Bush is represented by four translations from Spanish and one from Portuguese; Margaret Jull Costa by three translations from Spanish and two from Portuguese. In CTPB all the source texts were published from the 1980s onwards. In CTMJC, the dates of publication of the source texts span more than a century, from 1880 to 1993. All the translations were published in the past 25 years.

The ideal reference corpus would have been a parallel corpus of translations of Spanish and Portuguese narrative prose into English by many different translators. Such a corpus was not available, but COMPARA, a bi-directional parallel corpus of Portuguese and English narrative, was thought to provide a reasonable point of reference (Frankenberg-Garcia and Santos 2003). A key feature of this corpus and one that made it possible to use it as a reference corpus for this particular study is that instances of foreign words are tagged and therefore easily retrievable. The main problem with using COMPARA as a control corpus was that all the translations are from Portuguese, while only three of the ten translations in CTPB and CTMJC are from that language. However, given that Spanish and Portuguese are closely related languages, and that the results to be compared would be those for the English texts only, this was not deemed a major drawback. Another difference is that the texts in COMPARA are extracts (30 per cent of the total number of words) and not full texts as in CTMJC and CTPB, but since the results are normalized, this was not a primary concern. In other respects, COMPARA is a very good source of comparative data. At the time when the analyses were carried out, it included 34 English translations (totalling 700,554 tokens) of 33 different Portuguese texts, and 24 Portuguese translations (675,466 tokens) of 22 different English texts.[2] The translations into English were carried out by 16 different translators and, with one exception, they were all published in the past 35 years. An important advantage of using COMPARA was that it is freely accessible via the World Wide Web through an online concordancer.

10.2 The Data

10.2.1 Highlighted Source Culture Lexical Items in CTPB

Foreign lexical items were chosen for investigation because a previous study of typographical markers in the translations had revealed that italics and quotation marks used to highlight foreign items were often omitted and added (Saldanha 2005, 2011). In the first instance, the study focused on the data retrieved during that previous study, that is, on foreign items that were highlighted using either italics or quotation marks.

It is important to note that not only single words are considered in the results, but also instances where two or more foreign words are introduced at a particular point in the source text or translation as one unit, even when, by other standards, they may not be considered a multi-word unit. An example is *el payo* in the sentence ‘His ancestral distrust of *el payo* would in this case explain his defensive attitude . . . ’ (BGTT).[3] Here *el payo* is counted as one item even if it is an article followed by a noun because it is introduced in the text as a single lexical choice, the representation of one 'concept' or 'idea' which the translator chose to convey in a foreign language. Considering it as two distinct foreign words would artificially inflate the number of foreign items in the text. Another case worth mentioning here is that of *contos de rêis* (a former Portuguese currency) and the shortened form *contos*; these were considered variants of a single lexical item.

In CTPB the number of italics highlighting foreign items that are added in the target texts (and absent in the source texts) is much higher than that of italics present in the source texts and omitted in the target text. Added italics highlighting foreign items are also more evenly distributed among the texts: they appear in all five translations. Italics (and in one case, quotation marks) omitted are a feature of only three of the five translations and 79 per cent of all omitted italics are accounted for by one text pair (Bush-Goytisolo), so they cannot be considered a consistent feature in CTPB.

In most cases the foreign items italicized in the translations are source language items (Spanish and Portuguese). There are 25 Spanish lexical items (43 occurrences) distributed among the four translations from that language, and eight Portuguese words (ten occurrences) in the translation of Buarque's text. Table 10.1 shows the distribution of source language words per translated text, including total number of occurrences, total number of different lexical items and normalized frequency.[4] Frequencies were normalized per 30,000, which is the approximate average length of all the texts included in CTPB, CTMJC and COMPARA. Normalized frequencies are the most reliable indicators because of the different lengths of the texts.

The results in Table 10.1 suggest that the use of highlighted foreign items, in particular, source language words, is a consistent feature in CTPB. The tendency to retain source culture lexical items is also reflected in the use of Catalan and Shuar words in the translations of the works by Goytisolo and Sepúlveda respectively. Those languages co-exist with Spanish in the source cultures and the fact that the Catalan and Shuar words in question are not

Table 10.1 Distribution of source language items in translated texts in CTPB

	Occurrences		**Distinct lexical items**	
Translation	**Total**	**Normalized frequencies**	**Total**	**Normalized frequencies**
BBTT	10	8.5	8	6.8
BGTT	18	6.5	14	5.0
BOTT	20	9.2	7	3.2
BPTT	2	5.5	2	5.5
BSTT	3	3.3	2	2.2
Total CTPB	**53**	**7.2**	**33**	**4.4**

italicized in the source texts might reflect their common use in the linguistic communities where the novels take place. Take, for instance, the case of *riera* in Example 1, which is retained in the target text. In the source text, it offers an example of the 'diluted' Spanish that the author describes as being the language spoken by his immediate family during his childhood.

(1) BGST: . . . se extravió al salir de la estación en el camino de la riera y llegó a casa turbada . . .
BGTT: . . . she left the station on the way to the *riera* and reached home flushed . . .

If we consider all source culture words, that is, if we add words of Shuar and Catalan origin to the list of Spanish and Portuguese words kept in the translations, we have an even clearer pattern (see Table 10.2) showing that Bush tends to punctuate his translations with linguistic items that belong to the source culture.

10.2.2 Highlighted Source Culture Lexical Items in CTMJC

In CTMJC neither the omission nor the addition of italics and quotation marks highlighting foreign items is a consistent feature across the translations. Omissions are concentrated in two text-pairs. Additions are a feature of four of the five translations, but 71 per cent are accounted for by one translation. Frequencies do not correlate with text-lengths, so uneven distribution cannot be attributed to varying sample sizes.

A shared characteristic of the two corpora is that the majority of italicized foreign words in the target texts are words that have been retained in the source language in the translation. In CTMJC, there are 38 occurrences of highlighted Portuguese words and five of highlighted Spanish words in the

Table 10.2 Distribution of source culture lexical items in translated texts in CTPB

	Occurrences		**Distinct lexical items**	
Translation	**Total**	**Normalized frequencies**	**Total**	**Normalized frequencies**
BBTT	10	8.5	8	6.7
BGTT	23	8.3	17	6.1
BOTT	20	9.2	7	3.2
BPTT	2	5.5	2	5.5
BSTT	13	14.2	6	6.6
Total CTPB	**68**	**9.2**	**40**	**5.4**

Table 10.3 Distribution of source language items in translated texts in CTMJC

	Occurrences		Distinct lexical items	
Translation	**Total**	**Normalized frequencies**	**Total**	**Normalized frequencies**
JCSFTT	2	1.5	2	1.5
JCSCTT	8	8.0	2	2.0
JCQTT	30	38.2	7	8.9
JCVTT	3	3.8	1	1.3
Total CTMJC	**43**	**9.4**	**12**	**2.6**

translations. However, apart from being concentrated mostly in one text (almost 70 per cent are accounted for by the translation of Eça de Queiroz's *O Mandarin*), those 43 occurrences represent in fact only 12 different lexical items (see Table 10.3). These results are quite different to those obtained in CTPB, where a total of 53 source language items represented 33 different forms distributed among the five translations (compare Table 10.1 and Table 10.3).

Even though the normalized frequency for all occurrences of source language items is higher in CTMJC than in CTPB, this does not take into consideration the fact that the number of *different* lexical items is considerably lower in CTMJC. Moreover, and this is actually concealed by the normalized totals offered in Table 10.1 and Table 10.3, the distribution is considerably more even across texts in Table 10.1 than in Table 10.3. If we compare the results for each text in the two corpora, we note that the normalized frequency for four of the texts in CTMJC is lower than the lowest frequency in all CTPB target texts (2.2 in BSTT).

It is worth noting that, because the total frequencies are normalized, if more texts were added to CTMJC and the number of different source language words increased along the same ratio as recorded in the five texts already included, this would make no difference to the results presented. Having read other translations by Jull Costa, I have no reason to suspect the patterns described here would change if more translations were added to the corpus.

The differences between the use of source culture lexical items in the translations by Bush and Jull Costa are not only quantitative but also qualitative in nature. The communicative function of these lexical items is described in more detail in Section 10.4 below, but it is worth pointing out here that five of the 12 different forms in CTMJC are names of currency: *real, tostão, conto* and *mil-réis* in Portuguese, and *real* in Spanish. The latter is used only once, but different forms of the Portuguese terms are repeated several times in the translation of Queiroz's text and account for 25 of the 30 occurrences. Bush, on the other hand, never italicizes currency terms. There are several instances of such terms in CTPB: five instances of *pesos* (three in Paz's text and two in Onetti's), four instances of *pesetas* (in Goytisolo's), and two instances of *sucres* (in Sepúlveda's). In all cases the same terms are used in the translations without being italicized.

In view of the above results, it seems reasonable to put forward the hypothesis that retaining source culture lexical items is a characteristic of Peter Bush's translations but not so of Margaret Jull Costa's. However, the foreign items analysed in this study were only those retrieved from concordances of italicized and quoted items; therefore, the possibility that Jull Costa also uses source language items but does not use any typographic devices to highlight them cannot be discarded. For this reason, a more exhaustive search for foreign words was needed to validate the above results.

10.2.3 Non-Highlighted Foreign Lexical Items in CTPB and CTMJC

Since foreign words had not been tagged in the corpora and cannot be identified purely on the basis of graphic characteristics, the only alternative to retrieve a comprehensive list was to use alphabetically ranked wordlists automatically created in Wordsmith Tools and identify foreign words manually. The manual retrieval of foreign words from wordlists is extremely time-consuming. Nevertheless, the possibility of producing word lists automatically made the process quick enough for it to be cost effective, in a way that reading through the novels pencil-in-hand would not have been (see Kenny 2001: 130–2 for a similar problem in identifying creative forms).

Apart from being time-consuming, the manual identification of foreign words is prone to human errors. In order to ensure the highest possible degree of precision and recall, the process of identifying a list of foreign items was carried out in stages, starting with a very inclusive approach, to maximize recall, and ending with a very strict and restricted filter, to maximize precision. A single list was created for all the target texts in both corpora so as to avoid any possible bias from the researcher's expectation that foreign words would be more common in one corpus than another. The first stage of retrieval, carried out by the researcher, involved extracting all foreign items with the exception of proper nouns, personal titles, foreign language quotations and titles of books, magazines, and so forth. The second stage involved the participation of a native speaker who went through the lists filtering out words that were undoubtedly of common use in English. Judging the degree of assimilation of foreign words in a language is a rather complex matter to be left to native-speaker's intuition, but there is no flawless method. The solution opted for in this study was to use inclusion in a standard, comprehensive, English dictionary (the *Collins English Dictionary*) as the ultimate criterion for considering a word as lexicalized in the English language. Although, as Peters (2004: 296) notes, 'dictionaries themselves wrestle with the problem and their conclusions are sometimes inscrutable', a dictionary still provides an informed and reliably independent standard against which the data can be assessed.

After filtering out all instances occurring in the *Collins English Dictionary*, the total number of foreign items in CTPB was 52 and in CTMJC 11. Some of those items were repeated, so in CTMJC the total number of instances of foreign items was 22 and in CTPB 90. Because Wordsmith Tools neutralize all typographical differences, foreign words retrieved from the word lists necessarily overlapped with those retrieved from concordances of italicized items. As a result, although these results are invaluable in the sense that they completely rule out the possibility that Jull Costa might be using more foreign words but without italics, they do not change substantially the picture obtained from the findings presented in the previous section concerning the use of highlighted foreign items; they simply add further supporting evidence. In CTMJC, 10 of the 11 items were italicized; in other words, the only new foreign word revealed is *meseta* (plateau). In CTPB, 46 of the 52 foreign words were italicized. Those written in roman type were one French word, *palafitte* (stilt house) and the following five Spanish words: *calles* (streets), *chicha* (an alcoholic drink), *chirimoya* (custard apple), *partituras* (music scores) and *prú oriental* (a drink).

10.2.4 Comparative Data

Data from COMPARA was used to provide a relative norm of comparison against which the results from both corpora could be assessed. The aim of this comparison was to determine to what extent the frequent use (in Peter Bush's case) or infrequent use (in Margaret Jull Costa's case) of source culture items distinguishes the work of these translators from that of others. Non-italicized words were not counted because they cannot be retrieved from COMPARA. A total of 177 instances of foreign words were found in COMPARA, representing 66 different lexical items (see Table 10.4). The normalized frequency is 3.8, lower than that in CTPB (4.4) and higher than that in CTMJC (2.6). If we include words of Shuar and Catalan origin that are not italicized in the source text but only in the translation in CTPB, then the normalized frequency for CTPB is 5.3. Although the differences are not striking, it is interesting that the results from CTPB and CTMJC point in different directions in relation to the results from COMPARA. It is also important to remember that normalized totals average out differences in the distribution across files. In CTMJC, almost 60 per cent of all occurrences of source language words are concentrated in one translation while one of the five texts contains no source language words. In COMPARA almost half (i.e. 13) of the 29 extracts used for this study contain no foreign words, and approximately a third (i.e. 23) of all occurrences are accounted for by one translation.

Table 10.4 Source language words used in COMPARA, CTMJC and CTPB

	COMPARA	CTMJC	CTPB
Number of words in corpus	516,743	136,534	221,987
Total number of occurrences of SL lexical items in TTs	177	43	53
Total number of different SL lexical items in TTs	66	12	33
Normalized frequency of different lexical items	3.8	2.6	4.4

It is important to point out that although a pattern of choice seems to be evident from the results obtained, this is not pervasive enough within each text to be considered a prominent stylistic feature of the individual texts themselves. It is only when we take several translations into consideration, and when these are compared to the work of other translators, that the choice appears as characteristic of the two translators. This is because the analysis focuses on one single and not very common stylistic feature; but it is expected that once a wider range of translator-specific features is revealed using similar methods, their collective impact on the text can be more easily determined.

10.3 Beyond Frequencies: The Communicative Function of Source Culture Items in Translation

Given the diversity of texts included in each corpus, it seems unlikely that it is source text characteristics that require the use of source language items in one corpus and not the other. Still, some could argue that it is the cultural specificity of the terms appearing in the source text translated by Peter Bush which requires the use of foreign lexical items. However, a detailed qualitative analysis of the communicative function of foreign words used in the texts suggests otherwise: first, not all source culture lexical items used in the translations are culture specific; second, the use of borrowings – that is, the verbatim transferring of the culture-specific item into the target text (Hervey and Higgins 1992: 31) – is only one possible way of dealing with culture-specific items in translation; and finally, there are subtle but revealing differences in *when* and *how* the two translators use even the *same* cultural borrowings.

10.3.1 Cases of Self-Referentiality

Source culture items in the translations can be roughly categorized into two types: cases of self-reflexiveness or self-referentiality (Hermans 1996), and culture-specific items. Self-referential and culture-specific items are not mutually exclusive. Cases of self-referentiality often involve words mentioned rather than used. These do not always pose a problem for the translator. A problem arises only when there is no equivalent in the target language (see Examples 9, 10 and 11 below), or when there is an explicit or implicit reference to the linguistic system the word or expression belongs to (see Example 3). In the first case, it is the cultural or linguistic specificity of the word that poses the problem and not the fact that it is mentioned rather than used, so these cases can be subsumed within the category of culture-specific items. The second case presents instances of what Hermans (1996: 29) calls cases of self-reflexiveness or self-referentiality involving the medium of communication itself. Hermans mentions as 'obvious cases' of self-referentiality those where texts 'affirm being written in a particular language', or 'exploit their idiom through polysemy, wordplay and similar devices'. In Example 2, the second instance of *reconocerlo* is used to refer to the word itself, but because there is no reference to the linguistic system that the word belongs to, or to other signifiers in that system, it can be easily translated by the English word 'recognize'. This is not so in Example 3. Here, the speaker explains (despite several interruptions by his interlocutor, which have been omitted here) that the word *mandarim* comes from the Portuguese *mandar*, and it is clear from the dialogue that this is the language spoken by his interlocutor, whose voice is also that of the narrator. Translating *mandar* with an English word would result in an incongruity, although a rendering that involves omitting the Portuguese word would still be possible (e.g. '[*Mandarin*]' comes from the Portuguese word for 'command'):

(2) JCVST: . . . y sonríe aliviada al reconocerlo a Alfredi. Y reconocerlo es la palabra porque el médico-taxista lleva puesta (mal) una barba postiza blanca.
JCVTT: . . . only to smile with relief when she recognizes Alfredi. And 'recognize' is the right word since the doctor-cum-taxi driver is at this point wearing a (clumsily applied) false white beard.

(3) JCQST: 'Mandarim' [. . .] É o nome que no século XVI os navegadores do seu país, [. . .] deram aos funcionários chineses. Vem do seu verbo [. . .] Do seu lindo verbo 'mandar' . . .
JCQTT: 'Mandarin' [. . .] It's the name the sixteenth-century navigators from your country [. . .] gave to Chinese officials. It comes from the verb [. . .] From that lovely verb of yours 'mandar' – to command.

Hermans (1996: 29) argues that cases of self-referentiality are likely to trigger the intrusion of the translator's voice in the narrative text. This is what happens in Example 3, where the translator's strategy involves retaining the source language reference. Sudden departure from the language of the translation is bound to remind the readers that they are reading a translation. In all the cases of self-referentiality involving the medium of communication itself in CTPB and CTMJC, both translators have opted for a solution that does disturb, to some extent, the illusion of transparency, but in rather different ways. In the two cases where a word is mentioned rather than used in CTMJC and reference is made to the linguistic system to which it belongs, Jull Costa intervenes to provide a gloss for the Portuguese. Those two cases are Example 3 above and the word *chá* which appears in the same translation and is rendered as the word for tea, 'chá' (in JCQTT). Bush, on the other hand, consistently leaves the words to stand for themselves in all cases, without providing glosses, as in Example 4, where a Catalan term is compared to a Spanish one:

(4) BGST: . . . la belleza misteriosa del término 'luciérnaga' frente a la grosería y miseria del 'cuca de llum' local
BGTT: . . . the mysterious beauty of the term *luciérnaga* as opposed to the miserable obscenity of the local *cuca de llum*

10.3.2 Culture-Specific Items

Culture-specific items reflect the absence – at least from the translator's point of view – of a target text item that, given the context, can perform the same function as that performed by the source item in the source text. Aixelá (1996: 58) defines them as

> textually actualized items whose function and connotations in a source text involve a translation problem in their transference to a target text, whenever this problem is a product of the nonexistence of the referred item or of its different intertextual status in the cultural system of the readers of the target text.

Paradoxically, the cultural specificity of an item is not determined by the culture the item belongs to, but by the culture it is absent from. The non-existence of, or different value assigned to, the given item in the target culture, however, also depends on the context in which it appears. The context is important because the polysemous nature of words means that the same item may be culture-specific in one context and not in others, depending on the specific function that it fulfils in each. The degree of repetition of a culture-specific term in a certain text is another factor that may influence the translator's decision. Repetition facilitates the success of cultural borrowing, because it gives the receivers of the translation the opportunity to absorb both the form of the expression and its cultural content (Ivir 1987: 38). An example is the Spanish word *señora*, which is used several times in one of the translations by Jull Costa and twice in a translation by Bush.

Señora is used to refer politely to an adult woman and – as opposed to *señorita* – generally implies that the woman is married. *Señora* can also be used as a rather deferential form of address, either as a vocative or as a title. It is a term commonly used by maids, cleaners and other domestic workers to address their employers when these are women. The word *señora*, depending on the context, can be rendered in English as 'woman', 'lady', 'Mrs', 'madam', and in most cases the translation is unproblematic, as in Example 5:

(5) JCSFST: La señora, aunque había bailado con él en los teatros de París . . .
JCSFTT: Although the lady had danced with him in the theatres of Paris . . .

However, in Valenzuela's *Bedside Manners* the word *señora* is used in such a way that it becomes a culture-specific element. Example 6 shows how the term is first introduced, as used by a maid, María, in order to address the main character in the story. Jull Costa opts for keeping the Spanish word. She does not provide a gloss or add any form of contextual information, but the word is included in the *Collins English Dictionary* and most English-speaking readers can be expected to be familiar with it. Besides, it appears in the text preceded by 'address as', which leaves clear its function in the text.

(6) JCVST: ella . . . ni se había dirigido a María al llegar, ni le había dicho su nombre ni le había hecho pedido alguno. María por lo tanto la llama Señora, y ella se siente bien como Señora, en la cama, sin ganas de moverse.
JCVTT: she hadn't even spoken to María when she got there, hadn't even introduced herself or asked her for anything. María therefore addresses her as 'Señora' and she enjoys being the 'Señora', lying in bed, with no desire to move.

Before the main character is addressed as *Señora* by the maid, the narrator refers to her as *una mujer* (a woman) but from that moment on she becomes the *señora* both in the source and target texts, and we never know her name. In the source text, apart from the first two instances (illustrated in Example 6) *señora* is not capitalized. In the target text, however, the word is used with initial capital, which highlights the fact that it is a title (see Example 7):

(7) JCVST: – . . . Como corresponde, señora. Por qué no me dejará tranquila, se pregunta la tal señora, . . .
JCVTT: . . . Exactly what you need, Señora. Why won't she leave me in peace? The Señora wonders, . . .

In this text, the term is used in such a way that it becomes imbued with connotations that effectively describe the woman's situation in relation to that of other characters: unlike them, she is a middle class woman who is seriously out of touch with the events unfolding around her. None of the available English equivalents mentioned above can be used in all the instances where *señora* is used (compare Example 6 and Example 7 with Example 5 above), and it is this 'gap' in the target language that makes *señora* a culture-specific item in this particular context.

Bush, like Jull Costa, translates *señora* as 'lady' or 'woman' in most cases, but he also keeps the Spanish word in one instance (Example 8). However, this is not a case where the term becomes a keyword through repetition and the context is not as informative as in Example 6: although it can be assumed that the average English-speaking reader will be familiar with this word, it may not be so clear why a woman may be 'annoyed' at being referred to as *señora.*

(8) BOST: Ossorio consiguió un tono agresivo para decir: – Si se puede hablar sin reservas me gustaría hablarte. O podemos salir. No conozco a la señora.
Sabía que la palabra señora iba a crispar a la mujer, . . .
BOTT: Ossorio managed an aggressive tone of voice. 'If we can talk here quite freely I would like to talk to you. Or we can go out. I don't know the *señora*.' He knew the word *señora* would annoy the woman; . . .

While in Jull Costa's translation the use of *señora* was heavily determined by its context, this case strikes us as a less obvious instance of cultural specificity; the word is not repeated outside the example, and 'lady' could also have been a valid choice. This example shows how the two translators can opt to reproduce the same source language word but under very different circumstances.

Another interesting example is provided by the two translators' choices when dealing with very similar culture-specific terms: formal and informal forms of address. In the novels by Onetti and Sá-Carneiro, translated by Bush and Jull Costa respectively, the formal and informal forms of address in Spanish and Portuguese are used self-referentially in various instances. Although both translators reproduce in the translations the actual source language forms (*usted* and *tú* in Spanish and *você* and *tu* in Portuguese), they generally differ in the choice of verb introducing them (see Example 9 and Example 10).

(9) BOST: Siempre, en todo caso, nos trataremos de usted.
BOTT: ' . . . we must always use *usted* to each other.'

(10) JCSCST: . . . eu e Ricardo não nos tratávamos por tu,
JCSCTT: . . . Ricardo and I never addressed each other as 'tu',

The Spanish *trataremos de* and the Portuguese *tratávamos por* both mean 'address as'. The lemmas in the two languages are cognates; surface differences in the examples are explained partly by the fact that the Spanish verb is marked for future tense, while the Portuguese verb is in the past tense. Bush translates the Spanish as 'use "usted" to each other' in all instances, whereas Jull Costa opts for the more explicit 'address each other as "tu" ' or 'call each other "tu" ', making clear that what is being discussed is a form of address (see also Saldanha 2008). Note that the idea of 'addressing' is also present in the Portuguese *tratar por*, and in the Spanish *tratar de*, so it is only when contrasted with the choice of 'use' in the translation by Bush that Jull Costa's lexical choice strikes us as more explicit. In Example 11, however, the solution adopted by Jull Costa involves two explicitating shifts: first, *você* is qualified as 'formal', and second, 'call each other' is added before *tu* in a place where the source text is much more vague.

(11) JCSCTT: E olha, fica combinado: de hoje em diante acabou-se o 'você'. Viva o 'tu'!
[Literally: And look, it's agreed: from today onwards there's no more 'você'. 'Viva' the 'tu'!]
JCSCTT: Look, from now on, we'll have no more of this formal 'você' business. From now on we call each other 'tu'.

The translation of forms of address shows how even in similar situations the two translators' approaches can differ in subtle but rather revealing ways. This difference is part of a more general trend in the way Bush and Jull Costa deal with culture-specific items. Jull Costa uses cultural borrowings only when strictly necessary and ensures that the reader is provided with enough information to work out the meaning of the foreign words. Sometimes the information provided by the context in the source text itself is deemed enough, as in Example 6. Other times, the translator adds relevant information, although not usually in the form of intra-textual glosses as such, but rather in the form of subtle contextual clues that facilitate comprehension while trying to make the translator's voice as unobtrusive as possible as in Example 11. In Example 12, in the source text, *guardia civil* is a collocate of 'color', and it is an unusual collocate for that node, an instance of what Kenny (2001: 134–41) calls 'creative collocation'. In the target text, Jull Costa borrows the Spanish term and, although she does not define it, she provides two important pieces of information, namely, that a *guardia civil* wears a uniform, and that this uniform is green. What is more, the metonymy in the source text is rendered as a simile in the target text, so the comparative element is also more explicit (see Weissbrod 1992).

(12) JCSFST: En un rincón había una montaña de botellas, color guardia civil, cubiertas de polvo.
[Literally: In one corner there was a pile of bottles, guardia-civil colour, covered in dust.]
JCSFTT: In one corner there was a pile of dusty bottles, green as a *guardia civil's* uniform.

The higher number and variety of culture-specific terms in CTPB are indicative of a willingness to let the source culture shine through in the translation. Although in some cases the source language words represent culture-specific items for which there is no close equivalent in English, for example, *capirinha*,[5] or *almogávares*;[6] in others it would have been possible to offer an English translation. See for instance Example 1 above. Another example is *tómbola*, which can be translated as 'tombola', but where Bush has chosen to use the Spanish spelling; another is *bachillerato*,[7] which despite being to some extent culture-specific, can be translated as 'secondary school' in certain contexts (Example 13).

(13) BGST: Apoyándonos uno en el otro, llegamos a concluir nuestro bachillerato paticojo sin demasiados tropiezos.
BGTT: By helping each other, we managed to finish our crippled *bachillerato* without too many mishaps.

Bush's strategy differs from that of Jull Costa's not only in that he is more likely to use cultural borrowings but in that he rarely adds information that will clarify the meaning of borrowed terms. In CTPB there are 36 different source language terms that can be considered culture-specific, and intra-textual glosses are provided for only three of them: *felipes* (glossed as 'Popular Liberation Front'), *equis* (rendered as 'equis-viper') and *chicha* (rendered as 'chicha beer'). If Catalan and Shuar culture-specific terms are also taken into account, then intra-textual glosses are provided for five of the 46 culture-specific items to be found in Bush's translations. The word *tzantzas* is glossed as 'the sloths' and the word *yahuasca* as 'yahuasca plant'.

Where no explicitation accompanies the cultural borrowing, sometimes the meaning is nonetheless clear from the context. However, contextual information is not always sufficient to make out the meaning of the foreign items, as in Example 14.[8] This situation never arises in relation to any of the culture-specific items found in CTMJC.

(14) BGST: . . . su apellido no es catalán y de probable ascendencia gitana. Su desconfianza ancestral del payo explicaría en este caso su actitud defensiva ante la vida . . .
BGTT: . . . his surname isn't Catalan and sounds gypsyish. His ancestral distrust of *el payo* would in this case explain his defensive attitude to life . . .

Finally, it is worth noting that the avoidance of source culture borrowings is in line with Jull Costa's general reluctance to use any foreign words, since she does not always carry across to the translation words of other foreign origins found in the source text. Of all instances of foreign words (excluding English words) in the source text, Jull Costa retains only 46 per cent. The following is an example where a French word in the ST is not kept in the target text:

(15) JCSCST marcamos *rendez-vous* para a noite seguinte, na Closerie,
JCSCTT we arranged to meet the following night at ten in the Closerie.

Bush, on the other hand, keeps 90 per cent of all instances of foreign words (excluding English words) in the source text, and occasionally uses French words as translations of Spanish and Portuguese words (see Example 16).

(16) BOST: . . . dio un paso en la luz mostrando de golpe su cara, como en un calculado efecto de teatro, . . .
BOTT: . . . and stepped forward into the light, suddenly revealing his face, like some premeditated *coup de théâtre*, . . .

By choosing to use a foreign form to represent a particular phenomenon, the writer/translator places the phenomenon outside – or at least removed from – the implied reader's experience of the world. By translating it, the writer is bringing the phenomenon within the realm of what is familiar to the implied reader. Therefore, in terms of stylistic significance, the use of source culture words, and in particular cultural borrowings, can be said to affect the ideational and interpersonal functions of the translation.

10.4 Beyond the Text: What Do These Results Tell Us about the Translators?

Meta-textual data (including interviews with the translators, pieces of academic writing by the translators and reviews of their translations) were used in Saldanha (2005) in order to describe the translators' *positions* and *projects*, and the *horizon of translation* (Berman 1995) and thus contextualize the results. The analysis of meta-textual data suggests that the most important factor in determining the translators' choice is their different conceptualization of their role as intercultural mediators, in particular in relation to their readership.

During (separate) interviews with the translators, they were asked what they thought of the use of foreign words in translations. Jull Costa explained that she uses them only when the concept does not exist in English; 'but it's very rare'. She does not like using them, and as justification she offers, simply: 'because I am a translator'. Bush, on the other hand, asked in return: 'how does one determine what the readers' reactions are like?' The tendency is to avoid referring the reader to a dictionary, but 'why not? Some readers like to look up things in the dictionary and find out what the meaning of the foreign word is'.

So, while Jull Costa sees it as her job to bridge the cultural or linguistic gap, Bush is more willing to let his readers make the effort themselves. Bush also argues that readers tend to be 'patronized'. This is consistent with a general disposition to challenge readers which becomes obvious in Bush's own discourse on translation (see, for example, Bush 1999).

I would not go as far as saying that Jull Costa and Bush have different readerships in mind. They both translate for an educated English-speaking readership that is prepared to read translated literature, including 'difficult' writers such as Goytisolo or Saramago. They differ, however, in terms of how far they will go to meet the audience on its own terms and their willingness to align themselves occasionally with the source culture and present translated language as the language of an 'out-group'. Bush's position is that 'although the translator will inevitably think about the eventual readerships for his translation, the reader he must translate for is himself, as no-one else will be so embedded in the struggle between original and nascent text' (Bush 2002: 23).

Jull Costa, on the other hand, does not want the person reading the translation 'to come across things which may distract them from their reading experience'; she sees it as her challenge 'to make them [readers] stop thinking that translations are not worth reading, that they are not, somehow, the real thing' (personal communication). Her strategy, however, is not to challenge her readers but to reach out to them. Jull Costa wants her translations to be acceptable in the terms established by the target culture, her translations are driven by a desire to make their reading a pleasurable experience, which is not interrupted by encounters with information, such as foreign words, that the readers cannot process in their own cognitive environment.

10.5 Conclusion

This chapter has presented an analysis of the use of foreign words in translations by Margaret Jull Costa and Peter Bush based on the methodology proposed by Baker (2000) and further elaborated on in Saldanha (2005, 2011). Jull Costa was found to use fewer source culture items and Peter Bush more source culture items than is common in translations from Portuguese into English. An analysis of the communicative function of foreign words in translation suggested that Bush is more likely than Jull Costa to use cultural borrowings as a strategy for dealing with culture-specific terms. Jull Costa avoids using lexical items that are unfamiliar to the reader; and when she does, she provides contextual information that facilitates their understanding. Foreign lexical items, especially if unfamiliar to the reader, can be expected to disturb the reading experience, reminding readers of the fact that they are reading a translation and increasing the difficulty involved in processing the information. The choice of more familiar and more explicit renderings, in contrast, helps to produce a more coherent text. This is in line with other results (Saldanha 2005, 2008, 2011, forthcoming) concerning the use of emphatic italics and of the optional connective 'that' after verbs of speech in the two corpora. When considered as a whole, the results present a coherent pattern of choice and, in the light of extra-textual information about the translators and translation, I argue that they reflect the two translators' different conceptualizations of their readership and of their role as intercultural mediators. Cumulative evidence of translators' stylistic profiles, along the lines of what has been presented here, will hopefully serve to highlight the distinctiveness of the literary translators' work and, in the long term, it may help to change the way style is seen in relation to translation.

Notes

[1] SMALL CAPITALS are used here and elsewhere to represent lemmas.

[2] COMPARA is an on-going project and texts are constantly being added to it. Up-to-date information on the corpus is available from the website: http://www.linguateca.pt/COMPARA/ (accessed on 13 March 2011).

[3] The texts from the two corpora are identified by an acronym consisting of the translator's initial (B for Bush, JC for Jull Costa), the author's initial (G for Goytisolo, for example) and then ST for source text and TT for target text.

[4] Note that, because one item appears in two different texts (*pensión*, in BOTT and BGTT), the source language items total 33 in this table.

[5] A typical Brazilian cocktail made with the local sugar cane rum.

[6] Specially trained Spanish soldiers famous for their role in the Christian re-conquest of Spain during the thirteenth Century.

[7] Qualification obtained when finishing secondary school in Spain and certain Latin American countries.

[8] *Payo*, according to the *Diccionario de la Real Academia Española*, can be a word used by gypsies to designate 'someone not belonging to their race' (my translation).

References

Aixelá, Javier Franco (1996) 'Culture-Specific Items in Translation', in Román Álvarez and M. Carmen-África Vidal (ed.) *Translation, Power Subversion*, Clevedon: Multilingual Matters, 52–78.

Baker, Mona (2000) 'Towards a Methodology for Investigating the Style of a Literary Translator', *Target* 12(2): 241–66.

Berman, Antoine (1995) *Pour une critique des traductions: John Donne*, Paris: Gallimard.

Bush, Peter (1999) 'Translating Juan Carlos Onetti for Anglo-Saxon Others', in Gustavo San Román (ed.) *Onetti and Others*, Albany, NY: SUNY, 177–86.

— (2002) 'The Act of Translation: The Case of Juan Goytisolo's *A Cock-Eyed Comedy*', in Keith Harvey (ed.) *CTIS Occasional Papers* 2, Manchester: CTIS, UMIST, 21–32.

Frankenberg-Garcia, Ana and Diana Santos (2003) 'Introducing COMPARA: the Portuguese-English Parallel Corpus', in Federico Zanettin, Silvia Bernardini and Dominique Stewart (eds) *Corpora in Translator Education*, Manchester: St. Jerome, 71–87.

Halliday, M. A. K. (1971) 'Linguistic Function and Literary Style: An Inquiry into the Language of William Golding's *The Inheritors*', in Seymour Chatman (ed.) *Literary Style: A Symposium*, London and New York: Oxford University Press, 330–65.

Hermans, Theo (1996) 'The Translator's Voice in Translated Narrative', *Target* 8(1): 23–48.

Hervey, Sándor and Ian Higgins (1992) *Thinking Translation: A Course in Translation Method: French-English*, London and New York: Routledge.

Ivir, Vladimir (1987) 'Procedures and Strategies for the Translation of Culture', in Gideon Toury (ed.) *Translation across Cultures,* New Delhi: Bahri Publications, 35–46.

Kenny, Dorothy (2001) *Lexis and Creativity in Translation: A Corpus-Based Study,* Manchester: St. Jerome.

Leech, Geoffrey and Michael H. Short (1981) *Style in Fiction: A Linguistic Introduction to English Fictional Prose*, London and New York: Longman.

Malmkjær, Kirsten (2003) 'What Happened to God and the Angels: An Exercise in Translational Stylistics', *Target* 15(1): 37–58.

Munday, Jeremy (2008) *Style and Ideology in Translation: Latin American Writing in English.* London and New York: Routledge.

Parks, Tim (1998/2007) *Translating Style: The English Modernists and their Italian Translations*, 2nd edn, Manchester: St. Jerome.

Peters, Pam (2004) *The Cambridge Guide to English Usage*, Cambridge: Cambridge University Press.

Saldanha, Gabriela (2005) 'Style of Translation: An Exploration of Stylistic Patterns in the Translations of Margaret Jull Costa and Peter Bush'. Unpublished PhD Thesis, Dublin: School of Applied Language and Intercultural Studies, Dublin City University.

— (2008) 'Explicitation Revisited: Bringing the Reader Into The Picture'. *Trans-kom* 1(1): 20–35.

— (2011) 'Translator Style: Methodological considerations'. *The Translator* 17(1): 25–50.

— (in press) 'Emphatic Italics in English Translations: Stylistic Failure or Motivated Stylistic Resources?' *Meta* 56(2).

Weissbrod, Rachel (1992) 'Explicitation in Translations of Prose-fiction from English to Hebrew as a Function of Norms', *Multilingua* 11(2): 153–71.

Chapter 11

A Link between Simplification and Explicitation in English–Xhosa Parallel Texts: Do the Morphological Complexities of Xhosa Have an Influence?

Koliswa Moropa

In this chapter, I begin with an overview of corpus-based research projects in the African languages of South Africa before exploring the link between simplification and explicitation, and attempting to establish whether it can be assumed that the concordial system of Xhosa has some bearing on the connection observed between simplification and explicitation in translated texts. Since the breaking up of sentences is a translation strategy from which the entire discussion in this chapter emanates, I discuss some of the morphological features of Xhosa linked to the noun class system, the grammatical concord which brings about agreement between words in a sentence. A brief overview of the concept of translation 'universals' is provided and it is followed by the analysis of simplification and explicitation strategies. The breaking up of sentences is first discussed, and thereafter I illustrate how, in turn, this strategy leads to explicitation strategies such as the insertion of an explicit demonstrative at the beginning of the second target language sentence, lexical repetition and the addition of explanatory information. Examples cited are taken from the *English-Xhosa Parallel Corpus.*

11.1 Research Using Parallel Corpora in African Languages

Since the dawn of democracy in South Africa and the advent of a multilingual language policy, there has been a significant increase in the demand for the translation of hegemonic languages (English and Afrikaans) into the previously marginalized languages (Xhosa, Zulu, Ndebele, Swati, Northern Sotho, Southern Sotho, Tswana, Tsonga and Venda) and vice versa. Although the volume of translation into African languages has increased substantially, the problem is that translated texts, due to a lack of terminology, are not of a uniformly acceptable standard. This has prompted a number of South African scholars to begin using parallel corpora as part of their research on terminology and standardization, as well as on typical translation strategies. In Moropa (2005, 2007) I provide an overview of the two most salient corpus projects by Madiba (2004) and Gauton and De Schryver (2004) in this regard, which are also discussed briefly here.

Madiba (2004), using the *Special Language Corpora for African Languages* (SPeLCAL), illustrates how parallel corpora can be used as tools for developing the indigenous languages of South Africa. The SPeLCAL project was born out of the need for language resources to support the implementation of South Africa's multilingual language policy adopted after the democratic changes of 1994. The purpose of SPeLCAL (Madiba 2004: 136) is twofold:

- To provide a language resource for the compilation of specialized dictionaries, terminology lists and glossaries in the 11 official languages of South Africa;
- To provide a resource for research in linguistic fields such as terminology, terminography, translation, language for special purposes (LSP) and second language teaching.

In a pilot analysis, Madiba (2004) used Multiconcord to analyse translation equivalents of terms such as 'act', 'legislation', 'rule', 'order' and 'law' in a parallel corpus of English-Venda texts of *The Constitution of the Republic of South Africa* (Republic of South Africa 1996).

Gauton and De Schryver (2004) demonstrate how special-purpose multilingual and parallel corpora can be used as a translator's tool in finding suitable term equivalents when translating technical texts from English to Zulu. In conducting the research, the following general-language corpora were used:

- The *University of Pretoria Zulu Corpus* (PZC), an electronic corpus of five million running Zulu words established at the University of Pretoria by De Schryver and Dlomo. The corpus comprises literary texts, religious texts, internet files and pamphlets (Case Study One)
- The *University of Pretoria Internet English Corpus* (PIEC), an electronic corpus of 12.4 million English words 'culled' from the internet by Gauton (Case Study Two)

In Case Study One, multilingual corpora were used to investigate terminology used in the translation of HIV/AIDS texts using the Key Words Function in WordSmith Tools. In Case Study Two, parallel corpora dealing with labour issues were used to investigate labour terminology in order to determine the usefulness of such corpora as a resource for the translation of technical texts into Zulu. The corpora were queried using the parallel concordancing software program ParaConc. Gauton and De Schryver (2004) conclude that the best terminological results were obtained when using parallel corpora in conjunction with parallel concordancing software and translation equivalents can easily be identified with or without sentence alignment. In such cases, parallel corpora function as translation memory.

Elsewhere (in Moropa 2005, 2007), I demonstrate in detail how corpus-based research can contribute to the analysis of term formation processes in Xhosa. Examples cited are taken from the *English-Xhosa Parallel Corpus*, which comprises the following technical texts and their translated Xhosa versions:

- Annual Report of the Department of Arts, Culture, Science an*d Technology* (DACST) for 1997
- *A Short Guide to the White Paper on Local Government* (1998)
- Annual Report of the Pan South African Language Board (PanSALB) for 2001/2002
- Three manuals dealing with the Promotion of Access to Information A*ct (PAIA)* (2003)

Moropa (2005, 2007) analyses the *English-Xhosa Parallel Corpus* by means of ParaConc, a parallel concordancing program which was developed by Michael Barlow for linguists and translators who wish to work with translated texts. For the successful operation of this software, the texts must be aligned (cf. Figure 11.1).

In this chapter, I again draw upon the *English-Xhosa Parallel Corpus* in examining the link between simplification and explicitation, and attempting to establish whether it can be assumed that the concordial system of Xhosa has some bearing on the connection observed between simplification and explicitation in translated texts.

The concordial system which seems to impact on Xhosa translated sentences is outlined in the following paragraphs. This feature is characteristic of Xhosa as well as other South African indigenous languages.

11. 2 The Concordial System in Xhosa

When Xhosa was rendered in written form by the missionaries in the 1800s, they observed that the language was characterized by a ‘peculiarity’ which caused it to differ in its grammatical structure from other languages. Although this uncommon feature was a subject of much thought, it continued to be enveloped in mystery until it was designated by Rev W. B. Boyce in 1834 as the principle of euphonic/alliteral concord. Today we simply refer to it as a concordial relationship. The concordial system can be described as a frequent

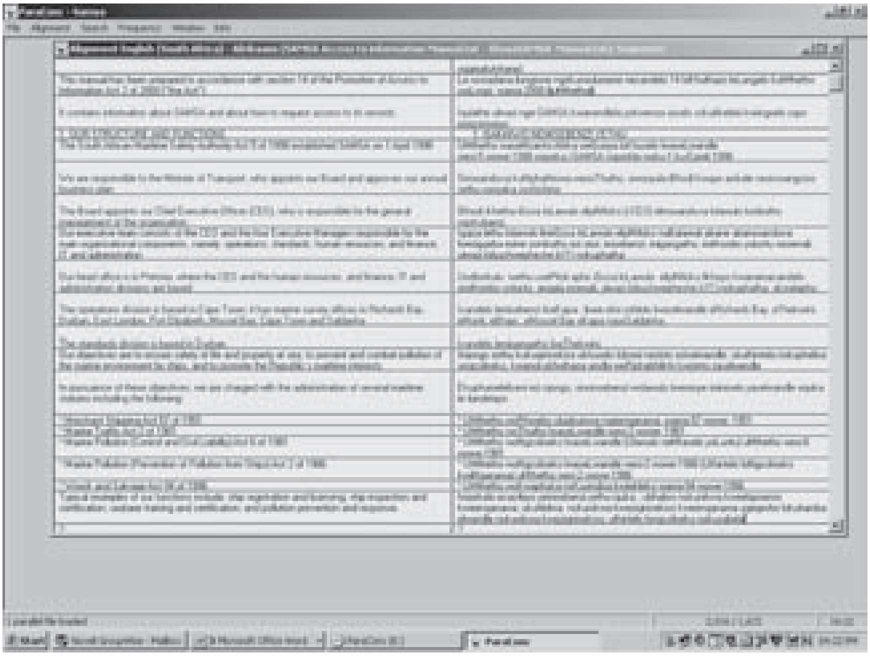


FIGURE 11.1 Aligned sentences in SAMSA *Access to Information Manual* (2003)

repetition of certain morphemes in the same sentence, and this promotes the euphony of the language (Moropa 2005: 82). The two sentences analysed below are extracted from aligned sentences in Figure 11.1.

(1) ST: We are responsible to the Minister of transport, who appoints our Board and approves our annual business plan.
TT: *Sinoxanduva kuMphathiswa wezoThutho owonyula iBhodi kwaye amkele nesicwangciso sethu sonyaka soshishino.*

(2) ST: Our executive team consists of the CEO and four Executive Managers responsible for the main organizational components, namely;
TT: *Igqiza lethu lolawulo lineGosa loLawulo eliyiNtloko naBalawuli abane abanoxanduva lwezigqeba ezine zombutho ezi zezi.*

In Example 1,[1] the noun *isicwangciso* 'plan' belongs to the *isi*-class and then the concords *se*- *so*- *so*- in the rest of the sentence bring about agreement between the noun and the qualifiers. In Example 2, the concordial relationship is brought about by *li*-, *ba* and *zi*-. The noun *igqiza* ('team') belongs to the

Table 11.1 Xhosa noun classes

Singular		**Plural**	
Class	Prefix	Class	Prefix
1	*um-*	2	*aba- (abalawuli)*
1a	*u-*	2a	*oo-*
3	*um-*	4	*imi-*
5	*i(li)- (igqiza)*	6	*ama-*
7	*isi- (isicwangciso)*	8	*izi- (izigqeba)*
9	*in-*	10	*izi(n)-*
11	*u(lu)-*	10	*izi(n)-*
14	*ubu-*		
15	*uku-*		
16	*pha-*		
17	*uku-*		

i(li)-class, *abalawuli* ('managers') belongs to the *aba*-class and *izigqeba* ('components') belongs to the *izi*-class (cf. Table 11.1: Xhosa noun classes).

Nouns in Xhosa are grouped together into various classes according to prefixes. Nouns that have the same prefix belong to the same class. Most of the classes occur in pairs, of which one is the singular and the other is plural. Table 11.1 is an illustration of the noun class system found in Xhosa. The nouns discussed above are put in brackets next to the relevant class.

The noun prefix is a very important feature of the Xhosa language, because all words which may stand in a special relation to a substantive are brought into agreement with it by the class concord. There are several concords such as subject concord, object concord, adjectival concord, possessive concord and relative concord. These concords are written together with that particular word because they are morphemes (Moropa 2005: 84). The subject concord does the work of a pronoun but morphologically it cannot be classified as a pronoun because it is a formative and not a word.

Examples 3 to 5, obtained by means of parallel concordance searches, are an illustration of the concordial relationship brought about by a class concord, and how this impacts on word frequency. The possessive stem -*thu* ('our') is used as an example. Each noun takes a different possessive concord, for example: *le*-, *we*- and *ze*-.

(3) *-thu* ('our') with class 5 noun = *lethu* (SAMSA Manual 2003):
ST: **Our** executive team consists of the CEO and the four Executive . . .
TT: *Igqiza* ***lethu*** *lolawulo lineGosa eliyiNtloko naBalawuli abane* . . .
ST: This means that if **our** information officer refuses . . .
TT: *Oku kuthetha ukuba, ukuba igosa* ***lethu*** *lolwazi liyasikhhaba* . . .

In Example 3, the possessive *lethu* qualifies class 5 nouns, *igqiza* ('team') and *igosa* ('officer') and the possessive concord is *le*-. It can also be mentioned that the qualifier in Xhosa comes after the noun.

(4) -*thu* with class 1a and class 3 nouns = *wethu* (SAMSA Manual 2003):
ST: **Our** head office is in Pretoria . . .
TT: *Undlunkulu* ***wethu*** *usePitoli* . . .
ST: Forms can be obtained from: ***Our** information officers . . .
TT: *Iifomu zingafunyanwa*: *Kumagosa* ***wethu*** *olwazi* . . .

In Example 4, the possessive *wethu* qualifies the class 1(a) noun, *undlunkulu* ('head office') and class 6 noun, *amagosa olwazi* ('information officers').

(5) -*thu* with class 10 nouns = *zethu* (SAMSA Manual 2003):
ST: **Our** objectives are to ensure the safety of life . . .
TT*: Iinjongo* ***zethu*** *kukuqinisekisa ukhuselo lobomi* . . .
ST: **Our** contact details
TT: *Iinkcukacha* ***zethu*** *zoqhakamshelwano*

In Example 5, the possessive *zethu* qualifies the class 10 nouns, *iinjongo* ('objectives') and *iinkcukacha* ('details').

It should be noted that, to get an idea of the overall frequency of a word in Xhosa, for example, the possessive 'our', the researcher will have to add all the different instances of *lethu, wethu, zethu* and so forth.

11.3 Using ParaConc as a Corpus Tool

As mentioned above, ParaConc was used as the tool for the analyses of the data in the corpus. A search method that seems to be particularly productive when one searches for possible translations and other associated words is the hot word tool in ParaConc.

11.3.1 Hot Words in ParaConc

Since Xhosa is written conjunctively, the hot word tool is very useful because it enables one to identify other possible translations of a word in a sentence. For example, in Figure 11.2, we find a dialogue box of possible translations of 'local government' as found in the *Short Guide* (1998). In the list of hot words *ulawulozidolophu* is the 'hottest' word. In other words, it ranks highest in the list of words associated with 'local government'. Barlow (2003: 34, see also Barlow 2009) points out that the figure can be taken as

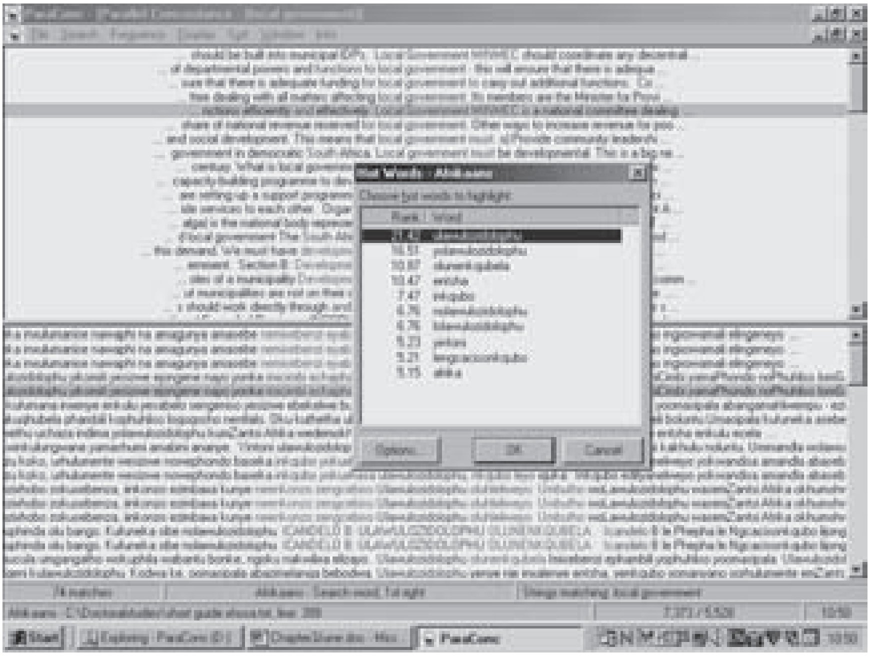


Figure 11.2 Hot word list

an 'approximate guide to the relative strength (hotness) of the different words'.

11.3.2 Paradigm Option in Hot Words

The paradigm option in hot words is used to boost the ranking of words whose forms are similar. To achieve this, we select 'OPTIONS' in hot words and tick the paradigm option, as shown in Figures 11.2 and 11.3.

Figure 11.4 provides an illustration of hot words associated with *-lawulozidolophu* using the paradigm option.

The information on hot words associated with *-lawulozidolophu* ('local government') in Figures 11.2 and 11.4 is summarized in Table 11.2 below. In Figure 11.2, where the paradigm option is not used, we get a list of four hot words that have the same stem *-lawulozidolophu* and in Figure 11.4, where the paradigm option is used, we get a list of 12 hot words associated with *-lawulozidolophu*.

In each example cited here, the stem *-lawulozidolophu* takes a different prefix or concord, and that leads to the formation of another part of speech.

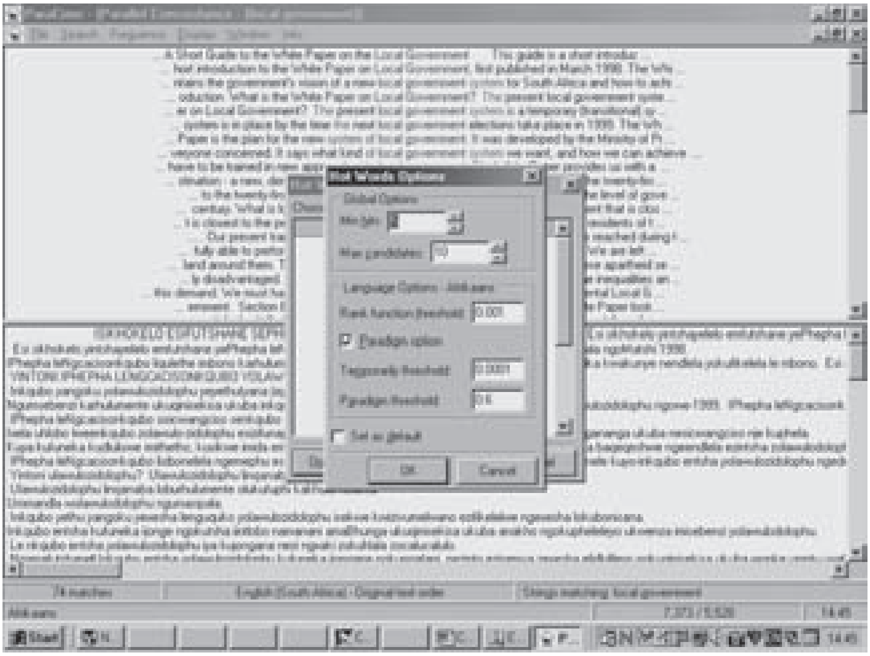

ParaConc - [Parallel Concordance - local government]
File Search Frequency Display Window Info
Hot Words Options
Global Options
Min hits
Max candidates
Language Options - Afrikaans
Rank function threshold 0.001
☑ Paradigm option
Tagging threshold 0.0001
Paradigm threshold 0.6
☐ Set as default
OK
Cancel
74 matches
English (South Africa) - Original text order
Strings matching: local government
Afrikaans
Start

FIGURE 11.3 Selecting the paradigm option

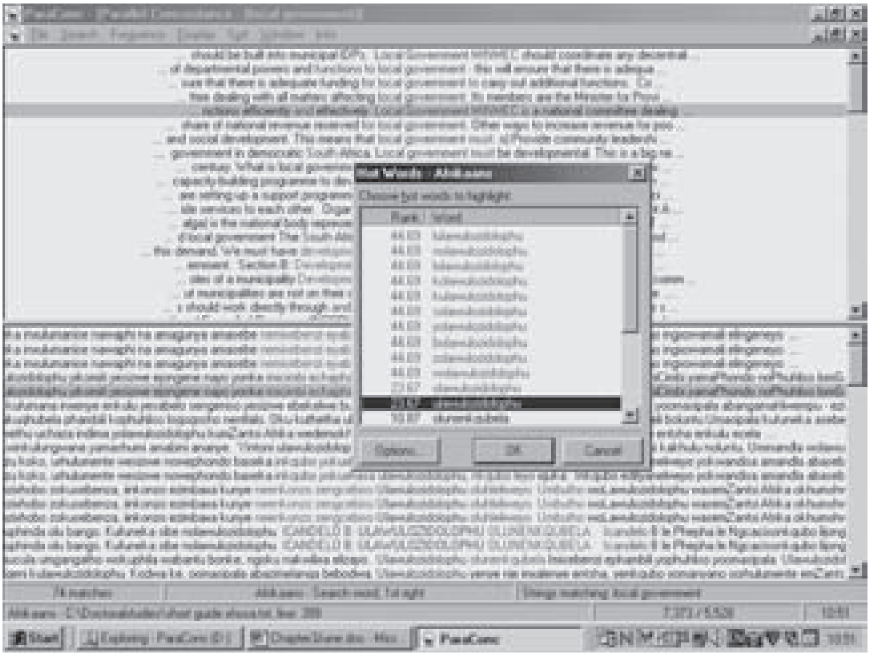

ParaConc - [Parallel Concordance - local government]
File Search Frequency Display Sort Window Info
Hot Words - Afrikaans
Choose hot words to highlight
Rank Word
Options...
OK
Cancel
74 matches
Afrikaans - Search word, 1st right
Strings matching: local government
Start
ParaConc

FIGURE 11.4 Hot word list of *-lawulozidolophu* using the paradigm option

Table 11.2 Frequency of -*lawulozidolophu* in *Short Guide* (1998)

Count	Percentage (%)	Word
19	0.3437	ulawulozidolophu
16	0.2894	yolawulozidolophu
6	0.1085	nolawulozidolophu
6	0.1085	lolawulozidolophu
4	0.0724	kulawulozidolophu
2	0.0362	bolawulozidolophu
2	0.0362	olawulozidolophu
2	0.0362	wolawulozidolophu
2	0.0362	zolawulozidolophu

For example, *yo-* is a possessive concord; therefore, *yolawulozidolophu* is a possessive formed from the noun *ulawulozidolophu,* as seen in Example 6:

(6) ST: Our present transitional system **of local government** . . .
TT: *Inkqubo yethu yangoku yexesha lenguquko* ***yolawulozidolophu*** . . .

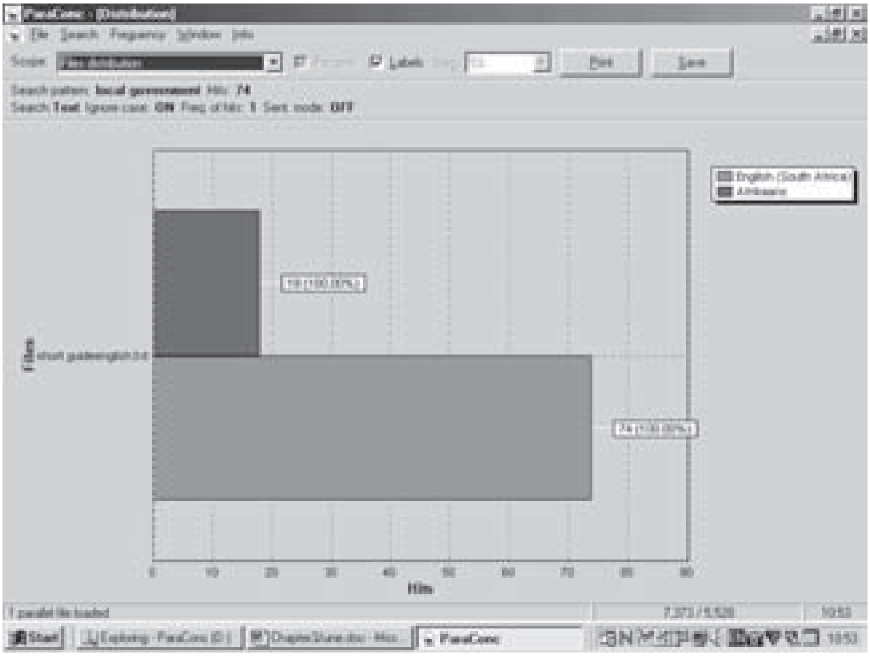
11.3.3 File Distribution

File distribution in ParaConc is another method that confirms the conjunctive nature of Xhosa. Here, a graph is used to illustrate the distribution of a word in parallel texts. For example, the graph below is an illustration of how the terms 'local government' and *ulawulozidolophu* are distributed in both the source and target texts. The noun *ulawulozidolophu* has 18 hits and *local government* has 74 hits. Other instances of *lawulozidolophu* do not appear in the graph because they are written with concords and therefore form other parts of speech like possessives, adverbs, locatives and so forth.

Now that the unique concordial system of Xhosa and how this feature causes this language to differ significantly from English has been explained, the next section can focus on investigating the close link observed between simplification and explicitation as universal features in translated texts. But first, different views on the concept of 'translation universals' are considered.

11.4 Universal Features of Translation

Searching for regularities in translation is not new, but the subject of translation universals remains controversial. While some scholars such as Laviosa-Braithwaite (1996) have found definite evidence of simplification as a universal feature of translation, other scholars such as Tymoczko (1998) maintain that



Search pattern: **local government** Hits: **74**
Search: **Text** Ignore case: **ON** Freq of hits: **1** Sort mode: **OFF**

Figure 11.5 File distribution of local government/*ulawulozidolophu*

the idea of making claims about universals is unimaginable, since it is impossible to capture translations from all times and all languages (Mauranen and Kujamäki 2004). Baker (1998) addresses the concern which may arise because of the focus on regularities in corpus-based translation studies. She states that a balance must be maintained between the general and the specific, and between the norm and the exception in corpus-based analysis of translation. Researchers should guard against studying the patterns of translated texts and then treating such patterns as rules.

According to Chesterman (2004: 3) a universal can be defined as 'a feature that is found (or at least claimed) to characterize all translations: i.e. a feature that distinguishes them from texts that are not translations. . . . a universal feature is one that is found in translations regardless of language pairs, different text-types, different kinds of translators, different historical periods, and so on'.

Chesterman (2004: 9–10) further points out that one of the problems regarding the so-called universals is representativeness. Since it is impossible to study all translations universally, we rely on samples, and when the sample is more representative, we become more confident about the validity of our results and claims, while our data may be unrepresentative in one way or another.

From the above discussion, it is clear that there is some debate regarding the use of the term 'universal' when referring to patterns that characterize translated texts, irrespective of language combinations. But these viewpoints do not dispute the general fact that, while original texts are individual in form, translated texts seem to have many common features.

11.4.1 Simplification Strategies

Simplification is an attempt by the translator to make the language or the message or both simpler. Baker (1996: 182) defines simplification as 'a translator's attempt to make things easier for the reader (but not necessarily more explicit)'. Toury (1995: 270) states that if the target text has a lower information load than the source text, it is because ambiguous information in the source text has been disambiguated, that is spelt out or made simpler in the translation process.

According to Laviosa-Braithwaite (1996), the first empirical studies to review and provide evidence for simplification in translation were the analyses carried out by Blum-Kulka and Levenston in 1983 of Hebrew–English translations. Laviosa-Braithwaite (1996) investigated simplification hypotheses in two subcorpora of the *English Comparable Corpus* (ECC), namely the Newspaper subcorpus of *The Guardian* and *The European* collections on the one hand and the Narrative sub-corpus on the other. Laviosa-Braithwaite (1997: 533) reports that evidence of at least three types of simplification have been found in translated texts, namely syntactic, stylistic and lexical. According to her, stylistic simplification involves breaking up long sequences and sentences in the source text and replacing elaborate phraseology with shorter collocations, reducing or omitting repetitions and redundant information, shortening overlong circumlocutions and leaving out modifying phrases and words. The types of simplification strategies which seem to be noticeable in the *English-Xhosa Parallel Corpus* are lexical and stylistic simplification, but only stylistic simplification is examined here, and particular attention is paid to the breaking up of sentences.

11.4.1.1 Breaking Up Sentences

One of the ways of simplifying the language used in translation is by breaking up sentences (Baker 1996: 181). In her investigation of simplification in the Newspaper sub-corpus of the *English Comparable Corpus* (ECC), Laviosa-Braithwaite (1996: 535) observed that translated articles have a lower mean sentence length than articles originally written in English. In other words, the overall length of sentences in the translations is shorter than the sentences in the source texts. Fabricius-Hansen (1999: 179) states that 'the information that needs to be encoded in the target text clause because of **grammatical reasons** [own emphasis] is often supplied unambiguously by the source text'. In the *English-Xhosa Parallel Corpus*, the strategy of breaking up sentences was used mainly by translators who translated the Annual Reports for DACST (1998) and PanSALB (2001/2002). These two annual reports are considerably longer than other texts in the corpus. In Table 11.3, examples from each source are provided showing clearly that some of the English sentences which were broken up by the translators are actually not long at all (e.g. Example 2).

A general observation is that the concordial system which makes Xhosa to differ in grammatical structure from English seems to influence the structure of Xhosa translated sentences. Since concords link up words in a sentence, it appears as if sometimes the translator, in her attempts to produce a meaningful sentence, finds it difficult to maintain the grammatical concord without splitting the sentence.

It was noted in the *English-Xhosa Parallel Corpus* that the breaking up of sentences (a simplification strategy), actually sometimes leads to explicitation. In another study, Pápai (2004: 145) identifies explicitation strategies in a parallel corpus of English–Hungarian texts. Pápai declares that if one considers the structural differences of English and Hungarian, Hungarian being an agglutinative language (just like Xhosa), one would expect that translations from English into Hungarian would result in implicitation rather than explicitation. Therefore, she sees the notion of explicitation as being closely linked to the notion of simplification in translation.

11.4.2 Explicitation Strategies

According to Shuttleworth and Cowie (1997: 55), explicitation can be defined as 'the phenomenon which frequently leads to TT stating ST information in more explicit form than the original'. The translator may add explanatory phrases, spell out implicit information or add connectives so that the text can flow logically and read easily. In other words, explicitation can be defined as the 'widening' of meaning. Pápai (2004: 145) defines explicitation in terms of process and of product. He writes that in terms of process, 'explicitation is a translation technique involving a shift from the source text concerning structure or content. It is a technique of resolving ambiguity, improving and increasing cohesiveness of the source text and also adding linguistic and extra-linguistic information.' In terms of product, 'explicitation is a text feature contributing to a higher level of explicitness in comparison with non-translated texts. It can be manifested in linguistic features used at higher frequency than

Table 11.3 Breaking up of sentences/sentence splitting

ST: English	**TT: Xhosa**	**Back translation**
DACST (1998):		
1. The Language Plan Task **Group** (known as LANGTAG) established by the **Minister** in November 1995 to advise him on the development of a National Language Plan for South Africa, submitted its final report to him on 8 August 1996.	1.1 ***Iqela*** *elisebenza nokuyila ulwimi (elaziwa ngokuba yiLangtag) lamiswa ngu**Mphathiswa** ngenyanga kaNovemba ka-1995, ukwenzela ukuba limcebise ngenkqubela-phambili engokuyila ulwimi lwesizwe loMzantsi Afika.*	1.1 The Language Plan Task **Group** (known as LANGTAG) was established in November 1995 by the **Minister** to advise him on the development of a National Language Plan for South Africa.
	1.2 ***Eli qela*** *lanika* ***uMphathiswa*** *ingxelo yalo yokugqibela ngomhla wesi-8 kaAgasti ngo-1996.*	1.2 **This group** submitted its final report to the **Minister** on 8 August 1996.
2. The stabilization of young people and children is a **priority** for the DACST in line with priorities of national government.	2.1 *Isebe lobuGcisa, iNkcubeko, iNzululwazi nobuChwepheshe (iDACST) line**njongo ephambili** yokudala uzinzo phakathi kolutsha.*	2.1 The **priority** of the Department of Arts, Culture, Science and Technology (DACST) is to stabilize the youth.
	2.2 ***Lo*** *ngumnqweno ophambili kaRhulumente.*	2.2 **This** is government's primary goal.
PanSALB (2001/2002):		
3. A National Lexicography **Unit** (NLU) for each of the official languages in South Africa had been established and registered as a section 21 company by March 2001.	3.1 *Kusekwe ama**Qumrhu** eSizwe okuBhalwa kwesiChazi-magama, iiNational Lexicography Units (iNLU) olwimi ngalunye olusemthethweni eMzantsi Afrika.*	3.1 National Lexicography **Units** (NLUs) have been established for each language in South Africa.
	3.2 ***La maqumrhu*** *abhaliswa phantsi kwecandelo lama-21 loMthetho weeNkampani ngeyoKwindla 2001.*	3.2 **These units** had been registered as a section 21 company by March 2001.

in non-translated texts or in added linguistic and extralinguistic information.' This means that the translator gives longer explanations or adds more information which does not occur in the source text.

The following section provides an analysis of explicitation strategies such as the insertion of an explicit demonstrative, the use of lexical repetition and the addition of explanatory information observed in Xhosa split sentences.

11.4.2.1 Insertion of an Explicit Demonstrative

The demonstrative is used to refer to a noun or pronoun that was previously mentioned. It is, therefore, important to note that the demonstrative is also governed by noun classes. This means that, because of the noun class prefixes, different demonstratives are used to denote the same position. The three positional types of demonstrative in Xhosa basically correspond to the English 'this', 'that' and 'that yonder'. The demonstrative is primarily used before a noun or pronoun to indicate position or time but it also has some secondary functions. It can be used before or after a noun to lay special emphasis on a person or thing, or to depict a certain meaning. Some of the secondary functions of a demonstrative in Xhosa, cited by Pahl et al. (1989: 695, quoted in Moropa 2005: 103) are:

- to refer to a person or thing just mentioned by the speaker/writer;
- to refer to a person or thing not mentioned prior to the reference, but usually clearly understood, and, if not qualified thereafter, used before the noun or pronoun or alone;
- to specify, single out, refer specifically to a person or thing, often in contrast to others used before or after the noun or pronoun, or alone;
- to emphasize an action pertaining to a person or thing; used after the noun or pronoun;
- to lay special emphasis on a person or thing; used both before and after the noun or pronoun thus emphasized.

When Xhosa translators break up sentences they always start the second sentence with a demonstrative as the noun referred to in sentence 2 is already mentioned in sentence 1. For example, in Table 11.4, all the sentences begin with the first position demonstrative *eli, lo, lo, olu, eli* and *la* respectively (the first sentence of each example has not been included in the table). Demonstratives *eli, lo, olu* and *la* denote the first position, but they differ because they refer to nouns which belong to different classes. *Eli* qualifies a class 5 noun *iqela* 'group', *olu* refers to class 11 nouns *uququzelelo* ('hosting') and *ukhuphiswano* ('competition'), *lo* refers to a class 3 noun *umdlalo* ('play') and *la* to a class 6 noun *amaqumrhu* '('units').

Some English–Zulu parallel texts were also examined because Xhosa and Zulu are both agglutinating languages, belonging to the Nguni language group. The aim was to establish if there were any similar patterns regarding the insertion of a demonstrative, repetition or addition in Zulu translated sentences as in Xhosa. The English–Zulu parallel texts which were manually searched are:

- *Giya!*, a newsletter of the KwaZulu-Natal Department of Arts, Culture and Tourism (Issue No. 3 March 2006a, and its Zulu version 2006b) downloaded from the Internet;
- The PanSALB *Annual Report* (2001/2002).

Table 11.4 Insertion of a demonstrative in Xhosa

TT: Xhosa	Back translation
1 ***Eli qela*** *lanika uMphathiswa ingxelo yalo yokugqibela ngomhla wesi-8 kaAgasti ngo-1996.*	1 **This group** submitted its final report to the Minister on 8 August 1996.
2 *Inxalenye* ***yolu ququzelelo*** *yayivela kwiDACST.*	2 Part of the funding for **this hosting** came from DACST.
3 ***Lo mdlalo*** *wawuyinxalenye yomsebenzi wokuphucula nosekelwe kwiGunya leNyaniso noXolelwaniso, kwaye wawuxhaswa ligumbi lezoPhando laseMarket Theatre kwisigaba salo sophando nophuhliso.*	3 **This play** was part of a project for development which was based on the Truth and Reconciliation Commission and supported by Market Theatre Laboratory in the research and development phase.
4 ***Olu khuphiswano*** *luphuhlisa lukhuthaze, lomeleze ulutsha oluneziphiwo kwicala lomdaniso.*	4 **This competition** develops and empowers the youth who are talented in dancing.
5 ***Eli qela*** *leqonga nelivela KwaZulu-Natala lazibalula ngezenzo zalo, lada lamenywa ukuba liye konwabisa abantu kumgcobo wokugqibela nakumzi wozakuzelwano loMzantsi Afrika eMaputo.*	5 **This** theatre **group** from KwaZulu-Natal as a result of its performance was invited to perform and entertain at the final gala event at the South African embassy in Maputo.
6 ***La maqumrhu*** *abhaliswa phantsi kwecandelo lama-21 loMthetho weeNkampani ngeyoKwindla 2001.*	6 **These units** had been registered as a section 21 company by March 2001.

Although the translated Zulu sentences in Table 11.5 are not examples of split sentences, they follow the same pattern found in Xhosa split sentences. For instance, in Example 1 (*kwa*)*lesi* ('of this') is added in sentence 1.1 and the noun *sakhiwo* ('complex') which was mentioned in sentence 1 is repeated in 1.1. In Example 2 *leli* ('this') is inserted in sentence 2.1, and *iBhodi* ('Board' which was mentioned in sentence 2 is stated again in sentence 2.1). In the translation of sentence 2 some information is added: 'its research' is translated as 'this board works hard conducting research'.

It is also worth mentioning that the use of the demonstrative to translate the English article 'the' is a common pattern in both Xhosa and Zulu. It is also sometimes used to translate the personal pronoun 'it'. When the demonstrative is used, the sentence becomes more explicit as the demonstrative qualifies that particular noun. It is assumed that the translators opt for this strategy because these languages do not have definite or indefinite articles. The examples cited in Tables 11.6 and 11.7 from the PanSALB (2001/2002) *Annual Report* show some similarities in the use of the demonstrative in Xhosa and Zulu.

Table 11.5 Insertion of a demonstrative + additional information in Zulu (*Giya*! March 2006)

ST: English	TT: Zulu	Back translation of second sentence only
1. The Durban Playhouse **complex** is having a R4.2 million refurbishment, designed to establish it as the Broadway of Africa.	1. ***Isakhiwo*** *se-Durban Playhouse sigixatshezwe ngezigidi zamarandi angu-4.2 sokushintsha indlela esibukeka ngayo sibe sezingeni elicokeme e-Afrika.*	
1.1 The make-over reflects the cultural diversity of KwaZulu-Natal, within an overall Zulu theme.	1.1 *Ukwakhiwa kabusha* ***kwalesi sakhiwo*** *kukhombisa izinhlobonhlobo zamasiko a-KwaZulu-Natali kanti sakhelwe phezu kwesiseko sobuZulu.*	1.1 The make-over **of this complex** shows the cultural diversity of KwaZulu-Natal, within an overall Zulu theme.
2. The Natal Sharks **Board** has been successfully protecting the province's bathing beaches from attack for more than 40 years.	2. *I-Natal Sharks* ***Board*** *isebenze kanzima kule minyaka engama-40 ivikela umphakathi ekungangenini engozini yokudliwa ngoshaka elwandle.*	
2.1 Its research into bather protection, with minimum impact on the environment is respected world-wide . . .	2.1 ***Leli Bhodi*** *lisebenza kanzima licwaninga ngempilo yasolwandle ukuze likwazi Ukuvikela abantu koshaka . . .*	2.1 **This board** works hard conducting research on maritime life so that it can protect people from sharks . . .

Table 11.6 Article 'the' = demonstrative in Xhosa and Zulu translated texts (PanSALB (2001/2002)

ST: English	TT: Xhosa	TT: Zulu
The following were among the issues dealt with during **the** conference.	***Le*** *ilandelayo yeminye yemicimbi eyayixoxwa (**ku**)**le** nkomfa.*	*Okulandelayo kungamanye amaphuzu adingidwa (**ku**)**le** ngqungquthela.*
The process of disseminating **the** document to the national language bodies . . .	*Indlela yokusabalalisa* ***olu*** *xwebhu kwimibutho yesizwe . . .*	*Kwase ke kuqalwa uhlelo lokusabalalisa* ***lo*** *mbhalo kuzindlafa ezihluka-hlukene*
The unit now falls under the auspices of the Board.	***Eli*** *qumrhu nalo sele lifakwe phantsi kwephiko leBhodi.*	***Leli*** *ziko manje selingaphansi kolawulo lweBhodi.*

In Table 11.7, the examples show that the pronoun 'it' is translated by means of a noun and a demonstrative in both languages.

In all these instances where the demonstrative has been inserted or added before a noun or used to translate the article 'the', it functions as a

Table 11.7 Pronoun = demonstrative + noun in Xhosa and Zulu (PanSALB 2001/2002)

ST: English	TT: Xhosa	TT: Zulu
It recently started an electronic database with information on the status quo regarding language policies of various organizations	*Kutsha nje* ***elo candelo*** *liqalise ukwakha uvimba we-elektronika nophethe ulwazi ngobume bemigaqo-nkqubo yeelwimi kwiinkampani ezahlukeneyo.*	***Lo mkhakha*** *usanda kuqala inqolobane esekhompyutheni, equkethe imininingwane ephathelene negxathu eselithathwe izihlangano ezehluka- hlukene malungana nomgomo wezilimi.*
Note: 'it' = pronoun	*elo candelo* = demonstrative + noun	*lo mkhakha* = demonstrative + noun

determiner or nominal qualifier. The grammatical function of a determiner is to determine or fix the limits of a noun or a noun phrase and, because it describes the noun in this manner, it has an adjectival function. According to Aitchison (1994: 37), determiners mark a simplification in grammar. Generally, determiners offer a range of options that can bring greater clarity and precision to a text.

11.4.2.2 Use of Lexical Repetition

In this case, lexical repetition refers to the repetitive use of a word in sentences where such sentences have been broken up in the translation. For example, in Table 11.8, nouns *iqela* ('group') and *uMphathiswa* ('minister') are repeated in sentences 1.1 and 1.2, whereas in the source text the words 'group' and 'Minister' are only used once. The same applies to sentences 3.1 and 3.2, 4.1 and 4.2, 5.1 and 5.2 where nouns *umdlalo* ('play'), *iqela* ('group') and *amaqumrhu* ('units') respectively are repeated in the translations. Sentences 2.1 and 2.2 differ slightly from the other examples as the repeated words are a verb and a noun: the verb -*ququzelela* ('hosted') in sentence 2.1 is repeated in the form of a noun -*(u)ququzelelo* in sentence 2.2. The translator might have repeated these words for clarity.

From Table 11.8, it can be seen that the English equivalents of these words appear once only in the source text. The translator's desire to explain meaning to the target reader, made her resort to lexical repetition each time in the second sentence instead of using subject/object concords or personal pronouns which would have fitted very well in the context – see Table 11.9.

Pápai (2004: 153) states that while 'translators want to create a clear and transparent target sentence, their aim can override the otherwise respected norm of translation, i.e. avoidance of repetition'. This seems to have happened here.

Table 11.8 Lexical repetition

Source text	Target text	Back translation
DACST (1998)		
1. The Language Plan Task **Group** (known as LANGTAG) established by the **Minister** in November 1995 to advise him on the development of a National Language Plan for South Africa, submitted its final report to him on 8 August 1996.	1.1 _**Iqela** elisebenza ngokuyila ulwimi (elaziwa ngokuba yiLangtag) lamiswa ngu**Mphathiswa** ngenyanga kaNovemba ka-1995, ukwenzela ukuba limcebise ngenkqubela-phambili engokuyila ulwimi lwesizwe loMzantsi Afika._	1.1 The Language Plan Task **Group** (known as LANGTAG) was established in November 1995 by the **Minister** to advise him on the development of a National Language Plan for South Africa.
group x 1 **Minister x 1**	1.2 _**Eli qela** lanika **uMphathiswa** ingxelo yalo yokugqibela ngomhla wesi-8 kaAgasti ngo-1996._ **iqela x 2** **uMphathiswa x 2**	1.2 **This group** submitted its final report to the **Minister** on 8 August 1996. **group x 2** **Minister x 2**
2. The ethnomusicology Department and UCT Ballet school **hosted** a Confluence conference in July 1997 that was partly funded by the DACST.	2.1 _Isebe lomculo wesintu neSikolo somdaniso webhaleyi seYunivesithi yaseKapa la**ququzelela** iNgungqthela ngeyeKhala kowe-1997._	2.1 The Department of Ethnomusicology and the School of Ballet of the University of Cape Town **hosted** a conference in 1997.
	2.2 _Inxalenye **yolu ququzelelo** yayivela kwiDACST._	2.2 Part of the funding for **this hosting** came from DACST.
3. *The Story Which I am About to Tell* by Mehlo Communications was a development **theatre** project on the Truth and Reconciliation Commission supported by the Market Theatre Laboratory in this research and development phase.	3.1 _Ibali Endiza Kulibalisa yayingu**mdlalo** weqonga wequmrhu iMehlo Communications._	3.1 *The story Which I am About to Tell* was a stage **play** by Mehlo communications.
theatre x 1	3.2 _**Lo mdlalo** wawuyinxalenye yomsebenzi wokuphucula nosekelwe kwiGunya leNyaniso noXolelwaniso, kwaye wawuxhaswa ligumbi lezoPhando laseMarket Theatre kwisigaba salo sophando nophuhliso._ **mdlalo x 2**	3.2 **This play** was a project for development which was based on the Truth and Reconciliation Commission and supported by Market Theatre Laboratory in the research and development phase. **play x 2**

4. South Africa was represented by Isithombe sikaShaka, a theatre **group** from KwaZulu-Natal which as a result of its performance, was invited to perform at the final gala event and at the South African embassy in Maputo.	4.1 *Ummeli woMzantsi Afrika yayili**qela** elibizwa ngokuba yi-Isithombe sikaShaka.*	4.1 The South African representative was a **group** known as Isithombe sikaShaka.
group x 1	4.2 ***Eli qela*** *leqonga nelivela KwaZulu-Natala lazibalula ngezenzo zalo, lada lamenywa ukuba liye konwabisa abantu kumgcobo wokugqibela nakumzi wozakuzelwano loMzantsi Afrika eMaputo* **iqela x 2**	4.2 **This** theatre **group** from KwaZulu-Natal as a result of its performance was invited to perform at the South African embassy in Maputo **group x 2**
PanSALB (2001/2002)		
5. A National Lexicography **Unit** (NLU) for each of the official languages in South Africa had been established and registered as a section 21 company by March 2001.	5.1 *Kusekwe **amaQumrhu** eSizwe okuBhalwa kwesiChazi-magama, iiNational Lexicography Units (iNLU) olwimi ngalunye olusemthethweni eMzantsi Afrika.*	5.1 National Lexicography **Units** (NLUs) have been established for each language in South Africa.
unit x 1	5.2 ***La maqumrhu*** *abhaliswa phantsi kwecandelo. lama-21 loMthetho weeNkampani ngeyoKwindla 2001.* **amaqumrhu x 2**	5.2 **These units** had been registered as a section 21 company by March 2001. **units x 2**

Table 11.9 Demonstrative + noun < Subject/Object Concord/Personal pronoun

TT: Demonstrative + noun	**TT: SC/ OC / Personal pronoun**
1.2 ***Eli qela* 'this group'** *lanika uMphathiswa ingxelo yalo yokugqibela ngomhla wesi-8 kaAgasti ngo-1996.*	1.2 ***Lam****nika ingxelo yalo yokugqibela ngomhla wesi-8 kaAgasti ngo-1996* (**It** gave **him** its final report on 8 August 1996.)
2.2 *Inxalenye* ***yolu ququzelelo*** **'this hosting'** *yayivela kwiDACST.*	2.2 **La**fumana inxalenye yenkxaso kwiDACST. (**It** received part of the funding from DACST.)
3.2 ***Lo mdlalo* 'this play'** *wawuyinxalenye yomsebenzi wokuphucula nosekelwekwiGunya le Nyaniso noXolelwaniso kwaye wawuxhaswa ligumbi laseMarket Theatre kwisigaba salo sophando nophuhliso*	3.2 ***Wawuyi****nxalenye yomsebenzi wokuphucula nosekelwe kwiGunya leNyaniso noXolelwaniso kwaye wawuxhaswa ligumbi laseMarket Theatre kwisigaba salo sophando nophuhliso* (**It was** part of a project for development which was based on the Truth and Reconciliation Commission and supported by Market Theatre Laboratory in the research and development phase.)

11.4.2.3 Adding Explanatory Information

The translator may sometimes add some information which is not found in the source text so that the message can be more easily understood by the target reader. According to Delabastita (1993: 36), addition as a translation strategy (i.e. the insertion of information in the TT that is absent in the ST) can partly be ascribed to translators' understandable 'concern for clarity and coherence, which prompts them to disentangle complicated passages, provide missing links, lay bare unspoken assumptions and generally give the text a fuller wording'. Sometimes, 'additions are due to conscious interventions of the translator who may believe, for example, that she can enhance the aesthetic qualities of her text by adding rhyme to an unrhymed ST, by using a more strongly metaphorical language, adding to the exotic flavour of the text and so forth'. Kruger (2000: 142) states that the complex structure of a ST may be an important constraint which causes the translator to add some information to the TT.

Fabricius-Hansen (1999: 176) is of the opinion that 'lexical and/or structural differences between the source and target languages compel the translator to consult not only the syntactic structure but also the semantic representation of the source string as well as the syntactic rules of the target language in order to create a semantically equivalent string which may even *need more information* [own emphasis] that is not contained in the source string itself'.

As illustrated in Table 11.10, additions such as 'not only water came out' and 'this competition' illustrate that the translator wanted the text to be explicit so as to get the message across. The translator made use of both addition

Table 11.10 Addition of explanatory phrases as well as lexical repetition

Source text	Target text	Addition	Repetition
1. In September 1912, an earthquake opened up a new spring, and fossil bones and stone artefacts were brought to the surface with the water.	1.1 *NgeyoMsintsi kowe-1912 kwabakho inyikima, kwaze kwavuleka umthombo.* 1.2 *Kulo mthombo* ***akuzange kuvele amanzi kuphela****, kwathi gqi nenqwaba yamathambo nezixhobo ezenziwe ngamatye*	***akuzange kuvele amanzi kuphela*** (‘not only water that came out’)	*umthombo* (‘spring’ x 2)
2. The DACST awarded a grant to FNB Vita Awards to continue with their role in encouraging, stimulating and developing the talent of young people in the dance arena.	2.1 *IDACST yafaka inkxaso-mali* ***ukhuphiswano*** *olukhuthazayo nolwaziwa ngokuba yiFNB Vita Awards.* 2.2 ***Olu khuphiswano*** *luphuhlisa lukhuthaze, lomeleze ulutsha oluneziphiwo kwicala lomdaniso.*	***Olu khuphiswano*** (‘This competition’)	*ukhuphiswano* (‘competition’ x 2)

and repetition: *umthombo* (‘spring’) in sentences 1.1 and 1.2, and *ukhuphiswano* (‘competition’) in sentences 2.1 and 2.2.

11.5 Conclusion

It can be concluded that the concordial system of Xhosa, a feature which distinguishes the structure of Xhosa from that of English, has a definite effect on Xhosa translations, since in order to maintain the necessary Xhosa grammatical concord, it is sometimes necessary to split sentences. The frequent use of the demonstrative at the beginning of the second sentence where sentences have been broken up, can be regarded as a common feature in translated texts. The use of the demonstrative tends to result in a more explicit sentence because of the qualifying function of the demonstrative. It has also been observed that the translator may repeat words or add some explanatory information in her attempts to expand the meaning of the original message to the target reader. The analyses demonstrate clearly that simplification and explicitation sometimes are very closely linked in translated texts in Xhosa. In an attempt to bring clarity and precision to a translation, translators sometimes make use of both strategies when translating a sentence.

Note

[1] In all examples, ST means source text and TT means target text.

References

Aitchison, James (1994) *Cassell Guide to Written English,* London: BCA.

Baker, Mona (1996) 'Corpus-Based Translation Studies: The Challenges That Lie Ahead', in Harold Somers (ed.) *Terminology, LSP and Translation Studies in Language Engineering: In Honour of Juan C. Sager,* Amsterdam: John Benjamins, 175–86.

— (1998) 'Investigating the Language of Translation: A Corpus-Based Approach', *Meta* 43(4): 480–5.

Barlow, Michael (2003) *Paraconc: A Concordancer for Parallel Texts* (Draft 3/03), Rice University, USA.

— (2009) *ParaConc and Parallel Corpora in Contrastive and Translation Studies,* Houston: Athelstan.

Blum-Kulka, Shoshana (1986) 'Shifts of Cohesion and Coherence in Translation', in Juliane House and Shoshana Blum-Kulka (eds) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies,* Tübingen: Gunter Narr.

Chesterman, Andrew (2004) 'Hypotheses about Translation Universals', in Gyde Hansen, Kirsten Malmkjær and Daniel Gile (eds) *Claims, Changes and Challenges in Translation Studies,* Amsterdam: John Benjamins, 1–13.

Delabastita, Dirk (1993) *There's a Double Tongue: An Investigation into the Translation of Shakespeare's Wordplay with Special Reference to 'Hamlet',* Amsterdam: Rodopi.

Department of Arts, Culture and Tourism (DACST) (2006a) *Giya*! Newsletter of KwaZulu-Natal Department of Arts, Culture and Tourism, March, Issue No. 3.

— (2006b) *Giya*! Iphephabhuku loMnyango Wezobuciko, Amasiko Nezokuvakasha Kwazulu-Natali, Ndasa, Ushicilelo Lwesithathu.

Fabricius-Hansen, Catherine (1999) 'Information Packaging and Translation: Aspects of Translational Sentence Splitting (German–English/Norwegian)', in Monika Doherty, *Studia Grammatica,* Berlin: Akademie-Verlag, 175–214.

Gauton, Rachelle and Gilles-Maurice de Schryver (2004) 'Translating Technical Texts into Zulu with the Aid of Multilingual and/or Parallel Corpora', *Language Matters, Studies in the Languages of Africa* 35(1): 148–61.

Kruger, Alet (2000) 'Lexical Cohesion and Register Variation: *The Merchant of Venice* in Afrikaans'. Unpublished D.Litt et Phil. Thesis, Pretoria, University of South Africa.

Laviosa-Braithwaite, Sara (1996) 'The English Comparable Corpus (ECC): A Resource and Methodology for the Empirical Study of Translation'. Unpublished PhD Thesis, Dept of Language Engineering, UMIST, Manchester.

— (1997) 'Investigating Simplification in an English Comparable Corpus of Newspaper articles', in Kinga Klaudy and Janos Kohn (eds) *Transferre Necesse*

Est. Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting, Budapest: Scholastica, 531–40.

Madiba, Ronald Mbulungeni (2004) 'Parallel Corpora as Tools for Developing the Indigenous Languages of South Africa, with Special Reference to Venda', *Language Matters, Studies in the Languages of Africa* 35(1): 133–47.

Mauranen, Anna and Pekka Kujamäki (2004) *Translation Universals: Do They Exist*? Amsterdam and Philadelphia: John Benjamins, 3–9.

Moropa, Koliswa(2004) 'A Parallel Corpus as a Terminology Resource for Xhosa: A Study of Strategies Used to Translate Financial Statements', in *Language Matters, Studies in the Languages of Africa* 35(1): 162–78.

— (2005) 'An Investigation of Translation Universals in a Parallel Corpus of English-Xhosa Texts'. Unpublished D.Litt et Phil. Thesis, Pretoria, University of South Africa.

— (2007) 'Analysing the English-Xhosa Parallel Corpus of Technical Texts with ParaConc: A Case Study of Term Formation Processes', *South African Linguistics and Applied language Studies* 25(2): 183–205.

Pahl, H., A. M. Pienaar and T. A. Ndungane (1989) *The Greater Dictionary of Xhosa, Volume 3*, Alice: University of Fort Hare.

Pan South African Language Board (PanSALB) (2002) *Annual Report/Umbiko wonyaka 2001/2002,* Pretoria: Afriscot Printers.

Pápai, Vilma (2004) 'Explicitation: A Universal of Translated Text?' in Anna Mauranen and Pekka Kujamäki (eds) *Translation Universals: Do they exist?* Amsterdam: John Benjamins, 143–64.

Republic of South Africa (1996) *The Constitution of the Republic of South Africa*, Pretoria: Government Printers.

Shuttleworth, Mark and Moira Cowie (1997) *Dictionary of Translation Studies*, Manchester: St. Jerome.

Toury, Gideon (1995) *Descriptive Translation Studies and Beyond*, Amsterdam: John Benjamins.

Tymoczko, Maria (1998) 'Computerized Corpora and Translation Studies', *Meta* 43(4): 652–9.

Chapter 12

Disfluencies in Simultaneous Interpreting: A Corpus-Based Analysis

Claudio Bendazzoli, Annalisa Sandrelli and Mariachiara Russo

12.1 Introduction

The aim of the present chapter[1] is to analyse two specific types of disfluencies in spoken language: mispronounced words and truncated (unfinished) words. The analysed material comes from the European Parliament Interpreting Corpus (EPIC), so the object of this chapter is spoken language produced by source language (SL) speakers and by target language (TL) speakers, that is simultaneous interpreters.

12.1.1 Main Features of EPIC

EPIC is one of the first examples of electronic corpora available in Interpreting Studies (further resources are discussed in Bendazzoli and Sandrelli 2009). It contains transcripts of speeches delivered at the European Parliament in English, Italian and Spanish and the simultaneous interpretations of each speech into the other two languages involved, covering all possible combinations and directions (Bendazzoli 2010; for information on the recordings and the transcription process see Bendazzoli et al. 2004 and Monti et al. 2005). The resulting EPIC corpus is structured into nine sub-corpora, three of which contain the source speeches and the other six contain the target speeches. Its structure makes it possible to carry out both parallel and comparable analyses (Shlesinger 1998; Laviosa 1998, 2002; Bowker and Pearson 2002), that is, to compare source texts (STs) with their interpretations (i.e. target texts, TTs) into the other languages, and to study the characteristics of the same language as spoken by source language speakers and by simultaneous interpreters working into that language from two different source languages. The structure of the corpus can be seen in Figure 12.1:

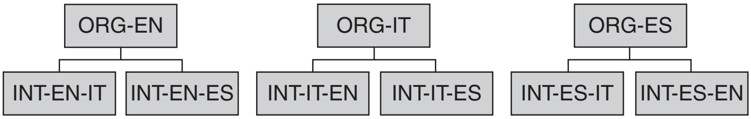


Figure 12.1 Corpus structure. (ORG = original speech, i.e. source text; INT = interpreted speech, i.e. target text; EN = English; IT = Italian; ES = Spanish)

Table 12.1 Corpus size

Sub-Corpus	No. of Speeches	Total Word Count	% of EPIC
Org-en	81	42,705	25
Int-en-it	81	35,765	20
Int-en-es	81	38,066	21
Org-it	17	6,765	4
Int-it-en	17	6,708	4
Int-it-es	17	7,052	4
Org-es	21	14,406	8
Int-es-en	21	12,995	7
Int-es-it	21	12,833	7
TOTAL	357	177,295	100

Currently, the size of the corpus is nearly 178,000 tokens. The various subcorpora differ in size (see Table 12.1 above), but EPIC is an open corpus and more transcripts will be added in the future.

EPIC is a machine-readable corpus, that is it can be explored by means of computational linguistics tools, and data can be automatically extracted and then analysed. Transcripts are part-of-speech(POS)-tagged and lemmatized, that is morphological information and lemmas are attached to each token in the form of a tag. This was done by using different taggers, depending on the language: *TreeTagger* (Schmid 1994) for English and Italian, and *Freeling* (Carreras et al. 2004) for Spanish (for details on the tagging process see Sandrelli and Bendazzoli 2006). The corpus was also indexed by using the *IMS Corpus Work Bench* (CWB) suite of tools (Christ 1994). This makes it possible to carry out simple or advanced queries through a dedicated web interface (see web references), in which it is also possible to set different parameters (e.g. mode of delivery, topic, type of speaker, etc.) in order to restrict queries. All the parameters available are specified at the beginning of each transcript in a header, that is, a sequence of fields providing information about the text file, the speaker (e.g. gender, political function, country) and the speech (e.g. topic, number of words, duration etc.).

All the materials are saved in formats that can run on a wide range of IT applications, enabling users to search them in various ways and use them both for research (see Sandrelli and Bendazzoli 2005; Russo et al. 2006) and for teaching purposes (Bendazzoli and Sandrelli 2005/2007; Russo 2010; Sandrelli 2010). In this chapter, as is explained briefly in Section 12.3, the transcription conventions used for mispronounced and truncated words were exploited for research on disfluencies. The specific hypotheses underlying the present study are presented in the following subsection.

12.1.2 Hypotheses and Focus

The focus of the present chapter is twofold, as it includes both quantitative and descriptive aspects. The basic aim is to provide information on the incidence of two types of disfluencies and outline their main features in the data under analysis. More specifically, all the occurrences of the two types of disfluencies were extracted and counted to ascertain whether they were more frequent in the source texts or in the target texts. Furthermore, the data were analysed to establish whether speakers and interpreters managed to repair their output, that is, whether mispronounced words were corrected and truncated words were completed. Our starting hypothesis is that simultaneously interpreted (SI) speeches are likely to contain a higher number of the two disfluencies and a lower number of corrections, owing to SI-related constraints, such as time pressure.

Descriptive data are provided in the more exploratory part of the chapter. Here the focus is on the main characteristics of the disfluencies under consideration and their possible causes. We looked at the presence of editing terms (verbal material added between mispronounced or truncated words and their repair), at the nature of each disfluency (single or serial), and at the point where the articulation of truncated words stopped (after the production of the very first phoneme or after several phonemes). The analysis also covers the possible causes behind each disfluency, classified according to specific categories that were identified by taking into account general classifications of speech errors in language production and previous studies on disfluencies in SI.

12.2 Theoretical Background

A review of the main literature on language production and on disfluencies (Fromkin 1973; Cutler 1982; Levelt 1983, 1989; Shriberg 1994; Akmajian et al. 1995; Fromkin and Rodman 1998; Eysenck and Keane 2000) was complemented by the study of available literature on cognitive processes of language production in simultaneous interpreting (Gile 1995; De Bot 2000) and on disfluencies in interpreted speeches (Pöchhacker 1995; Tissi 2000; Petite 2003, 2005; Van Besien and Meuleman 2004). The theoretical background thus sketched enabled us to produce operational definitions of the disfluencies that are the object of the present chapter.

12.2.1 Disfluencies and Spoken Language Production

Speech production is generally considered harder to study than comprehension because of the difficulty in devising suitable experimental tasks to reveal production mechanisms. Only the output of the speech production process is easily accessible for study: therefore, a large part of what is known about speech production has been surmised from the analysis of speech and of speech errors in particular.

It is difficult to estimate the frequency of speech errors in ordinary conversation. Some sources calculate about 2 per cent (Deese 1978, 1980 in Fromkin and Berstein-Ratner 1998), others indicate a much lower proportion, 1/1000 (Bock and Loebell 1988, in Akmajian et al. 1997). This is because different authors used different definitions of errors. One of the most comprehensive categorizations was suggested by Magno Caldognetto et al. (1982), who talked about non-fluencies, including silent pauses and disfluencies. Tissi (2000) adapted this classification for SI and divided disfluencies into filled pauses (vocalized hesitations, vowel and consonant lengthening) and interruptions (repeats, restructuring and false starts). Tissi specifies that repeats include repetitions of phrases, whole words or parts of a word, whereas restructuring includes corrections of phonological lapses and of formulation and content errors, as well as structure reformulations. According to Tissi (2000: 114), false starts 'occur when the speaker interrupts an utterance and begins a new one without having completed it'. Under this classification, speech errors are categorized as interruptions.

According to the available literature (above all Fromkin 1973 and Cutler 1982), speech errors may be classified as phonological, syntactic and semantic, as can be seen in Table 12.2.

Empirical data on speech errors in conversation have shown that word blends and substitutions never produce ungrammatical strings: this means that speech is planned prior to articulation. An example of planning is the fact that the grammatical class of a word is determined before the word is uttered: indeed, 99 per cent of semantic substitutions involve a word belonging to the same word class (e.g. a noun is replaced by another noun). Furthermore, word substitutions include both substitutions with another word with a similar semantic content and substitutions with a phonologically similar word: this indicates that the mental lexicon is organized both semantically and phonologically. Thirdly, errors involving parts of morphologically complex words

Table 12.2 Main types of speech errors

Phonological errors	sound exchange errors, spoonerisms*	you have hissed all my mystery lectures (= you have missed all my history lectures)
	anticipation errors	a leading list (= a reading list)
	perseveration errors	a phonological fool (= a phonological tool)
	malapropisms	she went to the groinecologist
	haplologies**	particularly > particuly
Semantic errors (lexical selection)	blends	semantic: a tennis athler (= player+ athlete)
		semantic and phonological: It's difficult to valify (= validate + verify)
	substitutions	phonological: at 4.30 we're adjoining the meeting (= adjourning the meeting)
		semantic: too many irons in the smoke (= in the fire)
	word exchange errors	I must let the house out of the cat
	morpheme exchange errors	he has already trunked two packs (= packed two trunks)
Syntactic errors	shifts	mermaid moves (mermaids move)

* In spoonerisms, the initial letter(s) is (are) switched within the same clause. Consonants are always replaced by consonants and vowels by vowels.

** Haplology involves the elimination of a syllable when two consecutive identical or similar syllables occur. Syllables are both medial and the structure of the two syllables is similar.

indicate that such words are assembled prior to articulation: e.g. *sesame street crackers* instead of *sesame-seed crackers* (Fromkin and Bernstein-Ratner 1998: 324). Fourthly, function words and affixes tend to be involved in shift errors (i.e. they are inserted in a different place in the sentence). Lastly, speech errors often reflect the subject's rule knowledge: a rule may be applied to an irregular form (*knowed* instead of *knew*), or may be applied wrongly (he *am* instead of he *is*).

In order to explain the above errors, Fromkin developed the first speech production model in 1971. Another was produced by Garrett in 1976, followed by a third by Levelt in 1989. Fromkin's Utterance Generator Model sees the production process as consisting of six successive stages in which the message has different forms of representation. Garrett's model is very similar to Fromkin's, although the labels suggested for the various stages are slightly different. Both models satisfactorily explain all types of observed speech errors.

However, they fail to take into account self-correction, which is very common both in everyday speech and in simultaneous interpreting. In order to explain self-correction, Levelt (1989) postulates the existence of a conceptualizer (for message generation), a formulator (for grammatical encoding and phonological encoding) and an articulator (for production of overt speech), and then adds a speech-comprehension system comprising an acoustic-phonetic processor (for phonetic representation), a parser (for phonological decoding and lexical selection, followed by grammatical decoding) and a discourse processing and monitoring component in the conceptualizer to complete the cycle. Basically, the speech comprehension system checks the subject's output and prompts repairs whenever necessary. It appears that the system tends to reject non-words, but accepts real words even though they are not the ones that the subject intended to utter (as shown by errors in the lexical selection stage).

In a paper on self-corrections, Levelt (1983) defines three components in any repair: the reparandum (i.e. the word to be repaired or corrected), the editing term (any verbal material produced between the error and the corrected version) and the reparatum (the repaired version considered correct by the speaker). The reparandum may be an unfinished word or a fully-uttered word: in either case the subject believes it inappropriate or incorrect and decides to replace it. Shriberg (1994) further specifies that the editing term is not always present. The space between the reparandum and the reparatum (interregnum) may contain empty pauses or repetitions of the problematic item. Moreover, the reparatum may be produced at a distance of a number of verbal items from the reparandum. Finally, Shriberg defines complex (or serial) disfluencies as multiple repairs in a serial relationship, in which the last item of the first disfluency is next to the first item of the following disfluency.

Levelt (1983) also specifies that repairs may be classified according to the stage of the production process, that is, they may be divided into pre-articulatory (or covert) repairs, mid-articulatory repairs and post-articulatory repairs. Another useful distinction is between error repairs and appropriateness repairs, the latter referring to those cases in which the speaker repairs an output that is formally correct but is considered stylistically unsatisfactory.

Levelt's model is consistent with Dell's spreading activation model of speech production (Dell 1986, in Fromkin and Bernstein-Ratner 1998), which sees the lexicon as organized into networks of words, based on semantic and phonological relatedness. When a concept is activated for production, all the lexical items sharing semantic or phonological properties with that word are activated as well. The subject chooses one of these multiple lexical candidates, which receives phonological activation for production. This mechanism explains slips of the tongue involving words sharing both phonological and semantic features (e.g. *arrested and persecuted* instead of *arrested and prosecuted*). It also explains why recently retrieved lexical items can induce a slip in the next segment.[2]

12.2.2 Language Production and Disfluencies in Simultaneous Interpreting

De Bot (2000) proposes an adapted version of Levelt's model (1989) to suit SI-specific features. The main difference between language production in ordinary speech and language production in simultaneous interpreting is that the interpreter does not generate the concept to be expressed in the TL, since s/he must convey the concept intended by the SL speaker. Presumably, the interpreter is influenced by the SL speech in all subsequent stages of TL production. The choice of grammatical structures and lexical items is not entirely free, because the interpreter works on the chunks of the SL speech as they become available. Therefore, the time factor places a constraint on the interpreter's syntactic and lexical choices in a way that is not to be found in ordinary speech production.

De Bot postulates that linguistic elements are labelled on a language basis and that the relevant language store is activated when an interpreter is working. When interpreting simultaneously, a language store (e.g. English) is activated for perception and comprehension, whereas the other store (e.g. Italian) is activated for language production. However, a further complicating factor is the translation element. The interpreter must search for the TL expression that conveys the meaning of the original best. Therefore, SI involves a mixture of vertical processing used to produce the most fluent TL version possible and horizontal processing that is necessary to reproduce keywords, technical terms, names, dates and so on.

Finally, the interpreter, like any speaker, monitors his/her own output thanks to the speech comprehension system. However, monitoring the TL output is made more difficult by the need to attend to the concurrent SL input. Gile (1995) describes this situation in his Effort Model: the different tasks involved in simultaneous interpreting compete for the interpreter's cognitive resources. Goldman-Eisler (1980) suggests that familiarity with the material to be translated facilitates simultaneous interpreting: highly frequent expressions for which a TL equivalent is readily available can be dealt with almost automatically, thus freeing up cognitive resources for TL monitoring. By contrast, when translating 'new' information, the effort to come up with an adequate TL translation competes with the need to monitor the TL output. Despite this difficulty, it has been shown in several studies that interpreters repair some of their errors. What follows is a brief overview of the most relevant literature on the subject.

In a study comparing the performances of experienced interpreters and trainees, Ilic (1990) showed that self-monitoring skills and the ability to repair errors increase with training and experience. Moreover, professional interpreters tend to notice and correct semantic errors more frequently than syntactic errors, whereas for beginners the opposite is true.

Pöchhacker (1995) carried out a study on slips of the tongue and structure shifts (false starts, lexical blends and syntactic blends) in a corpus of 145 texts. He hypothesized that interpreters would produce more slips and shifts than SL speakers, because of the simultaneity of the listening, production and monitoring tasks. The most common disruption in both STs and TTs was false starts.[3] Interpreters were found to produce more lexical and syntactic blends, whereas SL speakers produced false starts, corrected slips and uncorrected slips more frequently. Overall, Pöchhacker's starting hypothesis was not confirmed, in that SL speakers were found to produce more slips than interpreters. However, a significant finding was that corrected slips were more frequent in STs, that is interpreters repaired their output less often than SL speakers.

Van Besien and Meuleman (2004) analysed a small corpus of speeches in Dutch and their simultaneous interpretations into English. They found that interpreters tried to correct the SL speakers' errors whenever possible, that is they did not reproduce their errors in the interpretations. If the SL speaker repaired his/her output, the interpreters translated only the repaired version. Furthermore, 'the number of successful translations of repairs is not significantly influenced by the type of repair (error repair vs. appropriateness repair), the moment of repair, or by the presence or absence of an editing expression' (Van Besien and Meuleman 2004: 80). However, in most cases of fresh starts (when the speaker radically changed the structure of the sentence), interpreters translated both the false start and the fresh start.

Petite (2003, 2005) adopted Levelt's classification of repairs and his distinction between error repairs and appropriateness repairs. She found plentiful evidence of the self-monitoring mechanism, since interpreters were found to carry out all types of repairs.

Finally, in a study on silent pauses and disfluencies in SI, Tissi (2000) found no direct correlation between disfluencies in the STs and TTs, but pointed out that disfluencies in the interpreted texts may be caused by certain characteristics of the STs: 'Even if there is no formal correspondence, some occurrences in TT are caused by the need to wait for new items, delayed because of a speaker's pause and/or interruption' (Tissi 2000: 121).

12.2.3 Operational Definitions: Mispronounced Words and Truncated Words

The focus of the present chapter is on truncated words and mispronounced words. By making reference to the various classifications presented above, it is now possible to better define them. First, under the categorization proposed by Magno Caldognetto et al. (1982), later adapted by Tissi (2000), both truncated words and mispronounced words are *disfluencies*, and more specifically, they belong to the *interruptions* category. Furthermore, truncated words could be categorized as false starts, according to Tissi's definition, whereas mispronounced words would be classified as restructuring, but only when they are repaired. In our study we need a broader definition in order to include all occurrences of mispronounced words, irrespective of whether the speaker manages to correct the error or not.

Therefore, we reverse Tissi's perspective, starting by looking at two specific types of disfluencies in order to see whether they are repaired by speakers and interpreters, and whether it is possible to identify a possible cause for such errors. However, there is no point in devising a new categorization for general descriptive purposes, as was pointed out by Pöchhacker (1995) after a brief overview of existing taxonomies of speech errors.

As regards the categorization of speech errors and slips of the tongue suggested in Table 12.2, reference is made to it in classifying the occurrences of mispronounced words found in our corpus (see Tables 12.5 and 12.6).

Finally, Levelt's description of the repair mechanism and its constituent parts (reparandum, editing term and reparatum) will be used to show how speakers and interpreters manage to correct their output (or otherwise). Shriberg's concept of complex or serial disfluencies will also be employed to describe instances of multiple disfluencies found in our corpus.

All of these observations come together in the analysis grid that has been developed to classify each and every occurrence of truncated words and mispronounced words extracted from the corpus. The following section gives an overview of the analysis grid with all its fields.

12.3 Methodology

Since EPIC is POS-tagged, it is possible to extract data automatically by means of specific procedures. Its nine sub-corpora of speeches with different characteristics, including language and type of speech (ST or TT), allow for the retrieval of separate lists of data for each sub-corpus and type of disfluency. Clearly, the data thus obtained automatically from the transcripts served only as a written basis to trace all the disfluencies in the speeches: as the object of our analysis is disfluencies in spoken language, the real data to be analysed are, in fact, the recordings. Since each token was accompanied by some context (50 words to the right and left of the query) and by the header of the text file it came from, it was comparatively easy to track down the specific passage in the audio and video clips held in EPIC.

All the data were analysed by means of a dedicated classification grid that aims to summarize all the relevant features of disfluencies highlighted in the theoretical overview. Two Microsoft Excel tables were created to collect information on the truncated words and mispronounced words extracted from EPIC.

Table 12.3 Fields in the grid for the analysis of mispronounced words (org-en sub-corpus)

Transcript	Speech reference code	Reparandum	Editing Term	Reparatum	Corrected?	Number of items	POS	Type of error	Possible Cause	Problems in INT-EN-IT	Problems in INT-EN-IT	Comments

Table 12.4 Fields in the grid for the analysis of truncated words (org-en sub-corpus)

Transcript	Speech reference code	Reparandum	Editing Term	Reparatum	Completed?	Number of items	POS	Affected part of word	Type of error	Possible Cause	Problems in INT-EN-IT	Problems in INT-EN-ES	Comments

The two grids are virtually identical. The first column contains each occurrence in context; the second column contains the reference code indicating the speech from which the truncated or mispronounced word is taken. The third column contains the disfluency, followed by any editing term(s) in the fourth column and the final version produced by the speaker in the fifth. The sixth column is used to record whether the speaker was successful in repairing his/her output or whether the error went undetected and uncorrected. The seventh field is used to distinguish between simple disfluencies and serial disfluencies. The part of speech affected by the disfluency is recorded in the eighth column. The following column is used to record the type of error. Moreover, only in Table 12.4 is there a specific column that indicates where the speaker has interrupted the articulation of the truncated word (affected part of the word). In Table 12.5, an additional category (approximation) has been added to the error classification that was presented in Table 12.2, in order to better reflect our data. It indicates those cases in which a speaker produces the reparatum by means of repeated attempts, often simply by repeating the same verbal item to be repaired. In the case of mispronounced words, we use the term approximation to describe those cases in which the speaker begins uttering a word, then interrupts him/herself and pauses, and

Table 12.5 Types of speech errors (the examples are real occurrences from EPIC)

PHONOLOGICAL ANTICIPATION
to implement the two thousand and two **antif- anti-trafficking** framework decision to catch the criminal gangs

LEXICAL ANTICIPATION
I mean comparing that with **nati- ehm public ehm investments nationally** it's it's it's very little

PHONOLOGICAL PERSEVERATION
the Council also received ehm reached a general ehm approach on the **apr- proposed** regulation

APPROXIMATION
but they are certainly not **te- te- terminal** ehm and they can in our judgment be addressed

OTHER
so for the next three years we have **autho- we have certification** of projects that have actually being carried out

Table 12.6 Causes of disfluencies (the examples are real occurrences from EPIC)

LEXICAL SHIFT
so for the next three years we have **autho- we have certification** of projects that have actually being carried out

SYNTACTIC SHIFT
the population **getti- the growing** of the population

ARTICULATION
so I think there are two aspects // first of all **w- we're** stressing those elements which we think will ease

STALLING
what we're doing with a reduction of **con- ehm contaminating** emissions

OTHER
is the microphone **wor-** // well thank you ehm President and ehm welcome to our Commissioner David Byrne

finally completes the word. The category 'other' has also been introduced to collect examples that do not fall into any of the other cases.

The next field in Table 12.4 is used to record possible causes for the error (listed in Table 12.6). Sometimes it is possible to see that the speaker wanted to make a lexical change (lexical shift) or to adopt a different grammatical structure (syntactic shift). In other cases, by listening to the recording it becomes obvious that the speaker was finding it difficult to articulate the item in question (articulation). In other examples the speaker was presumably employing the stalling strategy, that is, producing verbal material to gain time and come

Table 12.7 General results in a comparable perspective

	Mispronounced		Truncated	
Epic sub-corpus	**No. of occurrences**	**%**	**No. of occurrences**	**%**
ORG-EN	20	0.046	323	0.756
INT-IT-EN	5	0.074	68	1.013
INT-ES-EN	19	0.146	115	0.884
ORG-IT	16	0.236	35	0.517
INT-EN-IT	119	0.332	221	0.617
INT-ES-IT	57	0.444	105	0.818
ORG-ES	24	0.166	46	0.319
INT-IT-ES	18	0.255	78	1.106
INT-EN-ES	102	0.267	282	0.740

up with a repaired output. Finally, there are cases in which it is impossible to indicate a source of error with any certainty (other).

The final field is used to record any comments on specific features of the segment under analysis which do not fall under any other field in the table.

The information recorded and classified by means of the grids has enabled us to study in detail all the occurrences of truncated and mispronounced words in the corpus. The data were evenly distributed among the three authors who compiled their own grids separately and then collated the results of their analysis.

12.4 Results and Discussion

The incidence of mispronounced and truncated words in the nine sub-corpora was calculated as a percentage of sub-corpus size, together with the overall number of occurrences of the two disfluencies under study (see Tables 12.7 and 12.8). Table 12.7 presents the data in a comparable perspective and Table 12.8 in a parallel perspective.

Table 12.7 shows that truncated words are much more frequent than mispronounced words in all sub-corpora. Furthermore, our results indicate that in the sub-corpora of interpreted speeches there is a higher incidence of both types of disfluency in comparison with the percentages found in the sub-corpora of speeches originally delivered in the same language. This is a general trend, irrespective of the language pair considered, and it corroborates an intuitive assumption: spontaneous speech production in the mother tongue is an easier task than simultaneous interpretation, even when this is performed into the mother tongue. Thus, this corpus-based analysis of disfluencies provides

Table 12.8 General results in a parallel perspective

	Mispronounced		Truncated	
Epicsub-corpus	**No. of occurrences**	**%**	**No of occurrences**	**%**
ORG-EN	20	0.046	323	0.756
INT-EN-IT	119	0.332	221	0.617
INT-EN-ES	102	0.267	282	0.740
ORG-IT	16	0.236	35	0.517
INT-IT-EN	5	0.074	68	1.013
INT-IT-ES	18	0.255	78	1.106
ORG-ES	24	0.166	46	0.319
INT-ES-EN	19	0.146	115	0.884
INT-ES-IT	57	0.444	105	0.818

further empirical evidence of the heuristic validity of Gile's (1995) Effort Model in explaining the complexity of the interpreting task: weaknesses in interpreters' performance may be caused by having to constantly balance ST listening and analysis, memory, TT production and coordination requirements.

If the results are presented in a parallel perspective, that is, if SL speeches are directly compared with their corresponding TTs in the other two languages, further interesting trends emerge (see Table 12.8).

The general trend concerning both mispronounced and truncated words seems to confirm the findings of the comparable analysis, that is the sub-corpora of interpreted speeches are generally characterized by a higher incidence of both types of disfluency than the sub-corpora of SL speeches. In particular, in the sub-corpora of speeches interpreted from English (INT-EN-IT and INT-EN-ES) the percentages of mispronounced words are similar. The same trend can also be detected in the sub-corpora of speeches interpreted between the two cognate languages in question (INT-IT-ES and INT-ES-IT), where the incidence of mispronounced words is higher than in their STs. This may be explained by the fact that, although the linguistic and phonological similarities between cognate languages may reduce the comprehension effort, at the same time production is not made easier because the risk of interferences is always present. The only exception to the general trend about mispronounced words concerns the English TTs (INT-IT-EN and INT-ES-EN), in which a lower percentage of mispronounced words was observed in comparison with their Italian and Spanish STs. The greater structural difference between a Germanic and a Romance language may enable interpreters to keep the two phonological systems separate, thus avoiding those pronunciation problems observed when interpreting between Romance languages.

As regards truncated words, the parallel analysis confirms that the incidence of this type of disfluency is greater in interpreted speeches, the only exception being the sub-corpus of speeches originally delivered in English. Indeed, the English SL speakers were the most affected by this type of disfluency.

These two exceptions (fewer mispronounced words in English TTs and fewer truncated words in interpreting from English STs) call a possible language-pair effect into play. Indeed, it was only in the language combinations in which English was involved, either as TL or as SL, that a different trend was consistently recorded.

Finally, in relation to truncated words, a ranking order can be seen in the sub-corpora of SL speeches, with the greatest frequency among English speakers (0.756 per cent), followed by Italian speakers (0.517 per cent) and finally Spanish speakers (0.319 per cent). This finding probably highlights differences in speaking styles: it would seem that English is generally spoken in a more 'faltering', 'staccato' way than Italian or Spanish. This also applies to English interpreters.[4]

The overall results were further broken down into two separate sub-sections, in order to analyse mispronounced and truncated words in more detail.

12.4.1 Mispronounced Words

This section includes results on self-corrections, presence of editing terms and number of items, and results on the types of error and their possible causes. A general overview of results is provided in Table 12.9 below.

Table 12.9 Features of mispronounced words (%)

	Corrected		Editing term		No. of items	
Sub-corpora	**Yes**	**No**	**Yes**	**No**	**Single**	**Serial**
ORG-EN	30	70	5	95	100	0
INT-EN-IT	11	89	3	97	98	2
INT-EN-ES	16	84	2	98	100	0
ORG-IT	0	100	0	100	100	0
INT-IT-EN	0	100	0	100	80	20
INT-IT-ES	17	83	11	89	89	11
ORG-ES	16.5	83.5	16.5	83.5	96	4
INT-ES-EN	16	84	5	95	100	0
INT-ES-IT	**7**	**93**	**3.5**	**96.5**	**100**	**0**

Mispronounced words were not frequently corrected. The trend is especially marked in all sub-corpora of interpreted speeches, irrespective of the language combination or direction (note the higher percentages of NO repair in the 'corrected' column). A possible explanation for the lack of repairs is that most of these disfluencies are not perceived by SL speakers and TL interpreters. As regards the latter, another possible reason is that the effectiveness of self-monitoring carried out in interpreting is reduced by the concurrent decoding, memorizing and encoding efforts.

This also applies to the trend concerning editing terms. There are very few editing terms (only 5 per cent across the 7 sub-corpora; 11 per cent in int-it-es and 16.5 per cent in ORG-ES). In our data, only Spanish speakers show a certain tendency to add verbal material before repairing their output.

As regards the number of items affected by disfluencies, Table 12.9 shows that mispronounced words are generally single words (≥ 80 per cent SINGLE vs. ≤ 20 per cent SERIAL). A similar pattern was observed for English and Spanish both as source languages and as target languages, with a very low percentage of serial disfluencies (between 0 per cent and 4 per cent). In interpreted speeches from Italian there is a higher incidence of serial occurrences: 20 per cent in the INT-IT-EN sub-corpus and 11 per cent in the INT-IT-ES sub-corpus. However, the two samples are very small in size (5 and 18 occurrences in the two sub-corpora, respectively – see Table 12.8).

Turning to error types, a fundamental role is played by the need to attend to a new SL input chunk while simultaneously producing a TL translation of a previous SL chunk, which may disrupt interpreters' TL production and cause interferences. Instances of such interferences have been found in EPIC, as can be observed in Table 12.10, in which phonological anticipations and

Table 12.10 Types of error for mispronounced words (%)

Sub-Corpora	Phonological anticipation	Approximation	Perseveration	Other
ORG-EN	**35**	**0**	**45**	**20**
INT-EN-IT	**13**	**2**	**26**	**59**
INT-EN-ES	**15**	**4**	**31**	**50**
ORG-IT	**12.5**	**0**	**50**	**37.5**
INT-IT-EN	**0**	**0**	**20**	**80**
INT-IT-ES	**28**	**0**	**22**	**50**
ORG-ES	**12.5**	**16.5**	**37.5**	**33.5**
INT-ES-EN	**5**	**0**	**26**	**68**
INT-ES-IT	**16**	**5**	**23**	**56**

perseverations can be seen to be the most frequent errors. No occurrences of lexical anticipations were observed.

Table 12.10 reveals that the most frequent types of error differ between STs and TTs. Interestingly, perseverations are more frequent in SL speeches than in interpreted speeches, while the general category 'Other' is more frequent in interpreted speeches than in SL speeches, including many cases of lexical and syntactic blends, malapropisms and incorrect pronunciation of proper names. It would seem that SL speakers who deliver their own self-planned and self-paced speeches tend to incur production problems confined to phonological carry-over effects. By contrast, interpreters are more likely to incur syntactic and lexical planning errors, which may be caused by insufficient *décalage* and/or cognitive overload.

The percentage of phonological anticipation is unevenly distributed in the nine sub-corpora. The only clear result is that this type of error is more frequent in English STs than in the related Italian and Spanish TTs. Once again, a language direction effect is detected, with both Italian and Spanish TTs being characterized by a lower incidence of phonological anticipation.

By contrast, in the sub-corpora of speeches interpreted from Italian into Spanish and vice versa, phonological anticipation is more frequent than in their respective STs. The same does not hold true for speeches interpreted from Italian into English and from Spanish into English. Once again, this points to the interference caused by linguistic affinity between the two Romance languages.

Approximation is hardly ever recorded in all the sub-corpora.

With respect to the possible causes of mispronounced words, articulation accounts for the overwhelming majority of cases in all sub-corpora (see Table 12.11). Some occurrences of stalling were recorded, but none of lexical shift, syntactic shift or 'Other'.

Table 12.11 Possible causes of mispronounced words (%)

Sub-corpora	Articulation	Stalling
ORG-EN	95	5
INT-EN-IT	96	4
INT-EN-ES	94	6
ORG-IT	94	6
INT-IT-EN	100	0
INT-IT-ES	67	33
ORG-ES	87.5	12.5
INT-ES-EN	95	5
INT-ES-IT	93	7

Interestingly, articulation problems are evenly distributed in the sub-corpus of English STs and in the two sub-corpora of TTs from English (95 per cent, 96 per cent and 94 per cent, respectively). However, the results concerning the three sub-corpora of Italian STs and their TTs in English and Spanish are less clear-cut, owing to the very small size of the sample. Finally, in the sub-corpora of Spanish STs and their Italian and English TTs articulation is also the main cause of error. However, a small percentage of stalling was also recorded in Spanish STs.

12.4.2 Truncated Words

Truncated words were more frequent than mispronounced words overall. Moreover, interpreters were found to produce a higher number of occurrences than SL speakers, with the exception of English STs. The results on the features analysed according to our grid can be seen in Table 12.12.

In general, truncated words tend to be completed, that is the *reparandum* is repaired. The trend is particularly marked in the sub-corpora of SL speeches (ORG-EN 88 per cent, ORG-ES 76 per cent, ORG-IT 66 per cent), while percentages in the sub-corpora of interpreted speeches are generally lower. This is in line with the idea that monitoring the output is part of the language production process, and that in SI there are cognitive demands that have a negative impact on monitoring. The sub-corpus of speeches originally delivered in Italian has a much lower percentage of repaired output (66 per cent), but this may be an effect of the small size of the corpus.

Let us turn to the second feature taken into account, that is, the possible presence of editing terms between the *reparandum* and the *reparatum*. As explained

Table 12.12 Features of truncated words (%)

Sub-corpora	Completed		Editing term		No. of items		Affected part	
	Yes	**No**	**Yes**	**No**	**Single**	**Serial**	**First ph.**	**More**
ORG-EN	88	12	22	78	86	14	50	50
INT-EN-IT	66	34	56	44	85	15	32	68
INT-EN-ES	58	42	54	46	93	7	27	73
ORG-IT	66	34	46	54	66	34	20	80
INT-IT-EN	94	6	34	66	69	31	53	47
INT-IT-ES	65	35	46	54	79	21	31	69
ORG-ES	76	24	43	57	85	15	13	87
INT-ES-EN	71	29	37	63	90	10	50	50
INT-ES-IT	68	32	68	32	91	9	17	83

previously, the presence of editing terms may be indicative of back-tracking on the part of the speaker/interpreter. In our overall results, English source speakers used editing terms less frequently than their Italian and Spanish colleagues, who inserted editing terms in nearly half of all the occurrences of truncated words (editing terms accounted for 22 per cent of occurrences of truncated words in ORG-EN, vs. 46 per cent in ORG-IT and 43 per cent in ORG-ES). In the sub-corpora of interpreted speeches, however, the frequency of editing terms is equal or higher than the frequency of editing terms in the corresponding STs, with the exception of TTs into English (ORG-IT 46 per cent vs. INT-IT-EN 34 per cent and ORG-ES 43 per cent vs. INT-ES-EN 37 per cent). Once again, a language direction effect comes into play, thus suggesting that interpreters working from the two Romance languages into English performed less back-tracking than interpreters working from English into Italian and Spanish or between two Romance languages.

As regards the number of truncated items contained in each occurrence, Table 12.12 shows that single truncated words are much more frequent than serial ones. This trend seems to be common across all languages and types of sub-corpora, albeit with percentage variations (single occurrences range from 93 per cent to 66 per cent). Finally, we examined where articulation is interrupted in truncated words, since production may stop after the first phoneme or may include a larger part of the affected word. In this respect, the differences found seem to be related to the language involved, rather than to the type of sub-corpus (i.e. of STs or TTs). Indeed, Spanish and Italian speakers and interpreters tend to interrupt words well beyond the first phoneme. By contrast, in the sub-corpora of English SL speeches and interpreted speeches the number of words truncated after the first phoneme almost equals the amount of truncated words in which articulation stops after producing more verbal material (exactly 50 per cent of cases in ORG-EN and INT-ES-EN, whereas in the INT-IT-EN corpus 53 per cent of 'first phoneme' cases were recorded vs. 47 per cent of cases in which the speaker uttered 'more' verbal material).

Let us now consider the types of error according to the categorization described earlier.

Approximation is the most frequent type of error in almost all sub-corpora. In particular, looking at Table 12.13 from a parallel perspective, this type of error seems to be more frequent in interpreted speeches than in SL speeches, with the exception of the English SL speeches. This peculiarity is mirrored by the results concerning the English TTs, in which the approximation percentages are also quite high. This seems to indicate that there is indeed a language family effect.

As regards the two types of anticipation considered in the present study, lexical anticipation tends to be more frequent than phonological anticipation, with few exceptions (ORG-ES and INT-IT-ES). This could mean that words are often truncated when speakers decide to change their speech plan (i.e. they

Table 12.13 Types of error for truncated words (%)

Sub-corpora	Phonological anticipation	Lexical anticipation	Approximation	Perseveration	Other
ORG-EN	2	6	73	5	14
INT-EN-IT	5	16	38	8	33
INT-EN-ES	1	16	37	5	41
ORG-IT	8.5	17	48.5	6	20
INT-IT-EN	4	18	65	7	6
INT-IT-ES	11.5	4	45	11.5	28
ORG-ES	17	11	36	8	28
INT-ES-EN	0	3.5	60	5	31.5
INT-ES-IT	2	7	65	1	25

are dissatisfied with lexical selection). This also applies to interpreters, who may need to reorganize their output as a consequence of further ST information becoming available. Finally, it should be noted that percentages in the ‘Other’ category are quite high, indicating a sizable number of other errors, such as blends, malapropisms, haplologies, and so on (though no exact calculation was made for each specific type, see Table 12.2). The incidence of these disfluencies is generally higher in the sub-corpora of interpreted speeches than in the sub-corpora of SL speeches, with only two exceptions to this rule (INT-IT-EN and INT-ES-IT).

Finally, the last set of parameters to be discussed in our analysis concerns the possible causes behind the occurrence of truncated words. The percentage results obtained for the five options available in our grid are shown in Table 12.14 below.

In all SL speeches the main cause of truncated words is articulation (48 per cent, 40 per cent and 61 per cent for English, Italian and Spanish, respectively). In the case of the Italian SL speeches, this equals the percentage of syntactic shifts. It is interesting to note that ‘stalling’ was not observed in either the Italian or the Spanish SL speeches, whereas it accounted for 29 per cent in English SL speeches.

The results concerning the TL speeches are less homogeneous. In the subcorpora of speeches interpreted from English into Italian and Spanish, there is a slightly higher percentage of shifts, a possible consequence of greater efforts and constraints in interpreters’ speech planning. However, this result was not confirmed in the speeches interpreted from Italian into English and Spanish, in which articulation is the main cause of truncated words (69 per cent and 50 per cent). It may be worth emphasizing that there is a small difference in the percentages of shifts, as these are slightly higher in Spanish TTs (21–29 per cent) than in English TTs (9–16 per cent). Once again, this result

Table 12.14 Possible causes of truncated words (%)

Sub-corpora	Lexical shift	Syntactic shift	Articulation	Stalling	Other
ORG-EN	13	9	48	29	1
INT-EN-IT	33	33	19	14	1
INT-EN-ES	40	26	10	23	1
ORG-IT	20	40	40	0	0
INT-IT-EN	9	16	69	6	0
INT-IT-ES	21	29	50	0	0
ORG-ES	22	17	61	0	0
INT-ES-EN	19	17	33	29	2
INT-ES-IT	31	22	32	11	4

suggests language family-related problems. Similarly, in the speeches interpreted from Spanish into Italian and English, articulation seems to be the most frequent cause of truncated words (33 per cent and 32 per cent), albeit to a lesser extent. Stalling comes second among the causes of truncated words in English TTs (29 per cent), whereas in Italian TTs lexical shifts are the second cause (31 per cent), followed by syntactic shifts (22 per cent). All in all, shifts appear to be slightly more frequent in the Italian TTs than in English ones, as was the case in Spanish TTs from Italian.

12.5 Conclusions and Further Developments

Our corpus-based analysis aimed at providing quantitative and descriptive data on two specific types of disfluencies, that is, mispronounced and truncated words. Their incidence and main features were investigated in EPIC, a corpus of SL speeches and their simultaneous interpretations into English, Italian and Spanish, in all directions and combinations (nine sub-corpora in total).

The comprehensive categorization presented above yielded a vast amount of data which have enabled us to test our initial hypotheses and to compare our results with those in the available literature on disfluencies in simultaneous interpreting.

Our first hypothesis was that simultaneous interpreters would produce more disfluencies than SL speakers. If results are read in a parallel perspective, the two types of disfluencies are more frequent in TTs than in their STs. However, there are two exceptions: English STs have more truncated words than their TTs in Italian and Spanish; and English interpreters seem to have fewer pronunciation problems than their Italian and Spanish speakers. In other words, interpreters working from a Romance into a Germanic language produced a more fluent output than when working between two Romance languages (with 'fluency' being defined as a low incidence of mispronounced and truncated words).

Furthermore, a comparable analysis of our results reveals that interpreted speeches, irrespective of the SL, are generally characterized by a higher percentage of both mispronounced and truncated words than SL speeches delivered in the same language.

Another hypothesis in this study was that because of SI-related constraints (such as time pressure), interpreters would not be able to repair their disfluencies. As regards mispronounced words, in general they were not corrected by either SL speakers or interpreters but the trend was especially marked in interpreted speeches, irrespective of the language combination and interpreting direction. As far as truncated words were concerned, we found that they were generally completed, more so by SL speakers than by interpreters.

Turning to the descriptive objective of our study, the data revealed the following patterns. As regards mispronounced words, very few editing terms were found. Only Spanish speakers and Spanish interpreters tended to add verbal material before repairing the mispronounced words. Mispronounced words generally occurred as single items.

With respect to truncated words, the most salient result about the use of editing terms is that it is less frequent in English SL speakers and interpreters. Single truncated words were more frequent than serial truncated words, with Spanish and Italian speakers and interpreters interrupting their output well beyond the first phoneme.

Further descriptive data were obtained on the possible types and causes of error. For mispronounced words, a comparison between STs and TTs highlighted two different trends. Interestingly, perseverations were more frequent in SL speeches than in interpreted speeches, while the general category 'Other' was more represented in interpreted speeches than in source speeches. Articulation problems are the main cause in all sub-corpora.

For truncated words, approximation was the most frequent type of error, with the exception of TTs from English STs, and error causes pointed to a 'language-family' effect, with Romance languages displaying a similar pattern irrespective of their being source or target languages.

Let us now consider the wider perspective offered by the literature on disfluencies in SI. Since the few studies available on disfluencies in interpreting used different definitions, a direct comparison of results is impossible. However, in some cases similarities in general trends can be identified. Our results on self-correction are in line with those obtained by Pöchhacker (1995) in his study on slips and shifts, that is SL speakers tend to repair their output more often than interpreters. However, while Pöchhacker found that SL speakers produced more speech slips and shifts than their interpreters, in our data this result was confirmed only in relation to truncated words in English STs versus their Italian and Spanish TTs. In all other sub-corpora, interpreters produced more truncated words than SL speakers. The same trend was seen in relation to mispronounced words.

On the basis of the data analysed in this chapter, further lines of investigation can be envisaged. For instance, the correspondence of disfluencies in STs and in the two related groups of TTs could be studied, so as to determine to what extent interpreters and SL speakers encounter similar difficulties in producing their speeches. Another element that is worth investigating is the part of speech affected by the disfluencies under consideration. The analysis grids used in the present work already include this kind of information: it would be possible to create frequency lists and explore them by applying corpus linguistics methods.

After the 'epic' effort required to create this trilingual machine-readable corpus, we hope that our contribution can help other scholars develop similar language resources and exploit existing ones, so that Corpus-Based Interpreting Studies may continue to grow as a sub-discipline.

Notes

[1] Although this chapter is the result of a joint effort, Claudio Bendazzoli is the author of Sections 12.1, 12.1.1, 12.1.2, 12.4.2; Annalisa Sandrelli of Sections 12.2, 12.2.1, 12.2.2, 12.2.3, 12.3; Mariachiara Russo of Sections 12.4 and 12.4.1. The conclusions (Section 12.5) were jointly drafted.

[2] Fromkin and Bernstein-Ratner (1998) give the following as an example: ask a friend to say the word *silk* several times. Then ask him/her to answer the following question very quickly, without thinking: "What do cows drink?" and they will probably say 'milk'.

[3] However, no specific definition of 'false start' is provided in Pöchhacker's (1995) study.

[4] A similar observation was made by Poyatos (1994: 45, 2002: 20) when describing the paralinguistic features of the English language. He emphasized the *staccato* effect as one of its intrinsic characteristics.

References

Akmajian, Adrian, Richard A. Demers, Ann K. Farmer and Robert M. Harnish (1995) *Linguistics, An Introduction to Language and Communication*, Cambridge, MA: MIT Press, 395–431.

Bendazzoli, Claudio (2010) 'The European Parliament as a Source of Material for Research into Simultaneous Interpreting: Advantages and Limitations', in Lew N. Zybatow (ed.) *Translationswissenschaft – Stand und Perspektiven. Innsbrucker*

Ringvorlesungen zur Translationswissenschaft VI (Forum Translationswissenschaft, Band 12), Frankfurt am Main: Peter Lang, 51–68.

Bendazzoli, Claudio and Annalisa Sandrelli (2005/2007) 'An Approach to Corpus-Based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus)', in Heidrun Gerzymisch-Arbogast and Sandra Nauert (eds) *Proceedings of the Marie Curie Euroconferences MuTra: Challenges of Multidimensional Translation – Saarbrücken 2–6 May 2005*. Available online at: http://www.euroconferences.info/proceedings/2005_Proceedings/2005_proceedings.html

— (2009) 'Corpus-Based Interpreting Studies: Early Work and Future Prospects', *Tradumatica 7. L'aplicació dels corpus linguistics a la traducció*. Available online at: http://webs2002.uab.es/tradumatica/revista/num7/articles/08/08art.htm

Bendazzoli, Claudio, Cristina Monti, Annalisa Sandrelli, Mariachiara Russo, Marco Baroni, Silvia Bernardini, Gabi Mack, Elio Ballardini and Peter Mead (2004) 'Towards the Creation of an Electronic Corpus to Study Directionality in Simultaneous Interpreting', in Nelleke Oostdijk, Gjert Kristoffersen and Geoffrey Sampson (eds) (2005) *Compiling and Processing Spoken Language Corpora, LREC 2004 Satellite Workshop, Fourth International Conference on Language Resources and Evaluation*, Lisbon: ELRA, 33–9.

Bowker, Lynne and Jennifer Pearson (2002) *Working with Specialized Language. A Practical Guide to Using Corpora*, London and New York: Routledge.

Carreras, Xavier, Isaac Chao, Lluís Padró and Muntsa Padró (2004) 'Freeling: An Open-source Suite of Language Analyzers', in Maria Teresa Lino, María Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds), with the collaboration of Carla Pereira, Filipa Carvalho, Milene Lopes, Mónica Catarino and Sérgio Barros, *Proceedings of the 4th International Conference on Language Resources and Evaluation*, ELRA: vol. 1, 239–42.

Christ, Oli (1994) 'A Modular and Flexible Architecture for an Integrated Corpus Query System', *COMPLEX '94*, Budapest.

Cutler, Anne (ed.) (1982) *Slips of the Tongue and Language Production*, Berlin: Mouton.

De Bot, Kees (2000) 'Simultaneous Interpreting as Language Production', in Birgitta Englund Dimitrova and Kenneth Hyltenstam (eds) *Language Processing and Simultaneous Interpreting*, Amsterdam and Philadelphia: Benjamins, 65–88.

Eysenck, Michael W. and Marck T. Keane (2000) *Cognitive Psychology: A Student's Handbook*, 4th edn, Hove and Philadelphia: Psychology Press (Taylor and Francis), 363–81.

Fromkin, Victoria A. (ed.) (1973) *Speech Errors as Linguistic Evidence*, The Hague: Mouton.

Fromkin, Victoria A. and Nan Bernstein-Ratner (1998) 'Speech Production', in Berko Jean Gleason and Nan Bernstein-Ratner (eds) *Psycholinguistics*, 2nd edn, Fort Worth: Harcourt Brace College Publishers, 309–39.

Fromkin, Victoria A. and Robert Rodman (1998) *An Introduction to Language*, Fort Worth: Harcourt Brace College Publishers.

Gile, Daniel (1995) *Basic Concepts and Models for Interpreter and Translator Training*, Amsterdam and Philadelphia: John Benjamins.

Goldman-Eisler, Frieda (1980) 'Psychological Mechanisms of Speech Production as Studied through the Analysis of Simultaneous Translation', in Brian

Butterworth (ed.) *Language Production, Vol. 1, Speech and Talk*, London: Academic Press, 143–53.

Ilic, Ivo (1990) 'Cerebral Lateralization for Linguistic Functions in Professional Interpreters', in Laura Gran and Christopher Taylor (eds) *Aspects of Applied and Experimental Research on Conference Interpretation*, Udine: Campanotto Editore, 101–10.

Laviosa, Sara (1998) 'The Corpus-Based Approach: A New Paradigm in Translation Studies', *Meta* 43(4): 474–9. Available online at: http://id.erudit.org/iderudit/003424ar

— (2002) *Corpus-Based Translation Studies: Theory, Findings, Applications*, Amsterdam and New York: Rodopi.

Levelt, Willem J. M. (1983) 'Monitoring and Self-repair in Speech', *Cognition* 14: 41–104.

— (1989) *Speaking: From Intention to Articulation*, Cambridge and London: MIT Press.

Magno Caldognetto, Emanuela, E. de Zordi and D. Corrà (1982) 'Il ruolo delle pause nella produzione della parola' [The role of pauses in speech production], *Il Valsalva-Bollettino Italiano di audiologia e foniatria* 5(1): 12–21.

Monti, Cristina, Claudio Bendazzoli, Annalisa Sandrelli and Mariachiara Russo (2005) 'Studying Directionality in Simultaneous Interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus)', *Meta* 50(4). Available online at: http://id.erudit.org/iderudit/019850ar

Petite, Christelle (2003) 'Repairs in Simultaneous Interpreting: Quality Improvement or Simple Error Correction', in Ángela Collados Aís, María Manuela Fernández Sánchez and Daniel Gile (eds) *La evaluación de la calidad en interpretación: Investigación*, Granada: Comares, 61–71.

— (2005) 'Evidence of Repair Mechanisms in Simultaneous Interpreting: A Corpus-Based Analysis', *Interpreting* 7(1): 27–49.

Pöchhacker, Franz (1995) 'Slips and Shifts in Simultaneous Interpreting', in Jorma Tommola (ed.) *Topics in Interpreting Research*, Turku: University of Turku, Centre for Translation and Interpreting, 73–90.

Poyatos, Fernando (1994) La comunicación no verbal: paralenguaje, kinésica e interacción, Madrid: Istmo.

— (2002) *Nonverbal Communication Across Disciplines*. Volume II: Paralanguage, kinesics, silence, personal and environmental interaction, Amsterdam and Philadelphia: John Benjamins.

Russo, Mariachiara (2010) 'Reflecting on Interpreting Practice: Graduation Theses Based on the European Parliament Interpreting Corpus (EPIC)', in Lew N. Zybatow (ed.) *Translationswissenschaft – Stand und Perspektiven. Innsbrucker Ringvorlesungen zur Translationswissenschaft VI (Forum Translationswissenschaft, Band 12)*. Frankfurt am Main [etc.]: Peter Lang, 35–50.

Russo, Mariachiara, Claudio Bendazzoli and Annalisa Sandrelli (2006) 'Looking for Lexical Patterns in a Trilingual Corpus of Source and Interpreted Speeches: Extended Analysis of EPIC (European Parliament Interpreting Corpus)', *Forum* 4(1): 221–54.

Sandrelli, Annalisa (2010) 'Corpus-Based Interpreting Studies and Interpreter Training: A Modest Proposal', in Lew N. Zybatow (ed.) *Translationswissenschaft*

— *Stand und Perspektiven. Innsbrucker Ringvorlesungen zur Translationswissenschaft VI (Forum Translationswissenschaft, Band 12).* Frankfurt am Main [etc.]: Peter Lang, 69–90.

Sandrelli, Annalisa and Claudio Bendazzoli (2005) 'Lexical Patterns in Simultaneous Interpreting: a Preliminary Investigation of EPIC (European Parliament Interpreting Corpus)', *Proceedings from the Corpus Linguistics Conference Series*, 1(1), ISSN 1747–9398. Available online at: http://www.corpus.bham.ac.uk/PCLC/

— (2006) 'Tagging a Corpus of Interpreted Speeches: The European Parliament Interpreting Corpus (EPIC)', in *Proceedings of the LREC 2006 Conference, Genova, Magazzini del Cotone 24–26 May 2006*. Genova: ELRA. Available online at: http://hnk.ffzg.hr/bibl/lrec2006/

Sandrelli, Annalisa, Mariachiara Russo and Claudio Bendazzoli (2007) *The Impact of Topic, Mode and Speed of Delivery on the Interpreter's Performance: A Corpus-Based Quality Evaluation*. Poster presented at the *Critical Link 5 Conference, Quality in Interpreting: A Shared Responsibility*. Parramatta – Sydney (Australia).

Schmid, Helmut (1994) 'Probabilistic Part-of-speech Tagging using decision Trees', paper presented at the *International Conference on New Methods in Language Processing*, 1994, Manchester, UK. Available online at: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf

Shlesinger, Miriam (1998) 'Corpus-Based Interpreting Studies as an Offshoot of Corpus-Based Translation Studies', *Meta* 43(4): 486–93.

Shriberg, Elizabeth Ellen (1994) 'Preliminaries to a Theory of Speech Disfluencies'. Unpublished PhD thesis, Berkeley, University of California.

Tissi, Benedetta (2000) 'Silent Pauses and Disfluencies in Simultaneous Interpretation: A Descriptive Analysis', *The Interpreters' Newsletter* 10: 103–27.

Van Besien, Fred and Chris Meuleman, (2004) 'Dealing with Speakers' Errors and Speakers' Repairs in Simultaneous Interpretation. A Corpus-Based Study', *The Translator* 10(1): 59–81.

Web References

EbS (Europe by Satellite): http://ec.europa.eu/avservices/ebs/schedule.cfm (accessed on 1 December 2010).

EPIC interface: http://sslmitdev-online.sslmit.unibo.it/corpora/corporaproject.php?path=E.P.I.C. (accessed on 1 December 2010).

European Parliament: http://www.europarl.europa.eu/ (accessed on 1 December 2010).

FreeLing: http://nlp.lsi.upc.edu/freeling/ (accessed on 1 December 2010).

IMS Corpus Work Bench: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/ (accessed on 1 December 2010).

TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (accessed on 1 December 2010).

Index