USING CORPORA IN CONTRASTIVE AND TRANSLATION STUDIES

Edited by Richard Xiao

Using Corpora in Contrastive and Translation Studies

Using Corpora in Contrastive and Translation Studies

Edited by

Richard Xiao

CAMBRIDGE SCHOLARS
PUBLISHING

Using Corpora in Contrastive and Translation Studies, Edited by Richard Xiao

This book first published 2010

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

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

ISBN (10): 1-4438-1755-4, ISBN (13): 978-1-4438-1755-4

CONTENTS

PART II: PARALLEL CORPUS DEVELOPMENT AND BILINGUAL LEXICOGRAPHY

PART III: CORPUS-BASED CONTRASTIVE STUDIES

ACKNOWLEDGEMENTS

This book is the result of collaborative efforts. A number of people have offered their generous help and support in bringing this book into a reality.

I would first like to take this opportunity to express my gratitude to all conference delegates, especially our keynote speakers Michael Barlow, Silvia Bernardini, Defeng Li, Hongwu Qin and Kefei Wang, as well as our presenting authors for their willingness to share their research outcomes at the 2008 conference of the international symposium Using Corpora in Contrastive and Translation Studies (UCCTS).

I am also grateful to the local organizing committee at Zhejiang University in China for their unfailing support, without which the UCCTS 2008 conference would not have become possible, and to our local postgraduate volunteers for their hard work, which has contributed greatly to ensure the success of the conference and to make the event more enjoyable.

The authors of the chapters included in this book deserve my thanks for their efforts in revising their manuscripts in line with reviewers' comments and suggestions after the conference. Many of them have incorporated new research outcomes in their updated contributions.

My special thanks go to the China National Foundation of Social Sciences for supporting the UCCTS 2008 conference on my project (grant reference 07BYY011), and also to John Zhenghua Ling for reading and commenting on the manuscript.

Last but not least, I thank my wife Lyn Zhang and our daughter Yina Xiao for their love and support, and also for their understanding when I could only spend very little time with them while I was working on this book.

Thank you!

—Richard
October 2009

CHAPTER ONE

INTRODUCTION

RICHARD XIAO

1. Introduction

This book contains a selection of papers from the 2008 conference of *Using Corpora in Contrastive and Translation Studies* (UCCTS), an international conference series launched to provide an international forum for the exploration of the theoretical and practical issues pertaining to the creation and use of corpora in contrastive and translation studies. The UCCTS 2008 conference, which took place at Zhejiang University in Hangzhou, China on 25-27 September 2008, is related, but not restricted to the following themes:

- Design and development of comparable and parallel corpora
- Processing of multilingual corpora
- Using corpora in translation studies and teaching
- Using corpora in cross-linguistic contrast
- Corpus-based comparative research of source native language, translated language and target native language
- Corpus-based research of interface between contrastive and translation studies

We have had the honour and pleasure of welcoming about 60 delegates from thirty-eight institutions in fourteen countries and regions. The languages covered in the papers presented at the conference range from English and Chinese to French, German, Dutch, Italian, Portuguese, Arabic, Persian, Japanese, Uyghur as well as Tok Pisin.

This book includes twenty-three peer-reviewed papers originally presented at the conference, which reflect the state of the art in using corpora in contrastive and translation studies in the international research community. The papers are grouped into three parts: Corpus-based Translation Studies, Parallel Corpus Development and Bilingual Lexicography, and Corpus-based Contrastive Studies.

2. Corpus-based Translation Studies

The first part of this book comprises ten papers focusing on corpus-based translation studies. The contribution by Marco Rocha presents a study on the translation into Portuguese of anaphoric demonstrative pronoun *this* on the basis of an English-Portuguese parallel corpus composed of literary texts and international law texts, as well as technical and scientific materials. The study attempts to show that the anaphoric demonstrative *this* plays a particularly important role in the definition of patterns for textual semantics in English, and that corresponding patterns in Portuguese branch out over a range of forms which may be predicted using the analytical framework proposed. It is expected that the definition of such patterns, which is possibly extensible to other Romance languages, will contribute to a better understanding of textual aspects in translation and also provide subsidies for the improvement of machine translation systems.

The chapter by Yan Ding, Dirk Noël and Hans-Georg Wolf is based on both parallel and monolingual corpus resources, in an attempt to establish the patterns in metaphor translation via a case study of the translation of metaphors related to *fear* between English and Chinese.

Yun Xia and Defeng Li report on the construction and use of specialized corpora in a comparative study of advertisements translated from Chinese into English and their counterparts in native English, aiming to show whether and how specialized comparable corpora can be used to inform pragmatic translation. The results show that differences between the translated texts and their comparable native texts in terms of informative and vocative functions are manifested in aspects like informativity, point of view, and general style. On basis of this study, the authors suggest that knowledge of how to compile and use corpora is an essential part of translational competence. Since pragmatic translation generally requires the translated texts to conform to the target culture in order to achieve the same communicative functions as the source texts, resources of this type will prove helpful in actual translation.

The contribution by Jun Miao and André Salem focuses on the most obvious of the translator's intervention in the translation process: the paratext like the footnotes added by the translator. Their study is based on a French-Chinese parallel corpus composed of the ten-volume complete work *Jean-Christophe* by Romain Rolland and its Chinese translation by Fu Lei, one of the greatest and most influential translators in China. A lexicometrical analysis of the translator's notes reveals that, contrary to the common belief that the translator's note is predominantly a medium of overcoming problems of untranslatability, Fu Lei uses the notes to achieve his declared goal of introducing his Chinese readers to Western culture, in addition to sharing with readers his views on history and, more generally, on mankind as a whole.

While the first three studies are concerned with translation, the chapter by Ernest Gao attempts to address the research question of how sufficient coherence is achieved in simultaneous interpreting (SI). Working within the theoretical framework of Idealized Cognitive Model (ICM), the study presents a quantitative analysis of SI coherence via coherence clues on the basis of a small corpus composed of transcripts of English-to-Chinese simultaneous interpreting.

The next four chapters are concerned with "translation universals", namely the linguistic features that are hypothesized to characterize all translations, which make the translated language different from the original target language. First, Gert De Sutter and Marc Van de Velde attempt to answer the question whether the underlying principles that guide language users to choose between different linguistic options differ between original and translated language, via an investigation of the factors governing PP (preposition phrase) extraposition in original and translated Dutch and German. Their investigation is based on a balanced corpus of literary Dutch and German containing excerpts of ten Dutch and ten German post-war literary works and their respective translations. The results show that typical translation phenomena, such as source language influence and normalization, also influence the subtle language-internal mechanisms that govern syntactic variation.

While there appears to be a consensus that translated language is different from native language, it is debatable whether the features uncovered on the basis of translational English and closely related European languages can be generalized to other translated languages. If such features are to be generalized as "translation universals", evidence from "genetically" distinct language pairs such as English and Chinese is clearly of critical importance. This part of the book includes two papers that investigate translated Chinese. Kefei Wang and Hongwu Qin base their study on a sizable bi-directional parallel corpus of English and Chinese, finding that 1) contrary to what is expected, translational Chinese has a higher type-token ratio and uses relatively longer sentence segments; 2) there is difference between original Chinese and translational Chinese in terms of distribution of word classes, that is, more function words and fewer content words are used in translated Chinese; 3) translational Chinese tends to exaggerate the compositional potentiality of some words or morphemes, which results in its significantly more frequent use of specific lexical bundles. It should be noted that some of these features of translated Chinese are contradictory to translation universal hypotheses.

While a parallel corpus approach is adopted in Wang and Qin's study, Richard Xiao, Lianzhen He and Ming Yue take a comparable corpus approach to the same issue. Their contribution introduces the newly created ZJU Corpus of Translational Chinese (ZCTC), which is a one-million-word balanced corpus created by following the same design of the Lancaster Corpus of Mandarin Chinese (LCMC, representing native Chinese), with the explicit aim of uncovering the potential features of translational Chinese. A comparison of the two monolingual comparable corpora of native and translated Chinese suggests that the core patterns of lexical features that Laviosa (1998) observes in translational English are generally also applicable in translated Chinese. The results also indicate that while mean sentence length may not be reliable as an indicator of simplification, the explicitation hypothesis is supported by the Chinese data. However, the normalization hypothesis is not supported as it may be specific to particular language pairs.

The chapter by Federico Gaspari and Silvia Bernardini illustrates an innovative methodology for investigating translational language by comparing the salient features of two interfacing forms of mediated discourse, namely non-native and translated language, focusing on the language pair English-Italian within the framework of translation universals. The proposed approach differs from classic approaches in two main ways. On the one hand, non-native written data are added to the equation while on the other hand the focus is restricted to a specific language pair, which is analysed in both directions. Their preliminary results appear to indicate that some of the features observed in translated language, and usually explained in terms of *translation* universals, are in fact also present in non-native production (cf. Granger's (1996: 48) observation of similarity between "translationese" and "learnerese"), suggesting that an extended notion of *mediation* universals might better capture the actual import of these shared phenomena.

The final chapter in the first part is contributed by Defeng Li and Chunling Zhang, which critically examines the progress, problems and prospects of corpus-assisted translation research in China. This examination has three aims: to identify major issues that have / have not been dealt with in corpus-assisted translation research (in China), to examine the research methods and designs employed in such studies, and to suggest future directions for corpus-assisted research in China. Though the recommendations are made with specific reference to English-Chinese translation and particularly in the Chinese context, they should also have implications for corpus-assisted translation research in general in other parts of the world.

3. Parallel corpus development and bilingual lexicography

The second part of this book includes five chapters that are concerned with parallel corpus development and bilingual lexicography. As is well known, parallel corpora are valuable resources in both translation research and bilingual lexicography. However, for a parallel corpus to be useful, it must be aligned at a certain level, for example, at document, paragraph, sentence, phrase, or word level, depending on the specific research questions or purposes and the ease with which alignment can be automated reliably. The first chapter in this part, which is contributed by Kim Gerdes, copes with alignment. His approach goes back to purely statistical distribution measures and tries to optimize them while aiming at simplifying the accessibility of the underlying system. Such a goal is less ambitious than many previous approaches as it only aims at an alignment on the paragraph level, but it is more ambitious as it is completely language independent and resource free because it only relies on basic common features that all bilingual texts have in common: similar distributions of many of the words. The free online tool that Gerdes has developed can give easy access to alignment even for distant and rarely considered language pairs.

The next two chapters introduce two new parallel corpora of a rarity. Samat Mamitimin and Umar Dawut present their project that develops a parallel corpus of Chinese and Uyghur, a Turkic language of Altaic family spoken by the Uyghur people in the Xinjiang Uyghur Autonomous Region of China. The authors elaborate on the corpus design, data collection, annotation and markup of the parallel texts as well as sentence alignment, which is followed by a discussion of corpus development tools and preliminary results. The new corpus represents great progress in developing linguistic resources of minority languages in China.

The contribution by Chong Zhu describes his project designing and developing a parallel corpus based on media subtitles. Multimedia translation, especially subtitle translation, has not received due attention, possibly because of an academic bias against the importance of media translation coupled with a lack of available parallel corpus resources for this kind of research. In this chapter, the author illustrates a way of building a parallel subtitle corpus, i.e. the Multi-Media Subtitle Corpus (MMSC), on the basis of his own experience, which is followed by a discussion of potential applications of such a corpus in translation studies.

The last two papers in the second part of this book are concerned with bilingual lexicography. While bilingual lexicons are useful resources for many purposes, building a dictionary by hand can be prohibitively time consuming and labour intensive. Unsurprisingly, researchers have been experimenting with automatic construction of bilingual dictionaries. The contribution by Sumithra Velupillai, Martin Hassel and Hercules Dalianis describes their three experiments in an attempt to improve automatic bilingual dictionary construction from unstructured corpora. The first experiment aims at creating parallel corpora and bilingual dictionaries from a multilingual website; the second tries to map a text in one language to a corresponding text in another on the basis of the frequency distribution of word initial letters; the third uses a memory-based machine learning technique with simple frequency features such as word, sentence and paragraph frequencies. These experiments have shown very promising results, but they need to be further developed and evaluated.

The final chapter in this part, which is contributed by Adriano Ferraresi, Silvia Bernardini, Giovanni Picci and Marco Baroni, describes two huge (ca. two billion words) "web corpora" of English and French, which are applied on a pilot experiment to a bilingual lexicography task focusing on collocation extraction and translation. The study has two purposes. The first is to provide qualitative evaluation of the Web corpora themselves, particularly the French one, for which no previous evaluation has been conducted. The second purpose is to ascertain whether corpora built automatically from the Web can be profitably applied to lexicographic work, on a par with more costly and carefully-built resources such as the British National Corpus (BNC). The results are very encouraging though, as the authors point out, further work is required to improve such Web corpora to better meet lexicographers' requirements, and to investigate the extent to which such automatically created free resources can be used in dictionary making.

4. Corpus-based contrastive studies

The third part of this book includes eight corpus-based contrastive studies. The contribution by Bart Defrancq addresses the similarities and differences between the *wh*-paradigms in concessive conditional clauses in English, French and Dutch. His study is based on both monolingual corpora to establish the usage and a parallel corpus to investigate what strategies translators develop to deal with the discrepancies between the languages involved.

Jianxin Wang presents a contrastive analysis of connectives in English and Chinese via a case study of *however* and its counterparts in two parallel corpora of the two languages, finding that contrastive relations tend to be expressed implicitly in Chinese but explicitly in English while Chinese contrastive connectors are generally used in sentence initial position but the positions of contrastive connectors can vary in English. The author suggests that such differences should be highlighted and given due attention in the teaching, learning and translation of the two languages.

The contribution by Hui Yin also focuses on English and Chinese, but in this study the author compares the so-called "satellites" in verb complexes (e.g. *out* in *fly out*) in two balanced corpora of the two languages. The results suggest that while English and Chinese are both satellite-framed languages, the satellites in the two languages are quite different in nature, which also have different frequency and distribution patterns.

Guiling Niu and Huaqing Hong present a comparative analysis of repetition patterns of rhetoric features such as sound, word and phrase in a multilingual context. Their study is based on a parallel corpus of English and Chinese advertisements in Singapore print media, in an attempt to investigate how different types of rhetorical figures are used in advertisements in the two languages.

The contribution by Masahiko Nose is a contrastive study of comparative constructions in English, Japanese and Tok Pisin – a creole language spoken in Papua New Guinea with a simple grammar and a lexicon mixed with English and other indigenous languages in New Guinea. This study represents an attempt to clarify the functional characteristics of Tok Pisin grammar by contrasting Tok Pisin with English and Japanese on the basis of the material in the biblical text *New Testament*.

Contrastive studies are intrinsically related with translation research on the one hand and with language teaching on the other. While "contrastive analysis" in a narrow sense is confined to cross-linguistic contrast, the term can be used in a broad sense to include contrastive studies of different languages as well as what Granger (1996, 1998) calls "contrastive interlanguage analysis" (CIA), which is also mainly concerned with comparison, for example, comparing the learner's interlanguage with the target native language, and comparing different interlanguages in terms of L1 background, age, proficiency level, task type, learning setting, and medium etc. (see Xiao 2007). The last two papers in this part are contrastive studies of the latter type. Carmen Dayrell's contribution investigates, on the basis of monolingual comparable corpora, the frequency and collocational behaviour of five sets of sense-related verbs in English scientific abstracts written by Brazilian graduate students, with the aim of uncovering potential differences between English abstracts written by Brazilian students and published abstracts of the same disciplines. The results reveal some differences in both frequency percentages of individual verbs and recurrent lexical patterning, which appear to be a result of L1 interference from Portuguese.

Similarly, the last chapter in this part is also concerned with English learners' academic writing, but the L1 background of such learners is Chinese. In this chapter, Yuechun Jiang and Zhiqing Hu present a contrastive study of reporting in English MA dissertations, which is based on a corpus composed of a random collection of English MA theses written by Chinese learners of English and native speakers of English, with each component amounting to 100,000 words. The results indicate that there are considerable similarities of usage as well as remarkable differences in the use of reporting between the two groups of learners.

I hope that readers will enjoy the latest developments in corpus-based contrastive and translated studies presented in this volume!

References

Granger, S. (1996), "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora", in K. Aijmer, B. Altenberg and M. Johansson (eds.) *Language in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund, March 1994*, 37-51. Lund: Lund University Press.

Granger, S. (1998), "The computer learner corpus: A versatile new source of data for SLA research", in S. Granger (ed.) *Learner English on Computer*, 3-18. London: Longman.

Laviosa, S. (1998), "Core patterns of lexical use in a comparable corpus of English narrative prose". *Meta* 43(4): 557-570.

Xiao, R. (2007), "What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English". *Indonesian Journal of English Language Teaching* 3(2): 1-19.

PART I

CORPUS-BASED TRANSLATION STUDIES

CHAPTER TWO

TRANSLATING ANAPHORIC *THIS* INTO PORTUGUESE: A CORPUS-BASED STUDY

MARCO ROCHA

1. Introduction

Since its revival in the 1980s, corpus-based research has grown quickly in size and sophistication. Corpus linguistics has become a fully developed approach to the study of languages with a well defined methodology. Once researchers realized that corpora need not be monolingual, the approach was expanded to parallel corpora, relying on similar techniques for search and retrieval. Gradually, corpus-based translation studies developed into what has now become a research paradigm in the field of translation studies (Tymoczko 1998, Olohan 2004).

This chapter presents initial results of cross-linguistic investigations on anaphoric demonstratives. It concentrates on the anaphoric usage of a single English-language demonstrative, *this*, and its renderings into Portuguese retrieved from a parallel English-Portuguese sample collected for the purposes of the present research, including three sources: COMPARA (2001), which contains literary texts and their translations; international law documents from the Organization of American States (OAS); and medical texts from the European Medicines Agency (EMEA). The remainder of the chapter is organized as follows: the second section briefly reviews the related works; the third section describes the methodology; the fourth section presents and discusses results; the fifth section summarizes implications of findings to textual semantics and machine translation, pointing out future developments.

2. The analytical approach

This short review of related works describes how the analytical approach adopted in the study was established. Three types of research are included in the review: work on anaphora; work on textual semantics; and work on corpus-based cross-linguistic research. An exhaustive review of literature on any one of these areas would be far beyond the scope of this chapter. Anaphoric phenomena have been studied within a wide variety of distinct frameworks, to such an extent that the very meaning of the term anaphora is a contentious issue. Research on human language technology added a great deal of new material to the literature on anaphora. Contrastively, there is a dearth of research on anaphoric demonstratives, perhaps because of the inextricably textual nature of this form of anaphor.

Cohesion in English (Halliday and Hasan 1976) is the well-known seminal work which has inspired a large amount of research on cohesion in texts. The authors analyze in detail the relationships – named cohesion ties – existing between lexical items in an instance of discourse. The concept of anaphora in much subsequent work is related to the notion of cohesion tie. The importance of referential chains was also demonstrated, showing how the repeated reference to a certain entity, by means of various linguistic devices, contributed to textual organization. Conversely, the phenomena of pronominalization and ellipsis could be understood more satisfactorily when approached with textual aspects in mind.

Halliday and Hasan divide cohesion ties into five classes, namely: conjunction, reference, substitution, ellipsis, and lexical cohesion. Of those, conjunction is the only one not included in the expanded concept of anaphora mentioned above. The lexical items covered by the category – such as *however*, *on the other hand* and *notwithstanding* – signal semantic relations between clauses or sentences they connect. These relations are an integral part of the way texts are organized, but they are not adequately characterized as anaphoric relations. The notion of an antecedent which must be identified for semantic interpretation does not explain the function of these items clearly. There is a degree of fuzziness in boundaries between the classes which is explicitly acknowledged by the authors, but it seems adequate to leave conjunctions out of the anaphora "world".

Botley (2000) designed an annotation scheme to analyze English anaphoric demonstratives. Each case in three different corpora was analyzed according to five features, namely: recoverability of antecedent, direction of reference, phoric type, syntactic function, and antecedent type. Possible values for recoverability of antecedent were: directly recoverable, indirectly recoverable, non-recoverable or not applicable, e.g., exophora.

The set of categories for direction of reference include anaphoric, cataphoric, and not applicable, that is, deictic or exophoric. Possible phoric types were referential, substitutional or not applicable. Syntactic function was classified as noun modifier, noun head or not applicable. Finally, demonstratives can be assigned to five distinct categories regarding antecedent type. These include nominal antecedent, propositional/factual antecedent, clausal antecedent, adjectival antecedent, and no antecedent.

Botley's approach bears many similarities to the one adopted in this study. Anaphoric demonstratives in a corpus have also been classified according to a set of properties thought to be relevant for modelling anaphoric relations. The features named recoverability of antecedent, direction of reference and antecedent type are closely mirrored by properties used to classify cases of anaphora in the analyzed sample of English originals. Different from the approach used here, Botley includes cases in which the demonstrative anaphor is a determiner, whereas this study does not consider these as anaphoric demonstratives.

Halliday and Matthiessen (2004) define semantics within the systemic-functional approach as one of the four strata in terms of which language is analyzed, the other three being context, lexico-grammar, and phonology-graphology. Within the approach, semantics subsumes aspects of what is usually called pragmatics, whereas some others are encompassed by the context stratus. Semantics is divided into three components: ideational semantics, concerning the propositional content; interpersonal semantics, accounting for exchange structure and expressions of attitude; and textual semantics, which deals with the way a text is structured as a message. This involves aspects of textual organization such as theme structure, given/new, rhetorical structure, and also cohesive devices, among which anaphoric relations are of primary importance.

In the present study, this approach to the understanding of text led to a concern with the different forms in which the anaphor *this* is used for referring. In particular, an effort is made to understand better the interaction between strata that allows adequate interpretation of references. Thus, the lexico-grammatical aspects are used as a starting point by including the property named grammatical function in the analytical approach. Moreover, the analysis tries to establish whether it is possible to use information on collocations in which the anaphor appears in order to improve understanding of how anaphoric references are integrated into semantic interpretation.

Dyvik (1998) discusses the possibility of using corpus-based approaches to translation studies as a basis for the study of semantics as such, thus not restricted to the actual analysis of translational phenomena. The author points out that translation data, as organized in a parallel translation corpus, may provide access to a “desirable multilingual perspective” for the study of semantics, which is mostly monolingual. Moreover, the sort of cross-linguistic meaning evaluation between expressions needed for translation is carried out as “a normal kind of linguistic activity”, instead of one based on theoretical analysis. These evaluations are externalized in “observable relations between texts”, thus contributing to “strengthen the empirical foundations of linguistics.”

By repeatedly creating and inverting mirror images of possible translations across two given languages – English and Norwegian – Dyvik is able to derive semantic representations for signs in each language, relying on sets of features assigned to the individual senses of a sign on the basis of these mirror images. The approach inspires the attempt in the present study to establish patterns of textual semantics by exploring the various possible translations of *this* into Portuguese, so as to explore the textual role of signs used to express the sort of referring carried out in English by the sign *this*. More specifically, it is expected that this exploration into translational textual semantics may uncover useful patterns for machine translation systems.

Santos (1998) analyzes perception verbs in English and Portuguese. The material includes texts originally written in both languages and their translations. Santos presents her material by first discussing their properties in each of the two languages separately. The study then proceeds to describe a number of translation pairs in detail, with particular attention to the translation of Portuguese *imperfeito* and *perfeito* tenses into English. Syntactic features, such as the presence of objects or verbal complements, are explored along with semantic features, like negation and habituality, searching for clues which might explain variation in choices made by translators.

Santos builds a picture of perception verbs in English and Portuguese by means of a detailed analysis of translation pairs, which points towards substantially different systems for the expression of perception in these two languages. Likewise, the present study examines translation pairs in search of patterns that may explain why cohesion devices differ in regard to the use of anaphora. Different from Santos, the approach used here focuses on tokens of a single anaphoric demonstrative of the English language, in texts originally written in English, in order to compare translation pairs. No analysis of Portuguese originals is carried out.

3. Methodology

Firstly, *this* tokens in a concordance extracted from the corpus were analyzed to select cases of anaphoric *this*, removing determiner tokens. COMPARA contained 361,852 English words when the sample was retrieved. The full concordance of *this*, as informed automatically, contained 1,033 tokens of *this* (0.28% of the corpus), out of which 1,000 were randomly selected by the query-handling interface in the COMPARA site. After removing determiners, 171 tokens remained, a proportion of 16.55%. There are 95,052 English words in the OAS corpus. There are 413 tokens of *this* in the corpus (0.43%), of which 43 tokens are anaphoric *this* (10.41%). In contrast, there are 30,580,774 English words in the EMEA corpus, but only 13,828,388 tokens in Portuguese, according to information in the OPUS site. The downloaded parallel Portuguese-English file contains 885,103 alignment units and 26,780,657 tokens in both languages. One would guess that there are 12,952,269 tokens in the English language, assuming the full Portuguese corpus was used in the alignment process. It is thus a large corpus with 35,374 tokens of *this*, roughly 0.27% of the guessed total of tokens. Averaging proportions of anaphoric *this* in the two other corpora (13.48%), there must be approximately 4,768 tokens of anaphoric *this* in the EMEA corpus. A random sample of 46 tokens of anaphoric *this* from the EMEA corpus was added to the sample, amounting to a total of 260 tokens analyzed according to the four properties described below.

3.1. Grammatical function

This property classifies each token of *this* into four standard grammatical categories related to their function in the sentence. Basically, these categories are subject, verb object and prepositional object, but the subject category was split into lexical verb subject and copular verb subject, as it has been perceived that the distinction was relevant for translational patterns. Each category is detailed below.

3.1.1. Lexical verb subject

The label is assigned to cases to which the standard notions of subject and lexical verb would apply (see Greenbaum 1996). One example is given below. The anaphoric demonstrative token is shown in bold.

(1) **This** includes life imprisonment if a young person is convicted of an offence for which an adult would get a life sentence.

As straightforward as the classification may seem, a few cases posed problems for the classification. Consider example (2) below.[1]

(2) "I don't know what I want," Alistair said drearily, "I just don't want all **this** to go on."

The anaphor is the head of a phrase which is the object of *want* and the subject of a non-finite clause *to go on*. Such tokens were classified as verb objects, not lexical verb subjects. The decision is, to a certain extent, arbitrary, but sets the function in relation to the main clause as a standard for double-function cases of the kind.

3.1.2. Verb object

This category applies the notion of transitivity as it usually appears in grammars of the English language. Thus, cases of pronominal *this* which function as direct or indirect objects of transitive verbs, as in example (3) below, are assigned this value.

(3) The plaintiff does not have to prove **this** "beyond a reasonable doubt", as in a criminal case.

Subjects of passive constructions were also classified as verb objects in the approach used in this study.

3.1.3. Prepositional object

This category is used to classify tokens of *this* within prepositional phrases, typically following immediately the preposition which is the head of the phrase, as in example (4) below (see Greenbaum 1996: 282).

(4) "Does Humphrey know about **this**?" I asked.

3.1.4. Copular verb subject

Tokens of anaphoric *this* as subjects of copular verbs were grouped in a separate category, since translational patterns are distinct in a way that is relevant to the study. One example is given below.

(5) But **this** was such a wonderfully small sigh, that she wouldn´t have heard it at all, if it hadn´t come quite close to her ear.

The distinction between subject and subject predicative was not seen as useful for the purposes of this investigation. In example (6) below, *this* can be classified interchangeably either as subject of an inverted construction or subject predicative. Grouping simplified the organization of data in tables.

(6) I fancy that the true explanation is **this**: It often happens that the real tragedies of life occur in such an inartistic manner that they hurt us by their crude violence, their absolute incoherence, their absurd want of meaning, their entire lack of style.

3.2. Reference type

Tokens of anaphoric *this* in the sample were also classified according to the way they relate to their antecedents, essentially regarding direction. Possible values were thus anaphoric, in strict sense, cataphoric and deictic. A number of classification difficulties arise in the process of assigning specific tokens to one of those categories. These will be discussed in detail further below. Firstly, the typical cases are presented.

3.2.1. Anaphoric reference

Anaphora, as understood in this study, is a textual relationship in which a given phrase – called the **anaphor** – depends on the identification of a typically preceding element in the text – called the **antecedent** – for semantic interpretation. Intuitively, one would expect this antecedent to be plainly visible in the text and not very distant from the anaphor, so as not to complicate identification, although years of investigation by the scientific community, particularly in the field of human language technology, have shown that this is often not the case, especially regarding demonstratives. Nonetheless, many typical cases of anaphora in the sense defined above do occur. One example is given below.

(7) The transmittal of documents shall in each case be subject to the decision of the Commission, which shall withhold the name and identity of the petitioner, if the latter has not authorized that **this** be revealed.

In the example (7), there is a specific noun phrase (*name and identity of the petitioner*) referred to by the anaphor. However, anaphoric demonstratives refer to textual antecedents very often, if compared to other forms of anaphor, and precise delimitation of the antecedent may be much more difficult, as in example (8) below.

(8) They heard him go banging cheerfully along the passage singing the theme tune for the World Cup, and **this** made them smile at one another, and the smiling suddenly made Lizzie feel rather vulnerable.

In the example (8), *this* may be said to refer to *hearing him go…tune for the World Cup*, or to *the fact that they heard him go…tune for the World Cup*, none of which is literally in the text. There is no principled way of deciding which is best, and whether this is an explicit or implicit antecedent. As a result, a degree of arbitrariness is unavoidable.

3.2.2. Cataphoric reference

The term cataphora is understood, for the purposes of this study, as reference to an element of the text which is still to be read. Example (9) illustrates tokens thus classified.

(9) "And know **this**, Berekiah Zarco - if you attempt to remove me from my home you will never find your uncle's murderer!"

The distinction between anaphoric and cataphoric references is not always as straightforward though. One token which blurs the dividing line is shown in example (10).

(10) She might have died the first death, of loss, but she would never, ever - and **this** she promised herself - die the second death, of forgetting.

Since the anaphor appears in a sentence between dashes, the textual antecedent is a discourse chunk which begins before the anaphor and is concluded subsequently. Tokens of this kind were classified as cataphoric, since the antecedent cannot be precisely identified before the chunk is fully read. Once again, there is a degree of arbitrariness. Anaphoric *this* referring to a discourse chunk may also require information given subsequently for the identification of this discourse chunk, as in example (11).

(11) When I announced that the lines about abortion had been cut from this week's script, she said, "Oh, good," and although she saw from my expression that **this** was the wrong response, she typically proceeded to defend it, saying that The People Next Door was too light-hearted a show to accommodate such a heavy subject - exactly Ollie's argument.

The actual antecedent of the anaphor is the utterance *Oh, good* preceding the demonstrative. However, the anaphoric noun phrase *the wrong response*, subject predicative in the copular sentence, plays a crucial role in the identification of the antecedent for the demonstrative, since the fact that *this* refers to a response is decisive in the antecedent identification process. It may be said that an implicit noun phrase head *response* is "revealed" by the subject predicative, forming a referring chain *oh, good*>*this*(response)>*the wrong response*. Still in processing terms, it may be argued that *this* refers cataphorically to *the wrong response*, using essentially syntactic information derived from the copular structure. Both anaphors would then refer back to the discourse chunk *Oh, good*. If this interpretation is accepted, then the anaphor might be classified as cataphoric. On the other hand, the phrase *Oh, good* precedes the anaphor in the text, a fact that requires no surmising on processing. This study classifies such cases as anaphoric.

3.2.3. Deictic reference

Deictic references are typically associated with pronominal demonstratives. These are references in which the antecedent can only be adequately identified in the situational context, and not in the text itself. OAS and EMEA documents contain no deictic references, as expected. Literary texts such as those held in COMPARA must provide the reader with all necessary information for interpreting the textual semantics involved, which precludes any identification of antecedents on the basis of situational data. However, characters move in a fictional setting which is revealed in the text by various means. Therefore, deictic references do occur if characters are seen as the processors of anaphoric relations. A definition of standards used for assigning cases to the deictic type is thus needed. One example is given below.

(12) Zoe held out the box. "Shall I heat **this**?"

For the characters, this is a case of deixis. However, it is also clear that the noun phrase *the box*, preceding Zoe's utterance, allows the reader to interpret the deictic reference by means of an antecedent explicitly introduced in the preceding text. A degree of inference is needed for the reader to understand the reference fully. On the other hand, if the information available to characters is the classification standard, a physical object in the environment where the dialogue occurs suffices. The matter is made more complex by tokens such as (13) below.

(13) "And **this** is Zoe."

This type of occurrence does not require any information provided by the narrator in order to be understood by the reader, since it amounts to a standardized pattern of deictic reference for introducing people and is readily decoded as *this* ***person*** *is Zoe*. Of course this is only possible because Zoe is known to be a person on the basis of visual situational information. Another relevant example is shown below.

(14) There was an air of discreet excitement in the room at the sight of the food, unfashionable, childlike, teatime food, resting on mats of decoratively pierced white paper which Dilys had brought down to Tideswell and made plain she expected

to be used. Judy, struggling to make the pecan squares and chocolate brownies that had been so much part of Caro's American repertoire, had said defiantly that her mother never used doilies. "But **this** is a funeral," Dilys said.

Example (14) also conveys a standardized form of deictic reference which points to the situation as a whole as antecedent. Processing for adequate semantic interpretation involves decodifying the utterance as *this event in which we are involved is a funeral*. Mention of a specific referent by the narrator is not required. Semantics in the preceding text conveys the idea of described situation, allowing adequate identification of the implicit antecedent, which is conclusively reinforced by the word *funeral*, a type of event. Example (15) below may also be classified within the same sort of processing strategy, a combination of linguistic patterns with situation description and lexical clues.

(15) "Is Professor Hogan somewhere? Or Mrs Hogan?" "Everybody gone home." "But **this** is their home," Philip protested.

The reference is also understood arguably without need of a specific physical object visible in the situation. The notion of *place*, associated with situations in general, is mostly available for reference at any time. Although the reference is deictic from the point of view of the characters, the reader does not need specific mention of a referent, since both preceding (*somewhere*, *home*) and subsequent (*their home*) texts provide a sufficient basis for the *this place* interpretation. It is therefore a combination of a linguistic pattern, in which the interpretation of an anaphoric token of *this* is potentially a deictic reference to the place where participants in a dialogue are, lexical clues pointing to a *this place* decoding, in particular the word *home*, and an unspecified but to a certain extent physically visible object in the situation.

Examples (12) - (15) were all classified as deictic. However, structurally similar tokens which draw exclusively on textual clues for the identification of the antecedent, for readers as well as characters, were not classified as deictic, since the situation, in the physical sense of the word, does not play a noticeable role in the processing required. Example (16) below was thus classified as cataphoric reference.

(16) So (and **this** is my conclusion) I am resigned to living as I have lived: alone, with my throng of great men as my only cronies - a bear, with my bear-rug for company.

3.3. Antecedent type

The third variable is dichotomic. Antecedents were classified as either explicit or implicit. The first category grouped those antecedents which were visible in the text. Antecedents that demanded some form of inference out of textual information for their identification were assigned to the second category. Example (17) is a token of anaphoric *this* with an explicit antecedent, whereas antecedent identification in (18) requires the sort of inference seen as characteristic of implicit antecedents.

(17) Aila knew that. He didn't keep anything from her. She knew some of the parents had complained about his having marched with the children over the veld to the blacks' school: a teacher should not be allowed to encourage such things. She knew that when the principal informed him of **this** it was a warning.

(18) She was startled. "Goodbye? Why, won't we be seeing each other again?" "Oh, we'll *see* each other," William said, "of course we will, but **this** is -- the end of this bit."

This variable caused a great deal of agony to the analyst, since it may be hard to ascertain whether the antecedent is implicit. Thus, in example (17), although the antecedent may be safely said to be visible in the text, the actual specification of the precise discourse chunk playing the role of antecedent is not straightforward. It is thought, however, that the classification of anaphoric *this* tokens according to the dichotomy defined above may prove useful for the analysis of translation choices and bring valuable insights into textual semantics.

3.4. Antecedent phrase structure

This property is also dichotomic, the two possible values assigned to cases being either **textual** or **nominal**. The latter classifies antecedents (example 19) that are "classical" cases of anaphora, in the sense that there is a preceding noun phrase antecedent for the anaphor. On the other hand, anaphors referring to discourse chunks had their antecedents classified as textual. In typical cases for demonstrative anaphors, the discourse chunk is a description subsequently referred to, as exemplified in (20).

(19) Citizens and permanent residents have the constitutional right to live or seek work anywhere in Canada. **This** includes the right to live in one province and work in another.

(20) The number of non-EU ADRs has risen sharply in recent years and **this** is expected to continue in 2002.

In example (19), the noun phrase *constitutional right* is the antecedent of the anaphoric demonstrative. In (20), the full preceding clause is the antecedent. Again, a number of tokens did not fit the distinction easily, such as example (21).

(21) If Morris had been pleased to describe the master of the house as a heartless scoffer, it is because he thought him too much on his guard, and **this** was the easiest way to express his own dissatisfaction -- a dissatisfaction which he had made a point of concealing from the Doctor.

The antecedent for the anaphor is the non-finite clause *to describe… heartless scoffer*, which is a discourse chunk that, at the same time, can function as a noun phrase, since it is a non-finite verb form. One way to solve the difficulty would have been to create a separate category to classify these cases, but this was considered unnecessary. Such antecedents were rare and thus grouped into the textual antecedent category.

3.5. The classification of translations

The aligned sample contained 260 tokens of *this* in English, but 280 tokens of Portuguese renderings, since 20 tokens in COMPARA had two distinct translations. Translation tokens were grouped into ten categories for classification, according to morphological criteria. Broadly, renderings 1 to 5 below reflect the Portuguese system of demonstratives and contractions. Portuguese demonstratives *isto* and *isso* are virtually equivalent in present-day usage. The distinction between the *isso*-group and the *aquilo*-group signals distance of the object referred to, the former being used when referents are near the speaker. It is in many ways similar to the *this*/*that* distinction in English, but not precisely, since there were a few tokens of *this* translated as *aquilo*. It is also true that the distinction in terms of distance is not strict, quite in the same way as in English. The distinction between *isso*/*aquilo* and *esse*/*aquele* has no correspondence in English. Both groups are mirrored as *this*/*that*.

The fifth rendering is typically used as a translation of *one*-anaphora, but it may occur as a rendering of *this* whenever the target text includes a relative clause linked to the anaphoric demonstrative. Patterns will be discussed shortly. Renderings from 6 to 11 are not classified as demonstratives in standard Portuguese grammar textbooks. Such renderings mean that target texts use different forms of anaphoric reference or rephrase the originals so as to make the use of an anaphor unnecessary. Rendering 12 is a relatively rare demonstrative.

1. Tokens translated as *isso*, *isto* and contractions, henceforth *isso*-group
2. Tokens translated as *aquilo* and contractions, henceforth *aquilo*-group
3. Tokens translated as *este*, *esse* and contractions, henceforth *esse*-group
4. Tokens translated as *aquele*, *aquela* and contractions, henceforth *aquele*-group
5. Tokens translated as *o*, *a*, *os*, *as* and contractions, used as demonstrative pronouns (use as articles is common, but not as translations of anaphoric *this*), henceforth *oadem*-group
6. Tokens with no corresponding word in target text, henceforth *omission*
7. Tokens translated as a non-pronominal noun phrase, henceforth NP
8. Tokens translated as *assim*
9. Tokens translated as object pronouns
10. Tokens translated as *aí*
11. Tokens translated as subject pronoun *ele* used as a prepositional object
12. Tokens translated as *tal*

Next section begins by presenting results for the English originals and then moves on to the cross-linguistic analysis.

4. Results and analysis

The presentation of results starts with frequency tables for each of the four properties cross-tabulated by source. Results and analysis of the cross-linguistic data are then presented, along with a discussion of aspects of textual semantics.

4.1. Frequency tables

Table 2-1 below shows frequencies for grammatical functions by source. Percentages are rounded in all tables. Percentages between round brackets along with frequency numbers refer to each source, thus to the totals for each column. The percentages in the last column to the right refer to the full sample.

Table 2-1. Distribution of grammatical functions by source

Category	COMPARA	OAS	EMEA	Total	%
Prep. object	42 (24.6%)	2 (4.65%)	3 (6.52%)	47	18.07
Verb object	57 (32.7%)	9 (20.43%)	8 (17.39%)	74	28.07
Lex. verb sbj.	11 (7.0%)	22 (51.16%)	26 (56.52%)	59	23.07
Cop. verb sbj.	61 (35.7%)	10 (23.25%)	9 (19.56%)	80	30.76
Total	171 (100%)	43 (100%)	46 (100%)	260	100.0

The distribution of grammatical functions for anaphoric *this* in the English originals could be said to be balanced, if totals are considered. Percentages in the last column of Table 2-1 vary from 18.07% to 30.76%, thus close to a 25% even distribution. However, numbers for each source vary widely. Prepositional objects are rare in both OAS and EMEA, but nearly 25% of COMPARA occurrences, which also shows higher percentages for verb objects. As a prepositional object, anaphoric *this* collocates most frequently with *like* (15 tokens, all in COMPARA), *of* and *with* (6 tokens each); *about* is the phrase-head preposition in 5 tokens.

Thus, the phrase *like this* amounts to 31.91% of the tokens classified as prepositional objects and should receive special attention.

Conversely, lexical verb subjects are rare in COMPARA but amount to more than half of tokens in OAS and EMEA. The assertive informational nature of text in law and medical documents seems to favour reference linked to new information expressed by lexical verbs.[2] The presence of colloquial forms in dialogues and reproduced thought of characters in literary texts seems to favour copular structures, the most common in COMPARA. Textual semantics typical of spoken language, such as evaluation (*is **this** a trick?*), descriptions of situational contexts (*Well, **this** is the very queerest shop I ever saw*), and specific elements in this context (*However, **this** was anything but a regular bee*) are often accomplished by means of *this* + *be*-form constructions.

Collocations involving forms of *be* are by far the most common for tokens of the anaphoric demonstrative as a copular verb subject. This holds for all text types. All but one token of *this* in this function (*this seems*) collocate with *be* forms, of which 49 are simple present tense (*is* and *isn't*), and 25 are simple past tense (*was* and *wasn't*). The remaining tokens are modal-plus-*be* forms. As an object, *this* collocates with a variety of verbs, but most notably with forms of *do* (11 tokens), *write* (4 tokens), *stop*, *hear* and *see* (3 tokens each). These collocations account for approximately 40% of the total for verb objects in COMPARA, with *do-this* forms alone adding up to slightly over 14%. Tokens appear predominantly as subjects of passive constructions – classified as verb objects ─ in EMEA (6 out of 8) and OAS (6 out of 9). Table 2-2 shows the distribution of reference types by source.

Table 2-2. Distribution of reference types by source

Category	COMPARA	OAS	EMEA	Total	%
Anaphoric	148 (80.7%)	42 (97.68%)	46 (100.0%)	236	90.77
Cataphoric	9 (7.0%)	1 (2.32%)	0	10	3.85
Deictic	14 (12.3%)	0	0	14	5.38
Total	171 (100.0%)	43 (100.0%)	46 (100.0%)	260	100.0

Anaphoric references are the most frequent type of references, with the other two possible classifications adding up to less than 10% of the total. All cases of deictic references are in COMPARA. This is also true of cataphoric references, except for a single token in OAS. This particular token was found in the speech of the secretary-general included in the corpus, not in an actual law document. Corpus data show that both cataphoric and deictic references are associated with spoken language. Table 2-3 presents results for antecedent explicitness by source.

Table 2-3. Distribution of antecedent explicitness by source

Category	COMPARA	OAS	EMEA	Total	Percent
Explicit	127 (74.27%)	43 (100.0%)	44 (95.65%)	214	82.30
Implicit	44 (25.73%)	0	2 (4.35%)	46	17.70
Total	171 (100.0%)	43 (100.0%)	46 (100.0%)	260	100.0

The distribution is strongly skewed towards explicit antecedents. There are no implicit antecedents in OAS, probably as a result of the need to be transparent in legal documents. There are two cases of implicit antecedent in EMEA, which are worth discussing.

(22) `Read all of this leaflet carefully before you start taking this medicine, even if` **`this`** `is a repeat prescription.`

(23) `Do not shake the vial, as` **`this`** `will cause foaming.`

The anaphor in (22) is a reference to *prescription*, which is not mentioned in the previous text. It is assumed that a person who is about to take the medicine had it prescribed by a doctor. However, the implicit antecedent is made explicit as a subject predicative in the copular clause of which the anaphor is the subject. It might be justifiably argued that this is not truly an implicit antecedent, but a cataphoric reference to the head of the noun phrase which appears in the same clause as a subject predicative. It could also be argued that the anaphor refers to the explicit antecedent *this medicine*. To the analyst's agony, the antecedent was classified as implicit. In (23), the verb in the preceding main clause is nominalized to become the implicit antecedent with the negation ignored. The adjustment was seen as too deep a transformation for classification as an explicit antecedent. Minor syntactic changes were seen as acceptable to classification as an explicit antecedent, but alterations of semantic content were not. The distribution of antecedent phrase structure by source is presented in Table 2-4.

Table 2-4. Distribution of antecedent phrase structures by source

Category	COMPARA	OAS	EMEA	Total	Percent
Nominal	50 (29.24%)	10 (23.25%)	15 (32.61%)	75	28.85
Textual	121 (70.76%)	33 (76.75%)	31 (67.39%)	185	71.15
Total	171 (100.0%)	43 (100.0%)	46 (100.0%)	260	100.00

The second dichotomic property in the model shows that the anaphoric demonstrative *this* has predominantly textual antecedents, different from personal pronouns and the "classical" concept of anaphoric reference. Proportionally, the tendency is stronger in the OAS data and weaker in EMEA, but percentages do not stray significantly from total percentages. As pointed out before, this raises the question of how to identify a textual antecedent precisely, especially in approaches based on computational processing.

A general pattern for anaphoric *this*, on the basis of the properties analyzed, is that the reference is anaphoric with an explicit textual antecedent. This is true in 86 cases of COMPARA, just over 50%, and holds for 33 tokens of OAS (76.75%) and 31 tokens of EMEA (67.39%). Literary texts seem again to allow a greater variety of usage, whereas legal and medical documents conform to the general pattern. Variation of reference types and antecedent explicitness is strongly concentrated on literary text tokens. As a result, a higher proportion of tokens in literary texts differ from the general pattern, although proportions for the textual/nominal variation are similar. The textual antecedent element, therefore, does not seem to be linked to text type, but rather to the use of anaphoric *this* in general.

4.2. Cross-linguistic analysis

The list of renderings in subsection 3.5 is used to present results for anaphoric *this* translations. An attempt is made to establish translational patterns associated with English collocations and categories in the variables. Findings are summarized in algorithmic form in the sense of instructions in steps on how to translate specific combinations of category and collocation. No actual processing tests in real-life computer systems were carried out. Table 2-5 shows frequencies for renderings by source with a view to include text type elements in the cross-linguistic analysis.

Table 2-5. Distribution of anaphoric *this* translations by source

Translation	COMPARA	OAS	EMEA	Total	Percent
isso-group	104 (54.45%)	26 (60.46%)	16 (34.78%)	146	52.14
aí	1 (0.52%)	0	0	1	0.36
aquilo-group	4 (2.08%)	0	0	4	1.42
Omission	28 (14.65%)	9 (20.93%)	10 (21.73%)	47	16.78
esse-group	19 (9.93%)	2 (4.65%)	4 (8.69%)	25	8.93
aquele-group	3 (1.57%)	0	0	3	1.08
Noun phrase	12 (6.28%)	3 (6.97%)	13 (28.26%)	28	10.00
o and variations	3 (1.57%)	0	0	3	1.08
assim	14 (7.38%)	1 (2.32%)	0	15	5.35
Object pronoun	3 (1.57%)	0	0	3	1.08
Subject pronoun	0	1 (2.32%)	0	1	0.36
tal	0	1 (2.32%)	3 (6.52%)	4	1.42
Total	191 (100.0%)	43 (100%)	46 (100%)	280	100.0

There are only four classes which appear in all three sources, namely, *isso*-group, omission, *esse*-group, and noun phrase (NP). Together these four classes account for 87.85% of translations and are the only renderings in EMEA, except for three tokens of *tal*. In OAS, there are also only three tokens that do not belong to these four classes. In all the sources, *isso*-group is the most frequent translation. Proportions are similar in COMPARA and OAS, going over 50%, but the percentage is much lower in EMEA, because NPs are almost as frequent as the *isso*-group class. Differently, omission is the second most frequent rendering in OAS, followed by NPs and *esse*-group. In COMPARA, omission is also the second most frequent choice, but *esse*-group comes third and *assim* ranks fourth, one token over NPs. The subsequent discussion tries to reveal motivations for these patterns on the basis of collocations and textual semantics, using the properties in the model as parameters.

There are only three tokens of *this* as a prepositional object in EMEA. Two tokens conform to the general pattern in all variables. The textual antecedent is the full preceding sentence in both (as in 24a below, which shows the English textual antecedent).[3] Both tokens are translated by generic noun phrases (*este facto* 'this fact') in (24b). Rendering as *isso*-group would be acceptable. The third one, shown in example (25), has a nominal antecedent, *diet*, which is the object of the verb in the preposed subordinate clause. This is retained in the target text, but the anaphor is omitted. Omission of subjects and objects is common in Portuguese. Some verbs apparently favour object omission. The standard *isso*-group translation might be seen as idiomatically inadequate in (25b).

(24a) Sudden onset of sleep during daily activities, in some cases without awareness or warning signs, has been reported uncommonly. Patients must be informed of **this** and advised to exercise caution while driving or operating machines during treatment with MIRAPEXIN.

(24b) Os doentes devem ser informados **deste**
The sick must be informed of-this
facto e aconselhados a redobrar a
fact and advised to double the
atenção ao conduzir ou utilizar
attention to conduct or utilize
máquinas…
machines…

(25a) If you are following a special diet for diabetes, you should continue with **this** while you are taking Tandemact.

(25b) Se estiver a fazer uma dieta especial
If be to do a diet special
para diabéticos, deve continuar
for diabetics, must continue
enquanto estiver a tomar Tandemact.
while be to take Tandemact.

In OAS, there are two occurrences of prepositional objects. One conforms to the general pattern, including the standard translation (*isso*), and the other has a nominal antecedent and is translated as *ele*, a subject pronoun (standard rendering of *he*) used as a prepositional object. It is the only token of *ele* as a translation for *this* in the whole sample. Subject pronouns as prepositional objects can only be used to refer to nominal antecedents, but *isso* would not be inappropriate, as it is used for both antecedent structures. Prepositional objects in COMPARA adhere to the general pattern regarding reference type, as there are no cases of cataphora and only one case of deixis. They are also predominantly textual (85.72%), but depart radically from the norm for having a high proportion of implicit antecedents. Translations by *assim* (10 tokens) are clearly associated with the most frequent collocation,[4] *like this* (15 tokens). The anaphor in the collocation is omitted in two Portuguese renderings in which the syntax is fully rearranged; it is translated by the preposition *como* (*like*) + *esse*-group in two cases and by *isso*-group in one case.

The attachment of the prepositional phrase seems to play a role in translators' choices. Even when this choice was different, *like this* could be translated as *assim* whenever attached to a verb phrase in source or target text. The basic translation correspondence also holds for tokens attached to noun phrases of generic reference, such as *things* and *anything*. For cases in which the noun phrase is semantically specific, such as *principles* or *pictures*, *like this* seems to be preferentially translated as *como* + *esse*-group, the variant of *esse* chosen according to agreement with the grammatical gender and number of the antecedent. The other collocations of anaphoric *this* as a prepositional object in COMPARA include *about this* (4 tokens), *at this* (2 tokens), *behind* (*all*) *this* (1 token), *beyond this* (3 tokens), *for this* (2 tokens), *in this* (1 token), *of this* (5 tokens), *than this* (1 token), *to this* (3 tokens), and *with this* (5 tokens).

The five tokens of the last collocation were translated as *isso*-group forms, as were four of the five tokens of the collocation *of this*. In one of them, the anaphoric demonstrative was omitted, along with the full phrase *the consequence of this*. The choice does not seem to reveal a pattern. Regarding tokens of *about this*, the Portuguese texts show four *isso*-group translations and one omission. The omission seems to be a translator's choice rather than a pattern. All tokens of *beyond*, *behind*, *for*, and *than* with *this* as a prepositional object are translated by *isso*-group. Anaphoric *this* as an object of *in* is translated by *aí*, an adverb of place mirrored as *there*. Although only one token was found in the sample, this is likely to be a pattern, since the translation by *isso*-group would not be appropriate. Two of the tokens of *to this* are translated by *isso*-group, omission occurs in one, in which the phrase *say to this* is translated as *responder* 'answer' and the prepositional phrase is omitted, probably a pattern for such usages of *say*. One token of *this* in the phrase *laugh at this* is translated as *isso*-group, whereas the second one is omitted. The translator chose to use *rir*, the Portuguese standard rendering of *laugh*, as an intransitive.

Translation patterns for anaphoric *this* as a prepositional object may be summed up in algorithmic form as follows:

1. If the collocation is *like this* attached to a verb phrase, translate as *assim*.
2. If the collocation is *like this* attached to a noun phrase, translate as *assim* if the noun phrase is semantically generic, such as *things*.
3. If the collocation is *like this* attached to a noun phrase, translate as *como* + *esse*-group if the noun phrase is semantically specific.
4. If the collocation is *in this*, the translation is *aí*.
5. If the collocation is *say to this*, the translation is *responder* and the anaphor is omitted.
6. If the collocation is *laugh at this*, the anaphor may be omitted and the verb translated as an intransitive.
7. If the collocation is *continue with this*, omit the preposition and the anaphor.
8. Translate by *isso* or by a generic noun phrase such as *este fato* or *este assunto* according to formality or specification requirements in all other cases.

Verb objects in EMEA show noticeable translational patterns. Translations by generic noun phrases are associated with textual antecedents, translation by *esse*-group appear for nominal antecedents, and *isso*-group renderings are used both for textual and nominal antecedents. Omission occurs when the translation uses the impersonal verb form *trata*-se 'concerns', which does not take an explicit subject. A degree of syntactic rearrangement is a recurring pattern in omissions of this kind. Verb objects in OAS that refer anaphorically to explicit textual antecedents show three distinct translations: *isso*-group, omission with syntactic rearrangement involving impersonal verb forms, and *assim*. There is one token with a nominal antecedent which was translated into a semantically specific summarizing noun phrase attached to a noun, thus changing the verb object into a noun phrase modifier. The *isso*-group cases reflect renderings which preserve the source language structure, often with the anaphor as the subject of a verb in the passive form. A token of *this* as the object of *deem* is translated as *assim*. There is probably a translational pattern associated with formal texts in this case.

Verb objects are translated by *isso*-group in 83.07% of the cases in COMPARA. There are two pairs in which the correspondent Portuguese word is a lexical noun phrase of a generic kind. These are *cena* ‘scene’ and *situação* ‘situation’. The translator’s choice is a consequence of the verbs to which *this* is linked as an object in the English text, namely, *describe* and *enjoy*. Their Portuguese translations, *descrever* and *gozar*, seem not to prefer *isso* as an object in standard usage, although this is possible with *descrever*. Plain omissions occur when *this* is the object of verbs *know*, *notice* and *write* in the original. Corresponding verbs in Portuguese (*saber*, *notar* and *escrever*) belong to the group of verbs that accept omission of the object well whenever it seems to be readily inferable from the preceding or subsequent text. Since there are also cases in which the omission does not occur and *isso*-group is used, it is hard to identify a pattern.

One of the cases of omission is the English phrase *made this an opportunity*, which is translated by *aproveitou a oportunidade*, for which a mirror translation would be *took advantage of the opportunity*, rendering the demonstrative unnecessary. It does not seem to constitute a pattern, but more tokens would have to be analyzed. Other omissions are in fact a consequence of translators’ choices involving syntactic rearrangements, which also do not seem to make up a pattern. There are three cases of *this* as a verb object in which the translation uses object pronouns instead of anaphoric demonstratives, apparently as a result of formality requirements.

Choices of translation for anaphoric-*this* tokens as verb objects may be summarized in algorithmic form, but variations are not easily associated with collocation data. However, the attempt is presented below.

1. If the Portuguese translation verb is *descrever* or *gozar*, translate as NP such as *cena* or *situação*.
2. If the English verb is *know*, *notice* or *write*, omission is possible.
3. If the English phrase is *make this*-IObj NP-Dobj, omission is possible, along with translation of *make* as a verb that collocates semantically with the NP.
4. If a more formal style in the target text is seen as adequate, use an object pronoun *o*, *a*, *os* or *as* as appropriate.
5. In all other cases, translate as *isso*-group.

Lexical verb subjects are the most frequent grammatical category in OAS and EMEA, but less frequent in COMPARA. Patterns for textual antecedents also include reference to the full preceding sentence, but there are references to the main clause in a compound sentence in which the anaphor occurs in the subordinate clause. There are two cases of “plain” omission, that is, the English structure is essentially retained in the target text with the anaphor omitted, but one rendering as omission involves changing a non-finite clause into a finite clause. The use of generic noun phrases, such as *esta situação* ‘this situation’ or *este facto* ‘this fact’, relates to textual antecedents, whereas domain-specific summary terms such as *estes sintomas* ‘these symptoms’ have nominal antecedents. All tokens refer anaphorically, except for one case of deictic reference, in which *this* is the subject of a *will do* phrase in the sense of *will work*. The sentence is preceded by the adverb of place *here* (*Here: this will do*), which signals, along with other subsequent clues, that the implicit antecedent is *this place*. The Portuguese translation omits the anaphor. The text type seems to be a crucial determinant of this sort of usage. Dialogues and narrator interventions in literary texts favour deictic references, whereas technical and legal texts do not, for obvious reasons. Translations as noun phrases seem to be interchangeable with *isso*-group, but the former reduces the degree of underspecification, as shown in (26b) below, in which *o quadro* ‘the picture’ is introduced as the subject. The *isso*-group rendering is from a second translation of the same book. The English text excerpt (26a) shows the antecedent:

(26a) He hated to be separated from the picture that was such a part of his life, and was also afraid that during his absence some one might gain access to the room, in spite of the elaborate bars that he had caused to be placed upon the door. He was quite conscious that **this** would tell them nothing.

(26b) Estava absolutamente convicto de que
Was absolutely convinced of that
o quadro nada revelaria a
the picture nothing would-reveal to
quem, porventura, o visse.
whom, perchance, it saw.

(26c) Estava perfeitamente convencido de que
Was absolutely convinced of that
isto nada revelaria a ninguém.
this nothing would-reveal to nobody.

It seems reasonable to say that (26b) reduces the vagueness of the English antecedent, whereas (26c) retains it. The tendency to reduce underspecification may be a characteristic of translated text, rather than a language-specific feature. The variation noun phrase/*isso*-group/omission requires further investigation in search of stable patterns. The algorithmic systematization is shown below.

1. If the reference is to a place or location, signalled by a preceding adverb of place, omit the pronoun in translation.
2. If a finite clause is translated as a non-finite clause, consider omission as fits syntax.
3. If the target text changes a textual referent to a nominal antecedent, translate as repetition of antecedent instead of anaphor.
4. If the antecedent is textual, prefer generic noun phrases such as *este facto* ‘this fact’.
5. If the antecedent is nominal, consider more specific summary terms such as *essa área* ‘this area’.
6. In all other cases, translate as *isso*-group.

Copular verb subjects are the most frequent type of grammatical function in COMPARA. Thus, anaphoric *this* is clearly associated with copular verbs in literary texts. Translations show more variety than for any other category. There are 20 *isso*-group renderings, 17 *esse*-group renderings and 16 omissions. This means that, except for two instances of *like this*, all *esse*-group translations come from this category. Therefore, explicit nominal antecedents linked to an anaphoric *this* subject in a copular construction favour *esse*-group renderings. Antecedents are usually subjects or objects in preceding main clauses or independent sentences. The preference also holds for introductions (see example 13) and collocations such as *this is the case*. On the other hand, adjectival subject predicatives favour translations by *isso*-group. Omissions occur interchangeably in favourable *isso*-group environments, but *esse*-group environments are not prone to omission variation, except for the *this is the case* collocation in a conditional clause, translated alternatively as *se for o caso* ‘if *be*-subjunctive the case’.

There are three translations as *assim* in which there is an indirect question as subject predicative in the English text, as in *if this was how I coped with…*. Translations as demonstrative *o* are associated with the relative pronoun *que* in clauses with a conclusive meaning in the target text. The constructions are a result of syntactic rearrangement in the translation of conditional clauses. There are two such occurrences in COMPARA. Translations as full noun phrases involve two cases of cataphoric reference and one case of implicit textual antecedent.

Renderings as *o/a seguinte* ‘the following’ can only work with cataphoric reference, often signalled in English by placing *this* in the final position in the clause. The implicitness is signalled in English by using *all this about* in questions such as *What’s all this about Joe?*, translated as *Que história é essa com o Joe?*. The use of the word *história* ‘story’ is the expected idiomatic solution. Differently, plain questions with *what* (*What’s this?*) are translated by *isso*-group (*Que é isso?*) and retain the English structure.

In OAS, there are ten tokens of *this* as copular verb subjects, and translations include *isso*-group, *esse*-group, NP and omission. Three tokens appear as *this is why*.... The collocation requires syntactic rearrangement in Portuguese, with the copular verb placed in the beginning of the sentence, followed by the preposition *por* and then the anaphor (*é por isso*/*isto que*...). This is the translation choice for two tokens. One is translated as a noun phrase (*foi por esta razão que*...), literally *was for this reason that*. The variation seems unmotivated. The collocation *this is where* is wholly omitted in the target text. The antecedent (*room outside the courtroom*) is present in Portuguese in the preceding sentence as in English. However, the emphasis on the place where the decision is to be made, present in the source text, is left out in the target text, which translates into *this is where they must decide* by *terá de se determinar* ‘it will have to be determined’.

There are three other cases of omission involving syntactic rearrangement in OAS. In one of them, the impersonal verb form *trata-se* is used instead of *this is*. The other two involve syntactic rearrangement in which the copular clause is changed into a prepositional phrase, thus rendering *this is the first time* as *pela primeira vez* ‘for the first time’. Translations by *esse*-group involve nominal antecedents which are the object of the verb in a preceding main clause or independent sentence. The copular structure includes a hyperonimic subject predicative, such as *This is the primary domestic legislation*… referring to an act.

There are nine cases of copular verb subjects in EMEA. Omissions occur in three tokens. Two of them occur in subordinate conditional clauses; the *trata-se* pattern is used in one of them, and the other retains the copular structure with the subject omitted. The third omission uses a lexical verb with the subject omitted. NP translations are associated with English collocations *this is the case* and *this is expected to*. Translators apparently choose NPs (*esta situação* ‘this situation’ and *esta tendência* ‘this tendency’) out of formality requirements. Copular structures are less formal in Portuguese. The third NP translation seems to aim at reducing underspecification with the use of *estes efeitos* ‘these effects’ instead of the anaphor. There is one translation by *isso*-group. The subject predicative is an evaluative adjective (*important*) applied to a textual explicit antecedent in an anaphoric reference, a recurrent pattern. There is one translation by *esse*-group. The subject predicative is a summarizing definite description expressed by a noun phrase. Favourable environments for each choice are confirmed. Below is a summary.

1. If the subject predicative is an indirect question with *how*, translate as *assim*.
2. If the subject predicative is an indirect question with *what* SUBJ *be*-form, translate as *assim*.
3. If the copular structure is a conditional clause that carries a conclusive meaning, such as *If this is obvious*, consider translating as *o que*....
4. If the reference is cataphoric with *this* in subject predicative position, translate as *o*/*a seguinte*.
5. If the collocation is ...*what is all this about*..., translate as *que história é essa com*...
6. If the subject predicative is a definite description, and the antecedent is nominal, translate by *esse*-group.
7. If the subject predicative is an adjective, translate as *isso*-group or omission.
8. If informality or undespecification is to be avoided, consider the *trata-se* pattern and a NP-lexical verb structure as options.

5. Conclusion and future developments

In spite of attempts to organize findings of the study in algorithmic form, no tests in actual computer systems were carried out. In fact, many instructions in the algorithms would not be trivial to implement in real-life systems. Nonetheless, translational patterns associated with grammatical categories and English collocations were detected. Other patterns were linked to the nominal/textual antecedent dichotomy, seen as a crucial aspect of anaphoric *this* in the approach. Cataphoric references seem to be associated with the position of the anaphor in the source text and with translation as a specific NP in the target text. Some aspects of textual semantics, such as evaluation, underspecification and formality, are difficult to measure, but appear to hold promise for future developments, combined with the other levels of information investigated in the study. It seems to be true that the standard *isso*-group translation for anaphoric *this* is seen by translators as somewhat too informal or conducive to underspecification for use in formal texts, although certain forms of omission with syntactic rearrangement are considered appropriate.

It is nonetheless undeniable that a substantial amount of ignorance is still a fact in textual semantics in general, both in monolingual and contrastive or translation studies. The interaction of ideational, lexicogrammatical and cross-linguistic aspects is not fully mapped and, to the extent that it is mapped, it is poorly understood. Anaphoric phenomena, especially those involving the interaction with textual semantics and cross-linguistic elements, also seem to require a lot more field work in the sense of analyzing corpus data. Perhaps, as advances are made, the routine of collecting and classifying tokens will become not so agonizing and time-consuming, allowing faster confirmation of pattern definitions. No matter how tentative and labourious, however, the approach may eventually pay off in terms of relevant findings. It seems also safe to conclude that accurate anaphora resolution, along with antecedent and reference type classification, would contribute substantially to the improvement of machine translation.

Notes

[1] All examples are taken from the sample collected.
[2] One typical example in the OAS corpus: “The Charter also provides that Parliament and the provincial legislatures must sit at least once a year. This ensures that our governments perform the work for which they were elected, and also that they will have to answer questions and explain themselves in public.”
[3] Glosses in English appear in italics below Portuguese forms.
[4] A description of the use of *assim* in Portuguese is beyond the scope of this chapter, but a single-word translation into English might be so, or, in cases such as those discussed in this chapter, like this.

References

Botley, S. (2000), *Corpora and Discourse Anaphora: Using Corpus Evidence to Test Theoretical Claims*. PhD thesis, Lancaster University.

Dyvik, H. (1998), “A translational basis for semantics”, in S. Johansson and S. Oksefjell (eds.) *Corpora and Cross-Linguistic Research: Theory, Method and Case Studies*, 51-112. Amsterdam: Rodopi.

Frankenberg-Garcia, A. and Santos, D. (2001), “COMPARA, um corpus paralelo de português e inglês na Web”, in *Cadernos de Tradução IX*, 61-79. Florianópolis: Universidade Federal de Santa Catarina.

Greenbaum, S. (1996), *The Oxford Grammar of the English Language*. Oxford: Oxford University Press.

Halliday, M.A.K. and Hasan, H. (1976), *Cohesion in English*. London: Longman.

Halliday, M.A.K. and Matthiessen, C. (2004), *An Introduction to Functional Grammar*. London: Arnold.

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

Santos, D. (1998), “Perception verbs in English and Portuguese”, in S. Johansson and S. Oksefjell (eds.) *Corpora and Cross-Linguistic Research: Theory, Method and Case Studies*, 319-342. Amsterdam: Rodopi.

Tymoczko, M. (1998), “Computerized corpora and the future of translation studies”. *Meta* 43(4): 652-660.

CHAPTER THREE

PATTERNS IN METAPHOR TRANSLATION: TRANSLATING FEAR METAPHORS BETWEEN ENGLISH AND CHINESE

YAN DING, DIRK NOËL, HANS-GEORG WOLF

1. Introduction and problem statement

Following the publication of Lakoff and Johnson's (1980) influential *Metaphors We Live By* it has become widely recognized in cognitive science and linguistics that metaphor is ubiquitous in language and cognition: complex and abstract concepts are often thought and talked about in terms of conceptually simpler and more concrete notions. A fairly recent outcome of this research is the realization that there is both cross-cultural similarity and variation in metaphor: some metaphors can be identified in a great many languages, while others are language-specific (Kövecses 2000, 2005). The reality of this makes the question of what happens to metaphors in the process of translating text a very relevant one.

To this question there are two sides, one linguistic, the other conceptual: a) how does cross-cultural variation in metaphor affect the translation of metaphorical expressions? and b) how does the translation of metaphorical expressions affect the metaphors they express? A number of studies approaching metaphor translation from a cognitive linguistic perspective have addressed either the one or the other of these aspects. Studies of the first kind include examples such as Mandelblit (1995), Maalej (2003) and Al-Zoubi *et al.* (2006), which all distinguish between a "similar mapping condition" (SMC) and a "different mapping condition" (DMC). In the SMC case the source language (SL) and the target language (TL) use an identical metaphor to conceptualize a particular notion; in the DMC case SL and TL conceptualize a particular notion using a different metaphor. While Mandelblit (1995) focuses on the translation process, using translators' reaction time as a parameter that reveals differences in the translation process in the SMC and the DMC situations, the studies by Maalej (2003) and Al-Zoubi *et al.* (2006) are product-oriented, offering several sets of examples that illustrate how translation products are dependent on SMC and DMC. What the three studies have in common is the conclusion that metaphoric expressions based on metaphors shared by SL and TL are more readily translatable than those based on metaphors that only exist in SL, as the translation of the latter involves a conceptual shift, i.e. a transfer from one way of conceptualizing an aspect of reality to another.

Examples of studies addressing the effect of translation on metaphor include Schäffner (2004) and Stienstra (1993). Their approaches are largely descriptive, focusing on how metaphors and metaphorical expressions are dealt with in actual translations. Schäffner (2004) identifies five types of metaphor translation in an investigation of translations of political texts between English and German:

> (1) a conceptual metaphor is identical in ST and TT at the macro-level without each individual manifestation having been accounted for at the micro-level;[1] (2) structural components of the base conceptual schema in the ST are replaced in the TT by expressions that make entailments explicit; (3) a metaphor is more elaborate in the TT; (4) ST and TT employ different metaphorical expressions which can be combined under a more abstract conceptual metaphor; (5) the expression in the TT reflects a different aspect of the conceptual metaphor. (Schäffner 2004: 1267)

The first type of metaphor translation, for instance, can be instantiated in the following case (Schäffner 2004: 1259-1260). In a speech delivered by the former German Chancellor, Helmut Kohl, a German sentence whose literal translation into English should be *The American forces in Germany are thus an important component of the transatlantic bridge* was actually translated as *The American forces in Germany are thus an important component of transatlantic friendship*, which at the first glance appears to suggest that a metaphor was deleted. However, according to Schäffner, *bridge* is only an individual manifestation of two macro-level metaphors, A STATE IS A PERSON and INTIMACY IS CLOSENESS, which she shows to be preserved in the translated text through a close analysis of the whole passage. In other words, the source text and the translated text make use of exactly the same macro-level metaphors, though, at the micro-level, the specific metaphorical expression using *bridge* is not rendered in the translation. Stienstra (1993) observed the same phenomenon when studying Bible translations into English and Dutch: the metaphor YHWH IS THE HUSBAND OF HIS PEOPLE is preserved at the macro-level, but its specific textual manifestation is not always accounted for in each individual case (cf. Schäffner 2004: 1261).

A drawback of the studies within the first of these two approaches is that they do not give enough attention to the authentic texts to show precisely how cross-cultural variation in metaphor can affect the outcome of the translation of metaphorical expressions. On the other hand, studies within the second approach, which do pay attention to translation products, do not try to relate the treatment that metaphorical expressions receive with cross-cultural variation in conceptual metaphor. The present chapter seeks to combine these two broad approaches through a corpus-based case study of FEAR metaphors in translations from English into Chinese. As such it also contributes to the study of the conceptualization and language of emotions, which has arguably taken pride of place within the now well-established "conceptual metaphor theory" paradigm (see, for instance, Geeraerts and Grondelaers 1995, Gevaert 2005, Györi 1998, Koivisto-Alanko 2006, Kövecses 1990 and 2000, Maalej 2007, Matsuki 1995, Yu 1995). We aim to describe the treatment that FEAR metaphors and their metaphorical expressions receive in actual translations and to explore the relation between this treatment and cross-cultural variation with relation to FEAR metaphors. Our specific research questions link up cross-cultural variation with translational practice:

1. Will expressions of a metaphor in SL be translated as expressions of the same metaphor in TL when the metaphor is shared by the two languages? In other words, will a metaphor in SL be preserved in translation if it is also available in TL?
2. If only SL has the metaphor, how will an expression of this metaphor be translated? a) literally, b) into an expression of the original metaphor in SL (and hence a novel one in TL), c) as an expression of a different metaphor, so as to at least retain the metaphorical nature of the language employed, or d) using other strategies? In other words, if a metaphor in SL is not shared with TL, what will happen to it in the process of translation? Will it disappear altogether or will it be preserved in some way?
3. If option c) in 2 occurs with any regularity at all, is there any discernible pattern in the cases where an expression of one metaphor is translated into an expression of another metaphor? In other words, is there any regularity in metaphor changes in the translation process?

As heralded by the title of the chapter, our approach to answering these questions will be a corpus-based one. We do not think this is much in need of justification: the state of the art in both metaphor and translation research is such that the advantages of a corpus-based approach to either of them no longer need to be advocated (for metaphor studies, see Deignan 2005 and Stefanowitsch and Gries 2006; for translation studies, see Baker 1995, 1999, Laviosa 1998 and Olohan 2004). The next section will illustrate which corpora were used and how they were used.

2. Methodology

A key methodological ingredient of our research is the corpus-based identification of metaphors and metaphorical expressions. This can broadly be said to consist of two consecutive methodological activities, or activity groups: 1) the retrieval from the corpora of lexical material that instantiates the target domain of the metaphors we are interested in and 2) determining a) which instances of this material constitute parts of metaphorical expressions and b) what metaphors these metaphorical expressions realize. After identifying the corpora used in section 2.1, we will outline the specifics of these activities in sections 2.2 and 2.3. Section 2.4 contains a note on the compilation of the frequency information that will contribute to answering the research questions detailed in section 1.

2.1. Resources

Four corpora were used in this study, which are all freely accessible online: the English-Chinese parallel corpus of the Hong Kong Institute of Education (PCHKIE for short, available at http://ec-concord.ied.edu.hk/paraconc/index.htm), the English-Chinese parallel corpus of Xiamen University (PCXU, http://www.luweixmu.com/ec-corpus/index.htm), the Lancaster Corpus of Mandarin Chinese (LCMC, available at http://corpus.leeds.ac.uk/query-zh.html) and the Chinese Internet Corpus (CIC, also available at http://corpus.leeds.ac.uk/query-zh.html). The two parallel corpora were used to identify English FEAR metaphors and their expressions and to establish how these are translated into Chinese. The monolingual Chinese corpora were used for the identification of idiomatic Chinese FEAR metaphors and expressions.

2.2. Target domain item retrieval

To be able to retrieve the metaphors relevant to this study from our corpora we adopted the Metaphorical Pattern Analysis method (MPA) proposed by Stefanowitsch (2006), who defines a "metaphorical pattern" as "a multi-word expression from a given source domain (SD) into which one or more specific lexical items from a given target domain (TD) have been inserted" (Stefanowitsch 2006: 66). MPA can retrieve a large number of metaphorical patterns by searching the target domain item in a corpus and identifying the metaphors associated with these metaphorical patterns. Since, in Stefanowitsch's approach, target domain items are always nouns referring to the target domain, it follows that, technically, not all metaphorical expressions of a metaphorical domain are also "metaphorical patterns" and that an approach based on identifying such patterns may not capture all metaphorical expressions, or all metaphors even, of a particular target domain. For instance, of the following expressions of the ANGER IS A HOT FLUID IN A CONTAINER metaphor, borrowed in part from Kövecses (2002: 96-97), only (1) to (3) are metaphorical patterns, because they contain the TD lexical item *anger*. (4) to (6), on the other hand, though clearly metaphorical expressions of the same metaphor, are not identified as metaphorical patterns and cannot be retrieved by the MPA method.

ANGER IS A HOT FLUID IN A CONTAINER

1. His pent-up anger welled up inside him.
2. My anger kept building up inside me.
3. He was bursting with anger.
4. She could feel her gorge rising.
5. We got a rise out of him.
6. He was angered to the point where his blood was starting to boil.

Stefanowitsch (2006: 69) has demonstrated, however, that MPA can nevertheless "identify metaphors more systematically and more exhaustively than non-corpus-based approaches."

Adopting the MPA approach we searched for the word *fear* in the two English-Chinese parallel corpora to retrieve expressions of FEAR metaphors in native English texts (NET) and the expressions matching them in the parallel translated Chinese texts (TCT), and searched for the word *kongju* in the Chinese corpora to extract Chinese FEAR metaphors from native Chinese texts (NCT). Altogether we extracted 203 instances of *fear* from the two parallel corpora,[2] 22 instances of *kongju* from the Lancaster Corpus of Mandarin Chinese and another 100 instances of *kongju* randomly exacted from the Chinese Internet Corpus. The reason we chose *kongju* as the Chinese target domain item is that an examination of the translations of *fear* in TCT showed that *fear* was translated as *kongju* in 51 of the 110 cases in which *fear* was part of a metaphorical expression in NET. Compared with *pa* (10 instances), *danxin* (9), *haipa* (9), *youlü* (8) and another nine words with a frequency of not more than three, this makes *kongju* the most typical equivalent of *fear* in the translation of metaphorical expressions of FEAR.

2.3. Metaphor identification

Metaphor identification is a problematic issue that has received considerable attention (Cameron 1999, Crisp 2002, Heywood *et al*. 2002, Semino *et al*. 2004, Steen 1999 and 2002, Pragglejaz Group 2007). Semino *et al*. (2004: 1272), for instance, list four major methodological problems in their metaphor identification and analysis process: how to draw a boundary between the literal and the metaphorical, how to precisely identify the tenor and the vehicle, how to extrapolate conceptual metaphor from metaphorical expressions, and how to extrapolate conventional metaphor from patterns of metaphorical expressions. Cameron (1999) advises that one way to raise the validity of research on metaphor is to work with precise criteria and to offer explicit decision making in metaphor identification. In the research reported here we adopt a metaphor identification procedure based on her work (Cameron 1999) and on work of the Pragglejaz Group (2007). The general principle of this procedure is that metaphors are identified by examining the verbs, prepositions and adjectives that the target items collocate with. The specific steps of the procedure are the following:

1. Read the expressions containing the target items to establish the meaning of the whole expression and each word.
2. Determine the verbs, prepositions and adjectives that collocate with the target domain items.
3. For each verb, preposition and adjective, determine whether it can collocate with words that denote a more concrete category of things in other contexts.
4. If it can, determine if FEAR contrasts with the more concrete category of things but can be understood in terms of it.
5. If yes, take the more concrete category as the source domain and formulate the metaphor.

Three points need to be specified with relation to this procedure. First, we identified not only general metaphors but also their specifications, if they have any. By ‘specification’ we mean subordinate level metaphors that highlight a particular aspect of the general metaphors. For example, it was found that the metaphor FEAR IS AN ENTITY has 7 different specifications in English, which we have formulated respectively as X FEELS FEAR WHEN X IS IN POSSESSION OF THE ENTITY, X FEELS FEAR WHEN THE ENTITY IS PRESENT, THE INTENSITY OF FEAR IS THE SIZE OF THE ENTITY, and so on.

Second, we consulted the *Collins English Dictionary* (CED, 3rd edition) and the British National Corpus (BNC) for English, and *Zhonghua Zaixian Cidian* (*Online Chinese Dictionary*, http://www.ourdict.cn/), the LCMC and the CIC for Chinese, in order to check whether the verbs, prepositions and adjectives combining with the target domain item can collocate with words that denote a more concrete category of things in other contexts. For example, we categorized *She has a great fear of fire* as an instance of X FEELS FEAR WHEN X IS IN POSSESSION OF THE ENTITY mainly on the basis of the fact that the first sense of *have* in the *CED* is “to be in material possession of”, illustrated by *He has two cars*. Because the *CED* does not have entries for *explain away, laugh away,* and *drive away*, we looked for these in the BNC to account for expressions like *explain the fear away*, *laugh the fear away*, and *drive the fear away*. The results showed that while *drive away* could collocate with words that denote concrete entities like *competitors*, *mosquitoes*, etc., *explain away* and *laugh away* could only collocate with words that denote abstract things like *difficulty, embarrassment,* etc. Therefore only *drive the fear away* was considered to be metaphorical and categorized as an instance of X FEELS FEAR WHEN THE ANIMATE BEING IS PRESENT, a specification of the metaphor FEAR IS AN ANIMATE BEING.

Finally, for additional help with metaphor identification and formulation, we also referred to work by Kövecses (1990, 2000), Stefanowitsch (2006) and Zhang (2000), who have all investigated either English or Chinese metaphors of FEAR.

2.4. Quantification

Subsequent to their identification, all metaphors and their specifications were tallied. Some metaphorical expressions were counted more than once because they instantiate two or more metaphors or specifications that do not conflict with each other. For instance, we counted the expression *I don’t have the smallest fear* twice, once as the expression of X FEELS FEAR WHEN X IS IN POSSESSION OF THE ENTITY and once as the expression of THE INTENSITY OF FEAR IS THE SIZE OF THE ENTITY. On the other hand, though many of the metaphors we identified can be organized into what Lakoff (1993) called “hierarchy structures”, in such a way that the source domain of one metaphor may logically include that of another, only the more specific metaphors were taken into account in such cases. An example is the inclusive relationship between FEAR IS AN ENTITY and FEAR IS A SUBSTANCE IN A CONTAINER. Though a substance in a container is definitely an entity, expressions of the second metaphor like *The child's convulsions filled us with fear* were not counted as expressions of the first, with respect to which it stands in a hyponymic relationship.

3. Results

As mentioned in section 2.2, the MPA of the NET corpus produced 203 *fear* expressions. 110 of these were established to be metaphorical. The number of metaphorical expressions in the matching translations is only 71, however. In the NCT corpus 85 out of a total of 122 *kongju* expressions were established to be metaphorical. All the general metaphors and specifications identified in the NET, TCT and NCT corpora are listed in Table 3-1, with the general metaphors indexed by Arabic numbers and the specifications of a general metaphor indexed by small English letters. In this table, N indicates the number of metaphorical expressions of a particular metaphor; n indicates how many of these English metaphorical expressions were translated into Chinese metaphorical expressions of the same metaphor,[3] i.e. how many times the metaphor was preserved; R indicates the frequency rank of a particular metaphor in NCT on a scale from 1 (most frequent) to 9 (least frequent), depending on the number of its expressions. Following each specification, the symbols [+], [–], [S] and [NS] further elaborate a specification, with [+] referring to the situation where X feels fear or feels more intense fear, [–] to the situation where X does not feel fear or feels less intense fear, [S] to the situation where fear itself is the logical subject of the metaphorical expression (i.e. either the grammatical subject of an active expressions or the agent of a passive) and [NS] to the situation where fear is not the logical subject of the metaphorical expression. In what follows, we will refer to particular specifications with labels like 2a[+][NS] for convenience. Table 3-1 also supplies an illustration of each metaphor. These are not attested examples but simplified expressions which only retain the key words and semantic relations that can manifest the metaphor. The illustrations in parentheses in Table 3-1 are simplified literal translations of Chinese metaphorical expressions which do not have corresponding metaphors in the NET.

Table 3-1. Metaphors in the NET, TCT and NCT

General metaphor and Specification			Illustration	NET		TCT	NCT	
				N	n	N	N	R
1. fear is a location								
a. x feels fear when x is in the location	+	NS	X lives in fear	5	1	2	3	6
	–	NS	To reason X out of fear	2	1	1	2	7
b. Others			X vacillates between fear and hope	4	3	3	0	
2. fear is an entity								
a. x feels fear when x is in possession of the entity	+	NS	X has fear	7	3	3	8	2
	–	NS	X has no fear	4	0	0	1	8
b. x feels fear when the entity is present	+	NS						
		S	Fear before X	1	1	1	0	9
	–	NS	X puts fear aside	3	0	0	0	9
		S	The fear is away	1	0	0	0	9
c. x feels fear when x is with the entity	+	NS	X views something with fear	1	1	2	4	5
	–	NS						
d. x feels fear when the entity exists	+	NS						
		S	There is fear	1	0	0	1	8
	–	NS	(to eliminate fear)	0	0	6	4	5
		S	fear disappears	5	3	5	3	6
e. x feels fear when the entity is lifted	+	NS	Something raises fear among X	2	2	3	3	6
	–	NS	Fear is laid	1	1	1	1	8
f. the intensity of fear is the size of the entity	+	NS	X has great fear	3	0	0	1	8
		S	X's greatest fear is …	1	1	1	0	9
	–	NS	X does not have the smallest fear	1	0	0	0	9
		S						

g. the intensity of fear is the amount of the entity	+	NS	(X has some of fear)	0	0	0	5	4
		S						
	–	NS	(to reduce X's fear)	0	0	0	2	7
		S						
h. the intensity of fear is the weight of the entity	+	NS	(X has heavy fear)	0	0	0	1	8
		S						
	–	NS	(to lighten X's fear)	0	0	0	3	6
		S						
i. x shows fear when the entity is seen	+	NS	(He shows fear on his face)	0	0	1	1	8
		S	Fear is manifest	1	0	0	0	9
	–	NS	X shows no tint of fear	1	1	1	0	9
		S						
j. Others			Underneath the fear there is hope	3	2	2	0	
3. fear is a substance in a container								
a. x feels fear when the substance is put into x's body container from outside	+	NS	Something fills X with fear	4	0	0	0	9
	–	NS	(to discharge the fear in X's heart)	0	0	0	2	7
b. x feels fear when the substance is present in x's body container	+	NS						
		S	His fear showed in his eyes.	3	3	5	11	1
	–	NS						
		S	(There is no fear in X's heart any more)	0	0	0	1	8
c. x feels fear when the substance is present in the container where x is	+	NS						
		S	Fear fills the car	1	1	1	1	8
	–	NS						
		S						

4. fear is an animate being								
a. x feels fear when the animate being is active	+	NS	Something awakes X’s fear	3	2	2	1	8
		S						
	–	NS	To quiet X’s fear	2	1	1	0	9
		S						
b. x feels fear when the animate being is present	+	NS						
		S	(The fear never leaves)	0	0	0	1	8
	–	NS	To drive X’s fear away	2	1	2	0	9
		S						
c. x feels fear when the animate being is present in the body container	+	NS						
		S	Fear creeps into X’s mind	1	0	0	0	9
	–	NS	To drive fear out of X’s mind	1	0	0	0	9
		S						
d. fear comes into being when the animate being is produced			Something breeds fear	2	2	2	7	3
e. Others			Fear seems to possess its own life	2	1	1	0	
5. fear is an opponent								
x feels fear when the opponent is having advantage	+	NS	X can not control fear	1	1	1	3	6
		S	X is besieged by fear	10	1	1	1	8
	–	NS	X conquers fear	10	8	8	4	5
		S						
6. fear is a supernatural being			X is haunted by fear	4	1	1	0	9
fear is a tormentor			Fear takes the marrow out of X	1	1	2	3	7
8. fear is a superior			X is kept silent by fear	1	0	1	2	7
fear is a disease			Fear is contagious	4	3	3	0	9
fear is a natural force			Fear sweeps over X	2	1	1	1	8

11. fear is a sharp object	Something is penetrated by fear	1	0	0	0	9
12. fear is a nuisance	X is disturbed by fear	4	3	4	4	5
13. fear is a foundation	Something is founded on fear	1	1	1	1	8
14. fear is a poison	Knowledge is the antidote to fear	1	0	0	0	9
15. fear is a legacy	Fear is the legacy of the Vietnam war	1	1	1	0	9
16. fear is a machine	X's fear has always operated	1	0	0	0	9
Total number		110	52	71	85	

Two sets of descriptive observations can be made from a first inspection of Table 3-1. The most general one in the first set is that translation leads to a reduction in the number of metaphorical expressions and metaphors. Only 71 of the original 110 English metaphorical expressions remain metaphorical after translation, and the numbers of general metaphors and specifications are reduced from 16 to 13 and from 15 to 14 in the NET and the TCT respectively. The metaphors that have been "lost in translation" are: 11. FEAR IS A SHARP OBJECT, 14. FEAR IS A POISON, and 16. FEAR IS A MACHINE. The metaphor specifications that have not been retained in the translations are: 3a. X FEELS FEAR WHEN THE SUBSTANCE IS PUT INTO X'S BODY CONTAINER FROM OUTSIDE, and 4c. X FEELS FEAR WHEN THE ANIMATE BEING IS PRESENT IN THE BODY CONTAINER. In addition, we can also observe a general decrease in the number of expressions of each metaphor. The most salient example is 5 [+][S], falling from ten occurrences in the NET to only one in the TCT. There is one case involving a noticeable increase, however, viz. 2d[–][NS], which does not occur in the NET but has six attestations in the TCT.

The second set of observations relates to cross-cultural differences in metaphors of FEAR between English and Chinese. First, both English and Chinese possess metaphors that are not shared by the other language. The NET has six general metaphors and two specifications that do not exist in the NCT, which are: 6. FEAR IS A SUPERNATURAL BEING, 9. FEAR IS A DISEASE, 11. FEAR IS A SHARP OBJECT, 14. FEAR IS A POISON, 15. FEAR IS A LEGACY, 16. FEAR IS A MACHINE, 2b. X FEELS FEAR WHEN THE ENTITY IS PRESENT, and 4c. X FEELS FEAR WHEN THE ANIMATE BEING IS PRESENT IN THE BODY CONTAINER. In contrast, the NCT only contains two specifications that do not exist in the NET, viz. 2g. THE INTENSITY OF FEAR IS THE AMOUNT OF THE ENTITY, and 2h. THE INTENSITY OF FEAR IS THE WEIGHT OF THE ENTITY.

Second, English and Chinese may differ greatly in the number of metaphorical expressions for a shared metaphor. Salient examples include 3b[+][S] (three cases in the NET, eleven in the NCT), 4d (two cases in the NET, seven in the NCT), and 5[+][S] (ten cases in the NET, one in the NCT). Metaphor 5. FEAR IS AN OPPONENT is quite interesting. In the NCT there are seven metaphorical expressions of 5[NS] where X is the logical subject that acts against fear, but there is only one metaphorical expression of 5[S] where fear is the logical subject that exerts influence on X. In the NET, on the other hand, there is an equal balance between the metaphorical expressions of 5[NS] and 5[S], eleven and ten cases respectively. This seems to indicate that Chinese, unlike English, usually uses this metaphor to conceptualize an attempt to control fear, but not the state of falling victim to fear.

4. Discussion

Here we will answer each of the three research questions formulated in section 1. As can be seen from the data, the answer to the first question is that the expressions of English metaphors are not necessarily translated as expressions of the same metaphors in Chinese even in cases when the metaphor is shared by the two languages. We speculate that this is because the degree of entrenchment of a metaphor in TL may have a stronger influence than the mere fact that the ST metaphor is also available in TL.[4] This is supported by the fact that metaphorical expressions of metaphors shared by SL and TL which have a high frequency rank in the NCT tend to be better preserved in the TCT. For example, the metaphorical expressions of 3b[+][S], 4d, and 5[–][NS], which occupy the first, the third and the fifth place in the NCT respectively, are all well-preserved. In contrast, only one of the ten expressions of 5[+][S] in the NET, which is a lowly ranked metaphor in the NCT, is retained as the expression of this metaphor in the TCT. Below, in our discussion of the third research question, we will offer further evidence supporting the claim that the degree of entrenchment of a metaphor can affect the preservation of metaphors in translations to a large extent.

An example that could invalidate our suggestion about the role of the degree of entrenchment of a metaphor is perhaps 2a[+][NS], the frequency rank of which is the second in the NCT, while only three out of the seven expressions of it in the NET are preserved in the TCT. However, a closer examination of these metaphorical expressions leads to an interesting finding that may explain why the preservation of 2a[+][NS] is hindered. Our data show that two instances of 2a[+][NS] are also metaphorical expressions of 2f[+][NS] in the NET. But the rank of 2f[+][NS] is rather low in the NCT and there is no occurrence of the same combination of 2a[+][NS] and 2f[+][NS] in the NCT.[5] In fact, these two instances, i.e. (1) and (2) were both translated literally:

1) She has a great fear of fire.

ta hen pai huo [PCXU]
she very fear fire
'She fears fire very much.'

2) She has a great fear of water.

ta hen pai shui [PCXU]
she very fear water
'She fears water very much.'

The fact that combinations of 2a[+][NS] and 2f[+][NS] got lost in translation is therefore quite in line with our suggestion about the role of the degree of entrenchment of TL metaphors. In other words, the translation of 2a[+][NS] confirms rather than disconfirms what was said in the previous paragraph.

The following observations are relevant to the second research question. At the most specific level, there are 18 English metaphors that do not exist in the NCT, viz. 2b[+][S], 2b[–][NS], 2b[–][S], 2f[+][S], 2f[–][NS], 2i[+][S], 2i[–][NS], 3a[+][NS], 4a[–][NS], 4b[–][NS], 4c[+][S], 4c[–][NS], 6, 9, 11, 14, 15 and 16. These metaphors have a total number of 31 metaphorical expressions in the NET. Ten of these were nevertheless translated into expressions of the same metaphor. For example, both the original English expression and the translated Chinese expression in (3) are instances of the same metaphor 9. FEAR IS A DISEASE, despite the fact that this metaphor does not appear in the NCT.

3) Fear can be contagious.

kongju hui ganran taren [PCXU]
fear can infect other people
'Fear can infect other people.'

Nine were translated into expressions of a different metaphor. For instance, the English expression of 1a[–][NS] in (4) was translated into an expression of 2d[–][NS].

4) He tried to reason himself out of fears.

ta shitu shuofu ziji xiaochu youlü [PCHKIE]
he try persuade himself eliminate fear
‘He tried to persuade himself into eliminating fear.’

Seven were translated literally, for example:

5) She walked in fear on the lonesome road.

ta yigeren zou zai lu shang, juede hen haipa [PCXU]
she alone walk on road on, feel very fearful
‘She walked alone on the road, feeling very fearful.’

Five were translated into metonymic expressions where the psychological effects of fear are used to stand for the emotion. This can be illustrated by (6) where the bodily action of trembling is used to refer to fear in the translated Chinese expression.

6) His fear of her has always operated, I know, when they were together.

wo hen mingbai, meifeng ta he ta zai yiqi
I very understand, whenever he and she be together
de shihou, ta jiu mianbuliao hunshen fadou. [PCXU]
MOD time, he would unavoidably all his body tremble
‘I understand quite well that he would tremble every time when he was with her.’

One might wonder how a metaphor can possibly be preserved in the TCT when it does not (appear to) exist in the NCT, as in (3). Two points can be mentioned in this respect. First of all, our NCT corpus undoubtedly does not cover all Chinese metaphors, so that we will fail to identify a number of metaphors existing in Chinese if we treat it as conclusive. An example is the metaphor 9. FEAR IS A DISEASE. Three of the four expressions of this metaphor found in the NET were translated into expressions of the same metaphor in the TCT, and these appear to be quite idiomatic Chinese. To confirm the existence of this metaphor in Chinese, we did a Google internet search for expressions containing the key words *kongju* 'fear', *manyan* 'spread' and *chuanran* 'infect', which produced a considerable number of examples from native Chinese texts. One of these is (7).

7) zuotian yatai shichang de dafudu bodong
Yesterday Asia-Pacific market MOD great fluctuation
biaoming shichang de kongju qingxu xiang chuanranbing
show market MOD fear mood like contagion
yiyang zhengzai kuosan
the same ASP spread
'Yesterday the great fluctuation of the Asia-Pacific market showed that the market's fear was spreading like a contagion.'
(http://blog.ce.cn/html/93/107593-71534.htm)

A second point is that metaphors can be borrowed. The fact that a metaphor has so far not been used in a language does not mean it is an unacceptable way of conceptualizing the target. Language users, and eventually languages, can borrow, or "calque", metaphors just as they can borrow words and "loan metaphorical expressions" are in principle not less plausible than loan words. It is an empirical fact, however, that metaphors that do not exist in the NCT are less easily preserved in translation: while 42 of the 79 metaphorical expressions of shared metaphors (53%) are translated as expressions of the same metaphor, this is only true of 10 of the 31 metaphorical expressions of unshared metaphors (32%).

Continuing the discussion of the second research question, and at the same time addressing the third question, we can note that there are 19 metaphorical expressions in the NET that are translated into expressions of another metaphor, nine of which are instances of shared metaphors and ten of which are expressions of unshared metaphors. A close examination of them shows that there *is* a pattern in the change of metaphors in translation. As can be seen from Table 3-2, it is always metaphors that have a low frequency rank in the NCT that are changed, and they are usually changed into metaphors that rank higher. The third research question can therefore be answered in the affirmative. Highly entrenched metaphors "obliterate" metaphorical expressions of lowly entrenched metaphors, so to speak. It is only natural, therefore, to return to our answer to the first research question, that highly entrenched metaphors are more easily preserved.

Table 3-2. Metaphor changes

NET		TCT	
Metaphor	Rank in NCT	Metaphor	Rank in NCT
1a[–][NS]	7	2d[–][NS]	5
1b		2c[+][NS]	5
2a[+][NS]	2	3b[+][S]	1
2b[–][NS]	9	2d[–][NS]	5
2b[–][NS]	9	2d[–][NS]	5
2b[–][S]	9	2d[–][S]	6
2i[+][S]	9	2i[+][NS]	9
3a[+][NS]	9	1a[+][NS]	6
4a[+][NS]	8	2e[+][NS]	6
4c[–][NS]	9	2d[–][NS]	5
5[+][S]	9	3b[+][S]	1
5[+][S]	9	7	7
5[+][S]	9	7	7
5[+][S]	9	8	7
5[–][NS]	5	2d[–][NS]	5
5[–][NS]	5	4b[–][NS]	9
6	9	12	5
6	9	12	5
14	9	2d[–][NS]	5

From Table 3-2, we can also make the following three observations pertaining to changes between particular metaphors:

1. 2b. X FEELS FEAR WHEN THE ENTITY IS PRESENT tends to be changed to 2d. X FEELS FEAR WHEN THE ENTITY EXISTS. There are five metaphorical expressions of 2b, three of which are translated into expressions of 2d.
2. 6. FEAR IS A SUPERNATURAL BEING exhibits a tendency to be changed to 12. FEAR IS A NUISANCE. Two of the four metaphorical expressions of 6 are translated into expressions of 12.
3. Two metaphorical expressions of 5[+][S], a specification of the OPPONENT metaphor, are translated into expressions of 7. FEAR IS A TORMENTOR.

These three observations seem to suggest that similarity can be a factor influencing a change of metaphors, in addition to the rank of entrenchment.

There are obvious points of overlap between the metaphors in each pair: if an entity disappears, it can be construed to no longer exist; a supernatural being can be a nuisance; an opponent can torture a person and a tormentor can be seen as an opponent. In other words, a change of metaphors is more likely to happen between more similar metaphors.

5. Conclusion and implications for metaphor translation studies

The main findings of this chapter can be summarized as follows. What will happen to a particular SL metaphor in translation, and what kind of treatment its metaphorical expressions will receive, is highly dependent on its entrenchment ranking in TL. If the metaphor occupies a high entrenchment rank in TL, it is more likely to be preserved in translation and its metaphorical expressions are more likely to be translated into metaphorical expressions of the same metaphor. If the metaphor is of a low entrenchment rank in TL, or not shared by TL, it is less likely to be preserved. It may either be translated non-metaphorically or changed into a metaphor that has a higher rank in TL.

This research has three implications. First, given that entrenchment matters in translation, a cognitive approach to translation studies should do more than distinguish between SMC and DMC (see section 1). For example, though 5[+][S] exists in both English and Chinese, it is frequently changed into other metaphors as a result of its low entrenchment rank in TL. The conceptual shift this involves was assumed to take place only in the DMC situation, i.e. the situation where SL and TL do not share a metaphor, but the present research offers a corrective to this assumption.

Second, since degree of entrenchment appears to play a significant role in metaphor translation, more quantitative analysis is needed to determine the degree of entrenchment of metaphors, which in turn calls for more corpus-based metaphor translation research.

Third, since not only metaphors but also their discrete specifications can differ in terms of entrenchment, it is essential to analyze metaphors at the more specific levels so that cross-cultural (and also diachronic) differences in metaphor can be captured more precisely.

In sum, a fine-grained, quantitative, corpus-based approach will greatly enhance research into metaphor translation.

Notes

1. ST and TT refer to the source text and the target text respectively.
2. As indicated above, in Stefanowitsch' (2006) method, target domain items are always nouns, so unless otherwise specified, the expressions we retrieved are those containing the noun forms of the lemma word, i.e. *fear* and *fears*.
3. In judging whether a metaphorical expression is translated as an expression of the same metaphor, we considered the most specific level only. In other words, for metaphors that have specifications, what we compared were the specifications rather than the general metaphors. For example, an expression of 2b[–][NS] is regarded as translated into an expression of the same metaphor if and only if its translated expression is still an expression of 2b[–][NS]. It is counted as translated into an expression of a different metaphor when its translated expression belongs to 2d[–][NS], though 2b[–][NS] and 2d[–][NS] are specifications of the same general metaphor.
4. We borrowed the term 'entrenchment' from Langacker (1987). As defined in Evans (2007: 73), it refers to "the establishment of a linguistic unit as a cognitive pattern or routine in the mind of an individual speaker." According to Langacker (1987: 59), linguistic structures and units fall along a continuous scale of entrenchment in cognitive organization, with the degree of entrenchment being closely related to the frequency of their occurrence, i.e. a linguistic structure or unit is more entrenched if it has a higher frequency of occurrence (see also Braine and Brooks 1995, Ambridge *et al*. 2008).
5. But there exists a different combination of specifications in the NCT, viz. the combination of 2a[+][NS] and 2g[+][NS], which seem to have a stronger tendency to combine with each other: five of the eight instances of 2a[+][NS] are also expressions of 2g[+][NS]; all five expressions of 2g[+][NS] are expressions of 2a[+][NS] as well.

References

Ambridge, B., Pine, J. M., Rowland, C. F. and Young, C. R. (2008), "The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument structure overgeneralization errors". *Cognition* 106: 87–129.

Al-Zoubi, M. Q., Mohammed, N. A.-A. and Al-Hasnawi, A. R. (2006), "Cogno-cultural issues in translating metaphors". *Perspectives: Studies in Translatology* 14(3): 230-239.

Baker, M. (1995), "Corpora in translation studies: An overview and some suggestions for future research". *Target* 7: 223-243.

—. (1999), "The role of corpora in investigating the linguistic behaviour of professional translators". *International Journal of Corpus Linguistics* 4(2): 281-298.

Braine, M. D. S. and Brooks, P. J. (1995), "Verb argument structure and the problem of avoiding an over general grammar", in M. Tomasello and W. E. Merriman (eds.) *Beyond Names for Things: Young Children's Acquisition of Verbs*, 352–376. Hillsdale, NJ: Erlbaum.

Cameron, L. (1999), "Identifying and describing metaphor in spoken discourse", in L. Cameron and G. Low (eds.) *Researching and Applying Metaphor*, 105-132. Cambridge: Cambridge University Press.

Crisp, P. (2002), "Metaphorical propositions: A rationale". *Language and Literature* 11: 7-16.

Deignan, A. (2005), *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins.

Evans, V. (2007), *A Glossary of Cognitive Linguistics.* Edinburgh: Edinburgh University Press.

Geeraerts, D. and Grondelaers, S. (1995), "Looking back at anger: Cultural traditions and metaphorical patterns", in J. R. Taylor and R. E. MacLaury (eds.) *Language and the Cognitive Construal of the World*, 153-179. Berlin: Mouton de Gruyter.

Gevaert, C. (2005), "The anger is heat question: Detecting cultural influence on the conceptualization of anger through diachronic corpus analysis", in N. Delbecque, J. van der Auwera and D. Geeraerts (eds.) *Perspectives on Variation: Sociolinguistic, Historical, Comparative*, 195-208. Berlin: Mouton de Gruyter.

Györi, G. (1998), "Cultural variation in the conceptualization of emotions: A historical study", in A. Athanasiadou and E. Tabaskowska (eds.) *Speaking of Emotions: Conceptualization and Expression*, 99-124. Berlin: Mouton de Gruyter.

Heywood, J., Semino, E. and Short, M. (2002), "Linguistic metaphor identification in two extracts from novels". *Language and Literature* 11: 35-54.

Koivisto-Alanko, P. and Tissari, H. (2006), "Sense and sensibility: Rational thought versus emotion in metaphorical language", in A. Stefanowitsch and S. Th. Gries (eds.) *Corpus-based Approaches to Metaphor and Metonymy*, 191-213. Berlin: Mouton de Gruyter.

Kövecses, Z. (1990), *Emotion Concepts*. New York: Springer-Verlag.

—. (2000), *Metaphor and Emotion: Language, Culture, and Body in Human Feeling*. Cambridge: Cambridge University Press.

—. (2002), *Metaphor: A Practical Introduction.* Oxford: Oxford University Press.

—. (2005), *Metaphor in Culture: Universality and Variation.* Cambridge: Cambridge University Press.

Lakoff, G. (1993), "Contemporary theory of metaphor", in A. Ortony (ed.) *Metaphor and Thought* (2nd ed.), 82-132. Cambridge: Cambridge University Press.

Lakoff, G. and Johnson, M. (1980), *Metaphors We Live By*. Chicago: University of Chicago Press.

Langacker, R. W. (1987), *Foundations of Cognitive Grammar* (Vol. 1): *Theoretical Prerequisites*. Stanford: Stanford University Press.

Laviosa, S. (1998), *The Corpus-based Approach: A New Paradigm in Translation Studies. Meta* 43(4): 474-479.

Maalej, Z. (2003), "Translating metaphor between unrelated cultures: A cognitive perspective". Unpublished paper, http://simsim.rug.ac.be /Zmaalej/transmeta.html

—. (2007), "The embodiment of fear expressions in Tunisian Arabic: Theoretical and practical implications", in F. Sharifian and G. B. Palmer (eds.) *Applied Cultural Linguistics: Implications for Second Language Learning and Intercultural Communication,* 87-104. Amsterdam: Benjamins.

Mandelblit, N. (1995), "The cognitive view of metaphor and its implication for translation theory", in M. Thelen and B. Lewandowska-Tomaszczyk (eds.) *Translation and Meaning, PART 3,* 482-495. Maastricht: Maastricht University Press.

Matsuki, K. (1995), "Metaphors of anger in Japanese", in J. Taylor and R. E. MacLaury (eds.) *Language and the Cognitive Construal of the World*, 137-151. Berlin: Mouton de Gruyter.

Olohan, M. (2004), *Introducing Corpora in Translation Studies.* London: Routledge.

Pragglejaz Group (2007), "MIP: A method for identifying metaphorically used words in discourse". *Metaphor and Symbol* 22(1): 1-39.

Schäffner, C. (2004), "Metaphor and translation: Some implications of a cognitive approach". *Journal of Pragmatics* 36: 1253-1269.

Semino, E., Heywood, J. and Short, M. (2004), "Methodological problems in the analysis of metaphors in a corpus of conversations about cancer", *Journal of Pragmatics* 36: 1271-1294.

Steen, G. J. (1999), "From linguistic to conceptual metaphor in five steps", in R. W. Gibbs and G. J. Steen (eds.) *Metaphor in Cognitive Linguistics*, 57–77. Amsterdam: John Benjamins.

—. (2002), "Identifying metaphor in language: A cognitive approach". *Style* 36(3): 386-407.

Stienstra, N. (1993), *YHWH is the Husband of His People: Analysis of a Biblical Metaphor with Special Reference to Translation*. Kampen: Kok Pharos.

Stefanowitsch, A. (2006), "Corpus-based approaches to metaphor and metonymy", in A. Stefanowitsch and S. Th. Gries (eds.) *Corpus-based*

Approaches to Metaphor and Metonymy, 1-16. Berlin: Mouton de Gruyter.

Stefanowitsch, A. and Gries, S. Th. (eds.) (2006), *Corpus-based Approaches to Metaphor and Metonymy*. Berlin: Mouton de Gruyter.

Yu, N. (1995), "Metaphorical expressions of anger and happiness in English and Chinese". *Metaphor and Symbolic Activity* 10(2): 59-92.

Zhang, H. (2000), "A comparative study on the formation and expression of Chinese and English emotion concepts". *Foreign Languages* 5: 27-32.

CHAPTER FOUR

SPECIALIZED COMPARABLE CORPORA IN TRANSLATION EVALUATION: A CASE STUDY OF ENGLISH TRANSLATIONS OF CHINESE LAW FIRM ADVERTS

YUN XIA, DEFENG LI

1. Introduction

Translation of pragmatic texts is essentially a decision-making process in which many people believe that the first priority is to achieve functional equivalence. Pragmatic texts generally belong to what Newmark (1988) defines as "informative" and "vocative" texts (Jia 2004: 3) and hence, the purpose of the overall translational action is to produce a target text that performs closest possible informative and vocative functions as the source text. For that purpose, the translator must be familiar with the genre conventions that the target text is to conform to. When the text conforms to the conventional patterns in the target language, the form of the translated text will not arouse the readers' attention and hence allows for an easier processing of the information contained in the text. On the other hand, when a text displays unfamiliar or unconventional form patterns, the audience may wonder why the translator has chosen these forms and whether they are meant to convey an extra amount of information (Nord 2007). This chapter reports on the construction and use of specialized corpora in a study of English translations of Chinese advertisements in comparison to original English advertisements, with the aim to show whether and how specialized comparable corpora can be used to inform translation evaluation.

2. Comparable texts and comparable corpora

The comparable corpus herein refers to a monolingual comparable corpus which “consists of a corpus of translations and comparable non-translations in the same language” (Olohan 2004: 35). The authentic, non-translated texts are chosen from the target language repertoire and represent the genre the target text is supposed to belong to. The Translational English Corpus (TEC), the first corpus of this kind, is designed and constructed under the direction of Mona Baker. The corpus consists of a collection of texts translated into English from a range of different source languages. It is usually used in comparison with a comparable set of non-translational texts taken from the British National Corpus (BNC).

The aim of most research using a comparable corpus of this kind is to capture “patterns which are either restricted to translated text or which occur with a significantly higher or lower frequency in translated text” (Baker 1995: 235), which can help find out about “the nature of translated text in general and the nature of the process of translation itself” (Baker 1995: 236). Baker (1996), for example, posits a number of features (e.g. explicitation, simplification, normalization) as “universal features of translation”. Following Baker, researchers such as Laviosa (1998), Olohan and Baker (2000), Puurtinen (2003), Tirkkonen-Condit (2004), and Hu (2007), among others, have made further efforts in developing ways to test as well as to propose new hypotheses about translation universals. Corpus-based comparative studies of translated versus non-translated texts have also proved useful in investigating translators’ styles (Baker 2000), and translation norms (Hu 2006) in specific socio-cultural contexts, which provide us with better insights into the nature of translation.

Translation of pragmatic texts, however, can also benefit from the comparable corpus-assisted approach in that the authentic, non-translated texts constitute an important type of “auxiliary texts” (Nord 2007:20) which can serve as a source of cultural and linguistic information for translators, and that it not only provides a translator with improved access to the appropriate conceptual and linguistic information of a specialized subject field as documented by experts in that field, but, more importantly, helps him learn to analyze the culture-specific features of textual and other communicative conventions in two cultures. This ability might be called “contrastive text competence” (Nord 2007: 19). In addition, the quantitative analysis of particular lexical and grammatical features in translated and non-translated texts makes it possible to investigate the characteristics of translated texts (i.e. the so-called ‘translationese’). It therefore constitutes an important means in translation evaluation, an important task in practical translation and translator training, which has often been made difficult by the subjectiveness in judging what constitutes an acceptable or adequate translation, as evaluators often differ in views on the criteria for good translation. We therefore would like to propose and show that specialized comparable corpora have a role to play in the assessment as well as improvement of translations.

3. Text selection

For the present study, a small corpus was built, consisting of ten English translations of Chinese law firm advertisements and ten non-translated English law firm advertisements. All texts are downloaded from the internet and saved in plain text format as required by the corpus-processing software Wordsmith Tools 4.0. The two collections are rather small, with only 3,143 tokens in the translated texts and 2,584 tokens in the non-translated texts. They are comparable in text category (advertisements) as well as primary functions (both intended to perform the same informative and vocative functions for receivers familiar with the English language and culture).

The data are lemmatized, hence enabling the treatment of inflected forms of a word as belonging to the same base form or lemma, and this is required for a more accurate word frequency analysis as well as keyword analysis. Our purpose is to find out how the two collections differ in their textual functions. However, as the size of the corpus is rather small, any conclusions are necessarily tentative, and the discussions that follow are mainly illustrative, which are intended to show the potential usefulness of a comparable corpus in translation evaluation.

4. Results and discussions

Wordsmith Tools 4.0 is used in the data processing to generate statistical descriptions of the two sets of texts. Such descriptions include type-token ratios, top-frequency lexical words, keywords, and word clusters, which will be elaborated and discussed particularly in relation to the evaluation of the translated advertisements in the remainder of this chapter.

4.1. General text style

The style of a text is one form of manifestation of its function as different styles might produce different effects in the addressees. For instance, a formal style often implies seriousness and authority while an informal style is usually preferred for casual occasions and implies intimate relationship between two parties. Corpus analysis can help to find out differences in text styles between translations and non-translations. As shown in Table 4-1 showing basic statistics generated using Wordsmith Tools, the number of sentences in both translated (TT) and non-translated texts (NTT) is almost the same, but the average sentence length of the translated texts is much higher than that of the non-translated texts. Generally, longer and more complex sentences indicate a more formal, serious and less emotional style whereas shorter and simpler sentences often give an impression of a concise, casual and dialogic style. This dialogic feature in the non-translated texts implies the addressers' intention to shorten the psychological distance between the potential addressees and themselves, which contributes to the 'appellative' or 'vocative' function of the texts.

Table 4-1. Type-Token ratio and sentence length

Text File	Overall	TT	NTT
Bytes	37,909	20,850	17,059
Tokens	5,727	3,143	2,584
Types	1,501	931	889
Type/Token Ratio	26.21	29.62	34.40
Ave. Word Length	5.42	5.45	5.43
Sentences	241	121	120
Sent. length	23.76	25.98	21.53

Lexical complexity also differs between the two collections of texts. Type/Token ratio (TTR) is a method of measuring the lexical complexity of a text, and the TTR of the translated texts is 29.62, much lower than that of the non-translated texts, indicating a higher degree of lexical repetition in the former and higher lexical variation in the latter. This is further supported by word frequency data shown in Table 4-2. The top six most frequent lexical words take up a greater proportion in the translated texts (9.26%) than in the non-translated texts (7.69%), which shows that these words are more frequently repeated in translated texts.

Table 4-2. High-frequency lexical words

	TT			NTT		
N	Word	Freq.	%	Word	Freq.	%
1	Firm(s)	77	2.45	Client (s)	53	2.05
2	Law	73	2.32	Attorney /Lawyer(s)	36	1.39
3	Legal	53	1.69	Firm(s)	33	1.28
4	Lawyer/ Attorney(s)	36	1.15	Law	28	1.08
5	Service	30	0.95	Legal	28	1.08
6	Client(s)	22	0.70	Service	21	0.81
	Total	291	9.26		200	7.69

4.2. Text informativity

The high-frequency lexical words usually characterize the aboutness of a text or corpus, and as Table 4-2 shows, the high-frequency words in both collections are almost the same except for a slight difference in the choice of *attorney* and *lawyer*. The non-translated texts tend to make a greater use of *attorney* than the translated ones, which, as we find, is mainly due to the greater proportion of U.S. texts in our collection. This leads us to believe that both types of texts are intended for the same informative function—both seem to focus on the same topics.

In order to make a closer examination of the two collections of texts, we also carry out a keyword analysis. In the Wordsmith Tools program, keywords are defined as words whose frequency is unusually high in comparison with some norm. Keywords were identified by comparing two word lists which were created using the program's wordlist tool, and log likelihood statistical test was used to calculate keyness. Keywords which are significantly more frequent in the study corpus than in the reference corpus are called 'positive keywords', and those significantly more infrequent ones are called 'negative keywords'. In our corpus study, the two subcorpora are compared with the non-translated subcorpus acting as a reference corpus, thus representing the norm of texts of the same kind, for the translated one. Such a comparison produces a list of words quantitatively most different in the two corpora (see Table 4-3). Proper nouns which denote the location of the firms and are thus less significant in the comparison, e.g. *Beijing*, *Louisiana* were eliminated. The results show that words significantly more frequent in the translated texts are *China*, *university(ies)*, *firm(s)*, *foreign* and *has*, with the last three in the table being negative keywords, i.e. words significantly less frequent in the translated texts.

Table 4-3. Keywords

N		Word	fr.	TT (%)	fr.	NTT (%)	Key-ness	P
positive	1	China	19	0.60	0		22.9	0.000002
	2	university (-ies)	18	0.57	0		21.6	0.000003
	3	firm(s)	77	2. 45	33	1.28	16.8	0.000041
	4	foreign	13	0.41	0		15.6	0.000077
	5	has	26	0.83	4	0.15	14.1	0.000176
negative	6	client(s)	22	0.70	53	2.05	20.3	0.000007
	7	we	13	0.41	41	1.59	21.4	0.000004
	8	our	19	0.60	92	3.56	68.9	0.000000

Table 4-4. High frequency 2-word clusters

	TT		**NTT**	
	2-word cluster	**Freq.**	**2-word cluster**	**Freq.**
1	law_firm(s)	45	our_client(s)	29
2	the_firm	26	law_firm(s)	12
3	real_estate	13	our_attorneys	9
4	legal_service(s)	11	the_firm	9
5	intellectual_property	7	our_firm	6
6	law_service	7	legal_services	4

According to this keyword analysis, there seems to be a contrast in the type of information emphasized in the two collections of texts. In the non-translated texts, the two parties of the communicative action—the addresser (here realized as *we*, *our*) and the addressee (*clients*)—occur significantly more frequent, thus foregrounding their close relationship. This finding is further supported by a word cluster analysis carried out on the corpus. Clusters are words which are found repeatedly together in each others' company in sequence. They represent a closer relationship than collocates, more like multi-word units or groups or phrases (Scott 2006). Table 4-4 shows the most frequent meaningful 2-word clusters in the two collections of texts.

As can be seen in the table, the most frequent meaningful two-word cluster in the non-translated texts is *our clients*, while that of the translated texts is *law firm*. In addition, the translated texts tend to stress a close relationship with the government and educational institutions, as can be seen through the concordance of two words *China* and *university* (see Appendices I and II).

The frequent collocation of the word *China* with government agencies and well-known universities, designed to create a better image of the enterprises, shows the addressers' intention to establish authority by stressing official authorization or permission, their close relationship with the government, and their social influence, as can be seen in examples 1a) and 1b).

> 1a) xxx Firm is a partnership law firm registered in Beijing with the approval from the direct government agency of **China**…
> 1b) xxx is able to maintain a close relationship with **China**'s legislative department and government.

However, to the addressees in the target English culture, a close relationship with the government has little to do with the efficiency and competitiveness of the enterprise. Furthermore, approval from the government agencies is a necessary condition for the legal operation of an enterprise, that is to say, the information is already presupposed and therefore redundant in the translations. According to Beaugrande and Dressler (1981), this type of information, which is already known to the readers, should be "first order of informativity", of less significance and even boring to the readers, thus indirectly affecting the vocative function as well as the acceptability of the translations.

The word *university* is mainly used in two contexts, one for describing the educational qualifications of the staff in the firm, the other for illustrating the enthusiasm of the said firm in supporting educational enterprises. Look at the two examples in 2a)-2b).

> 2a) The head partners and lawyers graduated from the top domestic **universities** such as Tsinghua **University**, Peking **University**…

> 2b) xxx firm donated 100,000 yuan to set up Scholarships in The East China **University** of Politics and Law.

While the participation of an enterprise in social welfare events serves to show their strength, competitiveness and public spirit, and promotes their image and reliability to some degree, the frequent mention of China's universities may add to the informativity of the text. It might even become a burden on the comprehension process, since the university names might not be part of the target text receivers' cognitive context and therefore will not activate their schemata. The following passage is a typical example.

> 3a) In 2003, xxx Firm donated 1,200,000 yuan to set up Teaching Awards and Scholarships in its name on its 10 anniversary in 6 law schools in Peking **University**, Qinghua **University**, China People's **University**, China **University** of Political Science and Law, Jilin **University** and Heilongjiang **University** respectively. In 2004, the firm donated 100,000 yuan to set up Teaching Awards and Scholarships in the First middle **school**, Hailun city, Heilongjiang province. In 2005, the firm donated 100,000 yuan to set up Scholarships in East China **University** of Politics and Law….

Readers from different cultural backgrounds are usually constrained by their knowledge structure and culture differences in their comprehension of textual information, which might be significant for the source text readers and completely the opposite for the target text readers. In practical translation, comparable corpus analysis may help the translator determine which kind of text will best suit the target audience, and whether a text needs to be adjusted in order to meet the new audience's information needs and cultural expectations. In the case of the above translated text, the translator might employ the technique of omission to make it more comprehensible to the target reader with a similar effect:

> 3b) xxx Firm has set up Teaching Awards and Scholarships in middle schools as well as law schools of many well-known universities in support of education in China.

4.3. Internal focalization vs. external focalization

The term 'focalization' is borrowed from the field of narratology where, as Genette (1980: 186) claims, it corresponds to the expression 'focus of narration', or the commonly used term 'point of view', referring generally to "the psychological perspective through which a story is told.

In the context of narrative fiction, it encompasses the narrative framework which a writer employs, whether this be first person or third person, and accounts for the basic position which is adopted in a story" (Simpson 1993: 4). Focalization might be internal, when narration is presented from the point of view within a character's consciousness, manifesting his or her feelings about and evaluations of the events and characters of the story, or external, when the story is described objectively from a position outside of any of the protagonists' consciousnesses. First person narration is considered as a typical criterion for internal focalization (Barthes 1975: 262).

Here in the context of advertising discourse, we take 'internal focalization' to mean description from the point of view of the enterprise as one party of the communicative event, whereas 'external focalization' refers to the point of view of the advertiser as an outsider engaged in an introduction of the enterprise.

Apart from the informative function, keyword analysis also suggests a difference in the vocative function of the two sets of texts as a result of a variation in focalization, as mainly shown in the different use of referential terms which are often good markers of focalization.

Concordance of the keyword *firm(s)*(see Appendix III) shows that 61 out of 77 occurrences of the word *firm(s)* in the translated texts collocate with the names of the enterprises or the determiner *the* to refer to the enterprises. This typical way of reference reflects the objective external point of view of the advertiser. That is, the advertiser's mentioning of an enterprise by its name or by the third person referential term *the firm* introduces an objective description about it, thus making the readers adopt an outsider's point of view in reading the text. So, the advertisement seems to be presented from some authority who is trying to convince the readers of its objectivity and reliability, resulting, however, in an increase in the psychological distance between the reader and the advertiser.

In contrast, there are only 33 occurrences of *firm(s)* in the non-translated texts, among which only 11 are used as direct reference to the enterprise, and six collocate with *our* as self-reference. First person reference is typically employed in the non-translated texts, as manifested by the abundant use of *our* and *we*, signalling internal focalization from the advertisers' point of view. The advertisers seem to be merged with the enterprises, and the 26 cases of collocation of *our* with *client(s)* imply the close relationship between the enterprises and their customers. This is in strong contrast with the translated texts in which *client(s)* mostly go with *its* or *the*. Compare the examples in 4) and 5):

> 4a) **xxx Law Firm** contrives to provide quality and tailored services to **its clients** with team forces… (TT)

> 4b) **We** also provide **our** clients with a range of private client services… (NTT)
>
> 5a) **The Firm** compensates **the clients** for the loss if **their** interests were damaged due to **its** lawyer's serious faults. (TT)
> 5b) For additional information on … please contact **us** and **we** will be pleased to assist you… (NTT)

The use of personal pronoun *we* and *our* implies the existence of 'you', which adds to the dialogic effect of the texts. In addition, the frequent use of *we* as well as *our* helps to transfer the reader subliminally to the internal point of view of the advertiser, resulting in the former's identification with the latter and therefore a decrease in the psychological distance between the two.

The primary principle determining any translation process is the purpose of the overall translational action (Nord 1997: 27). Although intended for the same ultimate purpose (to appeal to text receivers), the two sets of texts demonstrate different perspectives in fulfilling this task, which is mainly due to the differing conventions of the same genre in Chinese and English contexts. Since genre conventions are mostly culture-specific, they play an important role in functional translation. If a target text is to be acceptable as representative of a target-culture genre, the translator has to be familiar with the conventions that the target text is to conform to (Nord 1997: 54), and should be able to determine whether a text needs to be re-targeted for a new audience in order to achieve functional equivalence.

4.4. Idiomaticity and appropriacy

Word clusters can be used as a measure of idiomaticity of translations, and help to locate unnatural expressions, since a word cluster that occurs highly frequently may imply that the cluster is formulaic and thus can enhance idiomaticity and fluency (Wray 2002). In the field of law practice, clusters such as 法律服务 'legal services', 知识产权 'intellectual property' may be considered formulaic chunks of language. However, the phrase 法律服务 is seven times translated as 'law service' (see Table 4-4), for which no match is found in the non-translated texts. A further query of the unit into the BNC still produces no hits.

Keyword analysis can also indicate what might be considered inappropriate language use in translation texts. As Table 4-3 shows, the word *has* occurs significantly more frequently in the translated texts, and

21 of the 26 instances are used as markers of the perfect aspect. This might be caused by the fact that English is predominantly a tense language, whereas Chinese is exclusively an aspect language (c.f. Wang 1985: 151, Li and Thompson 1981: 184-237). The frequently used aspect marker 了 *le* to indicate a change in state might lead the translator to translate it into a perfect aspect in English.

5. Conclusions

In this chapter we set out to demonstrate how we might use a comparable corpus as an aid to assess Chinese-English translation of pragmatic texts. We believe that small specialized corpora can be effective in resolving issues pertinent to genre-specific languages. The corpus does not need to be sophisticated in terms of syntactic or semantic tagging. In fact it can be structurally very simple, text-only corpora, but they can nevertheless be very useful in the actual decision-making process in translation.

Translators may compile and use such a corpus to look for stylistic information (through general statistics), special field terms (through wordlist) and idiomatic expressions (through word cluster analysis). Corpus analysis techniques like keyword analysis may also help in translation evaluation, that is, they can help translators to find out how and to what extent their translations differ from non-translations in the informative and vocative functions. If translators fail to find solutions to these types of information needs, a mismatch may arise between translators' competence and their performance (see also Varantola 2003). On the basis of the statistical data gathered with the assistance of corpus tools, it is possible to probe into the possible causes (e.g. extra-linguistic factors) for these problems and improvement can be made if such factors are taken into serious consideration and constructive measures are adopted.

Compared with the traditional method of mainly subjective assessment, a computer-aided approach has obvious advantages. As the World Wide Web provides practically unlimited access to electronic texts that can be used in compiling individual disposable corpora for translation, there is now less concern about the cost efficiency of building a corpus. In fact, we believe that the knowledge of how to compile and use corpora for translational and evaluative purpose should be an essential part of modern translational competence and should therefore be an important component in the training of prospective professional translators.

References

Baker, M. (1995), "Corpora in Translation Studies: An overview and some suggestions for future research". *Target* 7(2): 223-243.

—. (1996), "Corpus-based translation studies: The challenges that lie ahead", in H. Somers (ed.) *Terminology, LSP and Translation: Studies in language engineering. In honour of Juan C. Sager*, 175–186. Amsterdam: John Benjamins.

—. (2000), "Towards a methodology for investigating the style of a literary translator?". *Target* 12(2): 241-266.

Barthes, R. (1975), "An introduction to the structural analysis of narrative". *New Literary History* 4(2): 237-272.

De Beaugrande, R. A. and Dressler, W. U. (1981), *Introduction to Text Linguistics*. London and New York: Longman.

Bowker, L. (2001), "Towards a methodology for a corpus-based approach to translation evaluation", *Meta* 46(2): 345-364.

Genette, G. (1980), *Narrative Discourse*. Oxford: Basil Blackwell.

Hu, X. (2006), *A Corpus-based Study on the Translation Norms of Contemporary Chinese Translated Fiction*. PhD thesis, East China Normal University.

—. (2007), "A corpus-based study of the lexical features of Chinese translated fiction". *Foreign Language Teaching and Research* (3): 214-221.

Jia, W. (2004), *A Functional Approach to Pragmatic Translation*. Beijing: China Translation and Publishing Corporation.

Laviosa, S. (1998), "Core patterns of lexical use in a comparable corpus of English narrative prose". *Meta*, 43(4): 557-570.

Malmkjær, K. (2004), "Translational stylistics: Dulcken's translations of Hans Christian Andersen". *Language and Literature* 13(1): 13-24.

Li, C. and Thompson, S. (1981), *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Newmark, P. (1988), *A Textbook of Translation*. New York and London: Prentice-Hall.

Nord, C. (1997), *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.

—. (2007), "Looking for help in the translation process - The role of auxiliary texts in translator training and translation practice." *Chinese Translators Journal* 1: 17-26.

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

Olohan, M and Baker, M. (2000), “Reporting ‘that’ in translated English: Evidence for subliminal processes of explicitation”. *Across Languages and Cultures* 1(2): 141-158.

Puurtinen, T. (2003), “Nonfinite constructions in Finnish children’s literature: Features of translationese contradicting translation universals”, in S. Granger, J. Lerot and S. Petch-Tyson (eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, 141-154. Amsterdam: Rodopi.

Simpson, P. (1993), *Language, Ideology, and Point of View*. London and New York: Routledge.

Tirkkonen-Condit, S. (2004), “Unique items - Over- or underrepresented in translated language?”, in A. Mauranen and P. Kujamaki (eds.) *Translation Universals: Do They exist*?, 177-186. Amsterdam: John Benjamins.

Varantola, K. (2003), “Translators and disposable corpora”, in F. Zanettin, S. Bernardini and D. Stewart (eds.) *Corpora in Translator Education*, 55-70. Manchester: St. Jerome Publishing.

Wang, L. (1985), *Modern Chinese Grammar*. Beijing: The Commercial Press.

Wray, A. (2002), *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Appendix I. Concordances of *China* in translations (19)

law firm conforms with the situation of	**China**'s	reform and opening to the outside worl
rosper in the rapid developing market of	**China**.	Initiative Law Firm is authorized to en
research regarding investment projects in	**China**,	as well as good fame and credit investi
val from the direct government agency of	**China**.	Gold Maxim Law Firm is composed of
become the best legal service provider in	**China**	and has already received recognition in
able to maintain a close relationship with	**China**'s	legislative department and government
e larger-scale comprehensive law firms in	**China**.	Since its establishment in 1999 in Gua
n Peking University, Qinghua University,	**China**	People's University, China University
ban infrastructure in People's Republic of	**China**	(herein after, PRC or China). During it
ch of the Public Policy Council under the	**China**	Policy Science Research Association.
e as a top international law firm is bright.	**China**'s	legal and social reforms, as well as its
aduated from top-tier law schools, both in	**China**	and overseas. More than 80% of them
nfidential references, Law Service Times,	**China**	Reform Daily, and China Consumptio
Peking University, Tsinghua University,	**China**	University of Political and Law, Natio
proval of Guangdong Bureau of Justice in	**China**.	Since our foundation, we have dealt wi
d among the top 20 Chinese law firms by	**China**'s	Ministry of Justice. In 2005, Tian Yua
international clients and governments in	**China**	for protection of their legal rights. The
lso, the lawyers gave the lectures in East	**China**	University of Polities and Law, Shang
nce with the laws of Peoples Republic of	**China**	and has been recognized and awarded

Appendix II. Concordances of *university(ies)* in translations (18)

he master degree of legal science and the	**university**	graduates of legal science constitu
p domestic universities such as Tsinghua	**University**,	Peking University, etc. Most of ou
ties such as, Tsinghua University, Peking	**University**,	etc. Most of our partners have stu
s appointed by the Law School of Peking	**University**	and the Law School of Qinghua U
niversity and the Law School of Qinghua	**University**	as a part-time advisor for Master
r Master degree law students and by Jilin	**University**	as an adjunct Professor. With over
0 anniversary in 6 law schools in Peking	**University**,	Qinghua University, China People
w schools in Peking University, Qinghua	**University**,	China People's University, China
sity, Qinghua University, China People's	**University**,	China University of Political Scie
versity, China People's University, China	**University**	of Political Science and Law, Jilin
ersity of Political Science and Law, Jilin	**University**	and Heilongjiang University respe
d Law, Jilin University and Heilongjiang	**University**	respectively. In 2004, the firm do
to set up Scholarships in The East China	**University**	of Politics and Law. In 2006, in or
to set up Scholarships in the Zhongshan	**University**.	In the future, the firm will donate
scholarships in the Law School of each	**university**	to reciprocate the society. Establis
awyers graduated from the top domestic	**universities**	such as, Tsinghua University, Pek
s gathered talents trained by well-known	**universities**	of the country. The qualification f
ee or above from law specialty of regular	**universities**	with lawyer certification; two rec

Appendix III. Concordances of *firm* in translations (20/77)

ave accumulated rich experiences. It is a law	**firm**	of high quality, which is capable of
erty and insurance etc. The office area of the	**firm**	are nearly 700 square meters and th
services. Guangdong Pengcheng Sunny Law	**Firm**	(original name "Shenzhen Sun
Firm, original name Shenzhen Sunny Law	**Firm**,	was established in October
rge-scale partnership firms in Shenzhen. The	**firm**	is located at 26/F, Xiwu Building,
through ten year's development, Sunny Law	**Firm**	has established a professional tea
mpany registration and so on. Since 2000, the	**firm**	has successfully handled several do
led hundreds of cases successfully. Suhu law	**firm**	conforms with the situation of Chin
n income and expenditure. In 1988, Suhu law	**firm**	was restructured into a law office
ice of the cooperative stock system. Suhu law	**firm**	has started from scratch, made ardu
takings and obtained continuous growth. The	**firm**	environment for the lawyers has b
ly, and the comprehensive strength of the law	**firm**	has been greatly strengthened. Suhu
firm has been greatly strengthened. Suhu law	**firm**	sets the "pursuit of the maximizatio
dly for the clients. For ten-odd years, the law	**firm**	has handled a great amount of larg
ust from the clients. The lawyers of Suhu law	**firm**	have been invited to act as their per
e Unites States, the headquarters of Suhu law	**firm**	in Nanjing has formed a network.
rm in Nanjing has formed a network. The law	**firm**	has more then 30 registered lawye
in hands with and offer assistance to the law	**firm,**	which has become a comprehensive
firm, which has become a comprehensive law	**firm**	with economic trade, finance an
elatively great abundance of honor. Suhu law	**firm**	has won the titles of "Civilized law

Appendix IV. Concordances of *firm* in non-translations (26/33)

uge, Louisiana Lawyers at the Babcock Law	**Firm**	are ready to help you with just
a Attorneys or have a question about our law	**firm**	or information about forming an L
ers can't help you, we will recommend a law	**firm**	that can. Baker & McKenzie is the
McKenzie is the world's leading global law	**firm**.	We have provided sophisticated l
t. The lawyers and other professionals in our	**firm**	are citizens of more than 60 countr
on. We are a family-owned and managed law	**firm**	ocated in Pittsburgh, PA. Our f
aged law firm located in Pittsburgh, PA. Our	**firm**	concentrates in Creditors' Rights,
P is considered a pre-eminent Vancouver law	**firm**.	We provide effective and practic
we will be pleased to assist you. McNair Law	**Firm**,	P.A. was founded in Columbia,
TO EXCELLENCE Without exception, our	**firm**	is committed to the pursuit of ex
the pursuit of excellence. All attorneys in our	**firm**	serve on various charitable, civic,
e is a simple, though multifaceted, key to the	**firm**'s	success: capability, capacity and
s of expertise and involve attorneys from the	**firm**'s	other areas of practice to provide
most talented professionals and to make the	**firm**	one of the finest places to work. Pi
our client service to new heights. Rose Law	**Firm**	is one of the largest and the oldest
l standards that have been established by our	**firm**	and the tradition of excellence whi
al representation comes from good people. A	**firm**	is known by its attorneys. Althoug
the most qualified lawyers to practice in our	**firm**,	technical competence is only part
of what makes a good lawyer or a quality law	**firm**.	Greatness requires depth, perspect
ess and litigation legal services. A mid-sized	**firm**	located in downtown San Diego's S
and financial services sector. In addition, the	**firm**	also provides substantial litigation
the key factors of the service provided by the	**firm**	is the depth of experience and level
ps and developing solutions to problems. The	**firm**'s	priorities are focused on the need
on Law Advisors is a Seattle-based boutique	**firm**	of seasoned corporate, transactio
e traditional "pyramid" structure. Instead, the	**firm**	is comprised primarily of experien
st, most respected, and most trustworthy law	**firm**.	Our clients can feel confident abo

CHAPTER FIVE

THE SPECIFICITY OF TRANSLATOR'S NOTES: TEXTOMETRICAL ANALYSIS OF THE FOOTNOTES IN FU LEI'S TRANSLATION OF *JEAN-CHRISTOPHE* BY ROMAIN ROLLAND

JUN MIAO, ANDRÉ SALEM

1. Introduction

This chapter focuses on an important figure of Franco-Chinese literary translation: Fu Lei (傅雷, 1908-1966). Thanks to him, Chinese readers have become acquainted with – and today still have access to – Western, mainly French, literature. The translations of Fu Lei are regarded as genuine literary masterpieces, and his style of translation is so remarkable that Chinese readers call it the “Fu Lei style”.

When reading Fu Lei's translation of *Jean-Christophe*, one is easily impressed by the abundance of prefaces and footnotes.[1] What is the link between his style of translation and the notes? In fact, the translator's note is an important issue in translation studies, because it is linked with many issues when we discuss the translator's legitimate place, his visibility and the issue of translatability.

Many scholars and researchers in translation studies have been engaged in studies and reflections on the subject of the translator's note. In the article *De l'érudition à l'échec: la note du traducteur*, Jacqueline Henry (2000) systematically examines many sides of this subject: linguistic, typographic, historical, and functional aspects of translator's notes. From stylistic and literary angles, Christelle Bahier-Porte (2005) provides her interesting studies on the notes in the Eastern stories *The Thousand and One Nights* and *The Thousand and One Days*. But it is a pity to note that these existing studies are often limited to citations of notes scattered in the works, and there are few studies using computational corpus-based methods, despite the increased importance of this approach in translation studies ever since its first introduction by Mona Baker (1993). We should also mention the article by Jennifer Varney (2005) which deals with taboo subjects in translation by examining translators' notes in a large corpus: Italian translations of Anglo-American fiction from 1945 to 2005. However, her work still uses traditional methods.

In this present chapter, we focus on an empirical study of the footnotes in Fu Lei's translation, inspecting in detail what the translator puts into the footnotes. To do this, our research relies on textometrical methods (also called "lexicometrical" methods), i.e. a division of the text into units and an analysis of juxtapositions and co-occurrences to capture the meaning of the text (cf. Lebart and Salem 1994). The corpus used in the present work is the source text and Fu Lei's translation of *Jean-Christophe* by Romain Rolland. This novel appeared for the first time sequentially in the *Cahiers de la Quinzaine* between February 1904 and October 1912. Mainly for this novel and the collection *Au-dessus de la mêlée*, Romain Rolland (1866-1944) received the Nobel Prize in Literature 1915. The electronic version of the whole work is available online.[2] The first complete publication of *Jean-Christophe* in Chinese, published between 1937 and 1941, was translated by Fu Lei using the original version of 1926 from the editor of *Librairie Ollendorff*. *Jean-Christophe* was edited many times in varying versions and unfortunately, we do not have access to the version that Fu Lei based his translation on. We note that our French corpus lacks the prefaces of the first edition of Romain Rolland for Volume 1 and Volume 4 that Fu Lei translated. Our electronic version in Chinese is also available online,[3] but we have corrected the flaws in this electronic version according to the 1998 paper edition by the Anhui Literature Publishing House which in turn is based on Fu Lei's complete re-translation collected in 1957 by the People's Literature Publishing House. The size of the whole corpus is about 1.2 million words.

In the processing of our parallel corpus, we used the automatic word segmentation tool *ICTCLAS* (Institute of Computing Technology, Chinese Lexical Analysis System, Chinese Academy of Sciences) for the Chinese part of the corpus.[4] Then, we made use of the *Alignator* program to realize a semi-automatic alignment of the French and the Chinese texts.[5] In order to facilitate the comparison of the French and Chinese texts, we changed all capital letters to the lowercase in the French text, and the Chinese punctuations to their European forms in the Chinese text, among some other adjustments in preprocessing.[6] We use the abbreviation "RR" to refer to the original French text, and "FL" for the Chinese translation by Fu Lei. Note also that we have not included in our present experiment the prefaces, which will be the subject of further research. We use two programs in our experiment: one is *Lexico3*,[7] and the other is *Alignoscope*.[8]

Besides the content of the footnotes, it is important to study their quantity, their locations, and their functions. We will address these issues one by one. Following this introduction, the second section will analyze the reason for which Fu Lei wrote the footnotes, and we will deal with this question in light of the theory of translation studies and especially in the examination of Fu Lei's personal experiences in the social context.

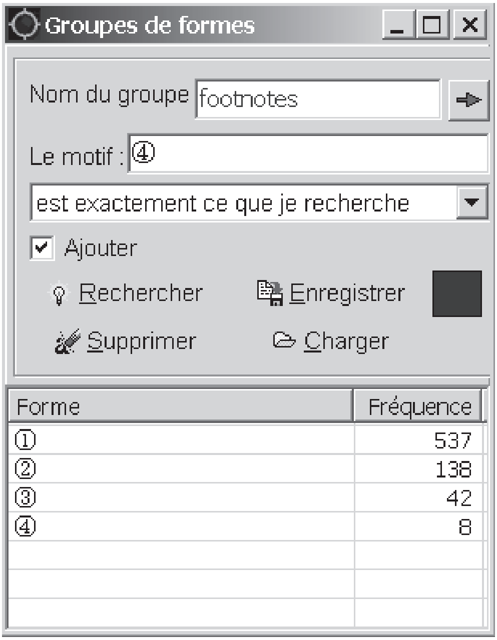
2. The results of the experiments

First of all, we intend to give the exact number of footnotes in Fu Lei's translation. It seems not very difficult to count the footnotes, at least for a small book when the footnotes are not too frequent. For example, Henry (2000: 232) found 23 translator's notes in the French translation for the English book *Small World* by David Lodge (1985), translated by Mauritius and Yvonne Couturier (1991). But it is cumbersome to search, record, and count all the footnotes scattered in *Jean-Christophe*, a book containing 10 volumes and 1801 printed pages.[9]

The electronic version of the corpus and *Lexico3* allow us to overcome the counting problem. The tool *Form Groups* (tags) allows us to combine occurrences of different graphics into one group. In order to find the footnotes, we create the group consisting of the symbols used for footnotes in the translation, i.e. ①, ②, ③, ④.[10]

We do not take into account the numbers used for the footnotes as they only differentiate between footnotes on a single page. Note also that the frequency count shown in Figure 5-1 counts the actual symbols and since the same symbol appears in the position to annotate and in the beginning of the corresponding footnote, each footnote is counted twice. So, the actual number of notes is half of the frequency shown in Figure 5-1. We might think that the frequency should be an even number, but the footnote marker ① has an odd frequency (537). This is due to the fact that at one occasion, Fu Lei refers inside a footnote to another footnote labelled ①.

By computing (536 +138 +42 +8) / 2 = 362, we obtain the precise number of footnotes that Fu Lei used in the translation of *Jean-Christophe*: 362, corresponding to about one note per five pages.



Nom du groupe: footnotes

Le motif : ④

est exactement ce que je recherche

☑ Ajouter

Forme	Fréquence
①	537
②	138
③	42
④	8

Figure 5-1. Form group of footnotes in Fu Lei’s translation

In order to tackle the textometrical analysis of footnotes, we collect, by seeking the footnote markers, and copy-pasting, all the footnotes in one file, called “note”. Then, we compare it with the text body without the footnotes. In the following, we present the basic properties of these files:

Table 5-1. General properties of footnotes in Fu Lei’s translation

Part	Tokens	Types	Hapax	Most frequent word	Frequency
Note	10,359	3,244	2,091	的	464
Text	572,127	19,055	6,182	的	47,454

Table 5-1 shows that the footnotes contain 10,359 word tokens, including 3,244 different word types and 2,091 hapaxes. With the total number of 362 footnotes, we can see that one footnote includes an average of 28.6 words. The most frequent word in the footnotes and in the text is both 的 (*de*, an auxiliary word), but there is a sharp contrast in its frequency in the text and footnotes, accounting for 4.48% and 8.29% respectively.

It is important to note that among the footnotes in Fu Lei's translation, there are also the original footnotes, translated by Fu Lei into Chinese. But for these notes, Fu Lei put "–原注" (*yuanzhu* 'original note') after each footnote to emphasize that it was the author who wrote it. This mark allows us to obtain easily the information about Romain Rolland's footnotes using the same method as before. We find that the original has 12 notes with 145 word tokens, 82 different types, and 59 hapaxes.

Thus, we discovered that the notes of the translator, excluding the author's original notes (10359 – 145) / 572127= 0.0178, occupy approximately 1.8% of the translation.

2.2. Where are the footnotes?

After this general quantitative analysis we will turn to the locations of the footnotes in the 10 volumes of the translation. How are they distributed? What words in the text have resulted in the use of footnotes?

2.2.1. Distribution of footnotes

To answer the questions mentioned above, we are launching, in a first step, the research on the footnote distribution in the translation, with the use of the form groups (①, ②, ③, ④) and statistics calculation (PCLC) in Lexico3 (Lamalle *et al*. 2003).[11]

As shown in Table 5-2, the number on the top of the table indicates the volume number. Under each volume number, there are two columns of numbers; the left one shows the frequency of each form (the footnote marker) and the right is each form's specificity.[12] The last row shows the total frequency of footnote markers in each volume.

It is easy to notice that volume 5 (entitled *La foire sur la place*) contains the highest number of footnotes, with 136 (206/2) occurrences. And volume 3 (entitled *Adolescence*) contains the least footnotes, with only 4 footnotes in the volume. So, Table 5-2 indicates that the distribution of footnotes is not equal among the volumes.

Figure 5-2 sketches the specificity of footnotes in each volume. The highest bar upturned indicates that volume 5 has the largest number of footnotes, with the specificity of about +31.[13] In contrast, the lowest bar downward suggests that volume 3 has the least footnotes, and its specificity is -21.

Table 5-2. Distribution of footnotes across volumes in Fu Lei's translation

Forms	01		02		03		04		05		06		07		08		09		10
①	38		20	-4	8	-14	89		136	+17	28		66		34	-4	50		68
②	2	-3	2	-3	0		36	+4	50	+13	2	-3	12		6	-3	8	-3	20
③	0		2		0		18	+5	18	+7	0		2		0		0		2
④	0		2		0		4		2		0		0		0		0		
Total	40		26		8		147		206		30		80		40		58		90

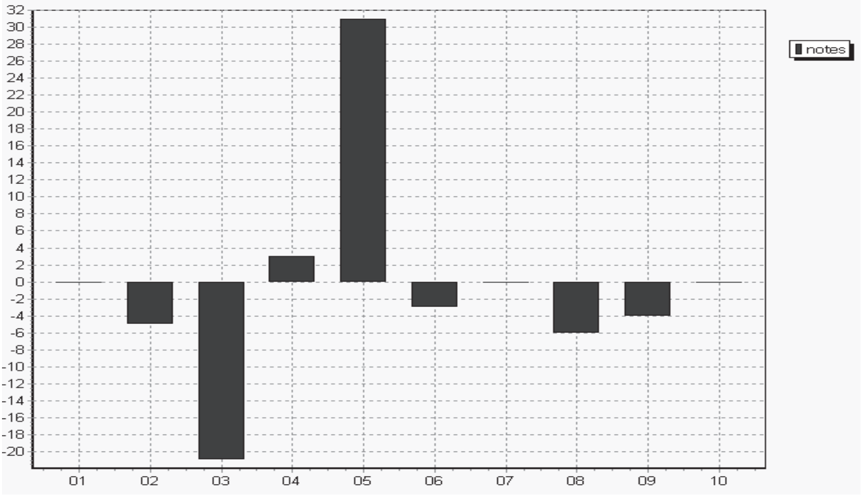


Figure 5-2. Specificity of footnote per volume in Fu Lei's translation

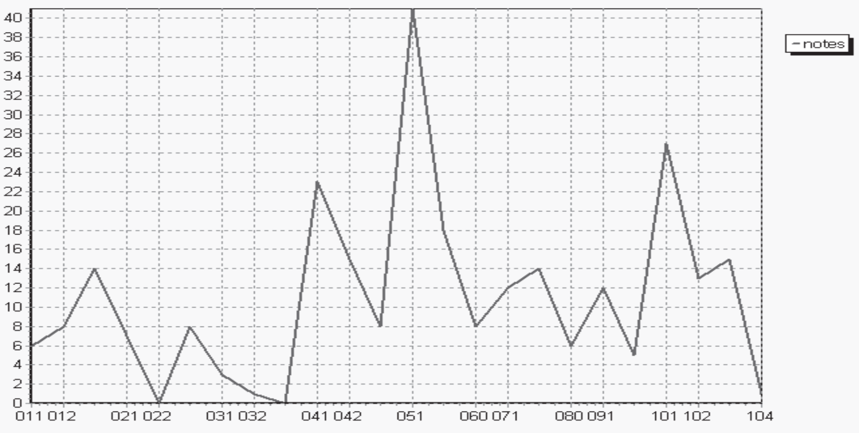


Figure 5-3. Footnote frequency per chapter in Fu Lei's translation

We can further refine the graph by taking into account not only the volumes but also the smaller unit, i.e. chapters of the novel. This is an easy thing to do in *Lexico3*.

Figure 5-3 shows in greater details the frequency of footnotes in the whole work separated by chapters. Note that some chapters have no footnotes at all, for example the second chapter of volume 2 (022, entitled *Otto*) and the third of volume 3 (033, entitled *Ada*). The highest number of footnotes is encountered in the first chapter of volume 5 (051, entitled *La foire sur la place*).

We cannot go into further details in the analysis of Figure 5-3, but we would like to emphasize that the unequal distribution clearly points to some specific needs at specific points in the translation. However, a purely lexicometrical analysis will not easily reveal these needs; one has to get back to the text and analyze the words in detail.

2.2.2. Concordance of footnotes

In a second step, we investigate to which words Fu Lei has added footnotes. At this point, we will briefly present how *Lexico3* gives direct access to the parts of text that are annotated with its included concordance of the form groups (①, ②, ③, ④).

For this experiment, we choose volume as the unit group, so, the footnote concordance results are listed and grouped by volume. As the size of the concordance page is limited, we show in Figure 5-4 only the beginning of the concordance of footnote markers in the first volume. It is also possible to specify the unit in which the results are presented. Finally, we can specify the size of the display window in number of occurrences (here we choose to present 50 words in the context). Note also that the system shows the total number of occurrences of the search term in the current volume. In volume 1, there are 40 occurrences of footnote markers. This means 20 footnotes in this area – the result corresponds to the number that we obtained in Table 5-2.

Let us see the first line of the result containing the footnote mark ①, which is for “高斯人” (g*aosiren* ‘Corsican’). Then, the next line indicates, after the marker ①, the content of the footnote for the word “高斯人” (g*aosiren* ‘Corsican’) in the precedent line: 按此系指拿破仑[...] (‘here refers to Napoleon’).[14] Thus, by examining the intervals lines (the first line indicates the word annotated, and the line after shows its content), we can study, in a systematic way, not only the annotated words in the text, but also the contents of the footnote.

In this way, the program gives easy access to all the footnotes in a large corpus.

2.3. What do the notes say?

Let’s now examine the contents of footnotes in Fu Lei’s translation of *Jean-Christophe*. It is important to point out how the use of the software is different from a traditional analysis based on careful reading by the researcher doing the translation studies. The analysis just by reading the texts will be subject to chance and to biases, as in the limit of some concrete citations from the text, it is very difficult for researchers to grasp the whole concept of the issue they deal with. Moreover, the constraints of different understanding levels of researchers add difficulty and deviation.

In fact, just as Berenson and Lazarsfeld (1984) pointed out, a typical text of intercultural communication requires a technique for an objective description of the systematic and quantitative content in the communication. We believe that it will be reasonable to address the translation problem with two methods: one is a quantitative approach focusing on the language units in the translation text; the other is a qualitative approach focusing on analysis in the textual and social context.

In the first step, we use the specific words of *Lexico3* to catch the content of footnotes through statistical data and mathematical calculations. In the second step, we want to make use of the automatic and simultaneous visualization provided by the tool *Alignoscope* to analyze the footnotes in the context.

Through the examination of the question of what the footnotes provide compared to the original content, we think we could reveal the translator's intention, and discern the role played by the footnote in the translation.

In order to do this, we proceed in three steps: 1) listing the specific vocabulary in the footnotes compared with the content of the text; 2) contextualization of footnotes in the simultaneous visualization of source-target texts; and 3) distinction of the types of the footnotes.

2.3.1. Specific words in the footnotes

We use again *Lexico3* to get the lists of positive and negative specific vocabulary in the footnotes in comparison with those in the translation text. In order to help readerse who do not know Chinese/French, we put English equivalents next to the Chinese/French words.

Table 5-3. Lists of specific and anti-specific words in the footnotes (extract)

Specific words in the footnotes						Anti-specific words in the footnotes					
Forms/SR		Note		Text		Forms/SR		Note		Text	
为	be, mean, consider	328	＊＊＊	1,079	＊＊＊	的	’s, of	464	＊＊＊	47,454	＊＊＊
①	①	269	＊＊＊	0		她	she, her	1	＊＊＊	8,078	＊＊＊
《	《	140	＊＊＊	261	＊＊＊	他	he, him	13	＊＊＊	17,654	＊＊＊
》	》	140	＊＊＊	261	＊＊＊	着	particle	3	＊＊＊	7,565	＊＊＊
之	’s, be	113	＊＊＊	469	＊＊＊	了	particle	8	＊＊＊	13,593	＊＊＊
法国	France	106	＊＊＊	502	＊＊＊	不	not, no	27	－42	9,348	＋42
世纪	century	105	＊＊＊	65	＊＊＊	是	be	48	－35	10,502	＋35
均	both, all	61	＊＊＊	3	＊＊＊	他们	they, them	2	－31	4,312	＋31
②	②	69	＊＊＊	0	＊＊＊	克利斯朵夫	Christophe	6	－30	4,830	＋30
此	this, the	65	＊＊＊	105	＊＊＊	把	marker of disposal sentence	3	－26	3,747	＋26
即	namely	44	＊＊＊	17	＊＊＊	到	arrive, to	4	－18	2,920	＋18

系	namely	36	＊＊＊	10	＊＊＊	很	very	1	－18	2,382	＋18
十九	19	34	＊＊＊	2	＊＊＊	你	you	6	－18	3,150	＋18
于	to , at	76	+49	361	-49	都	all, both	9	－16	3,281	＋16
^	separator between names	78	+48	425	-48	这	the, this	2	－16	2,247	＋16
此处	here	26	+47	0		自己	self	5	－15	2,591	＋15
近代	contemporary	27	+44	4	-44	要	want, must	4	－14	2,357	＋14
歌剧	opera	36	+44	35	-44	得	must	5	－14	2,496	＋14
指	refer, indicate	41	+42	72	-42	就	already, at once	6	－13	2,445	＋13
至	to, until	33	+40	33	-40	什么	what	1	－13	1,743	＋13
以	according to, use, with	66	+40	363	-40	想	want, would like	3	－13	1,978	＋13
时	time, period	49	+40	157	-40	在	at, be doing	68	－13	7,987	＋13
③	③	21	+38	0		里	inside, in	1	－13	1,751	＋13
称	name	27	+35	22	-35	我	I, me	15	－13	3,507	＋13
及	and	27	+33	29	-33	跟	and, with	1	－12	1,532	＋12
乃	be, so	20	+33	29	-33	说	say, talk	15	－11	3,254	＋12

故	so, thus	24	+33	15	-33	会	can, could	2	－11	1,689	＋11
与	and, with	89	+32	993	-32	和	and, with	7	－11	2,279	＋11
耶稣	Jesus	24	+32	17	-32	去	go	5	－11	1,996	＋11
神话	myth, fairy tale	22	+31	12	-31	来	come	9	－10	2,311	＋10
其	its, his, her	35	+31	100	-31	因为	because	1	－10	1,339	＋10
亦	namely	22	+31	11	-31	一个	a (an)	13	－10	2,712	＋10
德国	German	54	+27	28	-27	使	make	2	－10	1,469	＋10
希腊	Greece	23	+27	28	-27	还	also, too, even	3	－10	1,629	＋10

Table 5-3 gives us some idea of what the footnotes are about. We mention here only some important information provided by this table. First, the most frequent word 为 (*wei* 'be/mean/consider', 328 times) in the first row of the table shows the abundance of the designation meaning in the footnotes. A quick overview of the following words in the list shows that the words like 乃 (*nai* 'be'), 即 (*ji* 'namely'), 系 (*xi* 'namely'), 亦 (*yi* 'also') – containing confirmatory meanings, are also very common in the footnotes. It is interesting to note that the modern Chinese word for confirmation 是 (*shi* 'be') appears much more often in the text than in the footnotes (10,502 vs. 48). Considering that 为 (*wei* 'be') and 乃 (*nai* 'be') are the archaic substitutes for *shi* 'be', it appears that for the confirmation, Fu Lei adopted the modern Chinese words in the translation text, while he chose to use classical Chinese in the footnotes. We think that Fu Lei did it in order to avoid too many words in the footnotes, since classical Chinese allows a more concise style.

It comes as no surprise that the footnote markers themselves also obtain high scores in the list. As they appear only in the note side, it is natural to obtain the high specificity.

Moreover, the high specificity of the Chinese punctuation marks "《》" between the footnote and the text shows clearly that there are many citations of book and article titles in the footnotes. Contrary to the French "chevron" quotes "« »", the Chinese punctuation marks "《》" are reserved for the names of books, articles, pictures, songs etc. This is a clear indication that the footnotes in Fu Lei's translation provide rich cultural information.

Further down the list, another punctuation mark at the 15th row of the table attracts our attention: "^". This symbol is in fact a correction of the original "•", which is the separator between the Chinese phonetic transcription of foreigners' family names and first names. Given that this punctuation is only found in Chinese, we have changed it to "^" during our preprocessing of the corpus in order to make Chinese and French corpus more comparable. We can conclude that the footnotes contain a lot of information about foreign persons. But, the main characters of the novel occur only rarely in the footnote. For example, "Christophe" (克利斯朵夫) appears only 6 times in the footnotes, making it even appear in the anti-specific table. Besides, the personal pronouns - indicating the relationship between people,[15] e.g. 她 (*ta* 'she'), 他 (*ta* 'he'), 他们 (*tamen* 'they'), 自己 (*ziji* 'self'), 你 (*ni* 'you'), and 我 (*wo* 'I'), are also used rarely in the footnotes. This can be explained by the fact that footnotes are self-contained texts that cannot use many pronouns referring to some previously defined entity, and footnotes deliver knowledge which is not centred on the text's story.

Moreover, we find another group of words in the high specificity list: names of countries like 法国 (*faguo* ‘France’), 德国 (*deguo* ‘Germany’), and 希腊 (*xila* ‘Greece’) are listed in the table with respective frequencies of 106, 54, and 23. These nouns are also used in combination with the genitive particle 的 (*de*) to form the corresponding adjectives *French*, *German*, and *Greek* etc. As the novel is set in France and Germany, we believe that Fu Lei added information about these Western countries - a world far away and strange for Chinese readers. Why Greece? Maybe Fu Lei explains a lot of things about Greek culture.

Let’s briefly mention some other interesting words in this table. The words 世纪 (*shiji* ‘century’), 时 (*shi* ‘time, period’) and 近代 (*jindai* ‘modern’) remind us that history is an important element in the footnotes. With 51 occurrences, 歌剧 (*geju* ‘opera’) is undoubtedly a topic discussed extensively, while 耶稣 (*yesu* ‘Jesus’) and 神话 (*shenhua* ‘myth, fairy tale’) show Fu Lei’s effort to introduce religion and Western mythology to China.

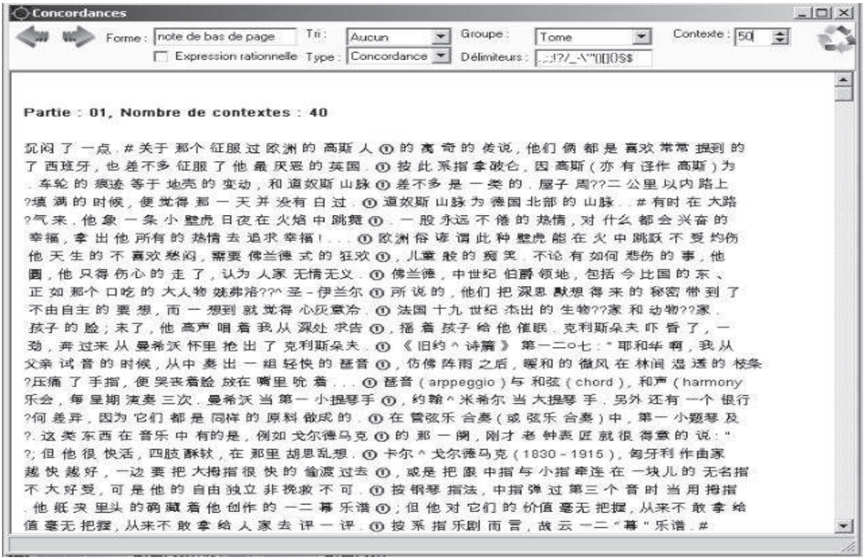
From this analysis, we can draw some conclusions as follows. Thanks to the textometrical methods, we saw that the footnotes in Fu Lei’s translation are characterized by a strong supply of information about the cultures, foreign personages, customs, and history. In this way, the social contexts of Western countries - a world where the novel is set - unfold to the Chinese readers. However, this is just our speculation based on our observations of statistical results which we will have to validate by providing concrete examples later.

2.3.2. Contextualization of footnotes

Here, we use the tool *Alignoscope* which displays simultaneously the source and target texts, in order to scrutinize the contents of footnotes in their context. In the Figure 5-4, we saw that *Lexico3* can quickly find the words annotated in the text and the content of footnotes, but now we would like to have direct access to the references of the original context.

By drawing attention to the importance of “re-contextualization” of each note in the target text, Varney (2005: 49-50) proposed to identify the elements of macro-structure of each note and assess their rhetoric strategy since “a comparative analysis of the target text segment signalled by the note and the corresponding segment in the source text will enable us to identify the problematic issue.”

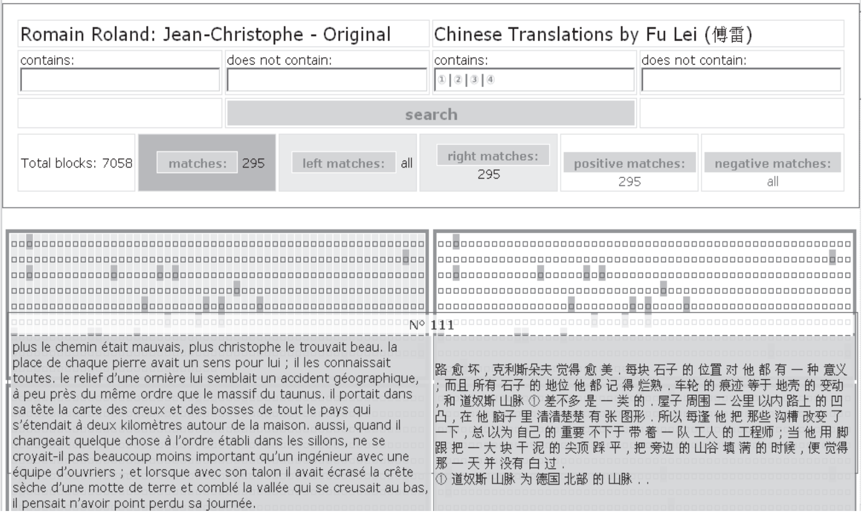
Following this idea, we are launching an experiment on all footnotes ①, ②, ③, ④ using *Alignoscope*.



Partie : 01, Nombre de contextes : 40

沉闷 了 一点 . # 关于 那个 征服 过 欧洲 的 高斯 人 ① 的 离 奇 的 传说 , 他们 俩 都 是 喜欢 常常 提到 的
了 西班牙 , 也 差不多 征服 了 他 最 厌恶 的 英国 . ① 按 此 系指 拿破仑 . 因 高斯 (亦 有 译作 高斯) 为
. 车轮 的 痕迹 等于 地壳 的 变动 , 和 道奴斯 山脉 ① 差不多 是 一 类 的 . 屋子 周??二 公里 以内 路上
?填 满 的 时候 , 便 觉得 那 一 天 并 没有 白 过 . ① 道奴斯 山脉 为 德国 北部 的 山脉 . . # 有时 在 大路
?气 来 . 他 象 一 条 小 壁虎 日夜 在 火焰 中 跳舞 ① . 一 股 永远 不 倦 的 热情 , 对 什么 都 会 兴奋 的
幸福 , 拿 出 他 所有 的 热情 去 追求 幸福 ! . . . ① 欧洲 俗 谚 谓 此 种 壁虎 能 在 火 中 跳跃 不 受 灼伤
他 天 生 的 不 喜欢 愁闷 , 需要 佛兰德 式 的 狂欢 ① , 儿童 般 的 痴 笑 . 不论 有 如何 悲伤 的 事 , 他
圈 , 他 只得 伤心 的 走 了 , 认为 人家 无情无义 . ① 佛兰德 , 中世纪 伯爵 领地 , 包括 今 比国 的 东 、
正 如 那个 口吃 的 大人物 姚弗洛??^ 圣 - 伊兰尔 ① 所 说 的 , 他们 把 深思 默想 得 来 的 秘密 带 到 了
不由自主 的 要 想 , 而 一 想到 就 觉得 心灰意冷 . ① 法国 十九 世纪 杰出 的 生物??家 和 动物??家 .
孩子 的 脸 ; 末了 , 他 高声 唱 着 我 从 深处 求告 ① , 摇 着 孩子 给 他 催眠 . 克利斯朵夫 吓 昏 了 , 一
劲 , 奔 过来 从 曼希沃 怀里 抢 出 了 克利斯朵夫 . ① 《 旧约 ^ 诗篇 》 第一二○七 : " 耶和华 啊 , 我 从
父亲 试 音 的 时候 , 从中 奏 出 一 组 轻快 的 琶音 ① , 仿佛 阵雨 之后 , 暖和 的 微风 在 林间 湿 透 的 枝条
?压痛 了 手指 , 便 哭丧着脸 放在 嘴里 吮 着 . . . ① 琶音 (arppeggio) 与 和弦 (chord) , 和声 (harmony
乐会 , 每 星期 演奏 三次 . 曼希沃 当 第一 小提琴手 ① , 约翰 ^ 米希尔 当 大提琴 手 . 另外 还有 一个 银行
?何 差异 . 因为 它们 都 是 同样 的 原料 做成 的 . ① 在 管弦乐 合奏 (或 弦乐 合奏) 中 , 第一 小题琴 及
?. 这 类 东西 在 音乐 中 有的是 , 例如 戈尔德马克 ① 的 那 一 阕 , 刚才 老 钟表 匠 就 很 得意 的 说 : "
?; 但 他 很 快活 , 四肢 酥软 , 在 那里 胡思乱想 . ① 卡尔 ^ 戈尔德马克 (1830 - 1915) , 匈牙利 作曲家
越 快 越 好 , 一边 要 把 大拇指 很 快 的 偷渡 过去 ① , 或是 把 跟 中指 与 小指 牵连 在 一块儿 的 无名指
不 大 好受 , 可 是 他 的 自由 独立 非 挽救 不 可 . ① 按 钢琴 指法 , 中指 弹 过 第三 个 音 时 当 用 拇指
. 他 纸 夹 里头 的确 藏 着 他 创作 的 一二 幕 乐谱 ① ; 但 他 对 它们 的 价值 毫无 把握 , 从来不 敢 拿 给
值 毫无 把握 , 从来不 敢 拿 给 人家 去 评 一 评 . ① 按 系 指 乐剧 而 言 , 故 云 一二 " 幕 " 乐谱 . #

Figure 5-4. Footnote concordance in Fu Lei’s translation (extract)



Romain Roland: Jean-Christophe - Original		Chinese Translations by Fu Lei (傅雷)	
contains:	does not contain:	contains: ①|②|③|④	does not contain:

Total blocks: 7058	matches: 295	left matches: all	right matches: 295	positive matches: 295	negative matches: all

N° 111

plus le chemin était mauvais, plus christophe le trouvait beau. la place de chaque pierre avait un sens pour lui ; il les connaissait toutes. le relief d’une ornière lui semblait un accident géographique, à peu près du même ordre que le massif du taunus. il portait dans sa tête la carte des creux et des bosses de tout le pays qui s’étendait à deux kilomètres autour de la maison. aussi, quand il changeait quelque chose à l’ordre établi dans les sillons, ne se croyait-il pas beaucoup moins important qu’un ingénieur avec une équipe d’ouvriers ; et lorsque avec son talon il avait écrasé la crête sèche d’une motte de terre et comblé la vallée qui se creusait au bas, il pensait n’avoir point perdu sa journée.

路 愈 坏 , 克利斯朵夫 觉得 愈 美 . 每块 石子 的 位置 对 他 都 有 一 种 意义 ; 而且 所有 石子 的 地位 他 都 记 得 烂熟 . 车轮 的 痕迹 等于 地壳 的 变动 , 和 道奴斯 山脉 ① 差不多 是 一 类 的 . 屋子 周围 二 公里 以内 路上 的 凹 凸 , 在 他 脑子 里 清清楚楚 有 张 图形 . 所以 每逢 他 把 那些 沟槽 改变 了 一下 , 总 以为 自己 的 重要 不下于 带 着 一 队 工人 的 工程师 ; 当 他 用 脚 跟 把 一 大 块 干 泥 的 尖顶 踩 平 , 把 旁边 的 山谷 填 满 的 时候 , 便 觉得 那 一 天 并 没有 白 过 .
① 道奴斯 山脉 为 德国 北部 的 山脉 . .

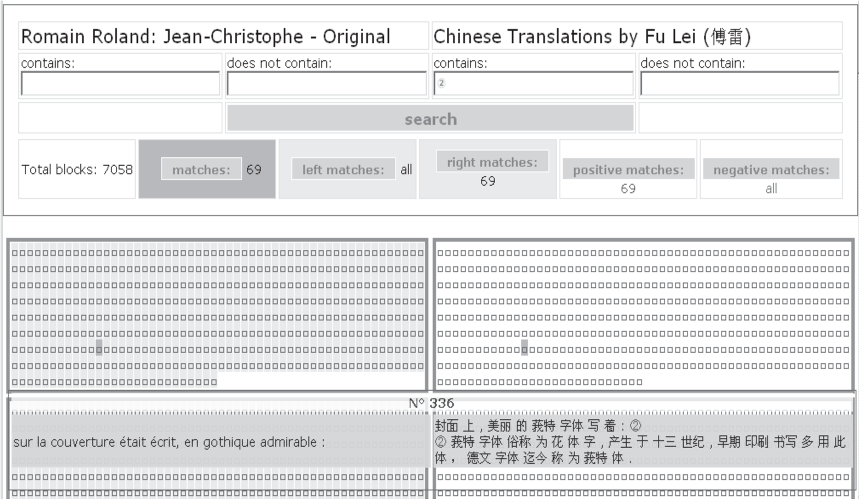
Figure 5-5. Simultaneous visualization of footnotes in source - target texts of *Jean-Christophe* (extract)

According to the information bar in the middle of Figure 5-5, we can see that there are a total of 7058 blocks throughout the parallel corpus. Since the complete parallel corpus (with the prefaces) is loaded in the online program *Alignoscope*, the result about the footnotes is slightly influenced: there are 4 footnotes of ① in the prefaces. But we will not deal with them in this present study. So, there are 291 (295-4) blocks which correspond to our quest for footnotes. We should bear in mind that the square in Figure 5-5 represents the paragraph unit of segmentation. And these paragraphs of the original and the translation are semi-automatically paired and marked by a marker (e.g. #) using the *Alignator* program.

Thus, we know that Fu Lei inserted footnotes in 291 paragraphs. As some paragraph blocks contain several footnotes at the same time, the number here is different from the result in Figure 5-1. In the following discussion, we will illustrate how the simultaneous visualization can help us to achieve the "re-contextualization" for the footnote.

Hovering over the shaded square displays a popup (No.111) with the two corresponding paragraphs, on the left is the original text,[16] and at its right side is its corresponding Chinese translation. It is easy to see that Fu Lei wrote a footnote ① for 道奴斯山脉 (*daonusi shanmai*). By referring to the source context at the left, we know that the footnote is annotating the Taunus Mountains. We equally see the footnote itself: "**道奴斯山脉**为**德国北部的山脉**" (the Taunus Mountains are a mountain range in northern Germany).[17] Although this footnote is simple, with only 7 Chinese words (segmented by the ICTCLAS program), it provides a brief geographic information for Chinese readers.

In fact, by a mere glance at the original context, we can see Romain Rolland described the poor state of the road Christophe took: he used a comparison between the relief of a rut and geography. In order to strengthen the effect of the description for the relief, Rolland wrote *à peu près du même ordre que le massif Taunus* 'roughly of the same kind as the Taunus mountain range'. Because France and Germany share a common frontier, a lot of French readers know the Taunus range. But this geographical knowledge is not obvious to Chinese readers who live far from Europe. Adding this footnote in translation, Fu Lei wants to transfer the geographical information to his Chinese readers in the hope that they can understand the original better.



Romain Roland: Jean-Christophe - Original

Chinese Translations by Fu Lei (傅雷)

contains:

does not contain:

contains: ②

does not contain:

search

Total blocks: 7058

matches: 69

left matches: all

right matches: 69

positive matches: 69

negative matches: all

N° 336

sur la couverture était écrit, en gothique admirable :

封面 上，美丽 的 莪特 字体 写 着：②
② 莪特 字体 俗称 为 花 体 字，产生 于 十三 世纪，早期 印刷 书写 多 用 此 体， 德文 字体 迄今 称 为 莪特 体．

Figure 5-6. Simultaneous visualization of footnotes ② in source - target texts of *Jean-Christophe*

In order to detail our examination of the footnotes, we give another example of the footnote ② in *Alignoscope*. As Figure 5-6 shows, 69 paragraphs in the Fu Lei's translation have the footnote marker ②. Look first on the Chinese side. Fu Lei put a footnote ② after 封面上, 美丽**的莪特字体写着** (‘On the cover, there was written in nice Gothic script’): ② 莪特字体俗称为**花体字**, 产**生于十三世**纪, 早期印刷书**写多用此体**, 德文字体迄今称为**莪特体** (‘The Gothic script, with a popular name of floral script, was created during the 13th century. The ancient press work adopted largely this kind of script. Until now, the German script is also called Gothic script’). Then, we return to the original context, where we find that the Chinese footnote is added to the word *Gothic* in the phrase *sur le couverture était écrit, en gothique admirable* (‘on the cover was written in admirable Gothic script’).

Significantly different from the Chinese script composed of Chinese characters, the Latin script is based upon alphabets. As for the Gothic script, it is a kind of script mainly reserved for use in printing. Aware that this information would be left unnoticed or not understood by most of Chinese readers, Fu Lei added this footnote to fill this cultural gap.

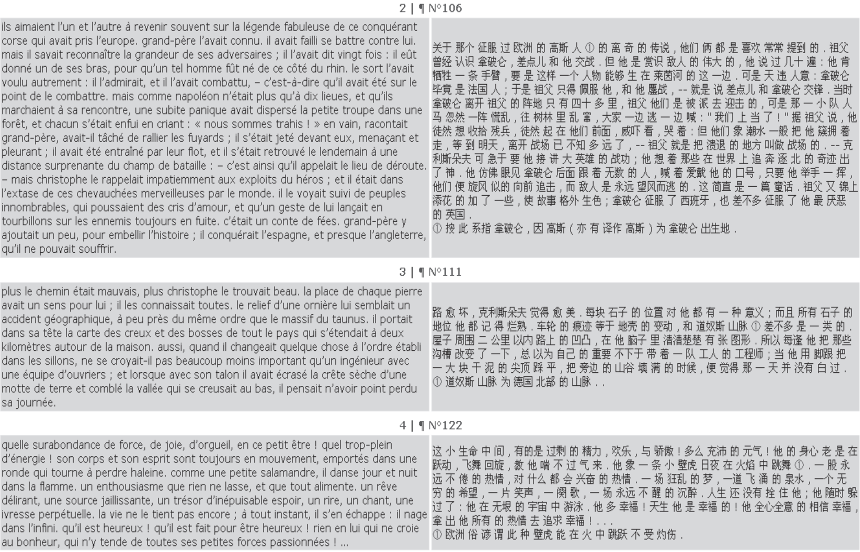
At the level of translation technique, there is a double interest in analyzing these footnotes in their context. First, it is about a translation technique: how to translate the term *Gothic script* into Chinese? In fact, the Gothic style in the architecture is often translated into **莪特式** (*geteshi* ‘Gothic style’). According to the translation text, Fu Lei kept the phonetic transcription 莪特 (*gete* 'Gothic') in Chinese,[18] then added 字体 (*ziti* 'font') or 体 (*ti* 'style') to clarify the meaning of the translated term. Second, it concerns the choice of words. In the footnote, Fu Lei wrote that the Gothic script has another popular name 花体字 (*huatizi* 'floral script'). In fact, the Gothic script is also called "black letter" in English.[19] In Chinese, 花体字 (*huaziti* 'floral script') and 黑体字 (*heitizi* 'black letter') are both used for the Gothic script. But Fu Lei chose to use the first instead of the second term. We do not know if it was Fu Lei who translated, for the first time, the Gothic script into Chinese as 花体字, but from this small example, we can assume that Fu Lei thought 花体字 'floral script' could highlight the ornamental feature of this writing.

A little further on, we meet once again a description about the Gothic script in the original (paragraph 325), but this time the author did not directly use the word *gothique* ('Gothic script'): *c'est le cahier était écrit à la main, de la grosse écriture du vieux, qui s'était spécialement appliqué. les en-têtes étaient ornés de boucles et de paraphes* ('This music score was handwritten, in the gross script of the old man, who had specially applied himself. The title was decorated with loops and ornaments'). At this time, Fu Lei translated it as 乐谱是手写的, 还是老人用他肥大的笔迹特别用心写的. 题目都用的花体字 ('The music score was written by hand; it was the old man who wrote it with special application in his big script. The titles all use the floral script'). A simple comparison shows that Fu Lei did not translate *loops and ornaments* word by word into Chinese, but a free translation with *floral script* – the ample information is located in the footnote of paragraph No. 326.

Of course, if we scrutinize the content of this footnote, we will find that in one point, the footnote is not very exact. The Gothic script became popular rather than arose in 13th century Europe. Surely, this mistake does not have a serious impact on the general meaning of the content of the footnote. In any case, this example illustrates that footnotes provide external knowledge to facilitate the Chinese reading of the original text, and at the same time, the information in the footnote becomes a part of the text of translation.

So, with the detailed examination of the footnotes in the context, we can evaluate their quality. Furthermore, in the examples cited above, we noticed that the word 为 (*wei* 'be/mean/consider') appears three times, and 德国(*deguo* 'Germany') once. This confirms that the information obtained in the Table 5-3 about the specific vocabulary items of the footnote is correct.

The software *Alignoscope* can also export the search results of all footnotes – a useful feature for doing an intensive review of all the footnotes in their context (Figure 5-7).



2 | ¶ N°106

ils aimaient l'un et l'autre à revenir souvent sur la légende fabuleuse de ce conquérant corse qui avait pris l'europe. grand-père l'avait connu. il avait failli se battre contre lui. mais il savait reconnaître la grandeur de ses adversaires ; il l'avait dit vingt fois : il eût donné un de ses bras, pour qu'un tel homme fût né de ce côté du rhin. le sort l'avait voulu autrement : il l'admirait, et il l'avait combattu, – c'est-à-dire qu'il avait été sur le point de le combattre. mais comme napoléon n'était plus qu'à dix lieues, et qu'ils marchaient à sa rencontre, une subite panique avait dispersé la petite troupe dans une forêt, et chacun s'était enfui en criant : « nous sommes trahis ! » en vain, racontait grand-père, avait-il tâché de rallier les fuyards ; il s'était jeté devant eux, menaçant et pleurant ; il avait été entraîné par leur flot, et il s'était retrouvé le lendemain à une distance surprenante du champ de bataille : – c'est ainsi qu'il appelait le lieu de déroute. – mais christophe le rappelait impatiemment aux exploits du héros ; et il était dans l'extase de ces chevauchées merveilleuses par le monde. il le voyait suivi de peuples innombrables, qui poussaient des cris d'amour, et qu'un geste de lui lançait en tourbillons sur les ennemis toujours en fuite. c'était un conte de fées. grand-père y ajoutait un peu, pour embellir l'histoire ; il conquérait l'espagne, et presque l'angleterre, qu'il ne pouvait souffrir.

关于 那个 征服 过 欧洲 的 高斯 人 ① 的 离 奇 的 传说 , 他们 俩 都 是 喜欢 常常 提到 的 . 祖父 曾经 认识 拿破仑 , 差点儿 和 他 交战 . 但 他 是 赏识 敌人 的 伟大 的 , 他 说 过 几十 遍 : 他 肯 牺牲 一 条 手臂 , 要 是 这样 一个 人物 能够 生 在 莱茵河 的 这 一边 . 可是 天 违 人意 : 拿破仑 毕竟 是 法国 人 ; 于是 祖父 只得 佩服 他 , 和 他 鏖战 , -- 就是 说 差点儿 和 拿破仑 交锋 . 当时 拿破仑 离开 祖父 的 阵地 只 有 四十 多 里 , 祖父 他们 是 被 派 去 迎击 的 , 可是 那 一 小 队 人 马 忽然 一阵 慌乱 , 往 树林 里 乱 窜 , 大家 一边 逃 一 边 喊 : " 我们 上 当 了 ! " 据 祖父 说 , 他 徒然 想 收拾 残兵 , 徒然 起 在 他们 前面 , 威吓 看 , 哭 着 : 但 他们 象 潮水 一般 把 他 簇拥 着 走 , 等 到 明天 , 离开 战场 已 不知 多 远 了 , -- 祖父 就是 把 溃退 的 地方 叫做 战场 的 . -- 克利斯朵夫 可 急于 要 他 接 讲 大 英雄 的 战功 ; 他 想 着 那些 在 世界 上 追 奔 逐 北 的 奇迹 出了 神 . 他 仿佛 眼见 拿破仑 后面 跟 着 无数 的 人 , 喊 着 爱戴 他 的 口号 , 只要 他 举手 一 挥 , 他们 便 旋风 似的 向前 追击 , 而 敌人 是 永远 望风而逃 的 . 这 简直 是 一 篇 童话 . 祖父 又 锦上添花 的 加 了 一些 , 使 故事 格外 生色 ; 拿破仑 征服 了 西班牙 , 也 差不多 征服 了 他 最 厌恶 的 英国 .
① 按 此 系指 拿破仑 , 因 高斯 (亦 有 译作 高斯) 为 拿破仑 出生地 .

3 | ¶ N°111

plus le chemin était mauvais, plus christophe le trouvait beau. la place de chaque pierre avait un sens pour lui ; il les connaissait toutes. le relief d'une ornière lui semblait un accident géographique, à peu près du même ordre que le massif du taunus. il portait dans sa tête la carte des creux et des bosses de tout le pays qui s'étendait à deux kilomètres autour de la maison. aussi, quand il changeait quelque chose à l'ordre établi dans les sillons, ne se croyait-il pas beaucoup moins important qu'un ingénieur avec une équipe d'ouvriers ; et lorsque avec son talon il avait écrasé la crête sèche d'une motte de terre et comblé la vallée qui se creusait au bas, il pensait n'avoir point perdu sa journée.

路 愈 坏 , 克利斯朵夫 觉得 愈 美 . 每块 石子 的 位置 对 他 都 有 一 种 意义 ; 而且 所有 石子 的 地位 他 都 记 得 烂熟 . 车轮 的 痕迹 等于 地壳 的 变动 , 和 道奴斯 山脉 ① 差不多 是 一 类 的 . 屋子 周围 二 公里 以内 路上 的 凹凸 , 在 他 脑子 里 清清楚楚 有 张 图形 . 所以 每逢 他 把 那些 沟槽 改变 了 一下 , 总 以为 自己 的 重要 不下于 带 着 一 队 工人 的 工程师 ; 当 他 用 脚跟 把 一 大 块 干 泥 的 尖顶 踩 平 , 把 旁边 的 山谷 填 满 的 时候 , 便 觉得 那 一 天 并 没有 白 过 .
① 道奴斯 山脉 为 德国 北部 的 山脉 . .

4 | ¶ N°122

quelle surabondance de force, de joie, d'orgueil, en ce petit être ! quel trop-plein d'énergie ! son corps et son esprit sont toujours en mouvement, emportés dans une ronde qui tourne à perdre haleine. comme une petite salamandre, il danse jour et nuit dans la flamme. un enthousiasme que rien ne lasse, et que tout alimente. un rêve délirant, une source jaillissante, un trésor d'inépuisable espoir, un rire, un chant, une ivresse perpétuelle. la vie ne le tient pas encore ; à tout instant, il s'en échappe : il nage dans l'infini. qu'il est heureux ! qu'il est fait pour être heureux ! rien en lui qui ne croie au bonheur, qui n'y tende de toutes ses petites forces passionnées ! ...

这 小 生命 中 间 , 有的是 过剩 的 精力 , 欢乐 , 与 骄傲 ! 多么 充沛 的 元气 ! 他 的 身心 老 是 在 跃动 , 飞舞 回旋 , 教 他 喘 不 过 气 来 . 他 象 一 条 小 壁虎 日夜 在 火焰 中 跳舞 ① . 一 股 永远 不 倦 的 热情 , 对 什么 都 会 兴奋 的 热情 . 一 场 狂乱 的 梦 , 一道 飞 涌 的 泉水 , 一个 无穷 的 希望 , 一 片 笑声 , 一 阕 歌 , 一 场 永远 不 醒 的 沉醉 . 人生 还 没有 拴 住 他 ; 他 随时 躲 过 了 : 他 在 无垠 的 宇宙 中 游泳 . 他 多 幸福 ! 天生 他 是 幸福 的 ! 他 全心全意 的 相信 幸福 , 拿 出 他 所有 的 热情 去 追求 幸福 ! . . .
① 欧洲 俗 谚 谓 此 种 壁虎 能 在 火 中 跳跃 不 受 灼伤 .

Figure 5-7. The exportation of footnotes in the *Jean-Christophe* corpus (extract)

All of the above show that the textometrical software greatly facilitates translation studies: On the one hand, it allows us to penetrate the paragraph where Fu Lei gave the footnote so that we study carefully the contents of the footnote in its context; and on the other hand, the simultaneous visualization makes it possible to find the original word annotated; and by scrutinizing source-target-texts, we can analyze the translator's motivation.

2.3.3. The types of footnotes

From the observations of footnotes in Fu Lei's translation, we attempt to give a summary of the contents of the footnotes in order to better examine the issue of adding footnotes in translation. We should remind the reader here that our synthesis is made by the analysis of the annotated words and by the simultaneous visualization. There are about seven types of footnotes as given in Table 5-4.

Such classification is not exhaustive; and as a lot of contents in the footnotes concern many fields, it is very difficult to make an exact distinction. Then, we can not cite all the typical examples in the translation. But it should be observed that the footnotes in Fu Lei's translation touch upon a range of social, cultural, historical, and linguistic domains.

However, the most remarkable type of footnotes is the commentary from Fu Lei. This is because Fu Lei provided not only the correlative information but also his personal opinions on the contents of the text, the author, and on the historical events. Even in the informative footnotes about places, works of art, Fu Lei now and then added his personal interpretations.

Clearly, through footnotes, we can see that as a literature translator, Fu Lei did not blindly follow the author, but rather he did his best so that his readers would get a better understanding of the essence of the original, and he used his personal aesthetic to comment on all: the oeuvre, the author, and the social affairs, etc. In brief, Fu Lei acts like a thinker who scans the esprit of the original through the surface of the words.

3. Analysis from the aspect of translation studies

If there are no notes in the text, the translation would seem purer and more "transparent", at least in appearance, because the footnote-style notes produce a visual effect on the translation: the extension of the page and its layout of the printed page changes. Our results above show that the footnotes add 10,214 words to the resulting translation.

Henry (2000: 235) notes that adding notes is in fact a translational notion which is linked with the space "already said and unsaid" in the text; precisely, it lies within the choice of the implicit and, most importantly, the explicit or not. But the issue arises from the incompleteness of knowledge (Lederer 1984), due to the unequal possession of knowledge between individuals; or in other words, there is a lack of "lexiculture" in some readers (Antoine 1998). In the present corpus, most of the Fu Lei's footnotes are informative and explanatory. They provide the knowledge of Western culture and of Western music in particular: a world almost completely unknown to the Chinese audience. Besides, many footnotes tell the stories of myths and religions, which create a course on culture for Chinese readers; and this information can increase enormously Chinese readers' understanding of the meaning of citations in the original texts and their underlying cultural concepts. Consequently, these footnotes compensate for the cultural loss.

Table 5-4. Seven types footnotes and examples

Subjects of footnotes	Content	Examples
Places	1. The geography of a place; 2. The geography in religion (The Bible); 3. The characteristic of a place; 4. The style of construction of one place or a building; 5. The name of a place.	Bloc 2978 克利斯朵夫被《圣经》中那股肃杀之气鼓舞起来了: 西乃山上的①, 无垠的荒漠中的, 汪洋大海中的狂风, 把乌烟瘴气一扫 而空. ① 西乃为 阿拉伯 半岛地名, 又为 山脉 名, 圣经载, 上帝于西乃山上授律于摩西. (The Sinai is an Arabic Peninsula, and it is also the name of the mountain. According to the Bible, the Sinai is the place where Moses received God’s command.)
Persons	1. The name, status and profession of somebody; 2. The story or anecdote of somebody; 3. Somebody’s ideas, or the ideas of one school.	bloc 4438 于是他姊姊不倦的叙述出征非洲的经过. 伟大的事迹, 可以和比查尔跟高丹士的故事媲美①. ① 比查尔与高丹士均十六 世纪时西班牙冒险家: 前者征服秘鲁, 后者征服墨西哥. (Pizarre and Cortes were two Spanish explorers of the 16th century, the first one conquered Peru, and the second Mexico.)
Translation techniques and linguistic knowledge	1. Explain the difficulty of translation and the method of translation; 2. Add the source of words, phrases and quotations; 3. Inform the figurative meaning and the original meaning;	Bloc 484 人家越想要他驯服, 做个循规蹈矩的德国小布尔乔亚①, 他越觉得需要摆脱羁绊. ① 布尔乔亚是法语 bourgeois (资产阶级)之译音, 在本书

	4. Explain the characteristic of source language; 5. Inform the content of the original text.	中, 多半系指中产阶级或市民阶层. (布尔乔亚 is the Chinese phonetic translation of the French word *bourgeois* (capitalist class); in this book, this word refers rather to the middle class and the citizen class.)
Social customs	1. The social customs (in terms of religion, law, publishing, social phenomenon, education system, military system, clothing, daily life etc) 2. The explanations for nations and peoples; 3. The cultural context; 4. The associations in the society.	Bloc 2293 华**特霍斯**可是对他一脸瞧不起的样**子**, 拿出尊严沉着的气派, 竭力在喧闹 声中表示不答应人家对他用这种口气, 教克利斯朵**夫**等他的消息; 一边把名片递给他 ①. 克利斯朵**夫**拿来扔 在他脸**上** , ① 西俗: 两人吵架时一造把名片递给 对造是表示愿意决斗. (Western custom: During a quarrel, giving a visiting card to another one means to launch a duel.)
Works of art	I. Music The musical instruments; The terms of music; The posts and the title in the field of music; The names of a piece of work, its content, its author, its styles (characteristics); The rules or custom in the field of music; II. Painting Explain the name of painting, its author, its date, and its characteristics; III. Play Explain the play (or opera), its content, its characters, the date of the staging, and its director;	Bloc 3138 下一天, 克利斯朵**夫**发现 所谓钢**琴**是件旧货店里买 来的破烂东**西**, 声音象吉他①; ① 吉他形似中提琴而略大, 共有六弦, 舞蹈音乐及民间音乐多用之. (Under the form of viola, but a little larger, the guitar has six strings, and is often used for dance music and popular music.)

	The types of plays; IV. Artistic status The status of the author of some works and his style; V. Mythical stories The mythical stories and characters, explain their significance; VI. Religious stories Stories and religious figures in the Bible; VII. Literary works The content of the work, its author, its style, and its social affects, etc.	
Translator’s comments	1. On the content of the original; 2. On the citation; 3. On the historical context in the original; 4. On the author or the other personages; 5. On the works of art.	Bloc 3559 在家里他有二十一个孩子, 十三个都比他死得早③, 其中一个是白痴; 其余都是优秀的音乐家, 替他来些小小的家庭音乐会, ③ 按所有罢哈的传记均称罢哈子女共 二十人 (前妻生七个, 后妻生十三个), 罢哈故世时 (1750 年) 尚生存者共有子女九人. 作者言其子女共二十一人, 有十三个比罢哈早故, 不知何所据. (According to all biographical books of Bach, Bach had twenty children (including 7 of his first wife, and 13 of his second wife). When he died (1750), there were 9 children alive. The author said that Bach had 21 children, and 13 died before him. We don't know on which information the author based his statement.)

Historical elements	1. Reminder of historical events; 2. The historical contexts of the work.	Bloc 2516 “那我们在中国已经实行过 了①.” ① 指 一九〇〇年八国联军入侵中国. (It refers to Eight-Power Allied Forces who invaded China in 1900.)

However, our research shows that the addition of footnotes in Fu Lei’s translation of *Jean-Christophe* is not limited to the introduction of the Western world knowledge; it also offers a way for the translator to accomplish his ambition to introduce new and good ideas to the readers. This manifests Fu Lei’s responsibility – a responsibility not only of a translator and but also as a thinker.

In one of Fu Lei’s letters (the 5 February, 1961) to his son Fou Ts’ong (傅聪),[20] Fu writes: “As you have often mentioned Hellenism in the art, I specifically translated the fourth chapter “La sculpture en Grèce” of H. Taine’s *Philosophie de l'art* into Chinese, amounting to over 60,000 words, and stapled the translation into a booklet. Although the original book has its English version, there are few notes on the myths, the historic events, and the anecdotes, which are in fact numerous in the original book. I don’t think you can read it completely. So, I put many notes besides my translation, hoping they will be useful for you.”[21] Without explanation, it is obvious to note from this extract, that Fu Lei added the notes in order to facilitate his son’s understanding of the original. Of course, in this case, Fu Lei did the translation of “La sculpture en Grèce” with the primary intention to help his son in his studies of Occidental art. But from this example, we can see that Fu Lei did his translation to help his readers to understand better the original text, and his son was the first reader of his translation in this instance.

In terms of legality or limit of using the translator’s notes, Henry (2000: 239) argues that it is a problem of the boundary between scholarship and failure, and it raises the problem of “moral contract” between the translator and the author, because the abuse of notes and badly used notes reveal the inability of the translator. But contrary to Henry’s concern that adding notes will cut the linearity of reading (Henry 2000: 238), we tend to agree to Genette’s (1987: 297) view: “[...] the optional reading (of the notes) therefore involves only some readers: those who are interested in any additional or digressive consideration. It is precisely the secondary importance of those considerations that justifies their rejection into a footnote.”[22] Placed in the margin of the text, the note in the category of paratext does not hinder the normal reading, as it is the reader who has the choice to read or not to read those notes (Bahier-Porte 2005). We agree that too many footnotes destroy the “clean face” of the original text, and they could eventually bore the reader. But keeping in mind that each footnote contains about 28 words and Fu Lei uses the succinct classic Chinese style, we see that Fu Lei has the goal to simplify his footnotes as much as possible.

Moreover, we think that the addition of footnotes is just for the purpose of ensuring the linearity of the content of the original text.

Because the other ways (see Henry 2000: 235-236), i.e. the addition of the equivalence in an interpolated clause, between commas or parentheses, are more likely to affect the fluidity or *qi* (气, 'spirit') - borrowing the Chinese term - of the original text. Anyway, we think we should examine the note issue with the translator's intention in the social context.

In fact, when Fu Lei returned to China in 1931, after four years of studies in France, China was suffering from wars. Realizing that "lack of moral strength, the Chinese today are living day by day as if asleep, or they have little to do, as if they were in an amorphous state",[23] he decided to translate three works of Romain Rolland: *Life of Tolstoi*, *Life of Michelangelo*, and *Life of Beethoven*. The goal he set with his translation was to encourage people like him in the middle of the war suffering of the time.

When Fu Lei translated *Jean-Christophe*, he expressed directly his admiration for this work in the preface of the translation: "*Jean-Christophe* is not a novel, - we should say: it is not just a novel; rather it is a glorious human epic. [...] I hope that the readers can open this treasure with the devotional mood."[24]

In the beginning, Fu Lei loved Romain Rolland's works just by his personal preference: He felt encouraged by the main characters of the novel. But from the moment he decided to translate Rolland's works, we might be tempted to say that it was rather a sense of social responsibility which pushed him to achieve the translation. Unsatisfied with his personal development, Fu Lei wanted his fellow Chinese citizens to wake up and to improve. It is under the circumstances of the time that Fu Lei chose to translate the works of Romain Rolland.

Anthony Pym (1993) noted the importance of the responsibility of the translator. He claimed, in his book *Pour une éthique du traducteu* (1993), that the sense of responsibility is the basic ethical question. In the paragraphs above, it is very clear that in the case of Fu Lei, because of this responsibility, Fu Lei wrote a lot of footnotes to facilitate the communication between the original (via his translation) and his readers. This reflects Fu Lei's attitude toward the triangle of the author, the reader and the translator.

Furthermore, we should mention some anecdotes about Fu Lei in order to investigate his intention to add the notes into his translation. The first job that Fu Lei got after his studies in France was as a teacher of History of Art and of French in the Institute of the Fine Arts in Shanghai. He edited with the director of the institute Liu Haisu (刘海粟) – also his friend, the *Album of World Famous Paintings* (《世界名画集 》), and the Album was published by Zhonghua Book Company in 1931. To meet his class needs, he translated also Paul Gsell's *Propos de Rodin sur l'art et les* *artistes* ('Rodin's Words on Art and Artists') for his students. Besides, there is another interesting thing: on the back of Fu Lei's visiting card was written in French "critique d'art" ('art critic'). Such things of Fu Lei were many, and these things made us think that there is surely some relation between Fu Lei's commentary notes in the translation and his professional insight. From an art critic's angle, Fu Lei incorporated his personal aesthetic interpretation of the content of the original text, but at the bottom of the page – where his intervention does not destroy the integrity of the original text.

In a word, there are two big reasons for Fu Lei to use footnotes. The first one is to reduce the cultural loss sometimes engendered by the process of translation. The second reason is to create a platform where he could communicate with his readers and discuss the work from a critique's point of view.

4. Conclusion

We have shown how the textometrical method can be used in translation studies. And this technique allows us to highlight the inherent characteristics of the text. The most important aspect of the method may be its productivity, since we can apply the same methodology to other texts. Of course, the application of this methodology and (in particular) the interpretation of the results require more reflection and verification in future research.

From our examination we learned that Fu Lei shared his knowledge and personal opinions with his readers in the footnotes he added. He was not a translator who followed behind the original author word by word, but rather he was someone who transferred, as much as possible, the content and the spirit of the text in the light of a literary critic. Thus, at the level of the technique, his translation seems very free but in fact it is very close to the essence of the original. This seems likely to be one of the characteristics of his translation style. Indeed, the question "What is Fu Lei's style?" is a too big question to answer here and it certainly needs further research of many years, but this question can serve as the indicator which leads us to ponder on the phenomena of translation as a whole in order to find useful ideas for the practice of translation.

Nowadays, with the rapid development of the exchange between countries thanks to fast transport, Internet, and media etc, the understanding of other cultures will be easier and easier. Maybe, the translator's informative note will be employed less and less, but it is still too early to conclude that in the future there will be less frequent use of notes in translation, because, with the independent consciousness of the translator, commentary notes may also increase. It is just our hypothesis, which awaits confirmation from systematic lexicometrical examination of translations in the future.

Notes

1. This chapter only addresses the footnotes, leaving an analysis of his prefaces to future research.
2. http://www.ebooksgratuits.com/ebooks.php
3. http://www.yifan.net/yihe/novels/foreign/yhklsdf/klsdf.html
4. http://www.nlp.org.cn/project/project.php?proj_id=6
5. http://elizia.net/alignator/alignator.cgi
6. For further information about the preprocessing, please refer to Miao and Salem (2008) and Miao and Gerdes (2008).
7. The *Lexico3* textometrical program is produced by the university research team SYLED-CLA2T (Systèmes Linguistiques Énonciation et Discours - Centre Lexicométrie et d'Analyse Automatique des Textes), founded in 1997. This software was originally developed by André Salem. See the website: http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/lexico3.htm for further details.
8. The program has been designed by Kim Gerdes, see the website: http://miaojun.net/alignoscope/
9. This counting refers to the edition by the Anhui Literature Publishing House, 1998.
10. There is no page in Fu Lei's translation which has more than 4 footnotes.
11. PCLC (Principales caractéristiques lexicométrique du corpus et de la partition) refers to the main lexicometrical features of the corpus and the partition. The PCLC function of *Lexico3* allows a quick visual comparison of the parts according to their most important textometrical characteristics (see Lamalle *et al*. 2003: 26-27).
12. The analysis of the specific serves to illustrate the frequency of each unit of text in all parts of the body. Here, this feature allows us to examine the distribution of footnotes related to the length of each volume.
13. The specificity of the footnotes in the volume is different from the accumulation of each footnote marker's specificity in the volume.
14. Nowadays, we use "科西嘉岛人" (*kexijiadao ren*) for "Corsian". There are differences between the place names in Fu Lei's translation and the place names used today.
15. About the employment of the personal pronouns in the parallel corpus *Jean-Christophe*, see Miao (2008).
16. This is the number of aligned paragraphs in the complete parallel corpus.
17. Unfortunately, the information Fu Lei provides is not entirely accurate. Indeed, the Taunus is a mountain range in Hessen, and it is bounded by the river valleys of Rhine, Main and Lahn, which are near the cities of Frankfurt and Wiesbaden. Hence, the Taunus is not a mountain chain in northern Germany, but rather in the south-west of the country.

18. Nowadays, we use 哥特 (*gete*) for *Gothic*.
19. See the wiki website: http://fr.wikipedia.org/wiki/%C3%89criture_gothique.
20. Fou Ts’ong is an internationally renowned pianist. He studied piano in many countries. During 1954-1966, Fu Lei and Fou Ts’ong exchanged a lot of letters, and these private letters are collected and published in *Fu Lei’s Home Letters,* by the San Lian Press. The first edition was in 1981.
21. “因你屡屡提及艺术方面的希腊精神（Hellenism), 特意抄出丹纳《艺术哲学》中第四篇“希腊的雕塑”译稿六万余字，订成一本。原书虽有英译本，但对其中的神话、史迹、掌故太多，尚无详注，你读来不免一知半解；我译稿均另加笺注，对你方便不少。” (Fu 2008: 158)
22. “[…] de lecture facultative elles ne s’adressent par conséquent qu’à certains lecteurs: ceux qu’intéressera telle ou telle considération complémentaire ou digressive dont le caractère accessoire justifie précisément le rejet en note” (Genette 1987: 297).
23. “顾精神平衡由足失却，非溺于精神而懵懵懂懂，即陷于麻痹而无所作为。”(Fu 2006: 462).
24. “《约翰·克利斯朵夫》不是一部小说，—应当说：不止是一部小说，而是人类一部伟大的史诗。[...]愿读者以虔敬的心情来打开这部宝典罢！” (Fu 1998: 5)

References

Baker, M. (1993), “Corpus linguistics and translation studies: Implications and applications”, in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology*, 233-250. Amsterdam: John Benjamins.

Berenson, B. and Lazarsfeld, P. F. (1984), *The Analysis of Communications Content.* Chicago and New York: University of Chicago and Columbia University.

Bahier-Porte, C. (2005), “Les notes dans les premiers contes orientaux”. *Féeries* 2: 91-108. URL: http://feeries.revues.org/document106.html.

Fu, L. (1998), *Collection of Fu Lei’s Translations*. Hefei: Anhui Literature Publishing House.

—. (2006), *Collection of Fu Lei’s Works: Volume of Letters.* Beijing: Contemporary World Press

Fu M. (ed.) (2008), *Fu Lei's Home Letters (A Selection with Annotation).* Tianjin: Academy of Social and Sciences Press.

Genette, G. (1987), *Seuils.* Paris: Éditions du Seuil.

Henry J. (2000), “De l'érudition à l'échec: la note du traducteur”. *Meta* 45: 228-240. URL: www.erudit.org/revue/meta/2000/v45/n2/003059ar.pdf

Lebart, L. and Salem, A. (1994), *Statistique Textuelle.* Paris: Dunod.

Lederer, M. (1984), “Implicite ou explicite”, in *Interpréter pour Traduire.* Lederer, M. and Seleskovitch, D. Paris: Didier Érudition.

Lamalle, C., Martinez, W., Fleury, S. and Salem, A. (2003), *Outils de Statistique Textuelle Lexico3*. Paris: Université de la Sorbonne nouvelle- Paris 3.

Miao, J. (2008), "Fu Lei en chiffre: Les pronoms personnels dans la traduction de Jean-Christophe". Paper presented at the International Symposium on Fu Lei and Translation, Nanjing, May 2008. http://miaojun.net/publications/FuLeienchiffres.junmiao.pdf.

Miao, J. and Gerdes, K. (2008), "Donner accès à l'oeuvre de Fu Lei". Paper presented at the International Symposium on Fu Lei and Translation, Nanjing, May 2008. http://miaojun.net/publications/donneraccesaFuLei.GerdesMiao.pdf.

Miao, Jun and Salem, A. (2008), "Comparaison textométrique de traductions Franco-Chinoises", in *Lexicometria*. http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicowww/navigations/Mult2.pdf

Pym, A. (1997), *Pour une Éthique du Traducteur.* Arra: Artois Presses Université, Presses de l'Université d'Ottawa.

Varney, J. (2005), "Taboo and the translator: A survey of translators' notes in Italian translations of Anglo-American fiction, 1945-2005", in *New Research in Translation and Interpreting Studies,* 47-57. Tarragona, Spain, 20-21 October 2006. URL: http://au isg.urv.es/library/papers/VarneyTaboo.pdf.

CHAPTER SIX

COHERENCE IN SIMULTANEOUS INTERPRETING: AN IDEALIZED COGNITIVE MODEL PERSPECTIVE

ERNEST WEI GAO

1. Introduction

The most spectacular and mysterious aspect of simultaneous interpreting (SI) is synchronicity. SI requires the interpreter to listen in one language while speaking in another with ear-voice span (EVS) or lag ranging from 2 to 4 seconds minimum (Paneth 1957) and from 2 to 10 seconds maximum (Oleron and Nanpon 1965). In addition, regarding speech rate, Seleskovitch (1965) suggested that an input rate of 100 to 120 words per minute was a comfortable one for interpreters, with 150 to 200 as the upper limit. This chapter explores, from the perspective of idealized cognitive model (ICM) of embodied Cognitive Linguistics (CL), how simultaneous interpreters maintain sufficient coherence in such a short and synchronic time span. In so doing, the study proposes a SI coherence model, which comprises coherence origin, global coherence and local coherence. A pilot empirical experiment is also presented in support of the model.

For the purpose of the present study, the three components of the model are defined from the perspective of embodied CL. Basically global coherence is a process of providing mental links and relations of ICM corresponding to the structure of the discourse, while local coherence primarily means the network of conceptual relations which underlie the discourse proper ranging from the whole to the part and from a chunk of utterances or neighbouring utterances to inside utterances. Coherence origin is a concept originally put forward in this research. Cognitive process starts with interactive embodiment with the actual world, and it is this interface between interactive embodiment and the reality that gives rise to the coherence origin. Accordingly the SI coherence begins with coherence origin (i.e. integrative embodiment), and threads through the global coherence (i.e. ICM) and the local coherence (i.e. concept and meaning via mental coherence).

In the remainder of this chapter, I will first introduce the concept of idealized cognitive model as the theoretic framework, and then describe the experiment in support of my proposal and finally present and discuss the results of the experiment.

2. The ICM approach

The ICM approach, developed in this study, is an interactive embodiment and mental coherence approach which draws on embodied CL (Lakoff 1987; Johnson 1987; Lakoff and Johnson1980, 1999; Langacker 2008; Wang 2006). Figure 6-1 shows how the present study develops its theme around ICM and how ICM is formed in the cognitive process.

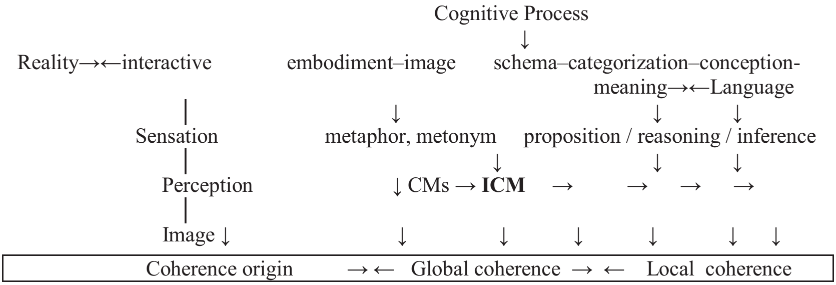

Cognitive Process
Reality→←interactive
embodiment–image
schema–categorization–conception-meaning→←Language
Sensation
metaphor, metonym
proposition / reasoning / inference
Perception
CMs → ICM
Image
Coherence origin →← Global coherence → ← Local coherence

Figure 6-1. ICM in cognitive process

CL claims that the cognitive process starts with interactive embodiment via sensation, perception, image to image schema, "a recurring, dynamic pattern of our perceptual interactions and motor programs that gives coherence and structure to our experience" (Johnson 1987: xiv). To access more complex and abstract concepts, metaphor and metonym are used to build up a model of representations of all relevant knowledge for a certain field, stored in brain and based on a group of relevant situations and contexts; this is how cognitive models (CMs) are built. CMs are designed for categorization and conceptualization pertaining to situations and contexts. And relevant CMs will constitute an abstract, unified and idealized model of understanding experience and knowledge in a certain field. They are idealized cognitive models (ICMs), from which concepts are organized and meanings or senses are made out of language. ICM is a "complex structured whole, a gestalt" (Lakoff 1987: 68), and also "experiential gestalt" (Lakoff and Johnson 1980: 81). ICM, the global pattern knowledge, determines the local coherence of concepts and relations via mental coherence.

It can be seen from the description above that ICM is interactive embodiment based on mental coherence in terms of discourse coherence. This study claims that ICM's interactive embodiment and mental coherence functions are two basic principles for discourse coherence and further uses the two principles as clues to explore and construct this theoretical framework. A second observation is that in terms of SI discourse analysis, the cognitive process involves the coherence origin, global coherence, and local coherence.

2.1. Interactive embodiment

2.1.1. Embodiment

CL emphasizes the role that the body plays in shaping the mind and believes that the nature of the human mind is largely determined by the form of the human body and that all aspects of cognition, such as ideas, thoughts, concepts, categories and languages are shaped by aspects of the body. These aspects include the perceptual system, the intuitions that underlie the ability to move, activities and interactions with our environment.

Lakoff and Johnson (1999) argue that the embodiment hypothesis claims that our conceptual and linguistic structures are shaped by the peculiarities of our perceptual structures. Based on embodied philosophy, CL uses bodily experience as the starting point, conceptual structure and meaning as the research focus and employs cognitive models and knowledge structure for a uniform and cross-disciplinary description of languages in order to reveal the cognitive patterns behind languages.

Taken together, we use the cognitive process to learn the world and its matters interactively, and make sense by categorizing, conceptualizing, reasoning and inferring. In time of perception and embodiment, body and space take the foremost positions, and then other basic CMs are gradually formed on the basis of interaction, giving rise to a variety of basic concepts and propositions, from which meaning is generated by inference and coherence is established.

2.1.2 Interaction

The embodied CL Interactionism View emphasizes the role of subjective initiative in understanding the world, and maintains that language is the outcome of subjective and objective interaction. Lakoff and Johnson (1999: 90) make remarks as follows:

> At the heart of embodied realism is our physical engagement with an environment in an ongoing series of interactions. There is a level of physical interaction in the world at which we have evolved to function very successfully, and an important part of our conceptual system is attuned to such functioning. Meaning comes, not just from 'internal' structures of the organism (the 'subject'), not solely from 'external' inputs (the 'objects') but rather from recurring patterns of engagement between organism and environment.

Discourse coherence stems from interactive embodiment, and it is achieved through interactive mental process when making sense of the whole discourse. The present research moves from interactive embodiment to mental coherence.

2.2. Mental coherence

Embodied CL maintains that discourse coherence should be explained plausibly from cognition. Givón (1990: 914) claims that "the grammatical devices that code referential coherence under various discourse conditions can be interpreted as mental processing instructions." It implies that discourse coherence could not be achieved by either cohesive devices or discourse structures but mainly by mental coherence or "the coherence in mental text" (Givón 1995:59).

Beaugrande and Dressler (1981: 85, 88) argue that discourse generation and apprehensive process should be explained by cognitive process and activation of relevant knowledge. McCarthy (1991: 27) also stresses "making these cognitive links in the text." He further claims that "if we take a text which is cohesive in the sense described above, we can see that a lot more mental work has to go on for the reader to make it coherent."

This study claims that mental coherence is ICM based in the process of understanding discourse; in other words, one intends to access embodied experience based ICM and background knowledge for global coherence, which further determines the local coherence. At this point, a question arises: what are the ICM constitutes?

2.3. ICM constitutes

ICM is constructed on the basis of the four CMs (Lakoff 1987: 68, 113; Wang 2006: 206): Propositional model, Image schematic model, Metaphoric model, and Metonymic model.

2.3.1. Propositional model

According to Lakoff (1987: 285), a propositional model is also defined as the actual mappings or projections of the external world without any need to employ imagination to specify elements, their properties, and the relations holding among them. Much of our knowledge structure is in the form of propositional models in that they are the most primary foundation for the following three CMs.

2.3.2. Image schematic model

As a pre-conceptual image, it is formed on the basis of interactive embodiment with the real world; it is more abstract and generalized than image. Image schematic models are the base for prototype, category, concept, CM and thought, and provide frame and structure for ICM. Image schema helps to construe our bodily experience by analogy and non-bodily experience by metaphor (Lakoff 1987: 453), and also indefinite CMs, categories, concepts, meaning and mental coherence. There are many properties of the schema (Lakoff 1987:272-275), five of which are explained as follows:

a) The container schema. Everything is either inside a container or out of it; an internal structure is arranged so as to yield a basic logic – P or not P.
b) The part-whole schema. The part-whole structure of other objects makes us get around in the world.
c) The link schema. As a string, rope, or other means of connection, structural elements are connected like LINK.
d) The centre-periphery schema. The centre defines the identity of the individual in a way that the peripheral parts do not. It gives rise to structural elements: an entity, a centre, and a periphery.
e) The source – path – goal schema. Structural elements show a source (starting point), a destination (end point), a path (a sequence of contiguous locations connecting the source and the destination), and a direction (toward the destination).

Propositional model and image schematic model are primary models for the others. The image schema theory is designed to explain not only how we perceive and organize the world cognitively but also how we establish discourse coherence.

2.3.3. Metaphoric model

CL argues that metaphors are conceptual, not linguistic. A metaphoric mapping involves a source domain and a target domain. The source domain is assumed to be structured by a propositional or image–schematic model. The mapping is typically partial; it maps the structure of the ICM in the source domain onto a corresponding structure in the target domain by cognitive entities through inference and different entities are linked to each other to access more understanding of the matter in question. Metaphoric models are turned into a cognitive model or inferential mechanism and play an important role in shaping categorization, conception, and inference.

2.3.4. Metonymic model

This means that the part easy to perceive in the same cognitive domain is used to understand the whole or another part. For instance, a typical member in a category is used to explain the whole category. Metonymic models are manifested in communication in that language could not be used for the complete representations of the reality by economy (see section 2.6.1).

Propositional models are objective while the other three are subjective. The first two models are designed to explain the contents and foundation of ICM while the latter two models are used as extensions of ICM. The four models are the foundation of the theoretical framework adopted in this study.

2.4. The cognitive world = ICM + background knowledge

On the basis of Lakoff (1987) and Lakoff and Johnson (1999), Wang (2006: 360) puts forward his theory of cognition world for discourse coherence, which falls into two parts as follows:

1) Idealized Cognitive Model (ICM).
2) Background knowledge means unconventional and specific knowledge, not necessarily universal or representative, either

> mutually knowledgeable or acquired through the online and immediate communication, changeable and dynamic in a specific context to compensate and confirm, or adjust, modify and even change the communicators' assumed knowledge and the current communication.

Therefore, the cognitive world refers to general knowledge acquired by cognition via interactive embodiment, internalized and stored in mind, either shared or accessed from the online discourse. Wang's (2006) idea of cognitive world helps to explain the conventional (ICM) and special or online (background knowledge) phenomena of coherence respectively. In communication, the reason why the conventionalized coherence is recognized depends primarily on the cognitive world: ICM + background knowledge. On one hand, the conventionalized phenomena and the prominent information (ICM) will have more chance to be activated and become topics, and will limit the scope of concepts to be activated; on the other hand, communication has complex scenarios due to the fact that it changes all the time, and it is not always predicable so that the background knowledge is used for accidental and online discourse.

At this point, a further question arises: how does ICM work in discourse? This research will try to find an answer in Zwaan's (2000, 2004, and 2005) Immersed Experiencer Frame.

2.5. The way ICM works in discourse: Immersed Experiencer Frame (IEF)

Zwaan (2000, 2004, 2005) establishes the IEF (Immersed Experiencer Frame) model for language comprehension, which views the comprehender as an immersed experiencer via interactive embodiment, and views apprehension of language input as indexing clues and representations of experience via meshing and sequencing trace. Meshing means the process to associate words and phrases in terms of the potentially interactive relations between the body and the objective matters to constitute a coherent action model. For instance, the word *chair* suggests its purpose of a seat on which one sits. To the comprehender, language is a set of clues to help simulate what is experienced in the sense of context. As immersed experiencers, both the speaker and the listener will have access to ICM for successful communication.

IEF divides comprehension processing into three basic elements: activation, construal and integration. The three elements correspond to the language in three processing units in terms of (1) language processing unit, (2) representation unit and (3) reference unit. This is illustrated in Table 6-1.

Table 6-1. The three elements corresponding to the language in three units

3 elements 3 units	Activation	Construal	Integration
Language processing	word / phoneme	clause / intonation	discourse
Representation	functional web	meshing net	sequencing web
Reference	object & action	event	event sequence

Just as Zwaan emphasizes (2004, 2005), the three elements of processing are not undertaken sequentially but overlapping to a large extent. The three elements corresponding to the language in three units are explained as follows.

2.5.1. Activation

Input words activate the functional web, which is activated when words are associated with referents. In other words, the experiential representation of the word and the relevant experiential representation of the referent are associated due to the co-occurrence. In this way, the word will not only activate its own experience (for instance, the word *chair* suggests its purpose for a seat in sense of audio-visual experience) but also its relevant experience of the referent (for instance, the word *chair* is used as a weapon to hit someone, involving the audio-visual and touch experience). In short, language input activates the traces and reconstructs the experience model.

2.5.2. Construal

Construal is the multi-functional web meshing operation for the psychological simulation of a specific event. The grammatical unit in the construal operation is the intonation unit. Zwaan (2004, 2005) sees language comprehension as attention adjustment for the events described. In construal, the event representations are generated through the initial activation of the functional web. Construal is an immediate and compounding process.

Every construal comprises a time frame and a space frame. In the time frame is a perspective, which normally means the perception of the protagonist. Every construal also includes focal entity, relation and background entity and the entities might have explicit features. Therefore, time, space, perspective, entities and features constitute the elements of construal.

2.5.3. Integration

Once an event representation is construed, it moves on to the next one. Previous relevant construed elements will be turned into a part of working memory, and affect the current construal along with the functional web activated by the current words. Integration means transfer from one construal to another following it. IEF assumes that the transfers are based on experience.

Zwaan (2004, 2005) views language comprehension as attention control of the event description. Therefore the agreement between language and experience, the prospective and retrospective overlapping on the basis of ICM, predication and coherence indexing clues will affect the integration from the beginning to the end.

The account of ICM above points to the question: what is the ICM role in establishing coherence? This question will be answered in the section that follows.

2.6. The ICM role in coherence

The ICM makes itself felt in three aspects: 1) activation of the cognitive world and filling in the missing links or default values, 2) mapping iconicity, and 3) achieving local coherence.

2.6.1. Activating the cognitive world and filling in the missing links or default values

Default values refer to values for “a slot that are used if no specific contextual information is supplied” (Lakoff 1987:116). In other words, the speaker could not present his idea fully in communication as he will have to make selection out of the whole information and mostly express verbally. Meanwhile the listener will turn to his ICM and background knowledge to access relevant information through activation of some elements of utterances, and further understand the whole discourse. The default values also show the communication feature of economy in that utterances connote more information than the literal meaning and default values help the listener to actively construct cognitive models for understanding discourse and to access various coherence clues.

Theoretically any word in a clause will tend to activate any information of ICM and turn out to be a starting point for the following utterances to achieve discourse coherence. In other words, because communication is restricted to mind and script, including social, cultural and experiential contexts and situations, some pieces of information are more prominent and salient than others, and tend to be more easily activated into the topics. This means that concepts exist at different levels in terms of salience, i.e. the concepts at frontier and remote level are less likely to be activated than those at prominent level. Therefore, discourse coherence is in right proportion with cognitive distance: the closer the cognitive distance in bridging the concepts in pairing utterances, the shorter the time to trace the propositional references.

2.6.2. Mapping iconicity

ICM gives rise to iconicity: similarities between the structure of language and the structure of the world. Iconicity plays the role of reference for both the objective world and language structures, including world structure, concept structure and embodied experiential structure. As Wang (2006) claims, discourse coherence, when a kind of mapping similarity between objective externals and mental internals is accessed, will emerge automatically in the chunk of discourse or a group of sentences. This is iconicity.

Iconicity, from CL perspective, stresses that language forms are outcomes of a variety of external and internal integrations of embodiment, cognition, semantics and pragmatics, etc.

Regarding the research on iconicity, this research focuses on iconic sequencing similarity between language sequence and thinking process. Several experiments of sequencing iconicity have been conducted and conclusions have been made (Clark and Clark 1968, Smith and McMahon 1970, Engelkamp 1974 as cited in Wang 2006:556, He and Ran 2006). The results of the experiment data conform to the principle of iconic sequencing: Sentences organized to natural sequence or the storage process are easier to access for relevant information in that it takes less amount of cognitive process; on the other hand, when sentence sequences go against the natural action sequence, the sequential information will not be accessed straightforward from memory in cognitive process in that it is difficult to convert language sequence into time sequence due to the mismatch, and it also requires more time and effort in inference. Sequencing iconicity is actions sequence and fits in the embodied cognition.

ICM, background knowledge, i.e. the cognitive world and iconicity help to establish global coherence among the utterances in actual communication, and also determines local coherence in discourse**.**

2.6.3. Determining local coherence

Wang (2006: 381-385), from cognitive perspective, puts forward three conditions for coherence, which this research terms as local coherence conditions: (1) conceptual bridge, (2) propositional reference, and (3) pragmatic inference. In CL, abstract concepts are transformations of concrete concepts derived from embodied experience and a proposition means “a continuous, analog pattern of experience or understanding, with sufficient internal structure to permit inferences” (Johnson 1987: 3-4).

Bridging reference is the cognitive and pragmatic phenomena to determine the referents or antecedents by inference for interpretation. And bridging inference is the process to associate the referents or antecedents for pragmatic interpretation.

By bridging reference and inference, the sense is made out of clauses and coherence is established. Inference plays a primary role in that the inference is made on the listener or reader’s part to arrive at an interpretation. Inference is seen as a process of filling in the missing links between utterances as non-automatic connections between elements in a text via prior knowledge representations used as a basis for deciding which missing links are, which are not, likely to be inferences and as filling in gaps or discontinuities in interpretation in the local discourse (Brown and Yule 1983: 257- 265).

Framed in contexts, communicators, on the basis of information provided by utterances, have access to their embodied experience based ICM to activate the conceptual elements and bridge them, trace the propositional references, and undertake pragmatic inference in the forthcoming and backward contexts, ultimately to construct a unified cognitive world for global and local coherence.

2.7. The ICM based methodology

So far, this chapter has provided CL theories relevant to coherence in SI from a top-down approach to coherence. The following will deal with specific methodological strategies for interpreting discourse coherence from a bottom-up perspective to show how the coherence is realized, or rather materialized in the discourse. The ICM based methodology is derived from cognitive grammar (Langacker 2008), including Cognitive Reference Point (CRP) and Trajector-Landmark.

2.7.1. Cognitive Reference Point (CRP)

According to the schema in section 2.3.2, our bodily experience tells us to locate the control centres in daily life and the control centre in the structural elements of discourse analysis. The world is conceived as being populated by countless objects of diverse characters. In the cognitive world are numerous mind entities and concepts, where once one concept in the cognitive world is determined a corresponding domain is to be laid out initially to "establish a mental contact" (as indicated by the dotting lines in Figure 6-2), locating a Cognitive Reference Point (CRP), and further narrowing the domain to access a concept. In language expression, normally once a descriptive scope is determined, it moves to the contents inside (Langacker 2008: 84).

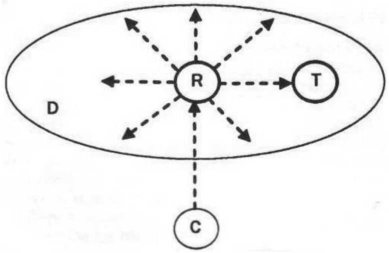


Figure 6-2. Reference-point relations
[Legends: C=Conceptualizer; R=Reference point; D=Dominion; T=Target]

This theory is applied to the analysis of local discourse coherence. In terms of a chunk of discourse, one theme, or theme + express are determined as reference points, serving as CRPs, followed by developing utterances and new information; and in this way discourse coherence is achieved. A discourse or a chunk of utterances, either spoken or written, will focus on one or more than one theme to gain the commanding role.

2.7.2. Trajector and landmark

The schema (section 2.3.2) also gives rise to the trajector (tr) and landmark (lm). CRP is further reduced into trajector and landmark to analyze individual utterances / clauses, as can be seen Figures 6-3a and 6-3b, which show a chunk of utterances where trajectors and landmarks are processed in a linear and constant form. When a relationship is profiled, varying degrees of prominence are conferred on its participants. The most prominent participants are called the trajector while some other participant is made prominent as a secondary focus, called a landmark (Langacker 2008:72).

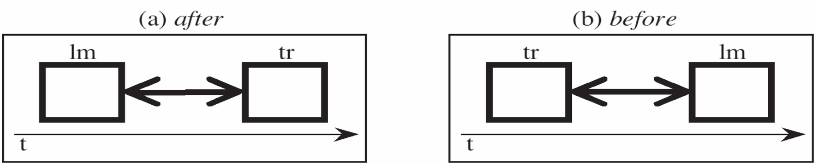


Figure 6-3a

Figure 6-3b

Figure 6-3. Trajector and landmark in clause and utterance

This research use image schema based CRP, trajector and landmark as coherence clues (refer to Figures 6-5 and 6-6 in section 3.3.2 for how they are used in this study).

2.8. Hypotheses

On the basis of the theoretical framework, the following hypotheses are formulated to address the research question in this study of how sufficient coherence in SI is achieved:

1) Coherence in SI stems from interactive embodiment, the coherence origin, and is established on the basis of ICM centred mental coherence;
2) ICM helps to fill in missing links and default values and access the cognitive world primarily in the form of propositional model and image schematic model in SI coherence;
3) ICM helps to realize local coherence.

This chapter has so far presented a complete, plausible and uniform SI coherence model. The section that follows will present the methodology and data used in this research.

3. The methodology

3.1. Materials

The corpus used in this study is based on a sample of English to Chinese interpretations from simulated conference session with transcripts for reference. The English discourse lasts 4.5 minutes with the transcript divided into 44 segments and 22 sentences (see the appendix for the transcript).

The local context is the occasion where Mr Craig, the sales manager from Hömedics in Guangzhou International Fair explains a new product, the Shiatsu Massage Cushion to the audience or potential customers. Mr Craig first greets the audience and introduces himself (S1-S3), then establishes the objective of promoting the product Shiatsu Massage Cushion (S4), and explains the meaning of Shiatsu (S5) and its purposes (S6). In order to achieve an interactive response from the audience, he asks the potential customers for their idea of the product (S7). Following this, he concentrates on his three points - structure (S8), fixing procedures (S9-15) and warnings of storage (S16-19) - before he ends his promotion by another strategy, namely concessional price (S20). The speaker playing the role of Mr Craig is a native English speaker, who has rehearsed before the interpreting.

This use of this corpus is justified in that the corpus data address the research question on the basis of following: (1) The subjects are placed in a real situation, physically experiencing the installation process and the purpose of the device. (2) On the basis of interactive embodiment subjects form an ICM, i.e. the subjects have built up propositional and image schematic model of the fixing procedures via mental space blending mostly in terms of frame and script and also map iconicity in the minds, i.e. the similarity between the structure, fixing steps of Shiatsu, in particular, the sequential iconicity is quite obvious. (3) The coherence clues are apparently displayed in the data in terms of CRPs, trajectors and landmarks. (4) This corpus is used for interpreting English for science and technology (EST) in this context and arguably in any other similar contexts also due to the fact that a large amount of EST is concerned with physical structures (physical descriptions), with the purpose of a device and how its parts work (functional descriptions), also with processes and procedures (procedural descriptions); EST discourse is similar to the natural patterns of time and space (rhetorical techniques) (Trimble 1985: 19). (5) The usefulness of this small corpus also supports Leech's (1991: 8-29) claim that size is not all-important. Small corpora may contain sufficient examples of frequent linguistic features. Also according to McEnery, Xiao and Tono (2006: 72), corpora that need extensive manual annotation or analysis involving human analysts, as in this case, are necessarily small.

The transcript of interpreting discourse, as one of the relevant sources of data, is used to describe and explain the facets of the coherence clues used in simultaneous interpreting. The transcript is annotated with top-down and bottom-up marks.

3.2. Subjects

There are three Chinese subjects of conference interpreting, one PhD (IA: Interpreter A) in conferencing interpreting, female, 30 years old, with 80 hours of actual conference interpreting experience, and two MSc trainees in conference interpreting, one male (IB: Interpreter B), the other female (Interpreter C), both 23 years old without any conference interpreting experience. The recruitment incentive is to compare the experienced simultaneous interpreter with non-experienced in the CL based theoretical framework of interactive embodiment and mental coherence. The three subjects are not in any physical contact with the product of Shiatsu.

The subjects' equipment was a standard laboratory with SI booths used for teaching simultaneous interpreting at Heriot-Watt University, UK. The subjects have interpreting lectures and slots in the lab, and they are therefore familiar with the environment which thus has no negative effect on them. The subjects' outputs were recorded on MP3, played back and timed on a PC. After the experiment, the subjects were provided with their respective outcomes with transcriptions as well as assessment.

3.3. The procedure

Immersed Experiencer Frame (IEF) (section 2.5) is used as guidance to SI coherence via the steps of activation, construal and integration. This study involves three experiment steps: 1) interactive embodiment experience to access SI coherence origin (section 3.3.1); 2) activation and construal for SI global and local coherence (section 3.3.2), and integration for SI coherence whole (section 3.3.3).

In order to trace down to the realizations or materializations of ICM, the two strategies discussed in section 2.7, namely cognitive reference point (CRP) and trajector-landmark, are used to mark the coherence clues.

3.3.1. Interactive embodiment: SI coherence origin

The experimenter brought the product of Shiatsu (see Figure 6-4) into the lab and organized the subjects to fix the Shiatsu parts according to the experimenter's oral instructions in Chinese. Then each subject sat on it and used the remote control to experience the massage. The subjects were guided to find out its physical structure, purpose, process, procedures and how its parts work. Having finished the step of interactive embodiment, the experiment moved to the second step.

Figure 6-4. The Shiatsu massage cushion

3.3.2. Activation and construal for SI coherence clues

The cognitive maps, as illustrated in Figures 6-5 and 6-6, are now filled (as reference only) with CRPs for topics and trajectors (slots) and landmarks (fillers), distributed to subjects for them to fill in with two cognitive reference points present to them i.e., the structure and fixing process, and the subjects were asked to keep a track of the mental path/clues (Lakoff 1987: 283) in the relational structure. In addition, Figures 6-5 and 6-6 show how the image schemas (section 2.3.2) work, including the container schema, the part-whole schema, the link schema and the source-path-goal schema. The cognitive slot-filler pair is illustrated in Figures 6-5 and 6-6.

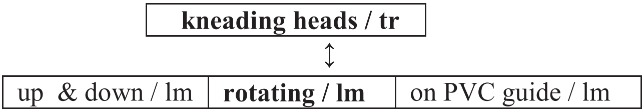


Figure 6-5. Shiatsu structures / CRP, trajector and landmark

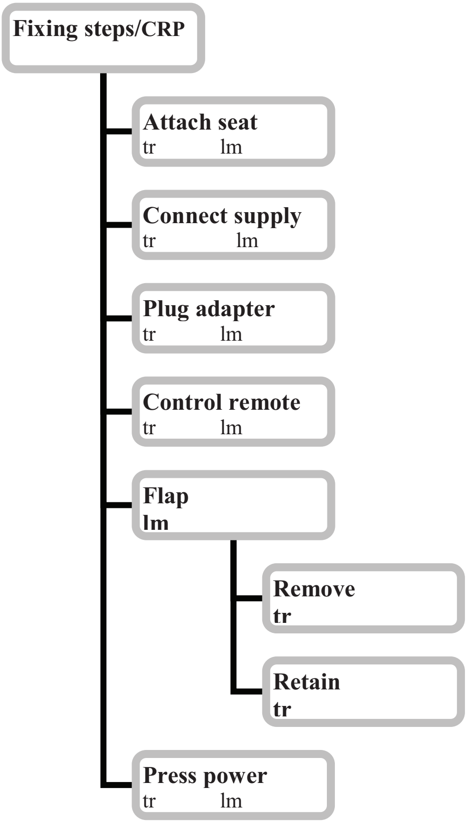


Figure 6-6. Fixing steps of Shiatsu / CRP, trajector and landmark

The subjects were given five minutes to mentally construe the items in the diagrams by themselves by rehearsing the speech in the speaker's role. The construal activities help the speaker and the interpreter to share the same pattern of image-schematic model in one domain, connected via their shared purpose on this occasion with shared structures and situations, chains of actions and events within the frame and script. Before interpreting, the interpreter can rely on mental mappings to activate the online speaker's corresponding structure for global coherence.

3.3.3. Integration for SI coherence whole

When subjects were sure that the recording was effective, the native English speaker started the speech with 120 words per minute. When the subjects finished, their recordings were collected and later on the interpreting outcomes were transcribed.

3.4. Data manipulation

In all the coherence and comprehension research assessment is made of the degree to which the response protocols – in this case, the subjects' translations – match or mismatch the input text (Ericsson and Simon 1984, Dillinger 1994). As a fundamental step of generating data from the observations, this is done in terms of CRP, trajector and landmark. CRP is used to determine the themes, the coherence clues, serving as a cognitive commanding role and setting the tone of coherence of the whole discourse. Trajector, landmark, concepts / slots and relations / fillers in the propositions are used as coherence clues to determine the prominent and secondary focuses inside clauses.

This study puts the SI coherence in the global and local discourse frame, observing SI coherence by combining ICM with the three local coherence conditions (section 2.6.3) via CRP (section 2.7.1), and trajector and landmark (section 2.7.2).

It is analyzed by cross-section in terms of three aspects: the embodied section, semi-embodied section and non-embodied section. The embodied section refers to the part in source language (SL) and target language (TL), which is interactive embodiment based sections, mapped and projected into the minds of the interpreter through mental space blending, ranging from sentences 8 to 15, segments 15 to 33 in the transcript (see the appendix). The semi-embodied section means the part in SL and TL, which is not directly or completely embodied but could be inferred or accessed by filling in the obvious missing links and also regarded as extensions of the embodied ICM, ranging from sentences 3 to 7, segments 4 to 14. The non-embodied section designates the part in SL, which is not embodiment based and remote from the core – embodied experience, including sentences 16 to 20, segments 34 to 43.

The descriptions above are based on comprehension; this study further observes the interpreting, that is, how local coherence is achieved in cognitive process and interpreting action and product via the cohesive clues, including anticipation, coordination, judgment and compensation (see section 4.2.4).

The attempt in the present study to measure the coherence clues using a small number of subjects made a more refined adaptation of this method necessary. The units of comparison are the individual slot-filler pairs that constitute each proposition by trajector (slot) and landmark (filler), and further, in which (trajector and landmark) the concept (slot) and relations (fillers) are marked rather than match entire propositions. That is, each slot-filler pair of each utterance in SL receives a score according to the degree of similarity between it and the segment of the subjects' response being analyzed, using the following ordinal scale of similarity:

- 0 if the slot-filler pair is not present in the segment, which is least similar.
- 0 if a sequence of the slot is mismatched with the SL. It is interesting to observe how any mismatch sequences affect the upcoming interpretation.
- 1 if there is a change in surface form of the filler without a change in meaning (paraphrase).
- 1 if the slot-filler pair appears in the segment verbatim, which is most similar.

For each of the experimental texts a database is constructed, using Microsoft Excel, in which each record (row) corresponds to a text proposition, and each field (column) corresponds to information about the cognitive linguistic properties of the text. This has made it easy to generate information about propositions with a given property (e.g. CRPs, trajectors and landmarks) once the raw data matrices are appended to these databases as new fields, generating dependent measures by performing calculation functions built into Microsoft Excel.

4. Results

We will see how sufficient coherence in SI is achieved by interactive embodiment, the coherence origin, ICM / global coherence, and local coherence through activation, construal and integration.

4.1. Activation and construal: coherence clues in the functional web

Quantitative analysis is undertaken according to the slot-filler pairs. Only words pertaining to the structure and fixing steps are considered to be acceptable, with one point for one slot, and then points are converted into a percentage (score ÷ n * 100).

The structure has four frames and the fixing steps have seven frames. The top two frames as cognitive reference points are not taken into account since they are already presented. The six steps are doubled because every frame has two slots in the form of a trajector and landmark. So there are fifteen points.

Table 6-2. Outcome of activation and construal of coherence clues

Slot-filler pairs / Interpreter	IA	%	IB	%	IC	%
15	14	93	13	86	11	73

Interactive embodiment is the coherence origin in SI coherence. The interactive embodiment, directly proportional to the SI coherence, makes itself felt in each section: activation, construal and integration / interpreting. Over the activation and construal, the subjects have a propositional and image-schema model of Shiatsu in terms of its structure and fixing steps via mental space blending. The three interpreting subjects, after their bodily experience with Shiatsu, filled the spaces to the frames and scripts with the coherence clues by 93%, 86% and 73% respectively. It is not surprising that the experienced IA does better than the inexperienced IB and IC but Pearson product-moment correlation coefficient shows 5% between IA versus IB and IC, which suggests that the three subjects share similar levels of cognitive competence. This also shows that they have the structure and fixing procedures mapped into their minds perfectly in terms of IA and basically of IB and IC and that the three trainees have activated the functional web, equipped with propositional and image-schema models, which give them access to global and local coherence in their interpreting.

In addition, the data of activation and construal also shows more specific and detailed cognitive processing properties in building up ICM to make sense and establish coherence. For instance, the data demonstrates that the prominent, dominant and central concepts will be activated more easily than the subordinate and obscure ones. In terms of the Shiatsu structure (Figure 6-5), the three subjects could access the action of kneading heads, "move up and down", but how (rotating) and where (on the guide) and furthermore, what material (PVC) of the guide are less activated. Only one (IA) of the three subjects recalls "rotating" out of the three points, which means that the experienced IA might be more observant than IB and IC. As a matter of fact, to be observant and ready to accumulate relevant knowledge in the situation and context is one of qualities for interpreters. This phenomenon suggests that interpreters should do their utmost to activate any nodes, even not so prominent or salient in the functional web so as to build up the detailed coherence.

The data of fixing steps also shows that the three subjects arrange the sequential order according to the natural sequence. This result means that the iconicity of the process is projected fully into the interpreters' minds and that subjects have fully entered into the speakers' mental spaces in that the subjects have natural chains of action stored in their minds, avoiding obstacles in generating the sequential discourse, which is in line with the conclusions of iconicity by cognitive linguists.

The data of activation and construal also shows that the interpreter subjects have put themselves in the position of the given speaker, being able to infer the speaker's thinking process, the purpose of the utterances as well as the emotions and attitudes. These observations in turn lead to other questions to be answered: How do the experienced and the inexperienced interpreters perform in the embodied section, the semi-embodied section and the non-embodied section for SI coherence? Are there any differences between them?

4.2. Integration: coherence clues between SL and TL

4.2.1. The embodied section

This includes coherence clues in the matching section in the frame and script of structure (S8) and fixing steps (S10-S15).

Table 6-3. The outcome of slot-filler pairs for embodied sections

Slot-filler pairs Interpreter	IA	%	IB	%	IC	%
15	14	93	15	100	14	93

Table 6-3 shows that the experiential structures are projected into the minds via the interactive embodiment by mental construal mostly in terms of propositional model and image schema model. It was felt in subjects' distribution of attention as they were able to effectively retrieve the relevant information from their memory storage and filling in the missing links and default values. The coherence clues (trajectors: verb/predicate and landmarks: noun / objects) were more closely matched than those in the activation and construal section respectively up to 14% (IB) and 21% (both IC and IA). It shows that the embodiment based frame and script underlies their performance for the global coherence; that when they have propositional model and image schema model, they could have easy access to the local coherence clues in form of trajector and landmark, concepts (slots) and relations (fillers); and that the subjects are free to bridge reference and inference for local coherence.

It is notable that the findings support Fillmore's (1982, 1985) observations of verbs, i.e. verbs are related with the whole situation, and could make some aspects more prominent and salient. In other words, it is the actions of the fixing procedure, which turn out to be the trajectors, the prominence and salience in the whole installation process to establish coherence. This finding challenges Wang's (2006) claim that trajectors are mostly animated living things or people as subject(s) in sentences; but here in this experiment, it is a series of actions, or verbs rather than nouns or animate living things or people that act as the trajectors in this technical context for the sake of coherence. Therefore it further supports Langacker (2000: 331, 359) in that the choice of subject and object is neither a logic nor a grammatical problem but rather an issue related to cognition construal, mental focus or prominent concept.

4.2.2. The semi-embodied section

This includes sentences 3 to 7, segments 4 to 14.

Table 6-4. The outcome of slot-filler pairs for semi-embodied sections

Slot-filler pairs Interpreter	IA	%	IB	%	IC	%
25	23	92	23	92	18	72

Table 6-4 shows that the interpreters could rely on propositional model and image schema via the mental space in terms of frame and script as the mappings and bridge reference and inference for coherence. Semi-embodied sections are extensions of the embodiment, around which the subjects could bridge reference and inference in the given context without effort. However, the semi-embodied sections are less fulfilled than the embodied sections to 1% (IA), 8% (IB) and 7% (IC). It implies that inference makes difference in SI interpreting, which is indicated in the subjects' inferential ability: IA and IB have a better inferring ability than IC.

4.2.3. The non-embodied section

This includes sentences 16 to 19, segments 34 to 43. In Table 6-5, non-embodied sections might be considered to be online background knowledge defined by Wang (2006, see section 2.4), which are unconventional and unpredictable.

Table 6- 5. The outcome of slot-pairs for non-embodied sections

Slot-filler pairs / Interpreter	IA	%	IB	%	IC	%
27	15	56	11	41	10	37

The data suggests that the online, changeable, always unpredictable communication is problematic for the interpreting subjects. Omissions occur up to 46% (IA), 49% (IB), and 63% (IC) for the three subjects respectively. It shows that the subjects without the embodied frame or script or knowledge of what is being interpreted are not able to follow the speaker as in the embodied sections or semi-embodied sections and that in this case they could not bridge the reference or inference for coherence.

This finding of the non-embodied section challenges Pőchhacker's (1992) claim that the interpreter as a low-knowledge individual, i.e. in technical conferences, will still manage to establish a sufficient degree of coherence as a basis for target text production, suggesting that Pőchhacker's observation is incomplete in this regard: without a comprehensive model of the text contents, the interpreter will not rely on the propositions to build up coherence as "models are assumed to be easier to remember than propositions because they are more elaborated and structured" (Setton 1999: 17).

The three subjects did not interpret the "230 volt AC" (S14) correctly but this so-called subject matter knowledge should belong to popular science, used in daily life. It implies that the interpreter should be equipped with enough knowledge of the subject matters or popular science to be professional and specialized in their interpreting career. Therefore, how to help interpreting trainees to build up the required knowledge or to comprehend the abstract concepts in EST by metaphor, metonymy and mental imagery is another question to be answered. In particular, how does one interpret a more complex and specialized subject matter? Furthermore, what is demarcation between abstract language and embodied language? Are there any hierarchical ranks among them? If so, what are the relations between them? The present study proves that the propositional model and the image schematic model are theoretically and empirically feasible in terms of frame and script but more research is needed for metaphorical and metonymic models.

4.2.4. Local coherence in cognitive process and interpreting action and product

The dataset described above covers the three sections: embodied section, semi-embodied section and non-embodied section. How do simultaneous interpreters, framed in ICM and based on inference, achieve local coherence in cognitive process and interpreting action and product in those sections? In what follows I will seek to answer this question by discussing four kinds of local coherence in cognitive process and interpretation and product: anticipation, compensation, judgment and coordination.

Anticipation (Table 6-6: IA, IB and IC):

Anticipation is normally divided into two types: linguistic anticipation based on the predictable or formulas or syntactic and extra-linguistic anticipation, including ICM and local coherence conditions.

Table 6-6 is taken from a semi-embodied section. The three subjects have a gestalt and whole mental structure in the form of propositional model and image schematic model of Shiatsu. When the speaker explained the term of Shiatsu in serial number 7 by *shi* meaning ‘finger’ and *stsu* standing for ‘pressure’, the three subjects could bridge the two concepts in serial number 8, while making pragmatic inference of the paired utterances by using 所以 ‘so, therefore’, the cause-effect connective device to make the serial numbers 7 and 8 coherent.

Table 6-6. Anticipation in SI coherence

7	S5: **As you know, Shi in Japanese means finger - and Shiatsu means pressure,**	**-大家可以知道，其实** As you know, in fact **Shiatsu 在日文里面意思就是指压** Shiatsu in Japanese means finger and pressure	**大家都知道，Shiatsu 在日语里** As you know, Shiatsu in Japanese **分别是手指和按压的意思** respectively means finger and pressure.	**@在日本来说 Shiatsu 是压力,另一个词叫手指** In Japanese, Shiatsu means pressure, and finger
8	**put them together and you've got Shiatsu;-meaning 'finger pressure'**	**比如，你把"手指"和"压"放在一起，就是指压的意思** For instance, you put finger and pressure together, (it) means finger pressure	**所以，放在一起就是指压按摩** So, put (them) together, (it) means finger pressure massage	**所以加在一起就是压力手指的意思** So put (them) together, (it) means pressure and finger

Compensation (Table 6-7: IA, IB and IC)

Compensation means supplementing and completing referents, which have been approximated to, or meanings of which have come too late for ideal integration.

Table 6-7 is a semi-embodied section. The speaker uses ellipsis in each of the serials from 12 to 14, but the three subjects augment and bridge the relevant referents in their respective interpretation, helping listeners achieve maximum relevance with least effort.

Table 6-7. Compensation in SI coherence

12	**How do you feel about it?** @	**你们感觉怎么样？** How do you feel like,	**那大家的感受是怎么样的？** How does everyone feel like?	+
13	**Comfortable?** +@	**舒服吗？** Comfortable?	**是不是很舒服呢？** Do you feel comfortable ?	++
14	**Yes,- I think you will agree.**+	**我相信你们一定感觉是很舒服的** I believe that you surely feel comfortable	**我想答案是肯定的。** I think the answer is positive	**你们刚才的个人体验是否非常的好？** Was your personal experience really good?

Table 6-8. Judgement in SI coherence

41	S20: **You could have it now at- 30 pounds** .	**你可以在爱丁堡约翰路易斯可以买到，六十九镑** You could get (this massager) in John Lewis, Edinburgh at 69 pounds	**你通过可以在这里爱丁堡 John Lewis 百货商场以六十九镑的价格购买** You could buy one at John Lewis supermarket at 69 pounds	**你在这里持三十镑，** You could buy it at 30 pounds
42	**In John Lewis, Edinburgh, it costs- 69 pounds**	**+但是，现在，只需要30镑** But now you could have one at only 30 pounds	**但是，在这儿，这里三十英镑的价格就可以购买** But here you could get it at 30 pounds	**然而在爱丁堡的 John Lewis 商店，69镑才可以买到** But you will have to pay 69 pounds at John Lewis, Edinburgh

Judgment (Table 6-8: IA, IB and IC)

Judgment is displayed in the interpreters' treatment of implausible input or deficit of background knowledge or rather ICM and the local coherence conditions. The interpreters are not visibly affected by speaker's implicated discourse.

Table 6-8 is a non-embodied section. The speaker here pauses between serial numbers 41 and 42 for 2 seconds, but the three subjects can discern the semantic contrast between the utterances by pragmatic inference and add connective device 但是 / 然而 'but / however' to keep the discourse coherent.

The opposite case of judgment ranges from non-embodied sentences 16 to 19, where the subjects show more stops, omitted importation, resulting in incoherent interpretations. It implies that the subjects are struggling for ICM and background knowledge, thus losing the links to achieve global coherence and ultimately ending up with failure in the four local coherence conditions.

Coordination (Table 6-9: IA, IB and IC)

Coordination centres its attention automatically by ICM based default on coordination between input and output, judging inputs while producing coherent speech.

Table 6-9. Coordination in SI coherence

6	**Two kneading heads,**	**它有两个导轨**+ It has two guides	**是通过两个按摩头** Two kneading heads	**有两个小转头** Two small rotating heads
7	**- rotating,**	**它们会上下移动**+ Travelling up and down	++	
8	**+ travel up and down along the PVC guides,**	**我们有两个按摩头可以顺着导板可以上下移动**+ Two kneading heads could travel up and down the guides	**这两个按摩头通过在固定在PVC的两个导轨上上下移动** The two kneading heads travel along the two PVC guides	**上下移动** move up and down

Table 6-9 is an embodied section. From this section of the corpus it can be seen that the three subjects are meeting the challenge of the complicated descriptions of the machine’s working mechanism, but they can coordinate by turning to the embodied image schema and manage to bring together the main information by bridging the concepts and propositions to achieve coherence.

5. Discussions

This study attempts to present a gestalt for coherence in SI by establishing a model: interactive embodiment (coherence origin) + ICM (global coherence) + local coherence, in order to address the research question of how sufficient coherence in SI is achieved.

Table 6-10. Total slot-filler pairs

Total slot-filler pair Slot-filler pairs sections	IA	%	IB	%	IC	%
Activation/construal (15)	14	93	13	86	11	73
Embodied (15)	14	93	15	100	14	93
Semi-embodied (25)	23	92	23	92	18	72
Non-embodied (27)	15	55	11	41	10	37
Total (82)	66	80	62	75	53	65

Table 6-10 shows that when the subjects have the interactive embodiment experience of Shiatsu, they have image schema of its structure and fixing steps in terms of frame and script. Therefore, they can fulfil the embodied section interpretation. During interpreting, they have easy access to the working memory to bridge reference and inference, i.e., filling the concepts / slots and relations / fillers in terms of trajector or landmark. Close to the embodied section, the semi-embodied section can be seen as an extension of the embodied section, which can be interpreted as if it was the embodied section even though it is actually not. But when they move further away from the embodied section to the non-embodied section, the three subjects start to confront difficulties resulting in long pauses and omissions of information. Therefore, embodiment is the origin of coherence, and ICM helps to set up the functional web, meshing net and sequencing events and actions, finally to achieve local coherence.

This observation also shows that the iconicity of the embodied process is projected into the interpreter's mind through embodiment. The subjects have natural chains of actions and events stored in their minds, accessing the sequential discourse, which agrees with the conclusions of iconicity by cognitive linguists: information organized to natural language sequence is easier to remember (Clark and Clark 1968, cited in Wang 2006: 555-556); and the subjects tend to analyze and store events according to the natural sequence to establish coherence (Smith and McMahon 1970, Engelkamp 1974 cited in Wang 2006: 555-556).

The results of the local coherence in cognitive process and interpreting action and product support the hypothesis: it is interactive embodiment and ICM that help the simultaneous interpreters to build up the links among TL speech. This point is demonstrated in the fact that the experienced (IA) and the inexperienced (IB and IC) interpreters process the output in the same way. This finding has significance for the SI training. It is not rational only to make SI trainees undergo so-called Devil's training in SI strategies; but instead, they should be provided with enough cognitive context, which can lead them into an interactive embodied process.

I have so far examined how simultaneous trainees perform in embodied, semi-embodied, and non-embodied sections, and how the three SI subjects achieve coherence in terms of cognitive process and interpreting action and product in this experiment. Based on the theoretical framework and the experimental sections presented above, this study proposes a SI coherence model, the fundamental structure of which is illustrated in Figure 6-7.

Figure 6-7 shows a 3 x 3 = 9 model, which indicates that there are three parts on parallel level and three parts on serial level. The three parts on serial level are embodied CL, SI coherence and IEF. Likewise, the three parts on parallel level are coherence origin, global coherence and local coherence. With a view to coherence origin, global coherence and local coherence respectively, each part has three modules on serial level respectively, that is, in terms of coherence origin, there are three modules: interactive embodiment based module, SI coherence origin module, and SI activation module respectively.

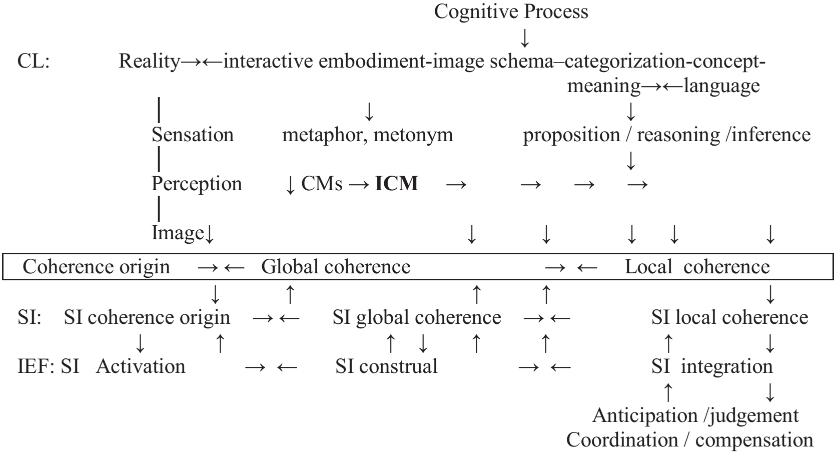

Cognitive Process
CL: Reality→←interactive embodiment-image schema–categorization-concept-meaning→←language
Sensation
metaphor, metonym
proposition / reasoning /inference
Perception
CMs → ICM
Image
Coherence origin →← Global coherence →← Local coherence
SI: SI coherence origin →← SI global coherence →← SI local coherence
IEF: SI Activation →← SI construal →← SI integration
Anticipation /judgement
Coordination / compensation

Figure 6-7. Fundamental structure of the SI coherence model

Next, global coherence has ICM based module, SI global coherence module, and SI construal module. Local coherence also has three modules: the conception, proposition, meaning and inference based module on the top module, and SI local coherence module and SI integration module accordingly. So this model comprises nine modules (3 x 3 = 9).

According to their embodied comprehension of the online speech, simultaneous interpreters access their cognitive world (ICM + background knowledge) to achieve the global coherence and use concept bridge, proposition reference and pragmatic inference to achieve sufficient coherence in SI via interpreting actions and products including anticipation, compensation, judgement and coordination.

Taken together, this model shows how the simultaneous interpreter manages to build up global and local coherence via ICM, the interactive embodiment and mental coherence approach.

6. Conclusions

In response to the three hypotheses proposed in section 2.8 of this chapter, the present research comes to the following conclusions.

1) SI coherence is based on interactive embodiment and mental coherence. The theories underpinning this conclusion are (a) “our experience is pre-conceptually structured at basic-level categorization” (Lakoff 1987: 269); (b) “basic level concepts consist not only of objects but of actions and properties” (Lakoff 1987: 270-271); (c) meaning

postulates themselves only make sense given schemas that are inherently meaningful because they structure our direct experience (Lakoff 1987: 273).

2) SI coherence is determined by embodied ICM globally and realized locally. The embodied ICM is in right proportion to SI coherence, and plays a role of default values, working through each SI section and also determines SI coherence. ICM could be used to compensate the language deficit to establish coherence in spite of close connection between SI and language proficiency.

3) With a view to ICM, the building blocks of cognition result from "basic level" interaction with the environment rather than the building up of "molecular" propositions from "atomic" concepts. Meaning and knowledge are "embodied" in the form of image-schematic cognitive models (Pöchhacker 1992). Regarding local coherence, one word could activate a series of experience or concept structure in the frame and further script specific and uniform knowledge structure or coherence schematization of experience for appropriate sequences of events in a particular context (Fillmore 1982: 124, 1985: 223).

The findings are hoped to contribute to SI teaching and materials development. It is not adequate for teachers only to focus on interpreting skills; instead, they must guide SI students to engage in interactive embodiment experience, particularly in terms of the applied fields, such as engineering and business interpreting, to build up ICM for the subject matters, in order to improve SI performances ultimately. In addition, interpreting textbooks should be written by integrating the SI skills with their practical experience to suit their professional preference, which forms a functional-web system of knowledge, namely ICM. Only in this way can SI students be trained to provide interpretations as coherent as they should be. In addition, these results suggest that simultaneous interpreters are made rather than born if they are trained according to the cognitive rules. Last but not least, the principle and tendency for training interpreters should be professionalization and specialization.

While these findings of the present study have answered some aspects of the research question, they also raise a series of new questions to be addressed: What underlies the relation between coherence and cohesive devices in SI? What kind of role do the cohesive devices play from cognitive perspective in SI coherence? Does the simultaneous interpreter adopt SI strategies and the cohesive devices consciously or unconsciously to establish coherence. In other words, are the capacities of SI strategies and cohesive devices trained or acquired naturally by the bilingual competence? Questions such as these clearly beg further research in the future.

Finally, this research has several limitations. Firstly, the corpus is small, with only three subjects of interpreting involved. Secondly, this research focuses on propositional and image-schematic models (i.e. the fixing procedure and structure of Shiatsu) for SI coherence, which are apparent and plausible in this SI experiment, but how the simultaneous interpreter copes with abstract concepts to achieve coherence is not addressed. How do professional simultaneous interpreters perform in this context? As a result, future and ongoing research will continue to investigate these areas of coherence in SI.

Notes

This study was carried out at the Department of Languages and Intercultural Studies, Heriot-Watt University as part of the author's PhD thesis. I thank my supervisors Professor Turner, Professor Perez and Dr Winters for their valuable advice and also the editor, Dr Richard Xiao, who has given suggestions and proof-read the whole chapter, but I am alone responsible for any shortcomings which remain.

References

De Beaugrande, R. and Dressler, W. (1981), *Introduction to Text Linguistics*. London: Longman.

Brown, G. and Yule, G. (1983), *Discourse Analysis*. Cambridge: Cambridge University Press.

Clark, E. and Clark, H. (1977), *Psychology and Language*. New York: Harcourt Brace Jovanovich.

Dillinger, M. (1994), "Comprehension during interpreting: What do interpreters know that bilinguals don't?", in S. Lambert and B. Moser-Mercer (eds.) *Bridging the Gap: Empirical Research in Simultaneous Interpretation*, 155–189. Amsterdam: John Benjamin.

Ericsson, K. and Simon, H. (1984), *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

Fillmore, C. (1982), "Frame semantics", in the Linguistic Society of Korea (ed.) *Linguistics in the Morning Calm* (Vol.1), 111-137. Seoul: Hanshin Publishing Co.

—. (1985), "Frames and the semantics of understanding". *Quaderni di Seamtica* 2: 222-254.

Givón, T. (1990), *Syntax: A Functional-Typological Introduction* (Vol.2). Amsterdam: John Benjamin.

—. (1995), "Coherence in text vs. coherence in mind", in M. A. Gernsbacher and T. Givón (eds.) *Coherence in Spontaneous Text*, 59-116. Amsterdam: John Benjamin.

He, Z. and Ran, Y. (2006), *Cognitive Pragmatics – Cognitive Study on Language Communication*. Shanghai: Shanghai Foreign Language Education Press.

Johnson, M. (1987), *The Body in the Mind – The Bodily Basis of Meaning, Imagination and Reason*. Chicago: The University of Chicago Press.

Lakoff, G. (1987), *Women, Fire, and Dangerous Things*. Chicago: The University of Chicago Press.

Lakoff, G. and Johnson, M. (1980), *Metaphors We Live By*. Chicago: The University of Chicago Press.

Lakoff, G. and Johnson, M. (1999), *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: Basic books.

Langacker, R. (2000), *Grammar and Conceptualization*. Berlin: Mouton de Gruyter.

—. (2008), *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

Leech, G. (1991), "The state of art in corpus linguistics", in K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics*, 8-29. London: Longman.

McEnery, T., Xiao, R. and Tono, Y. (2006), *Corpus-Based Language Studies*. London and New York: Routledge.

McCarthy, M. (1991), *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press.

Oleron, P. and Nanpon, H. (1964), "Recherches sur la traduction simultanee". *Journal de Psychologie Normale et Pathologgigue* 62: 73-94.

Paneth, E. (1957), An Investigation into Conference Interpreting. M.A. dissertation. London University.

Pöchhacker, F. (1992), "From knowledge to text: Coherence in simultaneous interpreting", in Y. Gambier and J. Tommola (eds.) *Translation and Knowledge* (SSOTT IV), 87-100. Turku: University of Turku Centre for Translation and Interpreting.

Seleskovitch, D. (1965), *Colloque sur l'enseignement de l'interpretation*. Geneva: AIIC (Association Internationale des Interpretes de Confeerence).

Setton, R. (1999), *Simultaneous Interpretation – A Cognitive-Pragmatic Analysis*. John Benjamin.

Trimble, L. (1985), *English for Science and Technology: A Discourse Approach*. Cambridge: Cambridge University Press.

Wang, Y. (2006), *Cognitive Linguistics*. Shanghai: Shanghai Foreign Language Education Press.

Zwaan R. (2004), "The immersed experiencer: Toward an embodied theory of language comprehension". *The Psychology of Learning and Motivation* 44: 35-62.

Zwaan R, and Madden C. (2005), "Embodied sentence comprehension", in D. Pecher and R. Zwaan (eds.) *The Grounding of Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, 224-245. Cambridge: Cambridge University Press.

Zwaan, R., Madden, C. and Whitten, S. (2000), "The presence of an event in the narrated situation affects its activation". *Memory and Cognition* 28: 1022-1028.

Appendix: The experimental text

S1: Good morning, ladies and gentlemen. S2: Welcome to our product promotion fair of Hömedics. I am Craig, sales manager. S3: I would like to recommend a good product to you. S4: Now look at **this**. This is a **Shiatsu Massage Cushion**. S5: As you know, Shi in Japanese means finger - and Shiatsu means pressure, put **them** together and you've got Shiatsu, meaning 'finger pressure'. S6: In this day and age of bad backs **due to** much office work, labour and sitting too long for your studies, Shiatsu Massager is the perfect relaxation gadget for every occasion. S7: Just now all of you have sat on it and have had a positive personal experience. How do you feel about it? **Comfortable? Yes, - I think you will agree.** S8: The **structure** of Shiatsu massager is simple. **Two kneading heads, rotating, travel up** and **down**, along the PVC guides, relieving pain and fatigue on your back. S9: I would like to talk about how to use this Shiatsu Massager. Briefly there are **six steps** you should remember: S10: **One**. **Attach** the **massage seat** to almost any chair, using the integrated **strap** at the back of the seat; ensure it is held firmly in place by adjusting the strap as necessary. S11: **Two**. **Connect** the **power supply lead** from the adapter with the corresponding lead in the side of the seat. S12: **Three**. **Plug** the **adaptor** into a 230V AC mains outlet T3 -and switch on. By the way, remember to finish steps 1 to 3 before switching the appliance on at the mains. S13: **Four**. Once seated, use the **remote control** to **operate** the appliance. Press the "Power" button once and select the desired massage zone- to start the massage. S14: **Five**. **For** an intense massage, **remove** the **flap** from the back of the cushion. **For** a gentler massage, **keep** the **flap** on and you can **soften** the **massage** further by placing a towel between your back and seat. S15: **Six**. **Press** the **"Power" button** for a second time to stop the massage. S16: Always remember to store it properly. Place the appliance in its box or in a safe, dry, cool place. S17: Water or any liquids that come into contact with the appliance are dangerous. S18: Avoid contact with sharp edges or pointed objects **which** might cut or puncture the fabric surface. S19: To avoid breakage, DO NOT wrap the power cord around the appliance. DO NOT hang the unit by the cord. S20: Ladies and gentlemen, **you could have it now at 30 pounds, in John Lewis, Edinburgh, it costs 69 pounds.** S21: If you are interested, feel free to try the Shiatsu massager again before you buy it. S.22: Thanks.

CHAPTER SEVEN

SYNTACTIC DIFFERENCES BETWEEN TRANSLATED AND NON-TRANSLATED DUTCH: A CORPUS-BASED IN-DEPTH ANALYSIS OF PP PLACEMENT

GERT DE SUTTER, MARC VAN DE VELDE

1. Introduction

With the advent of the corpus linguistic methodology in Translation Studies, an increasing amount of evidence has emerged that language use in translated texts differs systematically from language use in non-translated texts. These differences occur at all linguistic levels, i.e. at the lexical (e.g. Laviosa 1998, Kemppanen 2004, Tirkkonen-Condit 2004), morpho-syntactic (e.g. Olohan and Baker 2000, Puurtinen 2003) as well as discursive (e.g. Teich 2003, Mutesayire 2004, Serban 2004) levels. As it is claimed that these differences occur independently of the source and target languages, translation scholars have begun to explore the internal mechanisms of the translation process itself as the cause of the observed differences between non-translated and translated languages. Four frequently researched internal translation mechanisms or principles, which are held to be universal, are explicitation, simplification, normalization and levelling out (cf. Baker 1993 and 1996; for an overview and critical review, see Mauranen and Kujamäki 2004).

The present chapter presents a follow-up corpus-based study of the use and function of PP (prepositional phrase) placement in translated and non-translated German and Dutch. In a previous study (De Sutter and Van de Velde in press), German PP placement was the object of study; and in the present chapter, the research focus is shifted to Dutch PP placement. PP placement has been studied widely in German and Dutch linguistics. However, an in-depth comparison of the usage and functional differences of PP placement in non-translated and translated language has not yet been conducted (see section 2 for more information on this type of variation).

In line with our research programme at the University College Ghent Translation Studies, we aim at supplementing and extending previous corpus-based translation research in three respects. First, by providing more evidence on usage differences between translated and non-translated texts, especially in less well-studied languages such as Dutch. By means of this kind of evidence, we are able to (at least partially) verify to what extent some of the above-mentioned translation principles operate in translated Dutch texts. Second, by more intensively exploiting the range of statistical techniques available to studies in the humanities. As results of corpus linguistic analyses are inevitably drawn from a sample of translated and non-translated texts, it is indispensable to question the representativity and validity of the results: to what extent would the results be the same if a different sample was drawn from the same population? By means of statistics, this issue can be tackled in a systematic, objective and reliable way. Third, by comparing the distribution of linguistic variants in non-translated and translated texts, and investigating the factors that influence this distribution. For PP placement in Dutch, this means that the frequencies of two competing positions in the clause will be analyzed and compared as well as the impact of language-internal and language-external factors that guide language users in choosing between one of the two positions. Research in variational linguistics has revealed that the choice between competing forms, whether they are lexical, morpho-syntactic or discursive in nature, is usually governed by different types of factors (e.g. Arnold *et al*. 2000, Grondelaers 2000, Gries 2003, De Sutter 2005, Diessel and Tomasello 2005). This type of in-depth analysis allows us to answer the questions (i) whether there are any distributional differences between non-translated and translated texts in terms of PP placement and (ii) to what extent the set of factors that influences a given variation phenomenon differs between translated texts and non-translated texts.

This chapter is organized as follows. Section 2 provides a characterization of the object of study and reviews previous research of the determining factors of PP placement. In section 3, the data that underlie the analysis is presented. The results are presented and discussed in section 4. More particularly, section 4.1 is devoted to the general distribution of the variants in translated and non-translated Dutch, sections 4.2, 4.3 and 4.4 measure the effect of definiteness, function and clause type respectively and discuss the results qualitatively against the background of what is already known about translation universals and the cognitive-functional mechanisms of Dutch PP placement. Finally, section 5 summarizes our findings, and explores their most important theoretical, methodological and analytic implications.

2. PP placement in Dutch

Dutch syntax is characterized by the so-called brace construction (Van de Velde 1973, Haeseryn *et al*. 1997). This is a discontinuous construction in which the verbal elements take fixed positions in the clause. In main clauses, the finite verb takes the first or second position (this is called the first pole of the brace construction), while the infinite verbs and/or verbal particles occur at the end of the clause (the second pole). If no infinite verbs or verbal particles are present in the clause, the second pole remains empty. The non-verbal constituents in the clause are distributed around these two verbal poles in the prefield (the position just before the first pole), the middle field (the position in between the poles) and the postfield (the position after the second pole). In example (1), the first pole is taken by the finite verb *begint* 'begins', and the second pole by the infinite verb *te praten* 'to talk'. The prefield is taken by the personal pronoun *hij* 'he', the middle field by the prepositional phrase *met Benting* 'with Benting', and the postfield by the prepositional phrase *over de politieke toestand* 'about the political situation'.

(1) [Hij]$_{\text{prefield}}$ [begint]$_{\text{1st pole}}$ [met Benting]$_{\text{middle field}}$ [te praten]$_{\text{2nd pole}}$ [over de politieke toestand]$_{\text{postfield}}$ [source text: J. Vandeloo]
'He begins to talk with Benting about the political situation'

The brace construction in Dutch subordinate clauses deviates somewhat from the brace construction in main clauses, as can be seen in example (2). Instead of the finite verb, the first pole in subordinate clauses is taken by the conjunction (*toen* 'when'); if not present, the first position remains empty. The second pole is taken by the verbal elements (finite, infinite verbs as well as verbal particles; *verzorgd werd* 'was taken care of'), which converge at or toward the end of the clause.

(2) [toen]$_{\text{1st pole}}$ [hij]$_{\text{middle field}}$ [verzorgd werd]$_{\text{2nd pole}}$ [door de dokter uit Neuveville]$_{\text{postfield}}$ [source text: F. Dürrenmatt]
'when he was taken care of by the doctor from Neuveville'

The position of prepositional phrases (shortened as PPs hereafter) in the Dutch brace construction is variable. Language users may put PPs either in the middle field or in the postfield.[1] In examples (1) and (2), which are repeated as example (3a) and (4a), the PPs *over de politieke toestand* 'about the political situation' and *door de dokter uit Neuveville* 'by the doctor from Neuveville' are placed in the postfield; in examples (3b) and (4b) they are placed in the middle field (PPs are in italics and the verbal poles are demarcated by pipes).

(3) a. Hij |begint| met Benting |te praten| *over de politieke toestand*
b. Hij |begint| met Benting *over de politieke toestand* |te praten|
'He begins to talk with Benting about the political situation'

(4) a. |toen| hij |verzorgd werd| *door de dokter uit Neuveville*
b. |toen| hij *door de dokter uit Neuveville* |verzorgd werd|
'when he was taken care of by the doctor from Neuveville'

PPs in the postfield (examples 3a and 4a) are traditionally said to be extraposed, i.e. placed outside of the brace. In the remainder of this chapter, we will refer to this as *PP extraposition*. PPs that are located in the middle field will be referred to as *PP in middle field*.

Previous empirical analyses have pointed out that the variation in PP placement is not at random, as different factors have been found to influence the choice (Jansen 1978, 1979; Braecke 1990). The most frequently mentioned factors are definiteness of PP, function of PP, clause type, heaviness of PP, stress, discursive proximity, register, gender and social class. Table 7-1 summarizes the effects of such factors as found in previous work on non-translated Dutch texts.

As the present study only focuses on the factors *definiteness*, *function* and *clause type*, we will only review the effects of these factors (Jansen 1978). Table 7-1 shows that in Jansen (1978) significant effects were found for *definiteness* and *function*, but not for *clause type*. More particularly, it was found that PP extraposition occurs more often when the PP is indefinite and when the PP functions as a complement. Jansen (1978) explains the effect of definiteness against the background of a given-new theory of information structure. Assuming that indefiniteness relates to new information and definiteness to given information, the preference of indefinite PPs for the postfield can be explained by the tendency to place new information at the end of the clause. The effect of function is less easy to explain: Jansen (1978) resorts to two possible explanations (one in terms of generative transformation rules and the other in terms of semantic content), but he concludes that future research is needed to bring more clarity to this matter.

Table 7-1. List of relevant factors affecting the choice between presence vs. absence of PP extraposition in Dutch

Factors	Effects
Definiteness of PP	Probability of PP extraposition increases when PP is indefinite (vs. definite) *[Jansen 1978]*
Function of PP	Probability of PP extraposition increases when PP functions as a complement (vs. adjunct) *[Jansen 1978]*
Clause type	There is no association between PP extraposition and clause type (main vs. subordinate clause) *[Jansen 1978]*
Heaviness of PP	Probability of PP extraposition correlates positively with the length of the PP *[Jansen 1978, 1979; Braecke 1990]*
Stress	There is no association between PP extraposition and accentuation of PP (accented vs. non-accented PP) *[Jansen 1978]*
Discursive proximity	Probability of PP extraposition increases when an anaphor in the next clause refers to the PP *[Jansen 1978]*
Register	Probability of PP extraposition increases in formal texts (vs. informal texts) *[Jansen 1978]*
Gender	Probability of PP extraposition increases when language user is male (vs. female language users) *[Jansen 1978]*
Social class	Probability of PP extraposition increases when language user belongs to lower class (vs. high class) *[Jansen 1978]*

3. The data

This investigation is based on a corpus of 10 non-translated and 10 translated Dutch novels. The non-translated novels were written by 10 different Dutch-speaking authors while the translated novels were translated from native German novels (cf. Table 7-2; the name of the translator is given in the square brackets). A full overview of the novels included in the corpus can be found in the appendix. In order to increase the comparability of the non-translated and translated texts, texts were chosen from 10 different authors, written in the same period (1950-1970) and belonging to the same type of register (fine literature). By doing so, we attempt to exclude a potential underlying influence of individual, register and temporal variation as much as possible.

Table 7-2. Overview of the corpus of literary Dutch

Non-translated Dutch	Translated Dutch
I. Michiels	F. Dürrenmatt [Boey]
J. Vandeloo	H. Böll [V.d. Plas]
P. van Aken	W. Hildesheimer [Etty]
J. Daisne	P. Weiss [Schuur]
L.P. Boon	G. Kunert [Salomons]
H. Haasse	U. Johnson [Cornips]
W.F. Hermans	G. Grass [Manger]
J. Hamelink	S. Lenz [Coutinho]
J. Wolkers	J. Lind [Coutinho]
H. Mulisch	I. Dachmann [Mulder]

From the corpus of literary texts, we selected the first 250 sentences of each of the 20 novels (resulting in a data set of 5,000 sentences). All PPs located in the middle field or in extraposition were then retrieved, which resulted in 2,318 entries. After the data were retrieved, they were annotated for the dependent variable (PP placement) and for the three independent factors under scrutiny: definiteness of PP, function of PP and clause type.

The software used for the statistical analyses to be presented below is R 2.7.0 (2008). It is important to note that the focus will be on the interpretation of the analyses, not on the technical details (cf. Agresti 1996), and that all statistical analyses used in the remainder of this chapter are not a goal in their own right; they are simply tools which enable us, based on our (limited) sample of 2,318 observations, to make reliable claims about the population of all clauses with PPs in the middle field or in extraposition. For all statistical tests performed in this study, the significance cut-off level is set at .05: all p-values smaller than .05 indicate statistical significance, while p-values larger than .05 indicate non-significance.

4. Results and discussion

We analyzed the data in four steps. In the first step, we checked by means of a chi-squared test whether PP placement differs significantly in non-translated vs. translated Dutch. Additionally, if a significant difference was detected, an odds ratio (O.R.) was computed. O.R.'s range from 0 to +∞, thereby indicating what the exact size and direction of the difference is: the value 1 signifies 'no statistical association', values < 1 and > 1 signify a negative and a positive association respectively (i.e. the direction of the effect); values farther from 1 represent larger effect sizes. Thus, while O.R. values of 1.75 and 4.65 both indicate positive associations, the latter value, however, shows a greater effect. This part of the investigation will answer the question whether translated and non-translated literary Dutch exhibit significantly different syntactic preferences.

In the next three steps, the effect of each of the three selected factors on PP placement is investigated: do definiteness of the PP, PP function and clause type affect the position of Dutch PPs? Here too, chi-squared tests and O.R.'s are computed to find out whether each factor has a significant effect and, if it does, what the size and direction of the effect is. If a certain factor has more than two values (cf. PP function below), we computed adjusted standardized residuals to find out which value has the largest impact on PP placement. This part of the investigation answers the question whether PP placement in translated and non-translated language is influenced by the same factors.

4.1. PP placement in translated and non-translated Dutch: General distribution

Table 7-3 shows that the majority of PPs are placed in the middle field, irrespective of whether they are translated or non-translated. Nevertheless, PPs occur more frequently in the middle field in translated Dutch (83.36%) in comparison with non-translated Dutch (78.69%). This difference is statistically significant ($\chi^2 = 23.74$, d.f. = 1, $p < .0001$), and yields an O.R. = 1.72 (C.I. = 1.38 - 2.13),[2] signifying that the probability of PPs occurring in extraposition is 72% as large in non-translated compared to translated Dutch. This result is analogous to what we found for German PP placement (translated German exhibits significantly more PPs in middle field than non-translated German; De Sutter and Van de Velde in press). Moreover, this finding provides additional support for the hypothesis that translated and non-translated texts show significantly different syntactic preferences (e.g. Olohan and Baker 2000).

Table 7-3. General distribution of PP placement in translated and non-translated Dutch

	PP in middle field	PP extraposition
Non-translated Dutch	78.69% (768/976)	21.31% (208/976)
Translated Dutch	83.36% (1159/1342)	13.64% (183/1342)

The question arises, then, why translated Dutch exhibits a clearer preference for PPs in the middle field. One of the answers might be interference of source language. Recall that the source language of the Dutch translated texts in our corpus is German, and that German, in general, displays a more outspoken preference for PPs in the middle field (Van de Velde 1973). The observation, then, that in translated Dutch PPs are more often placed in the middle field, could be interpreted as a reflection of the German preference for middle field placement. In other words, translators are (either consciously or unconsciously) influenced by the syntactic preferences in the source language, as a consequence of which translated Dutch moves away from the syntactic preferences of non-translated Dutch.

However, if we compare the frequencies in translated and non-translated Dutch without taking into account the source language, another explanation must be considered, viz. normalization. As mentioned before, middle field position is the typical, dominant position for PPs in non-translated Dutch (78.69%). Hence, we can argue that the even higher frequency of PPs occurring in the middle field position in translated Dutch (83.36%) can be interpreted as an “exaggerated”, normalized use of this position. In other words, the translated texts “exaggerate features of the target language […] to conform to its typical patterns” (Baker 1996: 183).[3]

Both explanations seem plausible at first sight and our data do not seem to be biased in support of either of these explanations. It appears to us that this explanatory conflict touches on current debates in contemporary corpus-based translation studies whether or not the source text and system should be involved when explaining differences between translations and non-translations in a given language (cf. e.g. Teich 2003). Obviously, a decisive answer cannot be given on the basis of this small case study, but it is nevertheless clear that translation scholars should deal with this issue in near future.

Apart from the explanations mentioned above, still other explanations must be taken into account as well. Consider for instance the functional principle of information distribution in the clause that has been proposed to explain variation in PP placement (see e.g., Haeseryn *et al*. 1997: 1365-1366). According to this principle, extraposition is used as a focus position, i.e. PPs in extraposition receive additional prominence. If this principle is active in both non-translated and translated language, it is not inconceivable that translators have the tendency to decrease or augment the prominence of PPs during the process of translation (by replacing PPs in the middle field to extraposition or vice versa). In other words, the superficial differences between non-translated and translated languages in PP placement may be due to interference or normalization, as suggested above, but it might also be the case that these differences are triggered by underlying functional principles that guide translators, as well as other language users, in choosing between linguistic alternatives. Hence, translation scholars need to realize that superficial differences might have an indirect explanation, as different types of factors and the principles they represent affect the choice language users make during language production. This underscores the importance of investigating the determinant factors of linguistic variation in general and syntactic variation in particular. In the following three sections, three frequently mentioned factors of PP placement are investigated in order to verify these ideas.

4.2. The effect of definiteness on PP placement

Table 7-4 shows the distribution of the two PP positions in terms of the definiteness of the PP. At first sight, one might think that the factor *definiteness* works in the same way in both non-translated and translated Dutch: in both varieties, indefinite PPs are extraposed more frequently than definite PPs. However, the difference in placement of definite and indefinite PPs in non-translated Dutch is marginal and not statistically significant ($\chi^2 = 0.65$, d.f. = 1, $p > .05$) – contrary to Jansen (1978). In translated Dutch, on the contrary, the difference *is* statistically significant ($\chi^2 = 31.12$, d.f. = 1, $p < .0001$; O.R. = 0.40, C.I. = 0.29 – 0.56). The O.R. indicates that the odds of indefinite PPs in extraposition are 2.5 (i.e. 1 / 0.40) times higher than the odds of definite PPs in extraposition.

Table 7-4. Distribution of PP positions in non-translated and translated Dutch in terms of PP definiteness

		PP in middle field	PP extraposition
Non-translated Dutch	**Definite PP**	79.31% (575/725)	20.69% (150/725)
	Indefinite PP	76.89% (193/251)	23.11% (58/251)
Translated Dutch	**Definite PP**	89.23% (920/1031)	10.77% (111/1031)
	Indefinite PP	76.85% (239/311)	23.15% (72/311)

These results suggest that the factor *definiteness* does not work in the same way in non-translated as in translated Dutch. By means of the Breslow-Day statistic, we can statistically confirm that this is the case (Br-D $\chi^2 = 10.25$, d.f. = 1, $p < .002$). This leads to the important conclusion, at least for the effect of the factor *definiteness*, that translated and non-translated Dutch are not only different in having other PP placement preferences (section 4.1), but also in the underlying influence of the factor *definiteness*.

One possible explanation for this difference might once again be found in source language interference: since the translated Dutch texts are translated from German texts, and the German source texts exhibit a significant difference in PP placement between definite and indefinite PPs (cf. De Sutter and Van de Velde in press), the translator might have transferred this difference into the Dutch translations. If this would turn out to be the case, this would imply that translators can not only be influenced by the superficial syntactic preferences of the source language, but also by the conditioning factors that determine these preferences.

Another explanation for this difference starts from the above-mentioned functional explanation: if extraposition can be considered a syntactic means for focalization, then it is more prone to be used by new information, which can be linguistically marked by indefinite determiners, than by given information (marked by definite determiners). If this functional principle indeed has played a role in determining PP placement, our results suggest that this principle is only active in translated texts. This implies that translators, possibly for reasons of explicitation, put indefinite PPs more often in the prominent extraposition. If future research would confirm this latter explanation, the different effect of definiteness in translated and non-translated Dutch provides convincing evidence that an explanation on the basis of superficial findings (cf. the general distribution in Table 7-3) does not tell the whole story, as underlying mechanisms governing the choice between competing variants are not taken into account: what looks like source language interference or normalization (cf. section 4.1) might turn out to be the result of other translation processes (such as explicitation) that operate at deeper levels of linguistic variation.

4.3. The effect of function on PP placement

For the present research, we distinguish between four functional categories, viz. two complement and two adjunct functions. PPs that are subcategorized by the main verb, i.e. indirect objects, agentive objects or prepositional objects (example 5), are classified as *complement 1*. Obligatory adverbial phrases of direction and place (example 6) are subsumed under the separate heading *complement 2*, as it is a transitional category in between complements and adjuncts. Syntactically, it belongs to the complement category, as it is subcategorized by the main verb while semantically, however, it shows strong connections to the adjunct category, describing the circumstances under which the event referred to by the verb takes place.

(5) De trouwceremonie moest aanleiding hebben gegeven *tot pijnlijke overwegingen* [source text: H. Böll]
‘The wedding ceremony had to give cause to painful considerations’

(6) Deze blinde fotograaf bleek te wonen *in het smalste huis van de stad,* dat tussen twee enorme, hoge herenhuizen inlag [source text: W.F. Hermans]
‘This blind photographer appeared to live in one of the most narrow houses in the city, which lied in between two enormous, high mansions’

Adjuncts are not subcategorized by the main verb; they operate as satellites, both syntactically and semantically (such as adverbial phrases of time or place). As a result, they can be attached to very different verbs (cf. Somers 1984, Storrer 2003). The *adjunct 1* category consists of adverbial phrases that modify the main verb of the clause (degree, direction, duration, manner, means, qualification; example 7). The *adjunct 2* category comprises adverbial phrases that modify the complete clause (time, concession, cause, goal, reason, modality, place, consequence; example 8).

(7) Ik had ondertussen rondgekeken *naar de schilderijen die iets beter waren dan deze uit de wachtzaal* [source text: L.-P. Boon]
'In the meanwhile, I looked around at the paintings, which were somewhat better than these in the waiting room'

(8) Dat de jongste *na een tweede ruzie in de namiddag* het huis had verlaten [source text: U. Johnson]
'That the youngest one had left the house after a second argument in the afternoon'

Table 7-5 summarizes the distribution of the functional categories across the two PP positions in non-translated Dutch. PPs functioning as indirect objects, agentive objects or prepositional objects (complement 1) occur most frequently in extraposition (34.15%), followed by clause modifying adjuncts (adjunct 2; 29.58%), verb modifying adjuncts (adjunct 1; 23.14%) and syntactically necessary adverbial phrases of direction and place (complement 2; 1.54%). The chi-square test shows that the overall distribution of the different functional categories as a function of the two PP positions is highly significant ($\chi^2 = 89.63$, d.f. = 3, $p < .0001$).[4] These findings run parallel to the results in Jansen (1978).

Table 7-5. Distribution of PP position in non-translated Dutch in terms of PP function

	PP in middle field	PP extraposition
Complement 1	65.85% (108/164)	34.15% (56/164)
Complement 2	98.46% (255/259)	1.54% (4/259)
Adjunct 1	76.86% (186/242)	23.14% (56/242)
Adjunct 2	70.42% (219/311)	29.58% (93/311)

The adjusted standardized residuals show that PPs functioning as syntactically necessary adverbial phrases of direction and place (complement 2) occur less often in extraposition than expected (resid = 9.06), whereas indirect objects, prepositional objects and agentive objects (complement 1; resid = -4.40) and clause modifying adjuncts (adjunct 2;

resid = -4.32) occur more frequently in extraposition than expected. In other words, this means that it is especially the deviant behaviour of the complement 2 category that causes the overall distribution to be statistically significant. Now, one could argue that the inclusion of the complement 2 category is questionable, as it appears to refuse almost any kind of variation. Redoing the chi-squared analysis without the complement 2 category, however, reveals that the statistical significance is retained ($\chi^2 = 6.16$, d.f. = 2, $p < .05$). This means, then, that even without the extreme category of complement 2, PP function has a clear effect on the choice of PP placement in non-translated Dutch. A partitioned chi-squared test moreover elucidates that the difference between the complement 1 and adjunct 2 is not statistically significant, so that these can be grouped together. The three functional categories that remain, then, are complement 1 / adjunct 2, adjunct 1, and complement 2.

The effect of PP function on PP placement may, at least partly, be explained by means of the functional principle of focalization. Complement 1, being a syntactically and semantically integral part of the clause, and hence referentially important for clause semantics, is most prone to the extraposition for focalization reasons (adding additional prominence to an already important part of the clause). The behaviour of complement 2, on the contrary, being almost completely resistant to the extraposition, may be explained by the semantic-syntactic principle of inherence (Haeseryn *et al*. 1997: 1245). This principle, which accounts for a lot of word order phenomena in Dutch and German, says that elements that have a close semantic link with the verb, such as necessary adverbial phrases of place or direction (complement 2), are placed preferably close to the left of the second pole (in the middle field). This principle might also explain why verb modifying adjuncts (adjunct 1), i.e. adjuncts that are closely related to the verb, are more resistant to extraposition than clause modifying adjuncts (adjunct 2).

The effect of PP function on PP placement in translated Dutch is summarized in Table 7-6. Compared with the situation in non-translated Dutch, several things are noteworthy. First, there is an overall lower amount of PP extraposition, which can be traced back to the general preference of translated texts to place PPs in the middle field (cf. section 4.1). Second, PPs functioning as indirect objects, agentive objects or prepositional objects (complement 1) occur most frequently in extraposition (27.48%), followed by verb modifying adjuncts (adjunct 1; 19.47%), clause modifying adjuncts (adjunct 2; 16.98%) and syntactically necessary adverbial phrases of direction and place (complement 2; 1.64%).

The overall distribution of the different functional categories is highly significant ($\chi^2 = 106.68$, d.f. = 3, $p < .0001$).

Table 7-6. Distribution of PP position in translated Dutch in terms of PP function

	PP in middle field	PP extraposition
Complement 1	72.52% (161/222)	27.48% (61/122)
Complement 2	98.36% (479/487)	1.64% (8/487)
Adjunct 1	80.53% (211/262)	19.47% (51/262)
Adjunct 2	83.02% (308/371)	16.98% (63/371)

The adjusted standardized residuals show that PPs functioning as syntactically necessary adverbial phrases of direction and place (complement 2) occur less often in extraposition than expected (resid = 9.66), whereas indirect objects, prepositional objects and agentive objects (complement 1; resid = -6.58), verb modifying adjuncts (adjunct 1; resid = -3.06) and clause modifying adjuncts (adjunct 2; resid = -2.21) occur more frequently in extraposition than expected. Redoing the chi-squared analysis without the complement 2 category shows that statistical significance is retained ($\chi^2 = 9.63$, d.f. = 2, $p < .01$). Thus, even without the extreme category of complement 2, PP function has a clear effect on the choice of PP placement in translated Dutch. A partitioned chi-squared test moreover indicates that the difference between the adjunct 1 and adjunct 2 categories is not statistically significant, so that these can be grouped together. The three functional categories that remain, then, are complement 1, adjunct 1/2, and complement 2.

As to the explanation of this effect, one can easily adopt the inherence principle to account for the behaviour of complement 2. The difference between complement1 and adjunct 1 / 2, on the other hand, can be explained by the same functional principles mentioned for the non-translated data: complement 1 has a greater tendency to appear in extraposition for focalization reasons than adjuncts 1 and 2.

Let us briefly sum up. Both in translated and non-translated Dutch, an effect of PP function was found (even though individual functional categories have been shown to affect PP placement in slightly different ways; e.g. adjunct 1 vs. adjunct 2). The conclusion, then, is that, contrary to the effect of definiteness in the previous section, the underlying influence of the determining factor *function* does not differ substantially between non-translated and translated Dutch.

4.4. The effect of clause type on PP placement

Table 7-7 shows the distribution of the two PP positions in Dutch in terms of the clause type in which the PP is located, i.e. main clause or subordinate clause.

Table 7-7. Distribution of PP position in translated and non-translated Dutch in terms of clause type

		PP in middle field	PP extraposition
Non-translated Dutch	**Subordinate clause**	83.78% (501/598)	16.22% (97/598)
	Main clause	70.63% (267/378)	29.37% (111/378)
Translated Dutch	**Subordinate clause**	84.71% (637/752)	15.29% (115/752)
	Main clause	88.47% (522/590)	11.53% (68/590)

As can be seen, PP placement in main and subordinate clauses differs substantially in non-translated and translated Dutch: in non-translated Dutch, extraposition occurs most frequently in main clauses (29.37% vs. 16.22% in subordinate clauses), in translated Dutch on the other hand, extraposition occurs most frequently in subordinate clauses (15.29% vs. 11.53% in main clauses). The difference in PP placement between PPs located in main clauses and PPs located in subordinate clauses is statistically significant (contrary to Jansen 1978), both for non-translated (χ^2 = 23.86, d.f. = 1, p < .0001; O.R. = 0.47, C.I. = 0.34 – 0.64) and translated Dutch (χ^2 = 3.98, d.f. = 1, p < .05; O.R. = 1.39, C.I. = 1.01 – 1.90). The O.R. for non-translated Dutch (0.47) indicates that the odds of PPs in extraposition are 2.13 (i.e. 1 / 0.47) times higher when PPs are located in main clauses (in comparison to subordinate clauses). In translated Dutch on the other hand, O.R. (1.39) indicates that the odds of PPs in extraposition are 39% times higher when PPs are located in subordinate clauses (in comparison to main clauses). The different behaviour of the clause type factor in non-translated vs. translated Dutch is statistically confirmed by the Breslow-Day test (Br-D $\chi^2 = 23.23$, d.f. = 1, $p < .0001$), so that we once more can conclude that the underlying determinants of syntactic variation in non-translated language need not be identical to the determinants in translated language.

Two questions arise at this point: first, why does the effect of clause type in translated vs. non-translated Dutch differ? Second, how must these effects be understood? An answer to the second question is not immediately obvious as it is unclear how the clause type difference could be related to differences in information distribution. Why, for instance, would main clauses be more prone to place PPs in extraposition than subordinate clauses? Of course, other explanatory devices than information distribution can be considered in order to account for the observed difference in clause type. Future research has to elaborate on this.

A possible explanation for the difference in the first question might once again be found in source language interference: since the translated Dutch texts are translated from German, and native German exhibits no significant difference in PP placement between main and subordinate clauses (cf. De Sutter and Van de Velde in press), the translator might have transferred this difference to some extent into the Dutch translation. If one compares the distributions of the translated Dutch data and the non-translated German data in De Sutter and Van de Velde (in press), it is remarkable how similar these distributions are.

5. Conclusions

The present chapter has provided evidence that PP placement in Dutch translated and non-translated texts differs significantly. First, it has been shown that the general distribution of the two competing PP positions displays a statistically significant difference between non-translated and translated Dutch, as the latter variety exhibits significantly less PP extraposition than its non-translated counterpart. Second, the analysis has revealed that three frequently mentioned factors governing PP placement do not always have the same effect on the choice between one of the PP positions in translated and non-translated Dutch texts:

- The factor *definiteness* affects PP placement in translated texts, but not in non-translated texts.
- The factor *PP function* affects PP placement in non-translated and translated Dutch in a similar way.
- The factor *PP function* affects PP placement in non-translated and translated Dutch, but in different ways.

This leads us to the descriptive conclusions (i) that syntactic differences between non-translated and translated texts do not only occur in English and other well-studied languages, but also in less well-studied languages such as Dutch; and (ii) that the factors that influence variation phenomena might differ too between translated and non-translated texts.

On a theoretical level, this chapter has shown that typical translation phenomena, such as source language interference and normalization, also influence these language-internal factors, even though it has to be admitted that it is not yet clear on what basis an explanation for empirical data has to be sought. More particularly, it has demonstrated that the explanation for the difference in PP placement between translated and non-translated Dutch depends on whether or not the source text is involved in the explanation. Finally, on a methodological level, we hope to have shown the usefulness of several types of bivariate and stratified statistical techniques in order to test hypotheses and distinguish reliable patterns from non-reliable ones.

Notes

[1] For the sake of completeness, it should be mentioned that PPs can also be placed in the prefield. However, in this study the structural variation is limited to the variation between middle field and postfield.

[2] Since an O.R. is computed on the basis of a sample, one needs to check whether this O.R. is representative of the larger population (i.e. all PPs in non-translated and translated Dutch that are placed in the middle field or in extraposition). To that end, a confidence interval (C.I.) is calculated. This shows the range of values in which one can find with 95% certainty the true population O.R. The closer the limiting values of a confidence interval, the more precise the population O.R. can be determined. If the value ‘1’ (‘no association’) is not in the interval, one can say with 95% certainty that the association between variables is significant.

[3] We would like to thank our colleague Bart Defrancq (University College Ghent Translation Studies) for drawing our attention to this alternative explanation.

[4] O.R.’s cannot be computed here, since the factor *PP function* has more than two levels.

References

Agresti, A. (1996), *An Introduction to Categorical Data Analysis*. New York / Chichester / Brisbane / Toronto / Singapore: Wiley.

Arnold, J.E., Wasow, T., Losongco, A. and Ginstrom, R. (2000), “Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering”. *Language* 76: 28–55.

Baker, M. (1993), "Corpus linguistics and translation studies: Implications and applications", in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: in Honour of John Sinclair*, 17–45. Amsterdam: Benjamins.

—. (1996), "Corpus-based translation studies: The challenges that lie ahead", in H. Somers (ed.) *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, 175–186. Amsterdam: Benjamins.

Braecke, C. (1990), "Uit de tang of ± prominent?" *Taal en Tongval* (special issue 3: dialect syntax): 125–134.

Diessel, H. and Tomasello, M. (2005), "Particle placement in early child language: A multifactorial analysis". *Corpus Linguistics and Linguistic Theory* 1(1): 89–111.

Gries, S. Th. (2003), *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London / New York: Continuum Press.

Grondelaers, S. (2000), *De distributie van niet-anaforisch er buiten de eerste zinsplaats. Sociolexicologische, functionele en psycholinguïstische aspecten van er's status als presentatief signaal*. PhD thesis, University of Leuven.

Haeseryn, W., Romijn, K., Geerts, G., Rooij, J. de and Toorn, M. C. Van den (eds.) (1997), *Algemene Nederlandse Spraakkunst*. Groningen / Deurne: Nijhoff / Plantyn.

Jansen, F. (1978), "Hoe krijgt een spreker zijn woorden opeen rijtje? Taalgebruiksaspekten van de 'PP over V' konstruktie", in J. G. Kooij (ed.) *Aspekten van de woordvolgorde in het Nederlands*, 70–104. Leiden: Ingen.

—. (1979), "On tracing conditioning factors of movement rules: Extraposition of PP in spoken Dutch", in M. Van de Velde and W. Vandeweghe (eds.) *Sprachstruktur, Individuum und Gesellschaft*, 83–93. Tübingen: Niemeyer.

Kemppanen, H. (2004), "Keywords and ideology in translated history texts: A corpus-based analysis". *Across Languages and Cultures* 5(1): 89-107.

Laviosa, S. (1998), "Core patterns of lexical use in a comparable corpus of English narrative prose". *Meta* 43(4): 557-570.

Mutesayire, M. (2004), "Apposition markers and explicitation: A corpus-based study". *Language Matters: Studies in the Languages of Southern Africa* 35(1): 54-69.

Olohan, M. and Baker, M. (2000), "Reporting *that* in translated English: Evidence for subconscious processes of explicitation?" *Across Languages and Cultures* 1(2): 141–158.

Mauranen, A. and Kujamäki, P. (2004), *Translation Universals. Do They Exist?* Amsterdam/Philadelphia: Benjamins.

Puurtinen, T. (2003), "Non-finite constructions in Finnish children's literature: Features of translationese contradicting translation universals?", in S. Granger J. Lerot and S. Petch-Tyson (eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, 141-154. Amsterdam: Rodopi.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna: http://www.R-project.org.

Serban, A. (2004), "Presupposition in literary translation: A corpus-based approach". *Meta* 49(2): 327-342.

Somers, H. (1984), "On the validity of the complement-adjunct distinction in valency grammar." *Linguistics* 22(4): 507–530.

Storrer, A. (2003), "Ergänzungen und Angaben", in V. Agel, L. M. Eichinger, H.-W. Eroms, P. Hellwig, H. J. Heringer and H. Lobin (eds.) *Dependenz und Valenz. Ein internationales Handbuch der zeitgenössischen Forschung*, 764–780. Berlin: De Gruyter.

Sutter, G. De (2005), *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen*. PhD thesis, University of Leuven.

Sutter, G. De and Van de Velde, M. (in press), "Determinants of syntactic variation in non-translated and translated language: A corpus-based study of PP placement in German". *International Journal of Translation*.

Teich, E. (2003), *Cross-linguistic Variation in System and Text*. Berlin: De Gruyter.

Tirkkonen-Condit, S. (2004), "Unique items - over- or under-represented in translated language?", in A. Mauranen and P. Kujamäki (eds.) *Translation Universals. Do They Exist?*, 177-184. Amsterdam/ Philadephia: Benjamins.

Velde, M. Van de (1973), "De Ausklammerung in het Duits en het Nederlands". *Studia Germanica Gandensia* 14: 119–142.

Verhagen, A. (1981), "Gaan fokusbepalingen echt altijd aan de fokus vooraf?", in W. Breekveldt and J. Noordegraaf (eds.) *Voortgang van het onderzoek in de subfaculteit Nederlands aan de Vrije Universiteit*, 56–61. Amsterdam: VUA.

Appendix

Note: The superscripted numbers after the year of publication indicate edition numbers.

Dutch non-translated texts

Ivo Michiels, *Het afscheid*. Amsterdam-Antwerpen 1960^{2} [1957^{1}].

Jos Vandeloo, *Het gevaar*. Brussel 1960.

Piet van Aken, *Zondaars en sterren*. In: Zes Vlaamse novellen. Rotterdam / ’s Gravenhage 1952.

Johan Daisne, *Lago maggiore*. Brussel-Den Haag 1965^{3} [1957^{1}].

Louis Paul Boon. *Maagpijn*. In: Boontje’s Twee spoken. Amsterdam 1956^{3} [1952^{1}].

Hella Haasse, *De ingewijden*. Amsterdam 1967^{1} [1957^{1}].

Willem Frederik Hermans, *De blinde fotograaf*. In: een landingspoging op Newfoundland en andere verhalen. Amsterdam 1966.

Jacques Hamelink, *Een schijndode maan*. In: Horror vacui. Amsterdam 1967^{2} [1966^{1}].

Jan Wolkers, *Een roos van vlees*. Amsterdam 1970^{16} [1963^{1}].

Harry Mulisch, *De diamant*. Amsterdam 1966^{7} [1954^{1}].

Dutch translated texts (source language: German)

Friedrich Dürrenmatt, *De rechter en zijn beul*. Leuven 1970 [translated by: J. Boey].

Heinrich Böll, *Meningen van een clown*. Amsterdam-Brussel 1967^{3} [translated by: M. van der Plas].

Wolfgang Hildesheimer, *Tynset*. Utrecht/Antwerpen 1967 [translated by: T. Etty].

Peter Weiss, *Afscheid van mijn ouders*. Amsterdam 1968^{2} [translated by: K. Schuur].

Günter Kunert, *Uit hoofde van de hoeden*. Amsterdam 1969, [translated by: D. en M. Salomons].

Uwe Johnson, *Twee kanten*. Amsterdam 1966 [translated by: T. Cornips].

Günter Grass, *Kat en muis*. Amsterdam 1969^{3} [translated by: H. Manger].

Siegfried Lenz, *De man in de stroming*. Amsterdam – Antwerpen 1965 [translated by: L. Coutinho].

Jakov Lind, *Landschap in beton*. Amsterdam 1963 [translated by: M. Coutinho].

Ingeborg Bachmann, *Alles*. In: 8 Duitse verhalen. Amsterdam 1966 [translated by: H. Mulder].

CHAPTER EIGHT

A PARALLEL CORPUS-BASED STUDY OF TRANSLATIONAL CHINESE

KEFEI WANG, HONGWU QIN

1. Introduction

Translational language (TL) retains, to varying degrees, some features of its source language (SL), and, as such, it is generally viewed as “a non-standard version of the target language that is […] affected by the source language” (Hopkinson 2007: 13). By “non-standard”, we mean the language used in translation is not as idiomatic and prototypical as it is in texts originally composed in the same language, for the former contains deviations from the general TL patterns, with SL being its source (Toury 1995: 208).

Features of translational language can be observed in many ways. According to Santos (1995: 60), we can go into 1) common properties of all translations, i.e. the universals of translation (Baker 1993), 2) properties of translations particular to a source language and target language, i.e. translationese, and 3) properties of particular translated texts (with the author and the translator taken into consideration). This study focuses on the second set of properties.

The use of parallel corpora in the study of translational Chinese has emerged in recent years in China, as has been done by Ke Fei (2003), and Qin and Wang (2004). However, studies in this field are rare and inconsistent, owing to the lack of reasonable methodology and appropriate tools. For a better inquiry into translational Chinese, this study attempts a multilevel analysis.

2. Translation universals

Baker (1993, 1998a) reports that all translational languages share some features, namely, i) simplification (including lexical simplification, syntactic simplification, and stylistic simplification); ii) explicitation (what is implied in the original text is made explicit, so are the cohesion markers); and iii) normalization (source-text textemes tend to be converted into target-language repertoremes and diversity is lost).

Baker (1998b: 225) does acknowledge that some stylistic features of the source text tend to be transferred to the target text in translation; however, her findings are made mainly on the basis of a monolingual comparable corpus, without duly taking into account the influence of source language (see Hansen and Teich 2001, Wu and Huang 2006, Huang and Wang 2006). Moreover, Baker's study revolves around shallow linguistic features such as word length, type/token ratio, sentence length, lexical density, etc.; as a result, many abstract features peculiar to the languages in the translation pair are marginalized.

Given the empirical property of language specific data on which the universals are based, it is questionable whether they are applicable to all translational languages, especially to languages of different families. For instance, it is of interest to find out whether they are applicable to translational Chinese. To observe their applicability, we employ a multilevel analysis approach. On the one hand, we make use of Baker's analytic techniques; on the other hand, we conduct micro-level analysis (e.g. part-of-speech distribution, compositionality and load capacity) with a view to arriving at an adequate description.

3. Normality, corpus and multilevel analysis

3.1. Normality

Chinese has its own typical linguistic patterns, which has been termed "normality" (Yu 2002: 151). Normality is not a set of rules, but a native speaker's intuition about the genral patterns of the language he/she speaks, an intuition that cannot be precisely measured or defined. However, we can assume its presence in a certain amount of original Chinese texts, because, compared with translated Chinese texts, the former is closer to the normality.

3.2. The corpus

The comparable corpus for use in this study comprises samples from the General Chinese-English Parallel Corpus (GCEPC) created by Beijing Foreign Studies University (BFSU), a Chinese-English bidirectional parallel corpus containing about 20 million English words and Chinese characters. GCPEC has four subcorpora, namely Chinese-English Literature, Chinese-English Non-literature, English-Chinese Literature, and English-Chinese Non-literature (Wang 2004). The Chinese texts taken from CE and EC subcorpora can form comparable corpus, and the same is true for English texts. Besides, GCEPC enables us to take account of the English source texts in analyzing translational Chinese.

The comparable corpus used in this study is composed of approximately 3.5 million English words and Chinese characters of samples taken from GCPEC.

3.3. Multilevel analysis

The study will first consifer features at the macro level (type-token ratio, word length, and sentence length); then it will turn to a micro-level analysis (including part-of-speech distribution, keywords analysis, compositionality and lexical bundles).

4. Macro-level description

The macro-level description mainly concerns type-token ratio (TTR), word length and sentence length that illustrate the difference between translated and original Chinese texts. To facilitate computations of TTR, word length and sentence length, the Chinese texts are tagged using ICTCLAS, a Chinese lexical analysis system developed by the Institute of Computing Technology of the Chinese Academy of Sciences. Table 8-1 shows the statistics of the data, which were generated with WordSmith Tools 4.0.

As shown in Table 8-1, original Chinese texts (OCT), translated Chinese texts (TCT) and English source texts (EST) are different in terms of type-token ratio, word length, sentence length (S length) and sentence segment length (SS length for short), which will be discussed in great detail below.

4.1. Type-token ratio

Generally, the larger the corpus is, the smaller the TTR. Considering the apparent difference in size between the subcorpora, we cannot count too much on a mere TTR calculation; however, we can use STTR (standardized type-token ratio) count.[1]

Table 8-1. Macro-level description of translated and non-translated Chinese texts

	Tokens	Types	STTR	Word length	S length	SS length
OCT (lit)	466,414	23,047	46.72	1.36	25.46	6.02
OCT (non-lit)	222,758	11,066	41.92	1.76	27.05	7.20
OCT	**689,172**	**28,437**	**45.19**	**1.49**	**25.95**	**6.35**
TCT (lit)	578,148	24,213	47.36	1.44	25.81	7.00
TCT(non-lit)	496,218	26,174	47.65	1.64	31.52	8.58
TCT	**1,074,366**	**36,354**	**47.49**	**1.53**	**28.27**	**7.65**
EST(lit)	546,632	22,409	43.21	4.26	16.76	6.79
EST (non-lit)	487,673	25,739	44.37	4.87	20.24	9.32
EST	**1,034,305**	**35,695**	**43.75**	**4.54**	**18.23**	**7.78**

A higher STTR indicates more different lexical items, while a lower STTR suggests that fewer specific words are used and more general ones are frequent (Westin 2002: 75). As is shown in Table 8-1, STTR for OCT is 2.3 percent lower than that for TCT (47.49 vs. 45.19); however, they are higher than that for EST (43.75). This difference suggests TCT, in comparison with OCT, is not so “simplified” in terms of lexical diversity. This may serve as counter evidence for lexical simplification which states translational language tends to use simple words for ease of understanding.

4.2. Word length

For English and many other alphabetic languages, word length is a way of measuring lexical specificity and diversity. For Chinese, however, word length can reflect idiomaticity of language use: in Mandarin Chinese, most words used in discourse are disyllabic and monosyllabic, but “monosyllabic words are most frequently used” (Lü 1980: 9).

The mean word length of TCT and OCT are similar, but the former is 0.04 longer than the latter. Moreover, in contrast with TCT (53.16% words being monosyllabic, and 41.72% words disyllabic), OCT uses more monosyllabic words (56.96%) and fewer disyllabic words (38.58%). This suggests that TCT is not as idiomatic as OCT in terms of word length, if the use of short word indicates Chinese nomality.

4.3. Sentence length and sentence segment length

Sentence length (S length) measures sentences beginning with capital letters (for English but not for Chinese) and ending in full stops, exclamation marks, question marks or colons. We hereby note that Chinese sentences are calculated in words rather than in characters.

The calculation of mean S length yields the following results: TCT uses longer sentences (2.32 more words on average) than OCT; in addition, TCT uses much longer sentences than EST (18.23 for EST, and 25.81 and 28.27 for OCT and TCT respectively). The following instances illustrate why Chinese uses longer sentences than English does.

(1) And Pinkerton--Pinkerton--he has collected ten cents that
he thought he was going to lose.

那么 平克顿 -- 平克顿 -- 他 一定 是 要 回来 一 角
then Pinkedun -- Pinkedun--he surely is claim back one cent

钱 的 老 账, 这 笔 钱 他 本来 以为
money DE old debt, this sum money he originally believe

没有 盼头 了 。 (16: 22)
not-have prospect PRT .

(2) You speak collectedly, and you--are collected.

你 这 话 倒还 有 自制力, 而 你
your this utterance nevertheless have forbearance, but you

也 确实 镇静 。 (7:10)
also really calm.

As an isolating language, Chinese usually resorts to lexical means to express what is expressed grammatically in English. For example, the relative pronoun *that* in sentence (1) is replaced by a noun phrase 这笔钱 'this sum of money', and the added expression 本来 'originally' serves to express the temporal meaning conveyed by the past tense verb *thought*.

The above expansion is to some degree compulsory, there are also optional ones. For instance, as can be seen in example (2), there is shift of part of speech in translation (e.g. from the verb *speak* to the noun phrase 这话 'this utterance'; from the adverb *collectively* to the verb phrase 有自制力 'have forbearance'). These conversions bring about the adding of demonstrative (这) and verb (有), and other additions in the translated texts. Furthermore, modality implied in the original is made explicit in translation (e.g. 倒 'nevertheless', 也 'also' and 确实 'really'). Of whatever type, amplification contributes to the augmentation, which makes sentences in TCT much longer than their counterparts in EST. This provides evidence in support of the explicitation hypothesis.

However, a simple S length calculation cannot reveal intra-sentence properties of a language, for the result it derives might be quite different if we take sentence segments into account.

As Chen (1994: 281) reports, "about 75% of Chinese sentences are composed of more than two sentence segments separated by commas or semicolons."[2] If this is true, sentence segment length (SS length) may tell us more about sentence building of the languages in question. For ease of data retrieval, we still use <s> and </s> tags to delimit sentence segments. The results (in Table 8-1) show that SS length in TCT and OCT is shorter than that in EST, and that SS length in TCT is significantly longer than that in OCT, which means SS length of TCT is very similar to that of EST.

Obviously, the results based on S length are different from those based on SS length, yet they are not in conflict. For example, a Chinese sentence can be longer than an English one, yet the former may contain more sentence segments than the latter. In other words, Chinese sentences contain more but shorter segments (clauses or phrases) than English sentences do, which explains why English, compared with Chinese, is shorter sententially but longer segmentally.

The S length of TCT supports explicitation hypothesis; however, the SS length of TCT, in comparison with OCT, does not support the normalization hypothesis, because TCT is more akin to EST than OCT in terms of SS length.

5. Part-of-speech distribution

5.1. Statistical results

Part-of-speech (POS) distribution partially reflects typological features of a language. For POS distribution analysis of Chinese and English, our Chinese data in TCT / OCT are tagged with ICTCLAS while the English texts in EST are tagged using Lancaster’s CLAWS tagger. In addition, to ensure the representativeness of normality, we single out literature subcorpora for analysis.

As can be seen in Table 8-2, EST has a lower frequency of verb use, about 4% lower than OCT. This difference confirms our belief that English is prominently ‘nominal’ while Chinese is more ‘verbal’ (see also Si 2002: 55-58, Shao 2005: 24). In addition, Table 8-2 displays difference in the use of pronouns, prepositions and conjunctions. On the one hand, TCT is very similar to EST but quite different from OCT in the use of pronouns, which suggests TCT subjects to the influence from EST; on the other hand, TCT uses much fewer prepositions and conjunctions than EST (their source texts), which means the influence of EST opon TCT is less obvious in the use of prepositions and conjunctions.

Table 8-2. POS distribution in OCT, TCT and EST (literary components)

Distribution POS		OCT (lit)		TCT (lit)		EST (lit)	
		Freq.	%	Freq.	%	Freq.	%
1	Verbs	110,391	23.64	133,762	22.93	108,340	19.88
2	Nouns	100,827	21.59	113,823	19.52	112,536	20.65
3	Adjectives	24,948	5.34	24,672	4.23	35,846	6.58
4	Adverbs	48,676	10.42	52,266	8.96	42,065	7.72
5	Pronouns	41,259	8.83	68,859	11.81	64,433	11.82
6	Prepositions	14,536	3.11	25,932	4.45	58,743	10.78
7	Conjunctions	9,687	2.07	15,252	2.61	39,304	7.21
8	Numerals	17,322	3.71	20,174	3.45	8,463	1.5
9	Classifiers	14,209	3.04	16,337	2.80	0	0
10	Particles	39,370	8.51	57,372	9.84	0	0
11	Articles	0	0	0	0	52,325	9.60
12	Determiners	0	0	0	0	17,132	3.14
Total		421,225	90.26	528,449	90.6	539,187	98.88

Figure 8-1 shows that TCT and OCT are very similar in the shapes of their distribution curves, but they are quite different from EST. That suggests, as far as POS distribution is concerned, the difference between TCT and OCT is smaller than that between TCT and EST, and translational Chinese largely conforms to the normality of Chinese.

In sum, it is clear that, in comparison with OCT, TCT uses fewer lexical words such as verbs, nouns, adjectives and adverbs, but more function words like pronouns, prepositions and conjunctions. In line with these facts, we have good reason to assume that POS distribution in TCT is susceptible to the influence of the English source language.

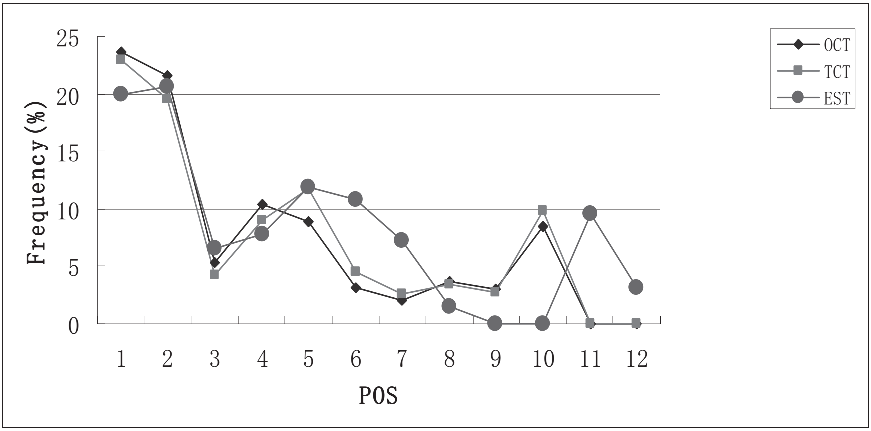


Figure 8-1. Part-of-speech distribution
Legends: 1=verb; 2=noun; 3=adjective; 4=adverb; 5=pronoun; 6=conjunction; 7=preposition; 8=numeral; 9=Chinese classifier; 10=Chinese particle; 11=English article; 12=English determiner

The above analysis suggests that TCT and OCT are similar in general POS distribution. However, it remains unclear if this is true of the distribution of specific lexical items. For a close look at the behaviours of lexical items, we employ keyword tool in WordSmith 4.0 to extract words that have statistically significant difference in frequency in the two subcorpora (TCT and OCT, the latter being the reference corpus). In what follows, we will observe the use of content words (in section 5.2) and function words (in section 5.3) in TCT and OCT.

5.2. Content words

5.2.1. Nouns

We find that TCT and OCT are different in the use of nominal expressions. Lexically, the frequency of words like 上帝 ‘God’ , 绅士 ‘gentleman’, 牧师 ‘priest’ are unusually high in TCT in comparison with OCT, while OCT makes a more frequent use of nouns like 当差 ‘lower official or servant in ancient China’, 洋车 ‘rickshaw’, 饺子 ‘*jiaozi*, dumpling’, 表姐 ‘cousin’, 姑奶奶 ‘sister of one’s paternal grandfather, but sometimes referring to speaker herself, arrogantly’, 旗袍 ‘chi-pao, a Chinese-style dress’, 夜壶 ‘chamber pot’, to name just a few.

Morphologically, morphemes like 兄 ‘brother’, 时 ‘time’, 堂 ‘cousin, mother’, 氏 ‘surname’, 斋 ‘studio’ are typically only found in OCT to form words like 令堂 ‘your mother’, 午时 ‘ midday hours’ and 白塔寺 ‘Baita Temple’.

Apparently, the difference in the use of common nouns is attributed largely to cultural and social differences between the two language communities.

5.2.2. Verbs

In TCT, we find that the following verbs are exceptionally frequent.

a) aspectual verbs: e.g. 开始 ‘begin’, 结束 ‘finish’;

b) ‘happen’ verbs: e.g. 发生 ‘happen’, 产生 ‘produce’;

c) ‘find’ verbs: e.g. 表现 ‘represent’, 发现 ‘discover’;

d) causative verbs: e.g. 让 ‘let, make’;

e) ‘judge’ verbs: e.g. 认为 ‘think’, 相信 ‘believe’, 感觉 ‘feel’;

f) psychological verbs: e.g. 害怕 ‘fear’, 怀疑 ‘doubt’;

g) others: e.g. 具有 ‘have’, 存在 ‘exist, there be’.

It should be noted that these words do occur in OCT, but their frequency is remarkably lower. The major reason for this difference is that their counterparts (verbs, prepositional phrases, and adjectives) occur much more frequently in EST.

We also find OCT uses more monosyllabic verbs than TCT. The frequency of verbs is unusually high in TCT compared to OCT, for example, 凑 ‘gather’, 搁 ‘place’, 甭 ‘don’t’, 傍 ‘depend on’, 嚷 ‘shout’, 嫌 ‘dislike’, 吵 ‘quarrel’, 捧 ‘hold in both hands’, 混 ‘mix; make trouble; lurk’, and 怔 ‘daze’. This means, from a monosyllable-idiomaticity correlation perspective, that TCT is less idiomatic than OCT.

5.2.3. Adjectives

As the keyword analysis shows, OCT frequently uses more monosyllabic adjectives than TCT, for example, 脆 'crisp', 高 'high', 贵 'expensive', 好 'good', 红 'red', 厚 'thick', 慌 'nervous', 紧 'tight', 老 'old', 俏 'charming', 小 'small', 饱 'full', and 苦 'bitter'. In contrast, TCT only frequently uses a small number of monosyllabic adjectives like 大 'big', 久 'long', 多 'many', and 快 'fast', but frequently uses more disyllabic adjectives. This is another evidence for the claim that TCT is less idiomatic than OCT in term of syllables.

5.2.4. Locative particles

In OCT, only 里 'inside', 外边 'outside', 内 'within' are more frequently used in comparison with TCT. In TCT, however, there are more such high-frequency locative particles including, for example, 以前 'before', 之前 'before', 之间 'between', 之后 'after', 之中 'in', 之外 'outside', and 周围 'around'. The unusual frequency of locative particles in TCT is a result of transfer from EST. Given that EST frequently uses prepositions such as *before*, *after*, *between*, *in*, *under*, *near*, and *around*, it is hardly surprising to find a rather frequent use of locative particles in TCT.

5.2.5. Adverbs

In TCT, some time-related adverbs such as 正 'in the process of', 已 'already', 已经 'already', and 一直 'always' are found to co-occur much more frequently with aspect markers *-zhe*, *-le* and *-guo* to unify their temporal features. The reason for this is that TCT has a strong tendency to explicitate by lexical means the perfective and imperfective senses inherent respectively in *have v-en* and *have been v-ing* constructions in English.

Modal adverbs express the speaker's attitude towards a proposition. In TCT, some modal adverbs are used with a strikingly high frequency, including 必须 'must', 或许 'perhaps', 竟然 'actually', 大约 'about', and 如此 'so' etc. In contrast, OCT frequently uses other words to perform the function, for example, 得 'have to', 兴许 'perhaps', 原来 'actually', 却 'actually', 来 'about', and 这么 'so'. It appears, then, that the difference lies not only in frequency, but in word choice as well: modal adverbs used in TCT seem to be more formal, far from being spoken and spontaneous as those used in OCT.

5.3. Function words

5.3.1. Pronominals

In OCT, the following pronominals show an unusually high frequency:[3] 大家 ‘all, everyone’, 她们 ‘they, them’, 怎 ‘what’, 怎样 ‘what’, 这 ‘this’, and 自己 ‘self’, etc.

In contrast, the following pronominals are unusually frequent in TCT: 她 ‘she, her’, 他 ‘he, him’, 他们 ‘they, them’, 它 ‘it’, 它们 ‘them’, 我 ‘I, me’, 我们 ‘we, us’, 那 ‘that’, 那儿 ‘there’, 那个 ‘that one’, 那时 ‘then’, 那种 ‘that kind of’, 这个 ‘this’, 这时 ‘at the time’, 这种 ‘this kind of’, 这样 ‘in this way’, 其他 ‘other’, 另 ‘the other’, 别的 ‘other’, 任何 ‘any’, 每个 ‘every’, and 一切 ‘all’. Apparently, in contrast with OCT, TCT exaggerates the use of first and third person pronouns, some demonstrative pronouns and some classifiers. In sum, the use of pronominals contributes to the uniqueness of translational Chinese. More analysis will be given in section 6.2.

5.3.2. Conjunctions

A major difference between TCT and OCT is found in the use of conjunctions. The statistics show that there are 15 exceptionally frequent conjunctions in TCT, including 不过 ‘but’, 但 ‘but’, 但是 ‘but’, 尽管 ‘though’, 或者 ‘or’, 不仅 ‘not only’, 而且 ‘moreover’, 另外 ‘in addition’, 哪怕 ‘even if’, 即使 ‘even though’, 如果 ‘if’, 然后 ‘then’, 因此 ‘so’, 于是 ‘upon that’, and 和 ‘and’. In contrast, only 7 conjunctions are used frequently in OCT, which are 可是 ‘but’, 并且 ‘and’, 况且 ‘besides’, 不但 ‘not only’, 所以 ‘so’, 假若 ‘if’, and 愈…愈 ‘the more…the more’. This difference in number and frequency only confirms our belief that logical relations are more implicit in Chinese than in English.

The POS distribution analysis made so far suggests that many features characteristic of TCT can be explained by EST interference. In short, compared with OCT, TCT uses more function words and more disyllabic words.

6. Compositionality

TCT may exaggerate the compositional potentiality of some morphemes or words in Chinese. The exaggeration can be observed in the following three ways.

6.1. Nominal morphemes

Take for example the Chinese morpheme 性 *-xing*, meaning ‘property’, which is similar to the nominal suffix *-ness* / *-ity* in English. It occurs 2.9 times per ten thousand words in original Chinese literary texts; in contrast, its frequency hits 5.2 in translated Chinese literary texts. Moreover, its diversity in composition increases in TCT. For example, there are 71 word types ending with *-xing* in translated Chinese literary texts, of which 42 are not used or rarely used in original Chinese literary texts, e.g. 独创性 ‘creativity’, 决定性 ‘decisiveness’, 可信性 ‘reliability’, 坚定性 ‘firmness’, 实质性 ‘substantiality’, and 强制性 ‘compulsiveness’. In the original Chinese literary texts, however, *-xing* is found more often to occur with monosyllabic words, forming words like 爽性 ‘straightforwardness’, 火性 ‘bad temper’, 牛性 ‘obstinacy’, 癖性 ‘natural inclination’, and 韧性 ‘tenacity’.

In translation, a translator consciously or subconsciously follows or imitates some features of a source language (Ke Fei 2005). A good case in point is the productive and diverse use of the *-xing* morpheme in TCT, where *-xing* is a translational equivalent of suffixes such as *-ity*, *-ness*, and *-dom* in English, hence it is more productive than it is in OCT. Interestingly, *-xing* is now a regular morpheme in Mandarin Chinese.

The phenomenon mentioned above is also true of morphemes such as 力 ‘force; ability’, and 度 ‘degree, extent’ in TCT, where they are very frequently used to form technical terms.

In a word, the high-frequency use of these morphemes in TCT suggests an imitation of their compositionality in English.

6.2. Composition of “Dem + Num + Cla”

The construction ‘demonstrative pronoun (Dem) + numeral (Num) + classifier (Cla)’ occurs more frequently in TCT than in OCT. Table 8-3 lists three constructions with such high frequency.

Table 8-3. Distribution of ‘Dem+ Num+ Cla’ constructions

	TCT		OCT	
	Frequency	Transitional probability	Frequency	Transitional probability
这一	1,010	0.035	334	0.019
这种	491	0.09	39	0.012
这件事	123	0.017	81	0.018

In OCT, the Dem-Num composition 这一 ‘lit. this one’ typically co-occurs with monosyllabic words such as 点 ‘point’, 条 ‘item’, 天 ‘day’, 次 ‘time’ and 年 ‘year’; but in TCT, its compositional potentiality is dramatically exploited so that it frequently co-occurs with disyllabic words like 问题 ‘problem’, 条款 ‘article, clause, item’, 事实 ‘fact’, 目标 ‘aim’, 领域 ‘field’, 计划 ‘plan’, 过程 ‘process’ and 观点 ‘viewpoint’. This exploitation contributes partially to high-frequency use of disyllabic words in TCT.

Another case in point is the Dem-Num phrase 这种 ‘this kind’, which is weak in compositionality in OCT where it almost exclusively co-occurs with 人 ‘person’. Even so, the combination 这种人 ‘this kind of person’ is far less frequent in OCT than in TCT (39 vs. 295). In TCT, 这种 ‘this kind’ is extremely productive, and it can yield diversified compositions, as can be seen in its combination with 药 ‘medicine’, 病 ‘disease’, 事 ‘matter’, 做法 ‘practice’, 想法 ‘idea’, 现象 ‘phenomenon’, 感觉 ‘feeling’, 情况 ‘situation’, 方式 ‘manner’, 方法 ‘method’, and many other nominals not frequently used in OCT.

The data in GCEPC shows that the frequent use of 这一 ‘this one’ and 这种 ‘this kind’ in TCT correlates with the frequent use of articles and demonstrative (such as *the / this / that*) in EST.

For the same reason, numeral-classifier phrases such as 一个 ‘an-individual’, 一件 ‘an-item’, 一位 ‘a-position’ and 一片 ‘a-slice’ occur more frequently in TCT than in OCT. The major reason for this difference is that indefinite articles in English strongly tend to be rendered into Chinese num-classifier phrases.

It should be noted that many Num-Cla-N bundles used in TCT are not always the direct translations of NPs in EST; in fact, some of them are renderings of pronouns or demonstratives. For example, 这件事 in TCT might be a translation equivalent of *it*, *this* or *that*, but not necessarily *the matter* or *the thing*.

(3a) **It** had happened at last.

这件 事 终于 发生 了。
This matter eventually happen LE

(3b) Nobody knows about **this** but us?

除了 咱们，没 人 知道 **这件 事** 吧？
except us, no man know this matter PRT

(3c) Humph! We'll see about **that**.

嗯，这件 事 我们 得 管一管 了。
Er, this matter we have to take care LE

(3d) There must 'a' been an angel **there**.

这件 事 一定 有 个 高手 在 帮 你的
This matter must have CLA master-hand ZAI help your

忙。
busy-work

The examples above demonstrate changes of cohesive devices in English-Chinese translation, i.e. a change from pronouns, demonstratives or demonstrative adverbs in English into NPs in Chinese translations. These changes reflect the difference between Chinese and English: the former employs more lexical devices to realize textual coherence. However, the changes are not frequently found in TCT; in fact, in comparison with OCT, TCT resorts more to the use of pronominals to achieve textual cohesion, which means translational Chinese is very similar to English source texts in the use of cohesive devices.

All in all, some grammatical devices typically used in English are prone to being imitated in English-Chinese translation, which in turn leads to the overuse of the compositionality of some morphemes, phrases and cohesive devices, thus making TCT less idiomatic than OCT.

6.3. Relative fixedness of some formulaic expressions

It is argued that some phrasal discourse markers such as comment clauses (e.g. *I think*, *it seems*) and conversational routines (e.g. *thank you*) are to some degree lexicalized, because they are relatively fixed usages (see Brinton and Traugott 2005: 67). In TCT, we do find some formulaic expressions that are rather frequently used. They might not be very “Chinese”, but they tend to be “institutionalized” and become rather fixed. Below are two examples.

The expression 随着时间的推移 ‘*lit.* with time’s move’ is now often used in Chinese, yet it is borrowed from English by imitating the structure of prepositional phrase *with the passage of time*. What interests us here is the fact that the Chinese expression can be used in translating many similar expressions in English, e.g. *as time went on*, *moment by moment*, *over time*, *as time drifted along*, *with a long-term time horizon*, *in the course of time*, *as time went by*, and even *eventually*.

Actually, the expression is now a regularly used expression in Chinese, and it has become even more popular than the very 'Chinese' expressions like 光阴荏苒 'time elapses quickly', 日复一日 'day after day', and 岁月流转 'with the passage of time'. Perhaps it is due to translation that the expression 随着时间的推移 has become an expression frequently used in Chinese.

Another example is the expression 是(不)可能[…]的 'It is (not) possible…', which conveys speaker's attitude toward a proposition. In TCT, this frequently used expression is a translation equivalent of many expressions in EST. Table 8-4 is a list of the possible equivalents.

Table 8-4. Expressions equivalent to 是(不)可能(的)

Equivalents	*possible*	*can*	*likely*	*will*	*might*	*probable*	*incapable*	Total
Frequency	41	10	5	2	2	1	1	64

In OCT, the expression 是(不)可能[…]的 appears only 17 times, all of which being found in political texts; in addition, it is usually put at the end of a sentence, hence not used frequently and diversely. In TCT, its use is diversified because it can appear at different positions in a clause (very similar to the use of 'possibility' expressions in English).

Furthermore, in TCT, the load capacity of 是不可能[…]的 is expanded. In the example (4), for instance, the capacity of […] is expanded to 34 Chinese characters (in the Chinese version, the highlighted parts are equivalent to the part following 'the impossibility that' in English):

> (4) …but had long since recognized the impossibility that any mission of divine and mysterious truth should be confided to a woman stained with sin, bowed down with shame, or even burdened with a life-long sorrow.
> …但从那以后，她早已承认了：任何上界的神秘真理的使命**是不可能**委托给一个为罪孽所玷污、为耻辱所压倒或者甚至为终生的忧愁而沉闷的女人**的**。

Through n-gram search in TCT, we find many other expressions behaving like the two expressions above, for example, 目的是为了 'for the purpose of', 在某种程度/意义上 'to some degree, in a sense', 是必要

的 ‘it is necessary that’, 一遍又一遍 ‘time and again’, 很久很久以前 ‘long, long ago’, 更确切地说 ‘precisely’, and 一般情况下 ‘generally’.

The fixedness of expressions in TCT suggests that once an expression enters into the target language through translation, it might become relatively fixed and frequently used in the translational language; it can even grow into a popular expression in the target language. TCT has a stronger tendency to use relatively fixed expressions to deal with diverse expressions (with same or similar functions) in EST, which provides support for the hypothesis of lexical simplification in translations, but at the same time weighs against the normalization hypothesis.

7. Conclusions

It can be seen from the discussions above that translational Chinese has the following features:

1) TCT uses fewer monosyllabic words than OCT does;
2) TCT tends to expand the normal load capacity of some Chinese constructions, which leads to longer sentence segments;
3) In comparison with OCT, TCT uses more function words;
4) TCT can change or expand the compositionality of some words or morphemes in Chinese.

These features do not fully support translation universal hypotheses.

Firstly, TCT uses more types and longer segments than OCT. This does not support lexical and syntactic simplification. Secondly, explicitation in TCT runs in parallel with implicitation, such as the implicitation of logical relations and co-reference devices. In this sense, explicitation is a relative notion. As far as English-Chinese translation is concerned, TCT is more explicit than OCT, but more implicit than EST. This relativity suggests that explicitation and implicitation co-exist in any translation pair; it is not always unidirectional. Finally, given that TCT exaggerates the compositional potentiality of some morphemes and words in Chinese, and that it has expanded the load capacity of some Chinese constructions, translational Chinese does not fully support the normalization hypothesis.

It can be concluded that the so-called translation universals might be a shifting phenomenon between specific languages or just some features characterizing local translated discourse. Considering that, they are by no means the corollaries applicable to all translational languages.

Notes

1. The STTR ratio is calculated by taking the average of the type-token ratios based on consecutive 1,000-word text chunks of the subcorpus in question (see Scott 2004).
2. A segment can surface as a clause or a phrase, with commas, semi-colons and colons being delimiters.
3. In Chinese linguistic literature, pronouns are classified as content words. However, considering that they form a closed set, we take them as function words.

References

Baker, M. (1993), “Corpus linguistics and translation studies: Implications and applications”. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, 233-250. Amsterdam: John Benjamins.

—. (1998a), “Réexplorer la langue de la traduction: une approche par corpus”. *Meta* 43(4): 480-485.

—. (1998b), *Routledge Encyclopaedia of Translation Studies*. London: Routledge.

Brinton, L. J. and Traugott, E. C. (2005), *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

Chen, H-H. (1994), “The contextual analysis of Chinese sentences with punctuation marks. *Literary and Linguistic Computing* 9(4): 281-289.

Hansen S. and Teich, E. (2001), “Multi-layer analysis of translation corpora: Methodological issues and practical implications”, in D. Cristea, N. Ide, D. Marcu and M. Poesio (eds.) *Proceedings of EUROLAN 2001 Workshop on Multi-layer Corpus-based Analysis*, 44-55. Iasi, Romania.

Hopkinson, C. (2007), “Factors in linguistic interference: A case of study in translation”. *Skase Journal of Translation and Interpretation* 2(1): 13-23.

Huang, L. and Wang, K. (2006), “Fanyi pubianxing yanjiu fansi” (Reflections on the corpus-based studies of translation universals). *Chinese Translators Journal* 5: 36-40.

Ke Fei (2003), “Hanyu ba zi ju tedian, fenbu ji Ying yi” (Chinese BA-construction and its English translation). *Foreign Languages and Their Teaching* (12): 1-5.

—. (2005), “Fanyi zhong de yin he xian” (Implicitation and explicitation in translation). *Foreign Language Teaching and Research* (4): 303-307.

Lü, S. (ed.) (1980), *Xiandai Hanyu Ba Bai Ci* (*Eight Hundred Words in Modern Chinese*). Beijing: The Commercial Press.

Qin, H. and Wang, K. (2004), “Jiyu yuliaoku de fanyi yuyan fenxi” (Parallel corpus-based analysis of translationese). *Modern Foreign Languages* (1): 44-52.

Santos, D. (1995), “On grammatical translationese”, in K. Koskenniemi (ed.) *Short Papers Presented at the Tenth Scandinavian Conference on Computational Linguistics,* 59-66. Helsinki.

Scott, M. (2004), *The WordSmith Tools* (v. 4.0). Oxford: Oxford University Press.

Shao, Z. (2005), *Han Ying Duibi Fanyi Daolun* (*Chinese-English Contrastive Analysis and Translation*). Shanghai: Huadong University of Technology Press.

Si, G. (2002), *Yi Dao Tanwei* (*Looking into Translation*). Beijing: China Translation and Publishing Corporation.

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

Wang, K. (ed.) (2004), *Shuangyu Duiying Yuliaoku Yanzhi yu Yingyong* (The Development of the Compilation and Application of Parallel Corpora). Beijing: Foreign Language Education and Research Press.

Westin, I. (2002), *Language and Computers: Studies in Practical Linguistics*. New York: Rodopi.

Wu, A. and Huang, L. (2006) “Guanyu fanyi gongxing de yanjiu” (On research of translation universals). *Foreign Language Teaching and Research* (5): 296-302.

Yu, G. (2002), *Yu Guangzhong Tan Fanyi* (*Yu Guangzhong Talking about Translation*). Beijing: China Translation and Publishing Corporation.

Chapter Nine

In Pursuit of the "Third Code": Using the ZJU Corpus of Translational Chinese in Translation Studies

Richard Xiao, Lianzhen He, Ming Yue

1. Introduction

Since the 1990s, the rapid development of the corpus-based approach in linguistic investigation in general, and the development of multilingual corpora in particular, have brought even more vigor into Descriptive Translation Studies (DTS) (cf. McEnery, Xiao and Tono 2006: 90-95). As Laviosa (1998a: 474) observes, "the corpus-based approach is evolving, through theoretical elaboration and empirical realization, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation." Presently, corpus-based DTS has primarily been concerned with describing translation as a product, by comparing corpora of translated and non-translational native texts in the target language, especially translated and native English. The majority of product-oriented translation studies attempt to uncover evidence to support or reject the so-called translation universal (TU) hypotheses that are concerned with features of translational language as the "third code" of translation (Frawley 1984), which is supposed to be different from both source and target languages.

As far as the English language is concerned, a large part of product-oriented translation research has been based on the *Translational English Corpus* (TEC), which was built by Mona Baker and colleagues at the University of Manchester (see Baker 2004). The TEC corpus, which was designed specifically for the purposes of studying translated English, consists of contemporary written texts translated into English from a range of source languages. It is constantly expanded with fresh materials, and had reached a total of ten million words by the year 2003. The corpus comprises full texts from four genres (fiction, biography, newspaper articles and in-flight magazines) translated by native speakers of English. Paralinguistic data such as the information about translators, source texts and publishing dates is annotated and stored in the header section of each text. The TEC corpus was created in such a way that parts of the *British National Corpus* (BNC) can be used as a comparable corpus, with matching composition and dates of publication.

The TEC corpus is perhaps the only publicly available corpus of translational English. Most of the pioneering and prominent studies of translational English, which have so far focused on syntactic and lexical features of translated and original texts of English, have been based on this corpus. They have provided evidence to support the hypotheses of translation universals in translated English, most noticeably simplification, explicitation, sanitization, and normalization (see section 2 for further discussion). For example, Laviosa (1998b) studies the distinctive features of translational English in relation to native English (as represented by the BNC corpus), finding that translational language has four core patterns of lexical use: a relatively lower proportion of lexical words over function words (i.e. significantly lower lexical density than in non-translated texts; see sestion 4.1), a relatively higher proportion of high-frequency words over low-frequency words, a relatively higher rate of repetition of the most frequent words, and a smaller vocabulary frequently used. This is regarded as the most significant work in support of the simplification hypothesis of translation universals. Olohan and Baker's (2000) comparison of concordances from the TEC and BNC corpora shows that the *that*-connective with reporting verbs *say* and *tell* is far more frequent in translational English, and conversely, that the zero-connective is more frequent in native English. These results provide strong evidence for syntactic explicitation in translated English, which, unlike "the addition of explanatory information used to fill in knowledge gaps between source text and target text readers, is hypothesized to be a subliminal phenomenon inherent in the translation process" (Laviosa 2002: 68). Olohan (2004) investigates intensifiers such as *quite*, *rather*, *pretty* and *fairly* in translated versus native English fiction in an attempt to uncover the relationship between collocation and moderation, finding that *pretty* and *rather*, and more marginally *quite*, are considerably less frequent in the TEC-fiction subcorpus; but when they are used, there is usually more variation in usage, and less repetition of common collocates, than in the BNC-fiction corpus. The observation that these moderating items are less frequently used in translated texts leads Olohan (2004: 142) to relate moderation to explicitation: "translators may remove or downplay elements of ‘moderation’, perhaps as part of a (non-deliberate) process of disambiguation or explicitation.”

Similar features have also been reported in the translational variants of a few languages other than English (e.g. Swedish). Nevertheless, research of this area has so far been confined largely to translational English translated from closely related European languages (e.g. Mauranen and Kujamäki 2004). If the features of translational language that have been reported on the basis of translated English are to be generalized as translation universals, it is of vital importance to find supporting evidence from non-European languages. Clearly, evidence from “genetically” distinct language pairs such as English and Chinese is arguably more convincing, if not indispensable. This motivates us to undertake a project that studies the features of translational Chinese.

This chapter first reviews previous research of the features of translational language (section 2). We will then introduce the newly created *ZJU Corpus of Translational Chinese* (ZCTC), which is designed with the explicit aim of studying translational Chinese (section 3). Section 4 presents a number of case studies of the lexical and syntactic features of translational Chinese while section 5 concludes this chapter.

2. Translation universals: A review

An important area of Descriptive Translation Studies is the hypothesis of so-called translation universals (TUs) and its related sub-hypotheses, which are sometimes referred to as the inherent features of translational language, or translationese. It is a well-recognized fact that translations cannot possibly avoid the effect of translationese (cf. Hartmann 1985; Baker 1993: 243-245; Teubert 1996: 247; Gellerstam 1996; Laviosa 1997: 315; McEnery and Wilson 2001: 71-72; McEnery and Xiao 2002, 2007). While Frawley (1984) recognized translated language as a distinct language variety, it is since Baker’s (1993) seminal paper that “the idea of linguistic translation universals has found a place at the centre of discussion in translation studies” (Mauranen and Kujamäki 2004: 1). Baker (1993) suggests that all translations are likely to show certain linguistic characteristics simply by virtue of being translations, which are caused in and by the process of translation. The effect of the source language on the translations is strong enough to make the translated language perceptibly different from the target native language. Consequently translational language is at best an unrepresentative special variant of the target language (McEnery and Xiao 2007). The distinctive features of translational language can be identified by comparing translations with comparable native texts, thus throwing new light on the translation process and helping to uncover translation norms, or what Frawley (1984) calls the "third code" of translation.

Over the past decade, TUs have been an important area of research as well as a target of debate in Descriptive Translation Studies. Some scholars (e.g. Tymoczko 1998, Malmkjær 2007, House 2008) are skeptical of translation universals, arguing that the very idea of making universal claims about translation is inconceivable, while others (e.g. Toury 2004) advocate that the chief value of general laws of translation lies in their explanatory power; still others (e.g. Chesterman 2004) accept universals as one possible route to high-level generalizations. Chesterman (2004) further differentiates between two types of TUs. One relates to the process from the source to the target text (namely "S-universals"), which requires a parallel corpus of source and target texts to investigate; the other ("T-universals") compares translations with native target-language texts, which requires a comparable corpus of translated and native target language to investigate. Mauranen (2007) suggests in her comprehensive review of TUs that the discussion of TUs should follow the general discussion on 'universals' in language typology.

Recent corpus-based translation studies have proposed a number of TUs, the best known of which include explicitation, simplification, normalization, sanitization and leveling out (or convergence). Other TUs that have been investigated include source language shining through (Teich 2001), under-representation, interference and untypical collocations (see Mauranen 2007). While individual studies have sometimes investigated more than one of these features, they are discussed in the following subsections separately for the purpose of this presentation.

2.1. Explicitation

The explicitation hypothesis is formulated by Blum-Kulka (1986) on the basis of evidence from individual sample texts showing that translators tend to make explicit optional cohesive markers in the target text even though they are absent in the source text. It relates to the tendency in translations to "spell things out rather than leave them implicit" (Baker 1996: 180). Explicitation can be realized syntactically or lexically, for instance, via more frequent use of conjunctions in translated texts than in non-translated texts.

For example, Chen (2006) presents a corpus-based study of connectives, namely conjunctions and sentential adverbials, in a "composite corpus" composed of English source texts and their two Chinese versions independently published in Taiwan and mainland China, plus a comparable component of native Chinese texts as the reference corpus in the genre of popular science writing. This investigation examines translation both as a product and as a process in an attempt to verify the hypothesis of explicitation in translated Chinese. In the product-oriented part of his study, Chen compares translational and native Chinese texts to find out whether connectives are significantly more common in the first type of texts in terms of parameters such as frequency and type-token ratio, as well as statistically defined common connectives and the so-called "translationally distinctive connectives" (TDCs). He also examines whether syntactic patterning in the translated texts is different from native texts via a case study of the five TDCs that are most statistically significant. In the part of the study that examines the level of influence of the source language on explicitation in the translation process, he compares translated Chinese texts with the English source texts, through a study of the same five TDCs, in an attempt to determine the extent to which connectives in translated Chinese texts are carried over from the English source texts, or in other words, the extent to which connectives are explicitated in translational Chinese. Both parts of his study support the hypothesis of explicitation as a translation universal in the process and product of English-Chinese translation of popular science writing.

Another result of explicitation is increased cohesion in translated text (Øverås 1998). Pym (2005) provides an excellent account of explicitation, locating its origin, discussing its different types, elaborating a model of explicitation within a risk-management framework, and offering a range of explanations of the phenomenon.

In the light of the distinction made above between S- and T-universals (Chesterman 2004), explicitation would seem to fall most naturally into the S-type. Recently, however, explicitation has also been studied as a T-universal. In his corpus-based study of structures involving NP modification (i.e. equivalent of the structure noun + prepositional phrase in English) in English and Hungarian, Váradi (2007) suggests that genuine cases of explicitation must be distinguished from constructions that require expansion in order to meet the requirements of grammar (see House's 2008 distinction between optional and obligatory linguistic choices). While explicitation is found at various linguistic levels ranging from lexis to syntax and textual organization, "there is variation even in these results, which could be explained in terms of the level of language studied, or the genre of the texts" (Mauranen 2007: 39). The question of whether explicitation is a translation universal is yet to be conclusively answered, according to existing evidence which has largely come from translational

English and related European languages (see section 4 for further discussion).

2.2. Simplification

Simplification refers to "the tendency to simplify the language used in translation" (Baker 1996: 181-182), which means that translational language is supposed to be simpler than native language, lexically, syntactically and / or stylistically (cf. Blum-Kulka and Levenston 1983; Laviosa-Braithwaite 1997). As noted earlier, product-oriented studies such as Laviosa (1998b) and Olohan and Baker (2000) have provided evidence for lexical and syntactic simplification in translational English. Translated texts have also been found to be simplified stylistically. For example, Malmkjær (1997) notes that in translations, punctuation usually becomes stronger; for example commas are often replaced with semicolons or full stops while semicolons are replaced with full stops. As a result, long and complex sentences in the source text tend to be broken up into shorter and less complex clauses in translations (a phenomenon that Fabricius-Hansen 1999 refers to as "sentence splitting"), thereby reducing structural complexity for easier reading. On the other hand, Laviosa (1998b: 5) observes that translated language has a significantly greater mean sentence length than non-translated language. Xiao and Yue's (2009) finding that translated Chinese fiction displays a significantly greater mean sentence length than native Chinese fiction is in line with Laviosa's (1998b: 5) observation but goes against Malmkjær's (1997) expectation that stronger punctuations tend to result in shorter sentences in translated texts. It appears, then, that mean sentence length might not be a translation universal but rather associated with specific languages or genres (see section 4.1 for further discussion).

The simplification hypothesis, however, is controversial. It has been contested by subsequent studies of collocations (Mauranen 2000), lexical use (Jantunen 2001), and syntax (Jantunen 2004). Just as Laviosa-Braithwaite (1996: 534) cautions, evidence produced in early studies that support the simplification hypothesis is patchy and not always coherent. Such studies are based on different datasets and are carried out to address different research questions, and thus cannot be compared.

2.3. Normalization

Normalization, which is also called 'conventionalization' in the literature (e.g. Mauranen 2007), refers to the "tendency to exaggerate features of the target language and to conform to its typical patterns" (Baker 1996: 183). As a result, translational language appears to be "more normal" than the target language. Typical manifestations of normalization include overuse of clichés or typical grammatical structures of the target language (but see section 4.4 for counter evidence), overuse of typical features of the genres involved,[1] adapting punctuation to the typical usage of the target language, and the treatment of the different dialects used by certain characters in dialogues in the source texts.

Kenny (1998, 1999, 2000, and 2001) presents a series of studies of how unusual and marked compounds and collocations in German literary texts are translated into English, in an attempt to assess whether they are normalized by means of more conventional use. Her research suggests that certain translators may be more inclined to normalize than others, and that normalization may apply in particular to lexis in the source text. Nevalainen (2005, cited in Mauranen 2007: 41) suggests that translated texts show greater proportions of recurrent lexical bundles or word clusters. Beyond the lexical level, there are a number of studies which explore grammatical normalization (e.g. Teich 2001, Hansen 2003).

Like simplification, normalization is also a debatable hypothesis. According to Toury (1995: 208), it is a "well-documented fact that in translations, linguistic forms and structures often occur which are rarely, or perhaps even never encountered in utterances originally composed in the target language." Tirkkonen-Condit's (2002: 216) experiment, which asked subjects to distinguish translations from non-translated texts, also shows that "translations are not readily distinguishable from original writing on account of their linguistic features."

2.4. Other translation universals

Kenny (1998) analyzes semantic prosody (i.e. a kind of meaning arising from collocation) in translated texts in an attempt to find evidence of sanitization (i.e. reduced connotational meaning). She concludes that translated texts are "somewhat 'sanitized' versions of the original" (Kenny 1998: 515). Another translation universal that has been proposed is the so-called feature of 'leveling out', i.e. "the tendency of translated text to gravitate towards the centre of a continuum" (Baker 1996: 184). This is what Laviosa (2002: 72) calls "convergence", i.e. the "relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features" that are discussed above.

'Under-representation', which is also known as the "unique items hypothesis", is concerned with the unique items in translation (Mauranen

2007: 41-42). For example, Tirkkonen-Condit (2005) compared the frequencies and uses of the clitic particle *kin* in translated and original Finnish in five genres (i.e. fiction, children’s fiction, popular fiction, academic prose, and popular science), finding that the average frequency of *kin* in original Finnish is 6.1 instances per 1,000 words, whereas its normalized frequency in translated Finnish is 4.6 instances per 1,000 words. Tirkkonen-Condit interprets this result as a case of under-representation in translated Finnish. Aijmer’s (2007) study of the use of the English discourse marker *oh* and its translation in Swedish shows that there is no single lexical equivalent of *oh* in Swedish translation, because direct translation with the standard Swedish equivalent *áh* would result in an unnatural sounding structure in this language.

Another feature of translational language is source language (SL) shining through, which means that “[i]n a translation into a given target language (TL), the translation may be oriented more towards the source language (SL), i.e. the SL shines through” (Teich 2003: 145). For example, Teich (2003: 207) finds that in both English-to-German and German-to-English translations, both target languages exhibit a mixture of TL normalization and SL shining through.

The above is not a comprehensive survey of translation universals. There are still some other features of translations that are often discussed in translation textbooks as strategies of translation (e.g. expansion), which will not be reviewed here.

3. The ZJU Corpus of Translational Chinese

As can be seen in the discussion above, while we have followed the convention of using the term ‘translation universal’, the term is highly debatable in the literature. Since the translation universals that have been proposed so far are identified on the basis of translational English – mostly translated from closely related European languages –, there is a possibility that such linguistic features are not universal but rather specific to English and / or genetically related languages that have been investigated. For example, Cheong’s (2006) study of English-Korean translation contradicts even the least controversial explicitation hypothesis.

We noted in section 2.1 that the explicitation hypothesis is supported by Chen’s (2006) study of connectives in English-to-Chinese translation of popular science books. Nevertheless, as Biber (1995: 278) observes, language may vary across genres even more markedly than across languages. Xiao (2009) also demonstrates that the genre of scientific writing is the least diversified of all genres across various varieties of

English. The implication is that the similarity reported in Chen (2006) might be a result of similar genre rather than language pairing.[2] Ideally, what is required to verify the English-based translation universals is a detailed account of the features of translational Chinese based on balanced comparable corpora of translational and native Chinese. This is the aim of our ongoing project *A corpus-based quantitative study of translational Chinese in English-Chinese translation*, which is funded by the China National Foundation of Social Sciences.

The project has two major parts. The first part aims to develop a translational counterpart of the *Lancaster Corpus of Mandarin Chinese* (LCMC), a one-million-word balanced corpus of native Chinese, while the second part undertakes a quantitative study of translational Chinese using a composite approach that integrates monolingual comparable corpus analysis and parallel corpus analysis as advocated in McEnery and Xiao (2002). The monolingual comparable corpus approach compares comparable corpora of translated language with the native target language in an attempt to uncover salient features of translations, while the parallel corpus approach compares source and target languages on the basis of a separate English-to-Chinese parallel corpus to determine the level of shining through, i.e. the extent to which the features of translated texts are transferred from the source language.

We have so far completed the first part of the project. The remainder of this section introduces the *ZJU Corpus of Translational Chinese* (ZCTC), while section 4 will present a number of case studies based on this corpus.

3.1. Corpus design

The *ZJU Corpus of Translational Chinese* (ZCTC) was created with the explicit aim of studying the features of translated Chinese in relation to non-translated native Chinese. It has modeled the *Lancaster Corpus of Mandarin Chinese* (LCMC), which is a one-million-word balanced corpus designed to represent native Mandarin Chinese (McEnery and Xiao 2004). Both LCMC and ZCTC corpora have sampled five hundred 2,000-word text chunks from fifteen written text categories published in China, with each corpus amounting to one million words. Table 9-1 shows the genres covered in the two corpora, together with their respective proportions. Since the LCMC corpus was designed as a Chinese match for the FLOB corpus of British English (Hundt, Sand and Siemund 1998) and the Frown corpus of American English (Hundt, Sand and Skandera 1999), with the specific aim of comparing and contrasting English and Chinese, the number of text samples and their proportions given in Table 9-1 are exactly the same as in FLOB and Frown.

Table 9-1. The genres covered in LCMC and ZCTC

Code	Genre	Number of samples	Proportion
A	Press reportage	44	8.8%
B	Press editorials	27	5.4%
C	Press reviews	17	3.4%
D	Religious writing	17	3.4%
E	Skills, trades and hobbies	38	7.6%
F	Popular lore	44	8.8%
G	Biographies and essays	77	15.4%
H	Miscellaneous (reports, official documents)	30	6%
J	Science (academic prose)	80	16%
K	General fiction	29	5.8%
L	Mystery and detective fiction	24	4.8%
M	Science fiction	6	1.2%
N	Adventure fiction	29	5.8%
P	Romantic fiction	29	5.8%
R	Humour	9	1.8%
Total		500	100%

The LCMC corpus has also followed the sampling period of FLOB / Frown by sampling written Mandarin Chinese within three years around 1991. While it was relatively easy to find texts of native Chinese published in this sampling period, it would be much more difficult to get access to translated Chinese texts of some genres - especially in electronic format - published within this time frame. This pragmatic consideration of data collection forced us to modify the LCMC model slightly by extending the sampling period by a decade, i.e. to 2001, when we built the ZCTC corpus. This extension was particularly useful because the popularization of the Internet and online publication in the 1990s made it possible and easier to access a large amount of digitalized texts.[3] While English is the source language of the vast majority (99%) of the text samples included in the ZCTC corpus, we have also included a small number of texts translated from other languages (e.g. Japanese, French, Spanish, and Romanian) to mirror the reality of the world of translations in China.

As Chinese is written as running strings of characters without white spaces delimiting words, it is only possible to know the number of word tokens in a text when the text has been tokenized (see section 3.2). As such, the text chunks were collected at the initial stage by using our best estimate of the ratio (1:1.67) between the number of characters and the number of words based on our previous experience (McEnery, Xiao and Mo 2003). Only textual data was included, with graphs and tables in the original texts replaced by placeholders. A text chunk included in the corpus can be a sample from a large text (e.g. an article and book chapter) or an assembly of several small texts (e.g. for the press categories and humour). When parts of large texts were selected, an attempt was made to achieve a balance between initial, medial and ending samples. When the texts were tokenized, a computer program was used to cut large texts to approximately 2,000 tokens while keeping the final sentence complete. As a result, while some text samples may be slightly longer than others, they are typically around 2,000 words.

Table 9-2. A comparison of ZCTC and LCMC corpora

Genre	ZCTC	Proportion	LCMC	Proportion
A	88,196	8.67	89,367	8.73
B	54,171	5.32	54,595	5.33
C	34,100	3.35	34,518	3.37
D	35,139	3.45	35,365	3.46
E	76,681	7.54	77,641	7.59
F	89,675	8.81	89,967	8.79
G	155,601	15.29	156,564	15.30
H	60,352	5.93	61,140	5.97
J	164,602	16.18	163,006	15.93
K	60,540	5.95	60,357	5.90
L	48,924	4.81	49,434	4.83
M	12,267	1.21	12,539	1.23
N	59,042	5.80	60,398	5.90
P	59,033	5.80	59,851	5.85
R	19,072	1.87	18,645	1.82
Total	1,017,395	100.00	1,023,387	100.00

Table 9-2 compares the actual numbers of word tokens in different genres as well as their corresponding proportions in the ZCTC and LCMC corpora.[4] As can be seen, the two corpora are roughly comparable in terms of both overall size and proportions for different genres.

Table 9-3. Level 1 part-of-speech categories

Level 1 POS category	Explanation
a	Adjective
b	Non-predicate noun modifier
c	Conjunction
d	Adverb
e	Interjection
f	Space word
h	Prefix
k	Suffix
m	Numeral and quantifier
n	Noun
o	Onomatopoeia
p	Preposition
q	Classifier
r	Pronoun
s	Place word
t	Time word
u	Auxiliary
v	Verb
w	Symbol and punctuation
x	Non-word character string
y	Particle
z	Descriptive adjective

3.2. Corpus annotation

The ZCTC corpus is annotated using ICTCLAS2008, the latest release of the *Chinese Lexical Analysis System* developed by the Institute of Computing Technology, the Chinese Academy of Sciences. This annotation tool, which relies on a large lexicon and the Hierarchical Hidden Markov Model (HMM), integrates word tokenization, named entity identification, unknown word recognition, as well as part-of-speech (POS) tagging. The ICTCLAS part-of-speech tagset distinguishes between 22 level 1 part-of-speech categories (see Table 9-3), which expand into over 80 level 2 and 3 categories for word tokens in addition to more than a dozen categories for symbols and punctuations.[5] The ICTCLAS2008 tagger has been reported to achieve a precision rate of 98.54% for word tokenization. Latest open tests have also yielded encouraging results, with a precision rate of 98.13% for tokenization and 94.63% for part-of-speech tagging.[6]

3.3. Corpus markup

The ZCTC corpus is marked up in Extensible Markup Language (XML) which is in compliance with the Corpus Encoding Standards (CES, see Ide and Priest-Dorman 2000). Each of the 500 data files has two parts: a corpus header and a body. The *cesHeader* gives general information about the corpus (*publicationStmt*) as well as specific attributes of the text sample (*fileDesc*). Details in the *publicationStmt* element include the name of the corpus in English and Chinese, authors, distributor, availability, publication date, and history. The *fileDesc* element shows the original title(s) of the text(s) from which the sample was taken, individuals responsible for sampling and corpus processing, the project that created the corpus file, date of creation, language usage, writing system, character encoding, and mode of channel.[7]

The body part of the corpus file contains the textual data, which is marked up for structural organization such as paragraphs (*p*) and sentences (*s*). Sentences are consecutively numbered for easy reference. Part-of-speech annotation is also given in XML, with the POS attribute of the *w* element indicating its part-of-speech category.

The XML markup of the ZCTC corpus is perfectly well-formed and has been validated using Altova XMLSpy 2008, a comprehensive editing tool for XML documents.[8] The XML elements of the corpus are defined in the accompanying Document Type Definition. The ZCTC corpus is encoded in Unicode, applying the Unicode Transformation Format 8-Bit (UTF-8), which is a lossless encoding for Chinese while keeping the XML files at a minimum size. The combination of Unicode and XML is a general trend and standard configuration in corpus development, especially when corpora involve languages other than English (cf. Xiao, McEnery, Baker and Hardie 2004).

4. Lexical and syntactic features of translational Chinese

This section presents four case studies of lexical and syntactic features of translational Chinese as represented in the new ZCTC corpus in comparison with the retagged edition of the LCMC corpus (see note 4). We will first verify Laviosa's (1998b) core features of lexical use in translational Chinese (sections 4.1 and 4.2), and then compare the use of connectives and passives in translated and native Chinese (sections 4.3 and 4.4).

4.1. Lexical density and mean sentence length

This section discusses the parameters used in Laviosa (1998b) in an attempt to find out whether the core patterns of lexical use that Laviosa observes in translational English also apply in translated Chinese. We will first compare lexical density and mean sentence length in native and translated Chinese, and then examine the frequency profiles of the two corpora in the following section.

There are two common measures of lexical density. Stubbs (1986: 33; 1996: 172) defines lexical density as the ratio between the number of lexical words (i.e. content words) and the total number of words. This approach is taken in Laviosa (1998b). As our corpora are part-of-speech tagged, frequencies of different POS categories are readily available.

The other approach commonly used in corpus linguistics is the type-token ratio (TTR), i.e. the ratio between the number of types (i.e. unique words) and the number of tokens (i.e. running words). However, since the TTR is seriously affected by text length, it is reliable only when texts of equal or similar length are compared. To remedy this issue, Scott (2004) proposes a different strategy, namely, using a standardized type-token ratio (STTR), which is computed every *n* (the default setting is 1,000 in the WordSmith Tools) words as the Wordlist application of WordSmith goes through each text file in a corpus. The STTR is the average type-token ratio based on consecutive 1,000-word chunks of text (Scott 2004: 130). It appears that lexical density defined by Stubbs (1986, 1996) measures informational load whereas the STTR is a measure of lexical variability, as reflected by the different ways they are computed.

Let us first examine the Stubbs-style lexical density in native and translational Chinese. Xiao and Yue (2009) find that the lexical density in translated Chinese fiction (58.69%) is significantly lower than that in native Chinese fiction (63.19%). Does this result also hold for other genres or for Mandarin Chinese in general as represented in the two balanced corpora in the present study?

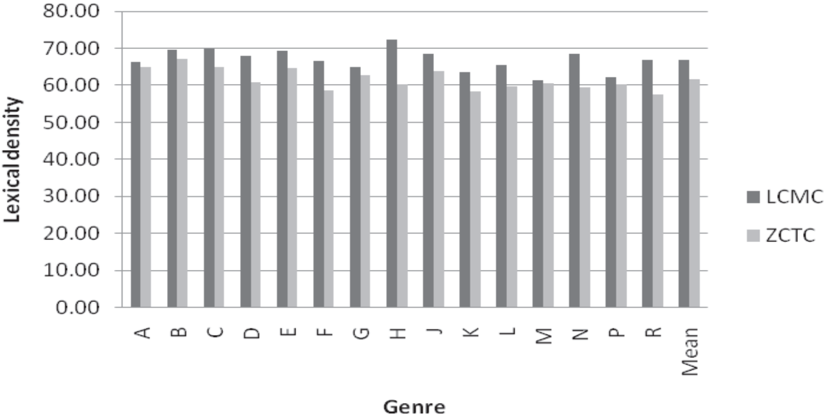


Figure 9-1. Lexical density in ZCTC and LCMC

Table 9-4. Mean differences in lexical density across genres

Genre	t score	Degree of freedom (d.f.)	Significance level (p)	Mean difference
A	-2.43	86	0.017	-1.446
B	-3.35	52	0.002	-2.180
C	-6.96	32	<0.001	-5.144
D	-8.07	32	<0.001	-7.307
E	-4.93	74	<0.001	-4.703
F	-9.79	86	<0.001	-7.934
G	-4.05	152	<0.001	-2.184
H	-9.61	58	<0.001	-12.21
J	-9.13	158	<0.001	-4.777
K	-5.64	56	<0.001	-5.193
L	-6.28	46	<0.001	-5.984
M	-0.44	10	0.667	-1.056
N	-13.66	56	<0.001	-9.122
P	-2.29	56	0.026	-1.987
R	-8.85	16	<0.001	-9.215
Mean	-4.94	28	<0.001	-5.342

Figure 9-1 shows the scores of lexical density in the fifteen genres covered in the ZCTC and LCMC corpora as well as their mean scores. As can be seen, the mean lexical density in LCMC (66.93%) is considerably higher than that in ZCTC (61.59%). This mean difference -5.34 is statistically significant (t = -4.94 for 28 d.f., p<0.001). It is also clear from the figure that all of the fifteen genres have a higher lexical density in native than translated Chinese, which is statistically significant for nearly all genres (barring science fiction M), as indicated by the statistic tests in Table 9-4. These findings are in line with Laviosa's (1998b) observations of lexical density in translational English.

However, if lexical density is measured by the STTR, then the LCMC corpus as a whole has a slightly higher STTR than ZCTC (46.58 vs. 45.73), but the mean difference (-0.847) is not statistically significant (t = - 0.573 for 28 d.f., p=0.571). This result is further confirmed by a closer look at individual genres (Figure 9-2). As can be seen, the differences for most genres are marginal. While some genres display a higher STTR in native Chinese, there are also genres with a higher STTR in translated Chinese. This finding extends Xiao and Yue's (2009) observation of translated Chinese fiction to Mandarin Chinese in general.

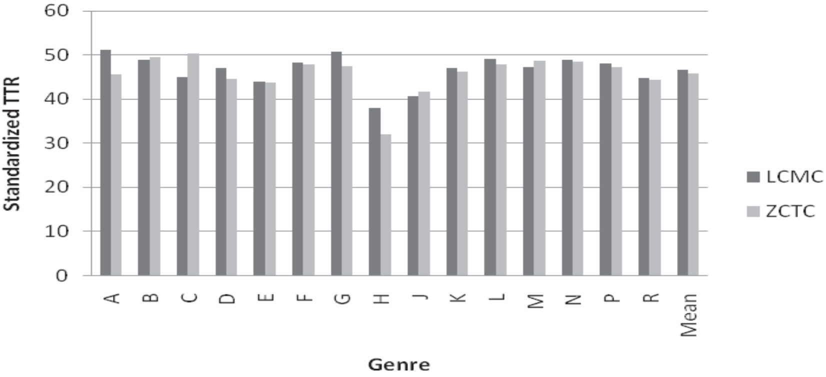


Figure 9-2. Standardized TTR in ZCTC and LCMC

In terms of lexical versus function words,[9] a significantly higher ratio of lexical over function words is found in native Chinese than in translated Chinese (2.08 vs. 1.64, t = –4.88 for 28 d.f., p<0.001, mean difference = –0.441). As can be seen in Figure 9-3, which gives the lexical-to-function word ratios in ZCTC and LCMC, all genres have a higher ratio in native

Chinese than in translated Chinese, and the mean differences for all genres other than science fiction (M) are statistically significant (see Table 9-5), especially in reports and official documents (H), adventure fiction (N) and humour (R).

Table 9-5. Mean differences in lexical-to-function word ratio across genres

Genre	t score	Degree of freedom (d.f.)	Significance level (p)	Mean difference
A	-2.60	86	0.011	-0.132
B	-3.34	52	0.002	-0.228
C	-6.84	32	<0.001	-0.497
D	-7.94	32	<0.001	-0.588
E	-5.05	74	<0.001	-0.438
F	-9.10	86	<0.001	-0.590
G	-3.98	152	<0.001	-0.167
H	-9.88	58	<0.001	-1.125
J	-8.96	158	<0.001	-0.435
K	-5.42	56	<0.001	-0.364
L	-6.23	46	<0.001	-0.431
M	-0.59	10	0.571	-0.097
N	-13.01	56	<0.001	-0.730
P	-2.34	56	0.023	-0.139
R	-8.46	16	<0.001	-0.664
Mean	-4.88	28	<0.001	-8.441

The result given in Figure 9-3 is similar to that illustrated in Figure 9-1 because both the ratio between lexical and function words and the proportion of lexical words in total word tokens measure informational load, i.e. the extent to which content words are used. This result is in line with Xiao and Yue's (2009) observation of translated Chinese fiction and further confirms Laviosa's (1998b: 8) initial hypothesis that translational language has a relatively lower proportion of lexical words over function words.

On the other hand, as noted in section 2.2, there have been conflicting observations of mean sentence length as a sign of simplification. Figure 9-4 shows the mean sentence length scores of various genres in native and translated Chinese. It can be seen that while native Chinese has a slightly greater mean sentence length, the mean difference between ZCTC and LCMC (–1.533) is not statistically significant (t = –1.41 for 28 d.f., p =

0.17). In both native and translated Chinese, genres such as humour (R) use relatively shorter sentences whereas genres such as academic prose (J) use longer sentences; in some genres there is a sharp contrast between native and translated Chinese (e.g. science fiction M) whereas in other genres the differences are less marked (e.g. academic prose J and press reportage A). It appears, then, that mean sentence length is more sensitive to genre variation than being a reliable indicator of native versus translational language.

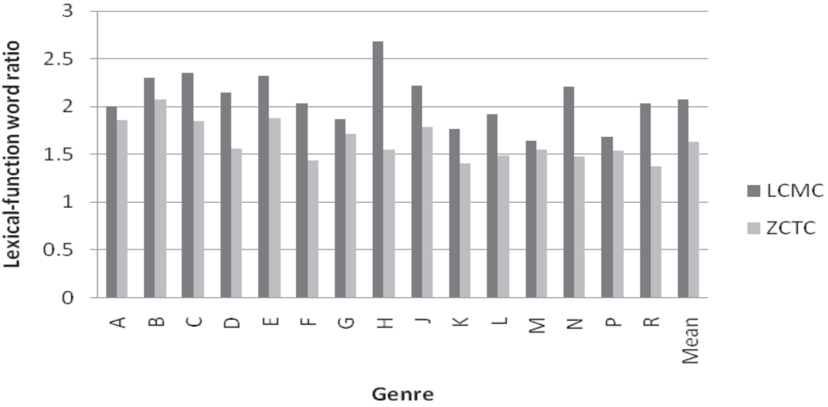


Figure 9-3. Lexical-to-function word ratios in ZCTC and LCMC

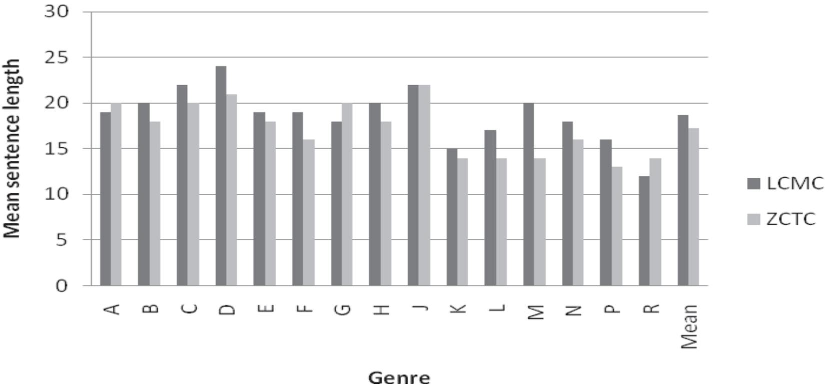


Figure 9-4. Mean sentence length in ZCTC and LCMC

4.2. Frequency profiles

Laviosa (1998b) defines “list head” or “high frequency words” as every item which individually accounts for at least 0.10% of the total occurrences of words in a corpus. In Laviosa’s study, 108 items were high frequency words, most of which were function words. In the present study, we also define high frequency words as those with a minimum proportion of 0.10%. But the numbers of items included can vary depending on the corpus being examined.

Table 9-6 shows the frequency profiles of translated and native Chinese corpora. As can be seen, while the numbers of high frequency words are very similar in the two corpora (114 and 108 respectively), high frequency words account for a considerably greater proportion of tokens in the translational corpus (40.47% in comparison to 35.70% for the native corpus). The ratio between high- and low-frequency words is also greater in translated Chinese (0.6988) than in native Chinese (0.5659). Laviosa (1998b) hypothesizes on the basis of the results of lemmatization that there is less variety in the words that are most frequently used. As Chinese is a non-inflectional language, lemmatization is irrelevant; and as noted earlier, the standardized type-token ratios as a measure of lexical variability are very similar in translated and native Chinese. Nevertheless, it can be seen in Table 9-6 that high frequency words display a much higher repetition rate in translational than native Chinese (3154.37 versus 2870.37).

Table 9-6. Frequency profiles of ZCTC and LCMC

Type	ZCTC	LCMC
No. of high-frequency words	114	108
Cumulative proportion	40.47%	35.70%
Repetition rate of high frequency words	3154.37	2870.37
Ratio of high-to-low frequency words	0.6988	0.5659

The above discussion suggests that the core lexical features proposed by Laviosa (1998b) for translational English are essentially also applicable in translated Chinese, which suggests that translated Chinese also demonstrates a tendency for simplification at lexical level. On the other hand, the mean sentence length is less reliable as an indicator of simplification in translational Chinese.

4.3. Connectives as a device for explicitation

Chen (2006) finds that in his Chinese corpus of popular science books translated from English, connectives are significantly more common than in a comparable corpus of original Chinese scientific writing; some connectives are also found to be translationally distinctive, i.e. significantly more common in translated texts. Chen (2006) concludes that connectives are a device for explicitation in English-to-Chinese translation of popular science books. Xiao and Yue (2009) also note that connectives are significantly more frequent in translated than native Chinese fiction. In this section, we will compare the two balanced corpora of translated and native Chinese in terms of their frequency and use of connectives in an attempt to find out whether the observations by Chen (2006) and Xiao and Yue (2009) can also be generalized from specific genres to Mandarin Chinese in general.

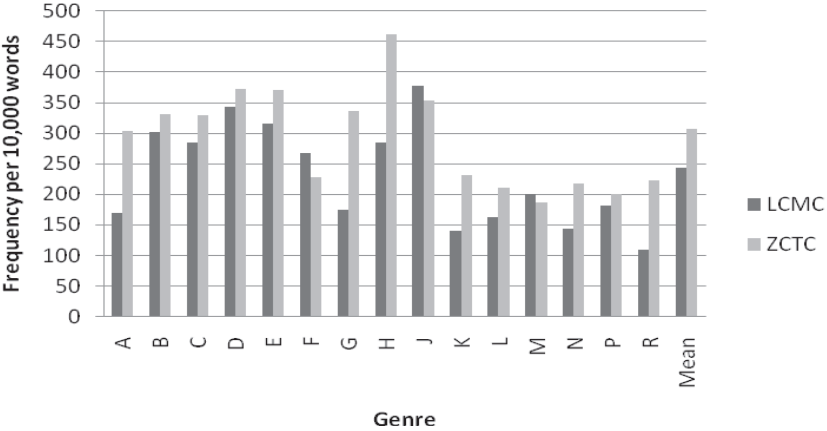


Figure 9-5. Normalized frequencies of conjunctions in ZCTC and LCMC

Figure 9-5 shows the normalized frequencies of conjunctions in the ZCTC and LCMC corpora. As can be seen, the mean frequency of conjunctions is significantly higher in the translational corpus (306.42 instances per 10,000 tokens) than in the native (243.23) corpus (LL=723.12 for 1 d.f., p<0.001). However, a genre-based comparison reveals more subtleties. Genres of imaginative writing (five types of fiction K-P and humour R) generally demonstrate a significantly more frequent use of conjunctions in translational Chinese,[10] a finding which supports Xiao and Yue's (2009) observation of literary translation. On the other hand, while conjunctions are considerably more frequent in most expository genres (categories A-J) in translated Chinese (particularly reports and official documents H and press reportage A), there are also genres in which conjunctions are more common in native Chinese (namely popular lore F and academic prose J).

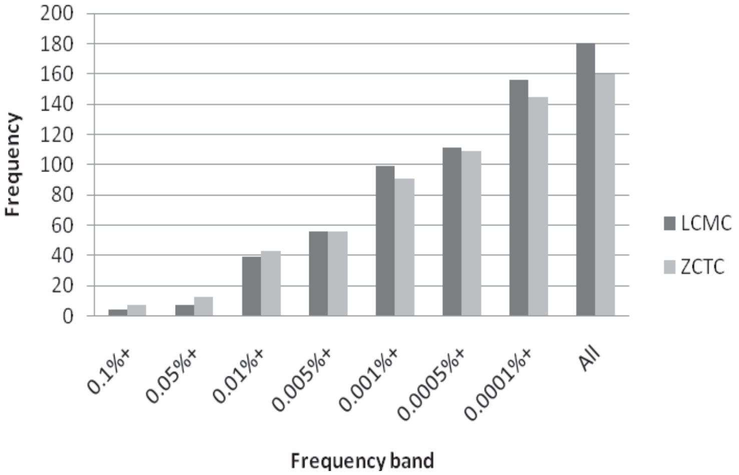


Figure 9-6. Distribution of conjunctions across frequency bands

Xiao and Yue (2009) find that a substantially greater variety of frequent conjunctions are used in translated fiction in comparison with native Chinese fiction. While this finding is supported by ZCTC and LCMC, the two balanced corpora yield even more interesting results. Figure 9-6 compares the frequencies of conjunctions of different frequency bands, as measured in terms of their proportion of the total numbers of tokens in their respective corpus of translational / native Chinese. As can be seen, more types of conjunctions of high frequency bands - i.e. with a proportion greater than 0.10% (7 and 4 types for ZCTC and LCMC respectively), 0.05% (13 and 7 types) and 0.01% (43 and 39 types) - are used in the translational corpus. There are an equal number of conjunctions (56 types) with a proportion greater than 0.005% in translational and native corpora. After this balance point, the native corpus displays a greater number of less frequent conjunctions of the usage band

0.001% and below. This finding further confirms our earlier observation of the use of high and low frequency words in translated Chinese (cf. section 4.2). It also provides evidence that helps to extend the explicitation hypothesis from English to Chinese and to generalize Chen (2006) and Xiao and Yue's (2009) observations from popular science translation and literary translation to the Mandarin language as a whole.

While the tendency to use conjunctions more frequently can be taken as a sign of explicitation, a closer comparison of the lists of conjunctions with a proportion of 0.001% in their respective corpus also sheds some new light on simplification. There are 91 and 99 types of conjunctions of this usage band in ZCTC and LCMC respectively. Of these, 86 items overlap in the two lists. Five conjunctions that appear on the ZCTC list but not on the LCMC list are all informal, colloquial, and simple, i.e. 以至于 'so…that…', 换句话说 'in other words', 虽说 'though', 总的来说 'in short', 一来 'first', which usually have more formal alternatives, e.g. 虽然 for 虽说 'though', and 总之 for 总的来说 'in a word'. In contrast, the 13 conjunctions that appear on the LCMC list but not on the ZCTC list are typically formal and archaic including, for example, 故 'hence', 可见 'it is thus clear', 进而 'and then', 加之 'in addition', 固然 'admittedly', 继而 'afterwards', 非但 'not only', 然 'nevertheless', and 尔后 'thereafter'.

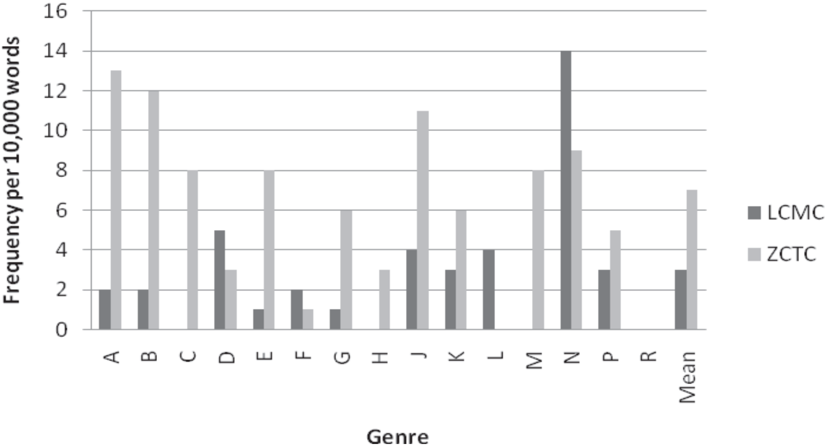


Figure 9-7. Distribution of informal conjunctions

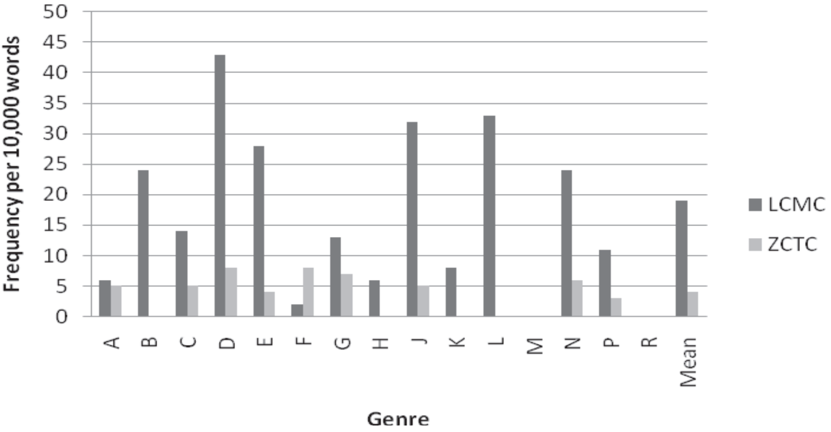


Figure 9-8. Distribution of formal conjunctions

This contrast is illustrated by the patterns of distribution of the two groups of conjunctions across genres. As can be seen in Figures 9-7 and 9-8, formal conjunctions of the second group are more common in nearly all native Chinese genres (barring popular lore F), whereas informal and simple conjunctions of the first group are more frequent in most translational Chinese genres (with the exceptions of religious writing D, mystery and detective stories L, and adventure stories N).[11]

These results appear to suggest that, in spite of some genre-based subtleties, translators tend to use simpler forms than those used in native language, thus providing evidence for the simplification hypothesis but against the normalization hypothesis.

4.4. Passive constructions

This section compares the distribution patterns of passive constructions in translational and native Chinese. Passives are of interest because of their different functions in Chinese and English. Moreover, as noted in section 3.1, the samples in our translational corpus ZCTC are mostly translated from English. In addition to a basic passive meaning, the primary function of passives in English is to mark an impersonal, objective and formal style whereas passives in Chinese are typically a pragmatic voice carrying a negative semantic prosody (Xiao, McEnery and Qian 2006: 143-144). As such, a comparison of the distribution patterns of passives in native and translated language will reveal whether translated Chinese demonstrates a tendency for normalization or whether passives are used in contexts where native Chinese would typically not use passives because of source language shining through.

While passives in Chinese can be marked lexically or syntactically (Xiao, McEnery and Qian 2006), we will only consider the "default" passive form marked by *bei* (被), which is also the most important and frequent type of passive construction in Mandarin.[12] Figure 9-9 shows the normalized frequencies of passives in the fifteen genres as well as their mean frequencies in the ZCTC and LCMC corpora. As indicated by the mean frequencies, passives are more frequent in translational Chinese, and the log-likelihood (LL) test indicates that the difference is statistically significant (LL=69.59 for 1 d.f., p<0.001, see Table 9-7). Since passives are nearly ten times as frequent in English as in Chinese (Xiao, McEnery and Qian 2006: 141-142), it is hardly surprising to find that passives are significantly more common in Chinese texts translated from English in relation to native Chinese texts (see Teich 2003: 196 for similar findings about passives in English-to-German translation), which provides evidence for source language interference (Toury 1995) or shining through (Teich 2001, 2003) but against normalization in translated Chinese in terms of passive use.

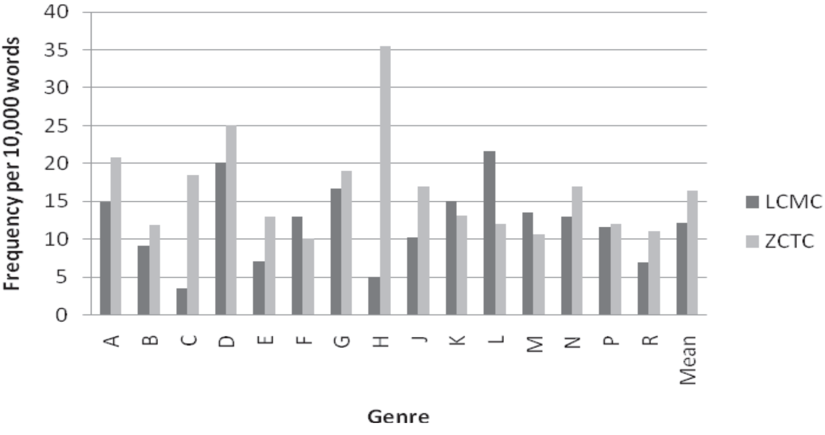


Figure 9-9. Distribution of passives in ZCTC and LCMC

Figure 9-9 also shows that there is considerable variability across genres. Table 9-7 gives the result of log-likelihood (LL) test for difference in each genre, with the significant results highlighted. A combined reading of Figure 9-9 and Table 9-7 reveals that in genres of expository writing such as press reportage (A), press reviews (C), instructional texts such as skills, trade and hobbies (E), reports and official documents (H), and academic prose (J), passives are significantly more frequent in translational Chinese. The contrast is less marked in genres of imaginative writing (K-R). In imaginative writing, significant difference is found only in the genre of mystery and detective fiction (L), where passives are significantly more common in native Chinese. The different distribution patterns of passives in translational and native Chinese provide evidence that translated Chinese is distinct from native Chinese.

Table 9-7. Log-likelihood tests for passives in ZCTC and LCMC

Genre	LL score	Significance level
A	**8.65**	**0.003**
B	1.83	0.176
C	**38.61**	**<0.001**
D	1.93	0.165
E	**13.29**	**<0.001**
F	3.17	0.075
G	2.16	0.142
H	**155.68**	**<0.001**
J	**27.75**	**<0.001**
K	0.88	0.347
L	**13.56**	**<0.001**
M	0.45	0.502
N	3.24	0.072
P	0.06	0.802
R	1.72	0.189
Mean	**69.59**	**<0.001**

Such patterns are closely related to the different functions of passives as noted earlier in Chinese and English, the overwhelmingly dominant source language in our translational corpus (cf. section 3.1). Since mystery and detective fiction is largely concerned with victims who suffer from various kinds of mishaps and the attentions of criminals, it is hardly surprising to find that the inflictive voice is more common in this genre in native Chinese. On the other hand, expository genres like reports and official documents (H), press reviews (C), and academic prose (J), where the most marked contrast is found between translational and native Chinese, are all genres of formal writing that make greater use of passives in English. When texts of such genres are translated into Chinese, passives tend to be overused because of source language interference or shining through (see section 2.4); that is, native speakers of Chinese would not normally use the passive when they express similar ideas. For example, the translated example 该 证书 就 必须 被 颁发 'this certificate then must PASSIVE issue' (ZCTC_H) is clearly a direct translation of the English passive *Then the certificate must be issued*. In such cases, a native speaker of Mandarin is very likely to use the so-called unmarked 'notational passive', i.e. the passive without a passive marker, which is very common in Chinese, as in 该 证书 就 必须 颁发 'this certificate then must issue'. This is clearly a case of source language shining through. It is presently not clear to what extent translated Chinese is affected by the source language in the translation process, which is part of our future investigation based on parallel corpus analysis in our project. However, available evidence of this kind does suggest that normalization may not be a universal feature of translational language (cf. section 2.3).

5. Conclusions

This chapter first provided a review of the state of the art of research in the so-called translation universals, namely the characteristic features of translational language. The limitations of the previous research in this area as revealed in our review led to the discussion of our project which is specifically designed to overcome such limitations. We also presented a new balanced corpus of translational Chinese created on this project which, together with a comparable corpus of native Chinese, provided a quantitative basis for our case studies of some lexical and syntactic features of translational Chinese.

Our case studies have shown that Laviosa's (1998b) observations of the core patterns of lexical features of translational English are supported by our monolingual comparable corpora of translational and native Chinese. Translational Chinese has a significantly lower lexical density (i.e. a proportion of lexical words) than native Chinese, but there is no significant difference in the lexical density as defined by the standardized type-token ratio. In relation to native Chinese, translated Chinese has a relatively lower proportion of lexical words over function words, a higher proportion of high-frequency words over low-frequency words, and a higher repetition rate of high-frequency words. All of these patterns point to the tendency in translated Chinese, as in translational English, for lexical simplification. Beyond the lexical level, our data shows that the mean sentence length is sensitive to genre variation and may not be a reliable indicator of simplification; but a comparison of frequent conjunctions in native and translational Chinese corpora appears to suggest that simpler forms tend to be used in translations. In spite of some genre-based subtleties, translational Chinese also uses conjunctions more frequently than native Chinese, which provides evidence in favour of the explicitation hypothesis. Our analysis of passives in the two corpora provides further evidence supporting the previous finding that translational language is affected by the translation process, though the extent of such influence is yet to be investigated. The source-induced difference between translational and native Chinese in their use of passives also suggests that source language shining through is more remarkable than target language normalization in translational Chinese.

On retrospection, we think that the features of translational Chinese explored in this chapter may not necessarily be translation universals but rather properties which are specific to English-to-Chinese translation due to translation shifts, because English takes up a predominant proportion of source texts in the ZCTC corpus. On the other hand, if more comparable corpora such as LCMC versus ZCTC are created and compared for other languages, especially those from different language families, the third code of genuine translation universals independent of translation pairs will be identified on the basis of corroborative evidence from a wide range of translated languages.

In this sense, we believe that the newly created *ZJU Corpus of Translational Chinese* (ZCTC) will play a leading role in the study of translational Chinese by producing more empirical evidence, and it is our hope that the study of translational Chinese will help to address limitations of imbalance in the current state of translation universal research.

Notes

* We are obliged to the China National Foundation of Social Sciences for supporting our project *A corpus-based quantitative study of translational Chinese in English-Chinese translation* (Grant Reference 07BYY011).

1. We are grateful to an anonymous reviewer who has pointed this out.

2. Chen (2006) compares a corpus of popular science books translated into Chinese from English and the science section of a native Chinese corpus.

3. Readers are reminded of this modification when they interpret the results based on a comparison of the LCMC and ZCTC corpora (see note 11). Those who are

interested in potential change during this decade in Mandarin Chinese are advised to use LCMC in combination with the UCLA Written Chinese Corpus (www.lancs.ac.uk/fass/projects/corpus/UCLA/), which is a native Chinese corpus of the LCMC family created by adopting the same sampling criteria to sample Chinese texts published in the 2000s (the same sampling period of the translational corpus ZCTC) for use in research of recent change in Chinese in the decade when the Internet started to become popular.

[4] The number of tokens given here for the LCMC corpus may be different from earlier releases, because this edition of LCMC has been retagged using ICTCLAS2008, which was used to tag the ZCTC corpus (see section 3.2 and note 6).

[5] See the official website of the ZCTC corpus (www.lancs.ac.uk/fass/projects/corpus/ZCTC/) for the full part-of-speech tagset as applied on the corpus.

[6] See the official website of ICTCLAS (www.ictclas.org) for the history and test results of the software tool. In order to ensure maximum comparability between translated and native Chinese corpora, a new version of the LCMC corpus has also been produced for use on our project, which is retagged using this same tool.

[7] Additional metadata information about the author (or source), the translator, the publisher (or journal and volume number), the year of publication, and the web link if any is given in an Excel spreadsheet accompanying the ZCTC corpus. This information is stored separately rather than in the corpus header so as to make the corpus header of the translational corpus ZCTC comparable with that of the native Chinese corpus LCMC.

[8] See Altova's official website (www.altova.com/xml-editor/) for a description of the tool and its latest update.

[9] In this study, we follow Xiao, Rayson and McEnery (2009) in treating adjectives (including non-predicate noun modifiers and descriptive adjectives), adverbs, nouns, and verbs as lexical words. Function words include the following POS categories: auxiliaries, classifiers, conjunctions, interjections, numerals and quantifiers, onomatopoeias, particles, place words, prefixes, pronouns, prepositions, space words, suffixes, and time words. Unclassified words and symbols and punctuations are excluded in our computations.

[10] Note that the difference in science fiction (M) is not significant (LL=0.641 for 1 d.f., p=0.423).

[11] A reviewer commented that the difference might result from different sampling periods of the translated and native Chinese corpora. This interesting comment led us to investigate frequent conjunctions in the UCLA corpus (see note 3). A comparison of frequent conjunctions in the LCMC, ZCTC and UCLA corpora shows that although ZCTC and UCLA share the same sampling period, frequent conjunctions in the UCLA corpus are still more similar to the native corpus LCMC.

[12] In addition to the default passive marker *bei*, there are a few other passive markers including for example, *gei*, *jiao* and *rang*. However, as passives with these markers typically occur in informal spoken genres while passives are eleven times as common in Chinese writing than in speech (Xiao, McEnery and Qian 2006: 136-137), only the default passives will be considered in this study on the basis of written corpora.

References

Aijmer, K. (2007), "Translating discourse markers: A case of complex translation", in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 95-116. Clevedon: Multilingual Matters.

Baker, M. (1993), "Corpus linguistics and Translation Studies: Implications and applications", in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology. In Honour of John Sinclair*, 233-250. Amsterdam: John Benjamins.

—. (1996), "Corpus-based translation studies: The challenges that lie ahead", in H. Somers (ed.) *Terminology, LSP and Translation Studies in Language Engineering: In Honor of Juan C. Sager*, 175-186. Amsterdam: John Benjamins.

—. (2004), "A corpus-based view of similarity and difference in translation". *International Journal of Corpus Linguistics* 9(2): 167-193.

Biber, D. (1995), *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Blum-Kulka, S. (1986), "Shifts of cohesion and coherence in translation", in J. House and S. Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, 17-35. Tübingen: Gunter Narr.

Blum-Kulka, S. and Levenston, E. (1983), "Universals of lexical simplification", in C. Faerch and G. Kasper (eds.) *Strategies in Interlanguage Communication*, 119-139. London: Longman.

Chen, W. (2006), *Explication Through the Use of Connectives in Translated Chinese: A Corpus-based Study*. PhD thesis, University of Manchester.

Cheong, H. (2006), "Target text contraction in English-into-Korean Translations: A contradiction of presumed translation universals?". *Meta* 51(2): 343-367.

Chesterman, A. (2004), "Beyond the particular", in A. Mauranen and P. Kujamäki (eds.) *Translation Universals: Do They Exist?*, 33-49. Amsterdam: John Benjamins.

Fabricius-Hansen, C. (1999), "Information packaging and translation: Aspects of translational sentence splitting (German-English/Norwegian)". In M. Doherty (ed.) *Sprachspezifische Aspekte der Informationsverteilung*, 175-214. Berlin: Akademie Verlag.

Frawley, W. (1984), "Prolegomenon to a theory of translation", in W. Frawley (ed.) *Translation: Literary, Linguistic and Philosophical Perspectives*, 159-175. London: Associated University Press.

Gellerstam, M. (1996), "Translations as a source for cross-linguistic studies", in K. Aijmer, B. Altenberg and M. Johansson (eds.) *Language in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund, March 1994*, 53-62. Lund: Lund University Press.

Hansen, S. (2003), *The Nature of Translated Text: An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations*. Saarbrücken: DFKI/Universität des Saarlandes.

Hartmann, R. (1985), "Contrastive textology". *Language and Communication* 5: 107-110.

House, J. (2008), "Beyond intervention: Universals in translation?" *Trans-kom* 1(1): 6-19.

Hundt, M., Sand, A. and Siemund, R. (1998), *Manual of Information to Accompany the Freiburg-LOB Corpus of British English*. Freiburg: University of Freiburg.

Hundt, M., Sand, A. and Skandera, P. (1999), *Manual of Information to Accompany the Freiburg-Brown Corpus of American English*. Freiburg: University of Freiburg.

Ide, N. and Priest-Dorman, G. (2000), *Corpus Encoding Standard—Document CES*. Available at: http://www.cs.vassar.edu/CES/ (9 December 2009).

Jantunen, J. (2001), "Synonymity and lexical simplification in translations: A corpus-based approach". *Across Languages and Cultures* 2(1): 97-112.

—. (2004), "Untypical patterns in translations. Issues on corpus methodology and synonymity", in A. Mauranen and P. Kujamäki (eds.) *Translation Universals: Do They Exist?*, 101-126. Amsterdam: John Benjamins.

Kenny, D. (1998), "Creatures of habit? What translators usually do with words". *Meta* 43(4): 515-523.

—. (1999), "The German-English parallel corpus of literary texts (GEPCOLT): A resource for translation scholars". *Teanga* 18: 25-42.

—. (2000), "Translators at play: Exploitations of collocational norms in German-English translation", in B. Dodd (ed.) *Working with German Corpora*, 143-160. Birmingham: University of Birmingham Press.

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

Laviosa, S. (1997), "How comparable can 'comparable corpora' be?". *Target* 9(2): 289-319.

—. (1998a), "The corpus-based approach: A new paradigm in translation studies". *Meta* 43(4): 474-479.

—. (1998b), “Core patterns of lexical use in a comparable corpus of English narrative prose”. *Meta* 43(4): 557-570.

—. (2002), *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam: Rodopi.

Laviosa-Braithwaite, S. (1996), *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. PhD thesis, University of Manchester.

—. (1997), “Investigating simplification in an English comparable corpus of newspaper articles”, in K. Klaudy and J. Kohn (eds.) *Transferre necesse est. Proceedings of the Second International Conference on Current Trends in Studies of Translation and Interpreting*, 531-540. Budapest: Scholastica.

Malmkjær, K. (1997), “Punctuation in Hans Christian Andersen’s stories and their translations into English”, in F. Poyatos (ed.) *Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media*, 151-162. Amsterdam: John Benjamins.

—. (2007), “Norms and nature in translation studies”, in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 49-59. Clevedon: Multilingual Matters.

Mauranen, A. (2000), “Strange strings in translated language: A study on corpora”, in M. Olohan (ed.) *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*, 119-141. Manchester: St. Jerome Publishing.

—. (2007), “Universal tendencies in translation”, in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 32-48. Clevedon: Multilingual Matters.

Mauranen, A. and Kujamäki, P. (2004), *Translation Universals: Do They Exist?* Amsterdam: John Benjamins.

McEnery, T. and Wilson, A. (2001), *Corpus Linguistics* (2^{nd} ed.). Edinburgh: Edinburgh University Press.

McEnery, T. and Xiao, R. (2002), “Domains, text types, aspect marking and English-Chinese translation”. *Languages in Contrast* 2(2): 211-229.

McEnery, T. and Xiao, R. (2004), “The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study”, in M. Lino, M. Xavier, F. Ferreire, R. Costa, R. Silva (eds.) *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*, 1175-1178. Lisbon, 24-30 May 2004.

McEnery, T. and Xiao, R. (2007), “Parallel and comparable corpora: What is happening?”, in M. Rogers and G. Anderman (eds.) *Incorporating*

Corpora. The Linguist and the Translator, 18-31. Clevedon: Multilingual Matters.

McEnery, T., Xiao, R. and Mo, L. (2003), "Aspect marking in English and Chinese". *Literary and Linguistic Computing* 18(4): 361-378.

McEnery, T., Xiao, R. and Tono, Y. (2006), *Corpus-based Language Studies*: *An Advanced Resource Book*. London/New York: Routledge.

Nevalainen, S. (2005), "Köyhtyykö kieli käännettäessä? Mitätaajuuslistat kertovat suomennosten sanastosta", in A. Mauranen and J. Jantunen (eds.) *Käännössuomeksi*, 141- 162. Tampere: Tampere University Press.

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

Olohan, M. and Baker, M. (2000), "Reporting *that* in translated English: Evidence for subconscious processes of explicitation?". *Across Languages and Cultures* 1(2): 141-158.

Øverås, L. (1998), "In search of the third code: An investigation of norms in literary translation". *Meta* 43(4): 557-570.

Pym, A. (2005), "Explaining explicitation", in K. Károly and Á. Fóris (eds.) *New Trends in Translation Studies*, 29-43. Budapest: Akadémiai Kiadó.

Scott, M. (2004), *The WordSmith Tools* (v. 4.0). Oxford: Oxford University Press.

Stubbs, M. (1986), "Lexical density: A computational technique and some findings", in M. Coultard (ed.) *Talking about Text. Studies Presented to David Brazil on His Retirement*, 27-42. Birmingham: English Language Research, University of Birmingham.

—. (1996), *Text and Corpus Analysis. Computer-assisted Studies of Language and Culture*. London: Blackwell

Teich, E. (2001), "Towards a model for the description of cross-linguistic divergence and commonality in translation". In E. Steiner and C. Yallop (eds.) *Exploring Translation and Multilingual Text Production: Beyond Content*, 191-227. Berlin: Mouton de Gruyter.

—. (2003), *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Berlin: Mouton de Gruyter.

Teubert, W. (1996), "Comparable or parallel corpora?". *International Journal of Lexicography* 9(3): 238-64.

Tirkkonen-Condit, S. (2002), "Translationese – A myth or an empirical fact? A study into the linguistic identifiability of translated language". *Target* 14(2): 207-220.

—. (2005), "Do unique items make themselves scarce in translated Finnish?", in K. Károly and Á. Fóris (eds.) *New Trends in Translation Studies. In Honor of Kinga Klaudy*, 177-189. Budapest: Akadémiai Kiadó.

Toury, G. (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 A. Mauranen and P. Kuyamaki (eds.) *Translation Universals: Do They Exist?*, 15-32. Amsterdam: John Benjamins.

Tymoczko, M. (1998), "Computerized corpora and the future of translation studies". *Meta* 43(4): 652-660.

Váradi, T. (2007), "NP modification structures in parallel corpora", in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 168-186. Clevedon: Multilingual Matters.

Xiao, R. (2009), "Multidimensional analysis and the study of world Englishes". *World English* 28(4): 421-450.

Xiao, R., McEnery, T., Baker, P. and Hardie, A. (2004), "Developing Asian language corpora: Standards and practice", in *Proceedings of the 4th Workshop on Asian Language Resources*, 1-8. Sanya, Hainan Island, March 25, 2004.

Xiao, R., McEnery, T. and Qian, Y. (2006), "Passive constructions in English and Chinese: A corpus-based contrastive study". *Languages in Contrast* 6(1): 109-149.

Xiao, R., Rayson, P. and McEnery, A. (2009), *A Frequency of Mandarin Chinese: Core Vocabulary for Learners*. London/New York: Routledge.

Xiao, R. and Yue, M. (2009), "Using corpora in Translation Studies: The state of the art", in P. Baker (ed.) *Contemporary Corpus Linguistics*, 237-262. London: Continuum.

CHAPTER TEN

COMPARING NON-NATIVE AND TRANSLATED LANGUAGE: MONOLINGUAL COMPARABLE CORPORA WITH A TWIST

FEDERICO GASPARI, SILVIA BERNARDINI

1. Introduction and overview

This chapter presents CONTE (the COrpus of Non-native and Translated English), a new type of resource which is being used in an ongoing research project looking into the salient features of two interfacing forms of mediated discourse, namely non-native and translated written language. This work focuses on the language pair Italian-English within the framework of translation universals, and adopts a novel approach hinging on a monolingual comparable perspective which diverges from research paradigms traditionally used in corpus-based translation studies.

After briefly reviewing the theoretical and methodological background to the project as a whole (section 2), the chapter explains the advantages offered by the novel research design adopted, describes the set-up of our first corpus, i.e. CONTE, the English monolingual comparable corpus (MCC) of non-native and translated texts, and provides an overview of the steps involved in designing and creating it (section 3). With the aim of illustrating the kinds of insights that our corpus can provide, both at the descriptive and methodological levels, a case study is presented focusing on the adverbial *therefore* (section 4). We conclude (section 5) by discussing the potential of CONTE as a research resource and outlining some of the applications and future developments planned.

2. Monolingual comparable corpora with a twist

The creation of the corpus that we describe in this chapter was motivated by an effort to approach the debate on translation universals from a new perspective. Research in this field was pioneered in the 1990s by Baker (1993) and Laviosa (1998, 2000, 2003), who produced groundbreaking results extending intuitions which had been gradually taking shape since the previous decade (Blum-Kulka 1986). Within this body of work it was hypothesized, and to some extent shown, that translations (mainly into English) tend to display features such as 'explicitation', 'normalization', 'levelling out', 'simplification', etc. on which a general consensus seems to have emerged within the translation studies community (see e.g. the overview in Laviosa 2002: 43-78).

However, some dissenting voices have been raised questioning the assumptions underlying this research agenda both in terms of objectives and methodology: criticism along these lines comes for example from Bernardini and Zanettin (2004), who focus on the limitations of the corpus-based perspective, and Salsnik (2007) who, following Toury (2004) and Chesterman (2004), takes issue with the misleading application of the notion of 'universal' to translational practice, which necessarily leads to generalizations suffering from weak empirical support (cf. also Malmkjær 2005).

Inspiration for our research project also came from work in the areas of Second Language Acquisition and English as a Second or Other Language: studies such as those presented in Færch and Kasper (1983), Blum-Kulka and Levenston (1983) and House and Blum-Kulka (1986) examine the properties displayed by non-native language production giving prominence to the notion of 'interlanguage' (Selinker 1972), which gained currency in particular throughout the 1970s and 1980s.

Since the 1990s, researchers working within the corpus-based paradigm have revived the interest in interlanguage-related issues, focusing especially on English. In particular, several studies have looked at the patterns of L2 written production by two different categories of language users with a range of mother tongues: learners at different levels of proficiency on the one hand, and people using the L2 for professional purposes on the other. The former investigations are conducted on the basis of learner corpora (e.g. Altenberg and Granger 2001, Nesselhauf 2004), while the latter make use of English as a Lingua Franca (ELF) corpora (e.g. Seidlhofer 2001, Mauranen 2003).

While studies that attempt to bring together these two research areas are few and far between, the hypothesis has been put forward that translation and non-native production may indeed share some common features. As early as the mid-1980s, Blum-Kulka (1986) suggested in a seminal article that explicitation strategies may be used both when translating and when writing a text in a foreign language; more recently Cardinaletti (2005: 60) has compared features of translated language with those typical of “language attrition”, hypothesizing that the source text affects the translator’s target language use as the L1 affects L2 production.

Within the corpus-based paradigm, research into translation universals (inspired by Baker 1993) has focused exclusively on translated texts vs. native speaker usage in the attempt to uncover phenomena that can be interpreted as translation universals. The aim of our project is to extend this paradigm, attempting a systematic search for common traits shared by translation and interlanguage. If such similarities were to be found, we would be in a position to extend Baker’s hypotheses about translation universals to language contact settings in general, and claim that such features are better explained in terms of *mediation* (rather than translation) universals. On the other hand, if no similarities were apparent, Baker’s claim that translation has its own unique properties would be reinforced.

Our method of testing Baker’s hypotheses involves A) comparing translations with original (non-mediated) L2 written production to see if similar features manifest themselves in non-native writing and translations for a given language; and B) considering both directions of a specific language combination (more on the corpus structure in section 3). The methodology and research design employed by Baker and her followers is thus extended by adding the extra dimension of L2 writing to the experimental setup, and refined to focus only on a specific language pair, thus avoiding the bias introduced by direction-specific effects within the language pair in question as well as potential effects due to different source languages.

This research agenda requires appropriate corpus resources, both for English and Italian. In the remainder of the chapter we focus in particular on CONTE, a MCC which consists of non-native and translated English texts. This is the first component of the pool of English / Italian corpus resources that we plan to set up and use in the context of our research into potential mediation universals.

3. The CONTE corpus

3.1. Design and construction

The CONTE corpus is part of a larger and more composite set of MCC resources, including non-native (NN) and translated (TR) texts in English and Italian, alongside (native) benchmark (BM) or reference corpora for the two languages (see Table 10-1; notice that the Italian component of the corpus is currently still under construction and will not be discussed in this chapter).

Table 10-1. Overall corpus structure (Italian MCC component under development)

	CONTE (English)		**CONTI (Italian)**	
	macro-typology	**domain**	**macro-typology**	**domain**
TR	translations from Italian into English	financial reports	translations from English into Italian	TBA
NN	direct written production in English by Italian native speakers	working papers in economics	direct written production in Italian by English native speakers	TBA
BM	written production by native speakers of English	commerce & finance, economics (BNC)	written production by native speakers of Italian	TBA

A number of design considerations guided the planning stages and the initial steps taken in the actual compilation of CONTE. When we started our work, we did not have in mind a specific text type or domain to focus on, as it appeared fairly difficult to identify a priori domain-matched texts in English which were translations from Italian and, on the other hand, others that had been written by native speakers of Italian directly in English.[1] As a result, we carried out extensive explorations of the World Wide Web looking for data that could be used in our research, and eventually decided to focus on texts related to economics, finance and business, which seemed to be easily available in substantial quantities for the non-native and translated English sub-corpora.

However, our extensive searches on the Web showed that no single text typology / genre exists which provides substantial amounts of freely available texts for both the non-native and the translated sub-corpora. Comparability was therefore established at the macro-topic level (economics / finance), and it was decided that the impact of genre / text type similarities and differences (if any) would be evaluated empirically through experience with the corpus. With regard to document availability, texts were sought that were meant to be widely circulated without being subject to particular restrictions in terms of copying, storing and further processing for research purposes, so as to reduce copyright problems. Lastly, we were also keen to avoid the complications and the time investment entailed by the need to scan paper documents, and therefore limited our data search only to online texts that were already available in digital format.

3.2. The non-native (NNENG) sub-corpus within CONTE

The non-native (NNENG) sub-corpus within CONTE features texts downloaded from the RePEc (Research Papers in Economics) online database,[2] a collaborative initiative which provides working papers, journal articles and software tools, currently listing over 500,000 documents and counting over 17,000 registered contributors, which claims to be the largest of its kind available on the Web. This resource seemed particularly attractive for our purposes because all RePEc materials are freely available, and are accompanied by links to the authors' institutional and personal home pages, alongside a list of contact details and information on their affiliation which proved helpful in establishing their language background. Only the working papers made available through the database were included in the corpus, while articles published in academic and scientific journals were disregarded because of likely copyright-related problems.

Working papers seemed particularly promising since they represent repositories of professional and academic writing which the authors themselves voluntarily circulate via semi-formal channels to disseminate their research findings and to encourage feedback and comments from other members of the academic community. As such, working papers would seem to be typically less subject to polishing / editing than published articles, and thus more likely to give us a direct insight into the linguistic habits and strategies of economists and financial experts who are

Italian native speakers writing in English in a professional and academic setting, while probably being less affected by the confounding effects of linguistic editing and revision by native speakers.

Authors in the database writing in English but whose names "sounded" Italian were identified and information relevant for establishing their language background, and in particular their status as native speakers of Italian, was sought on the Web. Very often consulting the authors' online CVs and résumés gave us the information we needed, if their working languages or those with which they were familiar were stated, for example explicitly listing Italian as their mother tongue (possibly along with a good, excellent, etc. knowledge of English and other languages). When the explicit indication of language background was missing, we considered e.g. citizenship, place of birth, institutions where they had completed school and university education, membership of professional bodies and organizations based in Italy, articles and monographs authored in Italian, etc. If a combination of these factors indicated beyond reasonable doubt that the candidate author concerned was a native speaker of Italian, his or her working papers were included in the corpus. Doubtful cases were discarded, as there was no shortage of suitable candidates.

Table 10-2. Details of the NNENG sub-corpus

Number of tokens	3,374,048
Number of types	149,830
Number of texts	410
Average number of tokens per text	8,229
Number of authors	195
Authors' gender	male: 153 female: 42
Authors with more than one text	125
Authors with ≥ 10 texts	3
Publication time span	1991-2008

When selecting the working papers for the corpus we tried to include a good spread of authors (i.e. ensuring diversity in terms of demographic and personal variables such as age, affiliation, level of academic seniority, etc.), although we did not see any problems with including more than one paper by particularly prolific writers. Regrettably, female authors are severely under-represented in the corpus (42 female vs. 153 male authors, or 21.5% of the total number of authors), as a consequence of the relatively low proportion of women writings contained in the database in the first place. All papers in the NNENG sub-corpus have single authors. Further information on the composition of the sub-corpus is provided in Table 10-2.

3.3. The translational (TRENG) sub-corpus within CONTE

Despite our efforts, we were not able to identify on the Web large enough numbers of working papers in economics which had been translated from Italian into English. For the TRENG component we therefore relied on financial statements and reports of well-known companies quoted on the stock exchange that had been translated from Italian into English and posted on the respective websites for consultation by shareholders, investors and other interested parties. These translations had one common feature: they all carried explicit notices warning their readers of their status as translations, and of the supremacy of the Italian source text for legal purposes.

Unfortunately, we were not able to document the translators involved in producing these English translations (it is quite possible that teams of more than one translator were employed for each translation, given the length of each document – see Table 10-3 for more details), as their identities are not disclosed, although in a few cases the name and contact details of the agency which took care of the translation project are provided. Ideally, we would have wanted to make sure that the translations into English had been done by native speakers of the target language, but regrettably no information is available on the texts which can help us to establish the identities or the language profiles of the translators. In an attempt to ensure as much variety as possible in our TRENG data, we selected no more than one financial statement or report per company, and as a result it is likely that these documents were translated by different (teams of) translators.

Table 10-3. Details of the TRENG sub-corpus

Size (number of tokens)	2,205,361
Number of types	86,027
Number of texts	39
Average number of tokens per text	56,547
Publication time span	2000-2007

Table 10-3 provides some details of the TRENG sub-corpus, showing that this component contains fewer texts (all substantially longer) than the NNENG counterpart, and a lower number of words.

3.4. Sampling strategy and data integrity

As far as the sampling strategy is concerned, for both the NNENG and the TRENG sub-corpus we opted for whole texts, following Sinclair's (1991: 19) suggestion that whole-text corpora are "open to a wider range of linguistic studies than a collection of short samples", and the widespread practice within corpus-based translation studies (e.g. Kenny 2001, Laviosa 1998). Omissions or deletions were avoided, as they would have been time-consuming, and might have introduced biases and potential inconsistencies. In particular, we considered and then dismissed the idea of expunging information in Italian, in-text verbatim quotations (in English) and material in the references / bibliography sections. While these might potentially represent sources of interference confounding the variables that we intended to investigate, the integrity of the texts under investigation was considered to be a priority; the filtering out of regularities found in parts of the texts not written by their main authors was therefore left for the corpus analysis rather than construction stage.

3.5. Benchmarking: the reference corpus

In order to support our investigations based on CONTE, we needed a reference corpus of native non-translated English for benchmarking purposes. For practical reasons, we decided to use a sub-corpus of the British National Corpus (World Edition) (Burnard 2007), selected so as to match as closely as possible the contents of the NNENG and TRENG sub-corpora. The 90-million-word written part of the BNC was designed according to two main concurrent criteria applied to the relevant texts, namely "domain" and "medium" (Aston 2001: 73). Domain roughly corresponds to subject matter (imaginative, arts, belief and thought, commerce and finance, and so forth). "Medium", on the other hand, covers five classes, i.e. book, periodical, miscellaneous published, miscellaneous unpublished, and to-be-spoken. Recognizing that corpus users may need finer descriptive categories, Lee (2001) provided an alternative arrangement using a more delicate categorization scheme identifying 46 "genres" for the written data. Following Lee's categories as presented in his "BNC Index",[3] we extracted the 112 texts that he grouped under the "W_commerce (commerce & finance, economics)" genre category as our reference corpus of native (British) English (BMENG, see Table 10-4 for more details).

Table 10-4. Details of the BMENG corpus (BNC commerce)

Number of tokens	3,759,366
Number of types	60,651
Number of texts	112
Authors’ gender	male: 54 female: 3 mixed: 6 unknown or n/a: 49
Average tokens per text	33,565
Publication time span	1985-1994

3.6. Corpus preparation[4]

After conversion of the original pdf files to plain text format and simple cleaning procedures through batch substitutions of non-alphabetic characters,[5] the texts were POS-tagged and lemmatized using the TreeTagger (Schmid 1994). Minimal metadata were then added to the texts (as attribute-value pairs in the “text” element preceding each text in the corpus), if these were judged to be potentially useful for on-the-fly sub-corpus selection. Thus, a typical text element in the TRENG sub-corpus contains the following information:

- id=“TRENG612”
- year=“2005”

This allows users to select for searching individual texts or texts published in/before/after a given year. The NNENG sub-corpus, on the other hand, contains slightly more information, i.e.:

- id=“NNENG018”
- year=“2006”
- author=“Antonio_Abatemarco”
- gender=“Male”

Thus, the non-native corpus also allows one to select or exclude from a search texts written by a given author or by males / females. Further metadata about the texts (e.g., source information) are available from a separate database. The corpus was then indexed with the CorpusWorkBench (Christ 1994), and made searchable with the associated Corpus Query Processor. At the time of writing, the corpus is available for searching via a remote Unix command line shell, though we hope to be able to make it available to the general public through a Web interface in the future.

4. Preliminary investigation: the adverbial *therefore*

Besides starting to shed new light on the hypothesized common ground between translated and non-native language (our long-term objective), initial investigations conducted on mediated English with CONTE have a methodological purpose. Given that corpus comparability is a tricky notion (Kilgarriff 2001), especially with reference to translated language (Bernardini and Zanettin 2004), we believe it safer not to assume comparability “by design”, i.e. based on external criteria, between our two corpora and the reference corpus we are currently employing (BNC commerce, section 3.5). Instead, we hope that, by accumulating results and constantly evaluating them, we can develop a better idea of the ways in which the corpora we are comparing resemble or differ from each other, and ultimately assemble data whose internal consistency or lack thereof may also tell us whether the comparability assumption is justified or not. At this stage, we take a largely serendipitous approach and look for broad trends rather than definitive evidence based on firm statistical grounds, which we leave for future more in-depth investigations.

As a first case study, we focused on the resultative adverbial *therefore*, one of several linking adverbials often used in written, especially academic, discourse “to signpost the logical and argumentative links between one part of the discourse and another” (Biber *et al*. 1999: 1046). We hypothesized this adverbial to be potentially overrepresented in our mediated corpora with respect to native English, as a consequence of either explicitation (Blum-Kulka 1986) or risk-avoidance (Pym 2008). A simple search for the lemma *therefore* in the three corpora shows that the hypothesis is not supported: the normalized frequency of the adverbial in the reference corpus is intermediate between the value found in the translated corpus and the one found in the non-native corpus, as shown in Table 10-5.

This finding is somewhat unsurprising, if one considers that the BNC sub-corpus employed for benchmarking purposes contains several different text types, including but not limited to academic prose, and that this adverbial is particularly frequent in academic language (Biber *et al*.

1999: 887). Thus, the much higher frequency observed in the non-native (academic) sub-corpus and the lower frequency in the translated corpus of financial statements / reports could be due to a text-typological difference unrelated to the mediation dimension, and the data at our disposal do not allow us to rule out this possibility.

Table 10-5. Frequency of *therefore* in CONTE

	BNC commerce		TRENG		NNENG	
	n.	n./M words	n.	n./M words	n.	n./M words
therefore	2,397	637.6	562	254.8	3,430	1,016.5
corpus size	3,759,366		2,205,361		3,374,048	

However, we can search for frequency data about patterns around *therefore*, and see whether the proportion of certain patterns to the total differs in the three corpora, and in particular whether translated and non-native English texts are more akin to each other than to original English. Here we focus on two patterns, namely "[punctuation mark] + therefore" and "[verb] + therefore". We can expect a search for the first pattern to return sentence / clause initial *therefore* and (mainly) medial (post-subject or post-verbal) *therefore*, as shown by the concordance in Figure 10-1:

Figure 10-1. Concordance of *therefore* preceded by a punctuation mark

```
ed out during the year <. Therefore> the actual saving rate    [NNENG]
te saving data are not <, therefore> , sufficient evidence to  [NNENG]
fected by the mutation <; therefore> by comparing the differ   [NNENG]
o realize new projects <: therefore> the problem is how the    [NNENG]
ted inflation of 1.9 % <. Therefore> the real rate i equall    [TRENG]
quarters . Performance <, therefore> , mirrored government     [TRENG]
enefits were suspended <; therefore> , no benefit has been     [TRENG]
tive figures presented <: therefore> , the comparative bala    [TRENG]
 clearing house system <. Therefore> all banks dealing in      [BNC Commerce]
ations . The telephone <, therefore> , saves time and gives    [BNC Commerce]
 or to create goodwill <; therefore> every letter should co    [BNC Commerce]
e demand at that price <: therefore> even those of them who    [BNC Commerce]
```

Quantitative data about the frequency of *therefore* in initial position (i.e. after a full stop, a colon or semi-colon) is shown in Table 10-6, while Table 10-7 gives results for the medial (i.e. post-comma) position. For each corpus, the third column of each table gives normalized frequency data per million words as a percentage of the total number of occurrences of the lemma *therefore*. As can be seen, the data for the two mediated corpora show similar trends, i.e., *therefore* used in initial position (Table

10-6) tends to be proportionally more frequent in mediated than non-mediated language, while medial *therefore* (Table 10-7) is proportionally slightly more frequent in non-mediated language.[6] Since the former is considered to be the unmarked position for linking adverbials (Biber *et al.* 1999: 891),[7] this observation might be explainable with reference to the normalization or (in Toury's words) "growing standardization" hypothesis, according to which "[in translation], textual relations obtaining in the original are often modified […] in favour of [more] habitual options offered by a target repertoire" (Toury 1995: 268). In other words, this might be an instance of normalization applying to both translated and non-native texts.

Table 10-6. The adverbial *therefore* in initial position

	BNC commerce			TRENG			NNENG		
	n.	n./M words	%	n.	n./M words	%	n.	n./M words	%
therefore	2,397	637.6	100	562	254.8	100	3,430	1,016.5	100
. therefore	314	83.5	**13.0**	98	44.4	**17.4**	1,141	338.1	**33.2**
; therefore	23	6.1	**0.9**	7	3.1	**1.2**	95	28.1	**2.7**
: therefore	7	1.8	**0.2**	2	0.9	**0.3**	29	8.5	**0.8**
corpus size	3,759,366			2,205,361			3,374,048		

Table 10-7. The adverbial *therefore* in medial position

	BNC commerce			TRENG			NNENG		
	n.	n./M words	%	n.	n./M words	%	n.	n./M words	%
therefore	2,397	637.6	100	562	254.8	100	3,430	1,016.5	100
, therefore	506	134.5	**21.0**	112	50.7	**19.8**	464	137.5	**13.5**
corpus size	3,759,366			2,205,361			3,374,048		

Focusing on the second pattern (a verb followed by *therefore*), translated and non-native texts are characterized by lower percentages if compared to the reference corpus (again, out of the total number of occurrences of the adverbial per million words, see Table 10-8). While there are 225 occurrences per million words of this pattern in the reference corpus, corresponding to 35.2% of the total occurrences of *therefore*, the percentage is lower in the translated sub-corpus (30.2%, or 77 occurrences per million words) and lower still for the non-native sub-corpus (17.2% or 175.4 occurrences per million words; remember that the adverbial is extremely frequent in the non-native corpus).

Table 10-8. "Verb + *therefore*" as a percentage of total occurrences of *therefore*

	BNC commerce			**TRENG**			**NNENG**		
	n.	n./M words	%	n.	n./M words	%	n.	n./M words	%
therefore	2,397	637.6	100	562	254.8	100	3,430	1,016.5	100
[verb] therefore	846	225.0	**35.2**	170	77.0	**30.2**	592	175.4	**17.2**
corpus size	3,759,366			2,205,361			3,374,048		

Table 10-9. "Verb + *therefore*" as a percentage of total number of verbs

	BNC commerce			**TRENG**			**NNENG**		
	n.	n./M words	%	n.	n./M words	%	n.	n./M words	%
total verbs	649,485	172,764.5	100	226,067	102,507.9	100	581,438	172,326.5	100
[verb] therefore	846	225.0	**0.13**	170	77.0	**0.07**	592	175.4	**0.10**
corpus size	3,759,366			2,205,361			3,374,048		

Furthermore, if we observe the (normalized) frequency of *therefore* immediately following a verb as a percentage of the corresponding normalized frequency of verbs in the corpus – rather than in comparison with the total occurrences of the adverbial, as given in Table 10-8 – we find that both mediated corpora have consistently lower values (see Table 10-9). In other words, the (slightly) higher frequency of post-verbal *therefore* in the reference corpus is not an effect of differences in verb frequency with respect to the two mediated corpora.[8]

5. Conclusion and further work

In this chapter we have presented a corpus-based research project whose long-term objective is the exploration of the hypothesis that non-native and translated language share similar features, and that these can be accounted for by the notion of mediation (rather than translation) universals. As a first step in this direction, the chapter has described CONTE, the monolingual English component of the corpus we are building, and presented a small-scale case study to illustrate the kinds of analyses for which it can be used, focusing on the behaviour of the resultative adverbial *therefore* in non-native and translated texts. For benchmarking purposes, we used a comparable corpus of native English derived from the BNC.

While this approach can yield promising results, and provide data to ascertain the validity of the corpus empirically, through accumulation and evaluation of results (Atkins *et al*. 1992, Hunston 2002: 28-30), we would like to compile reference corpora to use as benchmarks that are more closely comparable by design to the corpora under investigation.

We realize the limitations involved in using a sub-set of the BNC selected on the basis of the categories identified by Lee (2001) as a source of original native writing for benchmarking purposes in our study. Although inspection of this dataset confirmed that the texts contained therein do indeed belong to the commercial, financial and economics domains, they represent a very diverse array of genres. As explained in sections 3.2 and 3.3, the NNENG and TRENG components of CONTE, on the other hand, consist exclusively of research working papers in economics and financial reports, respectively; this text typological discrepancy between our mediated corpora and the original native English corpus is likely to have an adverse impact on data comparability and most likely on the reliability of our analyses.

We intend to overcome these problems by refining our corpus design in the near future, replacing the data derived from the BNC according to

Lee's categorization with more closely genre-matched and domain-controlled native corpora to be used for more rigorous benchmarking. Work is already underway to compile two separate tailor-made sub-corpora of original working papers in economics and financial reports written by native speakers of English. These are bound to be more closely comparable to our NNENG and TRENG data, insofar as they guarantee (pair-wise) homogeneity in terms of text type, enabling us to neutralize the confounding variables introduced by genre variability in the current set-up. Further analyses will thus be carried out utilizing these new ad-hoc corpus components instead of the BNC sub-corpus, and it will be interesting to see how this different corpus set-up will affect our findings. In particular, we expect these more fine-tuned native corpora to make patterns stand out more clearly and to make searches less labour-intensive.

After analyzing a wider range of phenomena for mediated English with CONTE, of which the case study presented in this chapter represents one example in terms of methodology and approach to the investigation, we intend to move on to explore if any comparable tendencies can be observed in the opposite language direction. The next step in our work is therefore going to be an investigation of similar phenomena with implications for the concept of translation (and mediation) universals for mediated (non-native and translated) Italian. We intend to replicate the corpus architecture described in section 3, using a MCC of Italian made up of the following three components: (i) translations from English, (ii) non-native texts written by authors with English as their mother tongue, accompanied by (iii) a suitable reference corpus for benchmarking purposes. For our preliminary investigations we plan to use a sub-corpus of the "La Repubblica" corpus (Baroni *et al.* 2004), subsequently developing, if necessary, fine-tuned reference corpora for non-native and translated production respectively.

We are currently surveying which texts are available in Italian that could offer a good level of comparability. The phenomena that one might investigate for Italian in an attempt to shed light on the 'mediation universals' hypothesis include the treatment of subject pronouns (differently from English, Italian is a pro-drop language; research in this area has been recently conducted from the points of view of developmental linguistics and Second Language Acquisition, see e.g. Serratrice 2005, 2007), the distribution of past-tense verbs (in particular the imperfect vs. the present perfect in the indicative mood), the use of definite articles and the pre- vs. post-noun positioning of attributive adjectives.

One extension that we are considering for our research would be to carry out broader analyses encompassing a parallel corpus component (clearly, this would be relevant only for the translated, not the non-native sub-component, for which no parallel texts exist). Although so far we have not taken this dimension into account, favouring a monolingual comparable approach, in the process of data collection for the translated component of CONTE we have also paid attention to source texts in Italian: whenever we were able to locate them, we kept a copy for future reference. Since the TRENG texts are translations into English of financial statements and reports of high-profile companies, parallel source texts in Italian were found and collected, though not processed at this stage. As a result, the option of using parallel data in more detailed investigations in the future is still open, particularly with a view to evaluating the impact of source-text effects on the translations.

Furthermore, for our research project we are currently restricting our focus to the English-Italian language pair in both directions, which should serve as a pilot investigation to establish the potential of our methodology. Looking at other language combinations in the future would clearly be essential to build a more accurate and comprehensive picture of mediation-related phenomena. This would involve creating corpora for other languages with structures similar to the one described in section 3, in order to check whether findings are consistent across different language pairs and if general patterns emerge.

As we have attempted to show in this chapter, the specialised English MCC which we have presented is a flexible research resource that can be deployed in a number of investigations adopting a variety of methodological set-ups and analytical approaches to uncover the features of mediated language. In the longer term, possible applications comprise comparing the language patterns typical of translated and non-native English in CONTE for certain phenomena against the patterns emerging in corpora of (spoken) English as a Lingua Franca (Seidlhofer 2001, Mauranen 2003) and learner English (Granger 2003), so as to deepen our understanding of (the similarities and differences between) different forms of language mediation.

Notes

* We would like to thank our colleague Adriano Ferraresi for very helpful comments on an earlier draft of this chapter.

1. We recognize that the strict differentiation between the notions of "native" and "non-native" speaker is an idealization, and we accept that the validity of these notions as rigorous descriptive categories is questioned in linguistics and

translation studies. However, due to space constraints and given the variables that we aim to isolate, in this chapter we have to rely on a broadly accepted intuitive understanding of these notions.

[2] Available at http://ideas.repec.org/i/eall.html [last accessed 28 July 2009].

[3] Available at http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/home/corpus_resources.htm [last accessed 28 July 2009].

[4] This section deals only with the preparation of the non-native and translational components of the English MCC, since the benchmark corpus was already available in a format adequate for the project.

[5] The data cleaning process is documented in the background information that comes with CONTE, so that its users are made aware of the (slight) interventions on the raw data.

[6] Notice that the percentages of initial *therefore* differ substantially between translated and non-native English (17.4 vs. 33.2% for occurrences following a full stop), possibly as a result of the text typological differences discussed above.

[7] In the case of *therefore*, at least in the BNC commerce sub-corpus, the unmarked position is not in fact the most frequent.

[8] Incidentally, this case study on *therefore* has also pointed at a general tendency displayed by both mediated corpora to make lighter use of modal verbs than the reference corpus, an intriguing finding deserving further study.

References

Altenberg, B. and Granger, S. (2001), "The grammatical and lexical patterning of MAKE in native and non-native student writing". *Applied Linguistics* 22(2): 173-194.

Aston, G. (2001), "Text categories and corpus users: A response to David Lee". *Language Learning & Technology* 5(3): 73-76.

Atkins, S., Clear, J. and Ostler, N. (1992), "Corpus design criteria". *Literary and Linguistic Computing* 7(1): 1-16.

Baker, M. (1993), "Corpus linguistics and translation studies: Implications and applications", in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*, 233-250. Amsterdam: John Benjamins.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. and Mazzoleni, M. (2004), "Introducing the *La Repubblica* corpus: a large, annotated, TEI(XML)-compliant corpus of newspaper Italian", in *Proceedings of LREC 2004*, 1771-1774. Lisbon: ELDA.

Bernardini, S. and Zanettin, F. (2004), "When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals", in A. Mauranen and P. Kujamäki (eds.) *Translation Universals: Do They Exist?*, 51-62. Amsterdam: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999), *Longman Grammar of Spoken and Written English*. London: Longman.

Blum-Kulka, S. (1986), "Shifts of cohesion and coherence in translation", in J. House and S. Blum-Kulka (eds.) *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*, 17-35. Tübingen: Gunter Narr.

Blum-Kulka, S. and Levenston, E. (1983), "Universals of lexical simplification", in C. Færch and G. Kasper (eds.) *Strategies in Interlanguage Communication*, 119-139. London: Longman.

Burnard, L. (2007), "Users' reference guide to the British National Corpus (XML edition)". Oxford: Oxford University Computing Services. Available online at http://www.natcorp.ox.ac.uk/XMLedition/URG/ [last accessed 28 July 2009].

Cardinaletti, A. (2005), "La traduzione: un caso di attrito linguistico", in A. Cardinaletti and G. Garzone (eds.) *L'Italiano delle Traduzioni*, 59-83. Milano: Franco Angeli.

Chesterman, A. (2004), "Hypotheses about translation universals", in G. Hansen, K. Malmkjær and D. Gile (eds.) *Claims, Changes and Challenges in Translation Studies*, 1-13. Amsterdam: John Benjamins.

Christ, O. (1994), "A modular and flexible architecture for an integrated corpus query system", in *Proceedings of COMPLEX'94*. Budapest, 1994. Available online at
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench
[last accessed 28 July 2009].

Færch, C. and Kasper, G. (eds.) (1983), *Strategies in Interlanguage Communication*. London: Longman.

Granger, S. (2003), "The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research". *TESOL Quarterly* 37(3): 538-546.

House, J. and Blum-Kulka, S. (eds.) (1986), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies*. Tübingen: Gunter Narr.

Hunston, S. (2002), *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

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

Kilgarriff, A. (2001), "Comparing corpora". *International Journal of Corpus Linguistics* 6(1): 97-133.

Laviosa, S. (1998), "The English Comparable Corpus: A resource and a methodology", in L. Bowker, M. Cronin, D. Kenny and J. Pearson (eds.) *Unity in Diversity? Current Trends in Translation Studies*, 101-112. Manchester: St. Jerome.

—. (2000), “TEC: A resource for studying what is ‘in’ and ‘of’ translational English”. *Across Languages and Cultures* 1(2): 159-178.

—. (2002), *Corpus-based Translation Studies. Theory, Findings, Applications.* Amsterdam: Rodopi.

—. (2003), “Corpus and simplification in translation”, in S. Petrilli (ed.) *Translation Translation*, 153-162. Amsterdam: Rodopi.

Lee, D. (2001), “Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle”. *Language Learning & Technology* 5(3): 37-72.

Malmkjær, K. (2005), “Norms and nature in translation studies”. *Synaps: Fagspråk, Kommuniksjon, Kulturkunnscap*. Bergen: Norges Handelshøyskole. 16: 13-19.

Mauranen, A. (2003), “The corpus of English as lingua franca in academic settings”. *TESOL Quarterly* 37(3): 513–527.

Nesselhauf, N. (2004), “Learner corpora and their potential for language teaching”, in J. McH. Sinclair (ed.) *How to Use Corpora in Language Teaching*, 125-152. Amsterdam: John Benjamins.

Pym, A. (2008), “On Toury’s laws of how translators translate”, in A. Pym, M. Shlesinger and D. Simeoni (eds.) *Beyond Descriptive Translation Studies. Investigations in Homage to Gideon Toury*, 311-328. Amsterdam: John Benjamins.

Salsnik, E. (2007), “Dagli universali traduttivi all’italiano delle traduzioni”, in C. Montella and G. Marchesini (eds.) *I Saperi del Tradurre*, 101-131. Milano: Franco Angeli.

Schmid, H. (1994), “Probabilistic part-of-speech tagging using decision trees”, in *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, 14-16 September 1994.

Seidlhofer, B. (2001), “Closing a conceptual gap: The case for a description of English as a lingua franca”. *International Journal of Applied Linguistics* 11(2): 133-158.

Selinker, L. (1972), “Interlanguage”. *International Review of Applied Linguistics* 10(3): 209-231.

Serratrice, L. (2005), “The role of discourse pragmatics in the acquisition of subjects in Italian”. *Applied Psycholinguistics* 26(3): 437-462.

—. (2007), “Referential cohesion in the narratives of bilingual English-Italian children and monolingual peers”. *Journal of Pragmatics* 39(6): 1058-1087.

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

Toury, G. (1995), *Translation Studies and Beyond.* Amsterdam: John Benjamins.

—. (2004), "Probabilistic explanations in translation studies: Welcome as they are, would they qualify as universals?", in A. Mauranen and P. Kujamäki (eds.) *Translation Universals: Do They Exist?*, 15-32. Amsterdam: John Benjamins.

CHAPTER ELEVEN

SENSE-MAKING IN CORPUS-ASSISTED TRANSLATION RESEARCH: A REVIEW OF CORPUS-ASSISTED TRANSLATION RESEARCH IN CHINA

DEFENG LI, CHUNLING ZHANG

1. Introduction

Baker's (1993) seminal article on applying corpus technology in translation research has been influential, thanks to the rise of the descriptive translation studies (Toury 1995). An increasing number of studies have since been carried out to examine translation per se both as a process as well as a product (Kenny 2001, Laviosa 2002, Granger *et al.* 2003, Mauranen and Kujamäki 2004, Hansen *et al.* 2004, Olohan 2004, Anderman and Rogers 2007, Williams 2007). Some studies focused on translation universals (Baker 1993), others examined translator's styles (Baker 2000; Bosseaux 2001, 2006; Li and Liu 2007; Winters 2004, 2007). Languages that have been studied include English, German, French and so on. It is against such a backdrop of avid attention in the West that translation researchers in the Chinese Mainland (China hereafter) have developed a keen interest in corpus-assisted translation research. Articles and monographs have been published and conferences devoted to this topic. The last couple of years have also seen production of several doctoral theses and hundreds of MA dissertations integrating corpus technology with translation studies. While an increasing number of researchers and graduate students are poised for more and further corpus-assisted translation research projects in China, a critical review of the related research seems to be warranted in order to find out what progress has been made over the past 15 years and what problems, if any, such research is facing.

More specifically, the purpose of such an examination is: to identify major translational issues that have – and have not – been dealt with in corpus-assisted translation research in the Chinese context; to examine the research methods and designs of these studies; and to suggest future directions for corpus-assisted research in China. It is also hoped that some of the pointers for future investigations, though based on research conducted in China, may also have wider implications for corpus-assisted translation research in general.

2. Collection of literature on corpus-assisted translation research in China

In order to gather together literature on corpus-assisted translation research in China, the present researchers made use of the China Journal Net (CJN), a searchable electronic database of academic articles published in Chinese journals of humanities and social sciences as well as sciences and technology.[1] Searches with keywords "语料库" (corpus), "翻译" (translation), "翻译共性" (translation norms), "翻译普遍性" (translation universals) and "译者风格" (translator's style) were conducted to identify journal articles on this topic. After weeding out articles which did not fall into the category of corpus-assisted translation research, and supplementing the list with numerous publications that the present researchers were aware of but were not included in the CJN database due to errors in the system, a list of journal articles was finally produced as the dataset for the present review. One edited volume and one PhD thesis were also found on corpus-assisted translation research. However, they were not included in this survey as much of the two books had already been published in the form of articles in academic journals, hence actually included in the data set of the present study. Although no claim will be made to have exhausted the literature on this topic, we can be fairly certain that very few related articles, if any, were missed out.

Three steps were taken in the analysis of the data. First, the articles, particularly the titles and the abstracts were scanned to identify the research issues that they covered. Second, the titles and the abstracts were reviewed again and sorted out according to the issues they each focused on. Third, one particular group of the data-based research articles were scrutinized in order to examine the research methods, particularly the research designs, and the new contributions of each study.

3. Findings

The CJN keyword search produced 94 journal articles on corpus related translation research published in journals between 1995 and 2008. They were then put into seven categories according to their foci (see Table 11-1). The largest proportion of articles are general surveys and reviews introducing corpus-assisted translation research that has been conducted in the West (36.2%), followed by articles on the use of corpora in translation teaching (17%) and then those on corpus design and methodological considerations in corpus-assisted translation research (16%). Data-based studies of specific translational issues account for a little more than 10%, and the use of corpora in dictionary making follows closely behind (9.6%).

Table 11-1. Types of articles on corpus-assisted translation research

Types of Article	No.	Percentage
General surveys and reviews of the use of corpora in translation research	34	36.2
Corpora in translation teaching	16	17.0
Corpus design and methodological considerations	15	16.0
Data-based studies of specific translational issues	11	11.6
Corpora in dictionary making	9	9.6
Corpora in machine translation	6	6.4
Book reviews	3	3.2
Total	94	100

3.1. General surveys and reviews

As seen in Table 11-1, over one third of the articles are general surveys and / or review articles on the use of corpora in translation studies. They are mostly summaries of corpus-assisted translation research carried out outside China. Their purpose is to introduce to the Chinese translation research community:

- use of corpora in translation research (e.g. Liao 2000, Huang and Fan 2004, Wang 2006, Zhao 2008);
- translational issues that have been studied, for instance, translation universals (e.g. Hu 2005, Wu and Huang 2006), translators' style, (e.g. Zhang 2002), translation criticism (e.g.

Xiao 2005, Wang 2007);
- a particular well-known corpus, for instance the TEC (e.g. Chen 2007)

It is perhaps worth noting that these 34 articles have considerable overlapping in their nature, scope, purpose, focus and contents and that much of the so-called research has been repeating itself. It is surprising to find that ten years after the publication of the first introductory articles on this topic, researchers are still writing more or less the same articles. For the sake of comparison, tabulated in Table 11-2 are abstracts of some articles published between 2000 and 2008. As can be seen, both the titles and the abstracts reveal the extent to which these articles are similar in scope and contents.

Given the fact that these articles were published over a period of eight years, one has to wonder why so many articles were devoted to repeatedly introducing corpus-assisted translation studies. Unless there was some strong resistance in China to this new field of research, which has so far been unheard of, such efforts seemed uncalled for.

Table 11-2. Titles and abstracts of some articles published between 2000 and 2008[2]

Title, name of journal, year of publication	Abstract
Brief survey of corpus-based translation studies abroad (Zhao 2008) Journal of Chongqing Jiaotong University (Social Sciences Edition)	Since the beginning of the 1990s, the study of translation through corpora has given rise to fully-fledged paradigm within the discipline of translation studies. It has been discovered that corpus plays a more and more important role in translation studies and translator training. It is attempted to give a general review of Corpus-based translation studies and analyze its theoretical and practical values, including translator training and machine translation so that Chinese readers can get a full view of this newly emerging academic field and broaden our horizons of translation studies in China.
TEC and TEC-based descriptive translation studies (Chen 2007)	The exploration of universal features of translation with the empirical study of translation process, think-aloud protocol,

Journal of Foreign Languages	Translog and the combination of induction and deduction has established a methodological foundation for corpus-based translation studies. The emergence of Translational English Corpus (TEC) was triggered by the incentive suggestions of descriptive translation studies by polysystems theorists and by Mona Maker’s systemic study of universal features of translation. This paper discusses the compilation of TEC and its major features, reviews TEC-based descriptive translation studies and proposes the potential ways of exploiting this corpus resource.
Corpus-based studies on universals of translation (Hu 2005) Journal of PLA University of Foreign Languages	Universals of translation, which are represented by simplification, explicitation, normalization, convergence and leveling out, etc, have been the most successful realm of corpus-based translation studies. This article introduces recent studies on universals of translation, and comments on their merits and limitations.
Study of western Translational English Corpus (Ding 2001) Journal of Foreign Languages	The new corpus resources have already had a profound effect on the development of the various areas of the linguistic research, particularly through the wide and firm empirical basis they provide for verifying hypotheses and for formulating theories. This paper aims to introduce western Translational English Corpus (shortened as TEC) and analyse how a corpus-based approach can be applied to the translation studies. It discusses TEC's generation, classification and its application. It also points out the advantages and difficulties that the study of TEC presents when attempting to answer questions arising specifically from within translation studies.

3.2. Corpora in translation teaching

How corpora can be used in teaching translation has also attracted much attention among Chinese translation researchers. With few exceptions, however, the 16 articles in this category have again mainly presented overseas research on and practices of using corpora in translator training. A couple of articles have offered thoughts about how corpora can be used in teaching translation in the Chinese context, but the ideas are yet to be implemented or tested. Surprisingly, over half a dozen articles published over a period of six years have almost identical titles, with some even published in the same years, for instance:

- Bilingual parallel corpus in translation teaching (Liu 2008)
- Parallel corpus in translation teaching: Theories and principles (Qin and Wang 2007)
- Preliminary discussion on corpus in translation teaching (Zhao 2007)
- Corpus in translation teaching (Li 2007)
- Application of corpus in translation teaching (Wu 2007)
- Preliminary discussion on corpus and practical translation teaching (Song 2006)
- New approach to translation teaching: Applications of bilingual parallel corpus (Lu 2005)
- Application of bilingual parallel corpus in translation teaching (Wang 2004)

Again, these similar titles show that many of the articles on using corpora in translation teaching simply summarize research on this topic published in the West; consequently, a considerable overlapping can again be found among these articles. One may again wonder why so much effort has been devoted to rather general discussions of how corpora can be used in translation teaching and why these academic journals keep publishing basically the same type of general discussions on the same topic over a period of nearly a decade.

3.3. Corpus design and methodological considerations

The third major area of corpus-assisted translation research that has attracted much attention in China is corpus design and methodological considerations in corpus building. These articles address mainly three types of questions:

- ‘representativeness’ of texts in corpus-building (Jiang and Jin 2007, Li 2007);
- designs of corpus, e.g. Chinese-Mongol (Shu *et al.* 2006), corpus of legal texts (Jiang 2005), corpus of medical texts (Chen and Shi 2005);
- technical problems in corpus building and exploration such as data extraction and retrieval (Wang 2000, Sun *et al*. 2008), and sentence alignment (Lin *et al*. 2008).

One third of the articles were actually not published in traditionally linguistics or translation studies journals. Rather, they appeared in science journals such as *Computer and IT*, *Computer Engineering and Application*, *Computer Applications*, *Standardization and IT*, and *Informatics*. This was expected since a number of studies focused entirely on technical aspects of corpus building such as extraction of data, sentence alignment algorithms and management of bilingual corpora. A check of the backgrounds of the authors also revealed that most of them were trained in computer technology rather than translation or language studies.

It might also be of interest to note that there have been efforts to build corpora involving languages of the Chinese ethnic minorities, for instance the Chinese-Mongol Corpus. In addition, several specialized corpora have also been developed, for instance, the corpus of legal texts (Jiang 2005) and the corpus of medical texts (Chen and Shi 2005).

3.4. Data-based studies of translational issues

The part that best represents the state of the art of corpus-assisted translation research in China should probably be the data-based studies of specific translational and linguistic issues. A total of 11 articles fall into this category. They deal with different translational issues, such as stylistic analysis in translation criticism (e.g. Mao and Qiu 2007, Duan and Li 2007), translation universals / norms (e.g. Hu 2007, Xu and Zhang 2006), corpora as the source of examples to illustrate translation of specific words and phrases or cultural concepts (e.g. Feng and Chen 1999, He 2008).

In fact, three of the above mentioned articles (He 2008, Wei and Lei 2007, Feng and Chen 1999) are not data-based empirical research in a real sense, because in these studies, corpus data are merely used as a source from which examples are drawn to illustrate methods of translating some phrases and expressions.

As mentioned at the beginning of the chapter, one of the major research purposes of the present study is to examine the research designs and methods used in corpus-assisted translation research. Since most of the studies in this group are data-based research, their research designs, particularly issues such as the structure of the corpus, text sampling techniques, research methods and the project reports are scrutinized.

It is found first of all that the description of the research and corpus design is in general rather sketchy. In most cases, only the name of the corpus is given, with little information provided on the decision process in constructing the corpus, e.g. what texts are included, how sampling is done, how the texts are marked up and annotated, why such a corpus thus constructed is the most appropriate for the research questions. There seems to be a ungrounded assumption among the researchers that people know about their specific corpora or that all corpora are built with the same considerations.

Secondly, with few exceptions, the corpora used in these studies tend to be very small in size, several of which include a single text and / or extremely short texts (e.g. Mao and Qiu 2007, Xu 2006). This, to a certain extent, has in effect defeated one of the most prominent advantages of corpus-assisted studies – processing of texts of a large number of words. By necessity, the conclusions drawn upon such small corpora are rather shaky and can be considered at best as working hypotheses to be confirmed or rejected with future further studies.

What was particularly problematic about these studies was that some of them only reported on the statistical results obtained with popular analysis tools such as Wordsmith without further making sense of the numbers. That is, the studies merely stopped at some statistical facts such as type / token ratio and sentence length. No attempts were made to explain how these facts came about and what they could tell us about the different factors at play in the process of translation. For instance, in studies comparing translators' styles, no explanations were offered as regards what caused the differences in the translation styles of the translators in question, socially, culturally and ideologically.

4. Discussions and implications

4.1. Originality in corpus research

As discussed earlier, one major problem in corpus-assisted translation research in China is repetition (not replication) of previous research. Many of the published articles are rather general, some even superficial, repetitive introductions to corpus technology in translation research and teaching. Only a few in-depth studies have been found. Such a phenomenon seemingly defying easy comprehension can be explained once it is put in the context of language studies in China over the past three decades.

In language studies in the Chinese mainland, there seems to be a general interest among language and translation teachers to write introductory review articles on frontier research efforts and initiatives in the West. Three reasons appear to be relevant. The first reason relates to the research tradition in arts and humanities, particularly in the areas of language studies in China. A number of well-known linguists of the older generation in China inadvertently set an undesirable example for the younger generations. As many people know, a number of our older linguists in China made their names simply by introducing linguistic theories from abroad to China. Typical of their "research method" back then was spending one or several years as visiting scholars in Western English-speaking countries, such as the United States, the United Kingdom and Australia, to collect as many books and articles on a particular topic as possible, and then summary-translating and editing them into books in Chinese when they returned, thus making their names after selling the books for thousands and thousands of copies. This was extremely prevalent in the late 1970s and early 1980s because back then, most Chinese academics of language studies were denied access to western linguistic theories. As a result, these introductory publications were actually welcomed and well received in China. Given the unique and special circumstances they lived in – China having been cut off from the rest of the world for decades, they could earn their credit like that because they satisfied an academic need in those years. However, for younger generations of Chinese linguists and/or translation scholars living today in an era of unprecedented globalization and international exchanges as well as easy access to the internet, we have no reason to equate research with mere introduction of new theories and ideas from outside China. Rather, we must view ourselves on a par with our western counterparts.

Secondly, many Chinese academic journals, including some of the top ones, seem to be interested in this kind of introductory articles summary-translating western linguistics and translation theories. Take TAP (think-aloud protocols) as an example. Translation process research using think-aloud protocols as a research method first started in the late 1970s and early 1980s in Germany. A search in CJN produced approximately half a dozen introductory articles that were published in top Chinese language journals such as *Foreign Language Teaching and Research, Chinese*

Translators' Journal, *Foreign Language Journal* and so on. In fact any one such article was more than enough. To make matters worse, no in-depth follow-up research has been conducted afterwards. It seems that some "researchers" were actually capitalizing and thriving on such introductory research since inclusion of articles published in these top journals in their CVs means job opportunities and /or promotions. We believe that these academic journals are actually sending out wrong messages to the research community by publishing these survey articles over and over again. The result was that many are keen to introduce major linguistic or translation theories into China but with little interest in any further in-depth original research to follow up.

One might want to note that novelty or originality of research does not necessarily mean developing a major new theory; however, it does mean that you are expected to add something new or previously unknown to the existing body of research, not just collect and summarize other people's work in your mother tongue. Novelty or rather originality can be expressed in different forms:[3]

- a new insight in an existing debate;
- the application of an established theory to a new area;
- an expression of disagreement with a certain position argued by another writer;
- an extension of a previously developed line of enquiry.

True genuine research is not about second-hand information, nor relay information from other sources. What really defines research is the novelty, which is like the spirit and soul of a piece of research. Whether it is qualitative or quantitative, the research has to present something novel.

Unfortunately, over the past 15 years of corpus related translation research in China, very few original studies have been carried out. To rectify the situation, Chinese corpus researchers should stop simple and superficial introduction, put themselves on a par with their overseas colleagues and start RESEARCHING. Even though English might be a handicap for some researchers and hence they perhaps have little chance to publish their research in international English journals, quality research in Chinese journals is just as respectable and desirable.

4.2. CATs or CBTs

One major observation made above of corpus-assisted translation studies in China is the repetition of studies of similar nature, design, and findings. For instance, many have examined the concept of translation universals suggested by Baker (1993), and stopped at the statistical results, appearing to prove the obvious or the proven. For instance, at a corpus conference we attended in Shanghai Jiaotong University in 2007, approximately 80% of the data-based studies were conducted in very similar ways: merely presenting the statistics generated with the software Wordsmith without further exploration of the constraints, pressures and motivations that influence the act and process of translation.

Tymoczko (1998) warned against the possible danger of pursuing scientific rigor as an end in itself through empty and unnecessary quantitative investigations. Baker (1993) is also concerned that research into the nature of the third code may not go beyond the study of recurrent linguistic patterns. Sinclair (2005) also cautioned us against the danger of a vicious circle of researchers constructing a corpus to reflect what they already know or can guess about its linguistic detail. This problem is unfortunately what the corpus-assisted translation research in China is facing at present. A typical corpus-assisted study today can be represented in Figure 11-1.

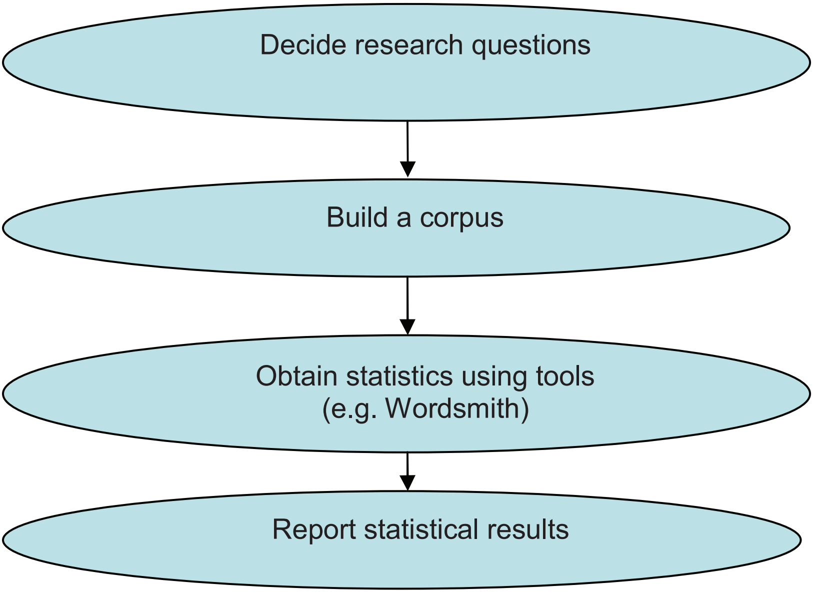

Decide research questions
Build a corpus
Obtain statistics using tools
(e.g. Wordsmith)
Report statistical results

Figure 11-1. The flowchart of a typical corpus-assisted translation research project in China

What is missing in the chart is the analysis after obtaining the statistical results. Such an analysis goes beyond the obvious and known and probes further into the causes of the phenomenon revealed through the statistics. A truly useful corpus-assisted translation study should therefore carry at least the following two features:

1. It does not merely present statistical results to prove the obvious and known;
2. It goes further to explain the statistical results by looking at the causes for such tendencies as revealed in the statistical results.

So, a useful and sensible corpus-assisted translation research project must have an additional part of sense making following the presentation of statistical results, as shown in Figure 11-2.

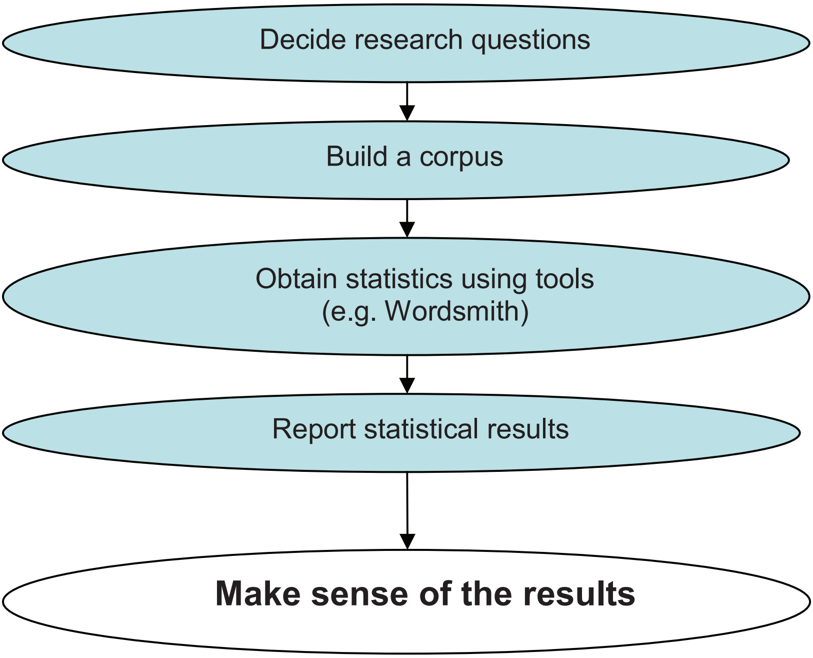

Decide research questions
Build a corpus
Obtain statistics using tools
(e.g. Wordsmith)
Report statistical results
Make sense of the results

Figure 11-2. The flowchart of a corpus-assisted translation study

Corpus-assisted translation research needs to be distinguished from corpus-assisted linguistics research. Translation is a much more complicated social activity, involving considerations of culture, context and socio-political factors. Simply providing the numbers does not tell much about the process of translation. What is really important and useful is the part of sense-making, essential to any corpus-assisted translation study. As Laviosa (1998: 1) points out, the purpose "is not merely to unveil the nature of the 'third code' per se, but most importantly, to understand the specific constraints, pressures, and motivations that influence the act of translating and underlie its unique language." So, the step of sense-making is to answer questions such as how and why the translation came about the way it did and what social, cultural and political effects it brought about in the target language, as shown in Figure 11-3.

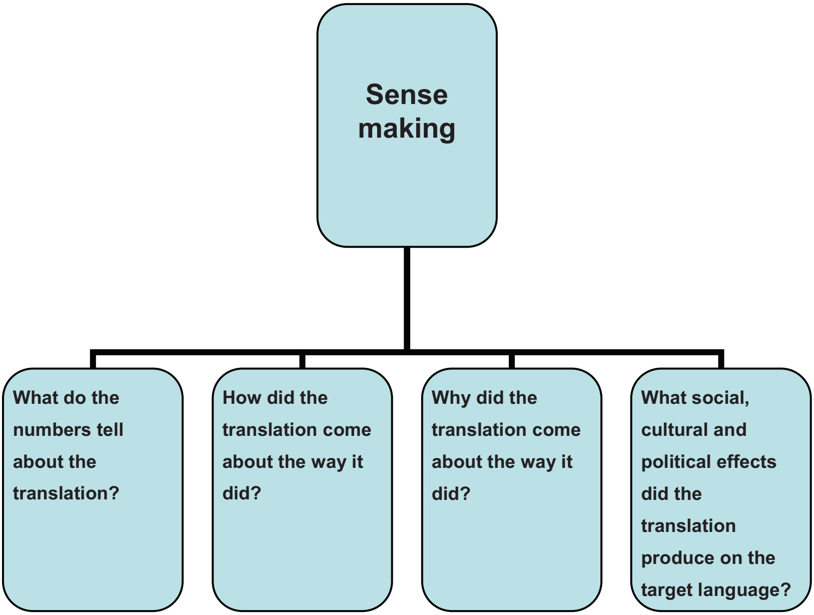

Sense making
What do the numbers tell about the translation?
How did the translation come about the way it did?
Why did the translation come about the way it did?
What social, cultural and political effects did the translation produce on the target language?

Figure 11-3. The flowchart of sense-making

Toury (1995) suggests a three-phase research methodology for systematic descriptive translation research. The description of the translation product should be accompanied by discussions of the wider role of sociocultural systems. The three-phase methodology was summarized by Munday (2001: 112) as follows:

1. Situate the text within the target culture system, looking at its significance or acceptability.
2. Compare the ST and the TT for shifts, identifying relationships between "coupled pairs" of ST and TT segments, and attempting generalizations about the underlying concept of translation.
3. Draw implications for decision-making in future translating.

Such an analysis will necessarily involve the study of paratexts to the texts in examination. Genette (1997) divides paratextual elements into two kinds: peritexts and epitexts. Peritexts appear in the same location as the text and are provided by the author or publisher. They include titles, subtitles, pseudonyms, forewords, dedications, prefaces, epilogues and framing elements such as the cover and blurb. (Genette 1997: 12) An epitext "is any paratextual element not materially appended to the text within the same volume but circulating, as it were, freely, in a virtually limitless physical and social space" (Genette 1997: 344). It can be the marketing and promotional material provided by the publisher, correspondence on the text by the author, and also reviews and academic and critical discourse on the author and text which are written by others. And by extension, it should include correspondence on the translation by the translator and / or the author and also reviews and academic and critical discourse on the translator and the translation written by others.

As such, a corpus approach is nothing more than a research tool for language studies or translation research. It does not provide any theoretical framework for our research. Instead it provides a tool and only a tool to analyze a large amount of language data otherwise not possible by human hand. Hence corpus-related research cannot be considered as a paradigm shift or new approach as suggested by Laviosa (1995). There is no such a thing as *corpus-based* translation studies (CBTs), but rather *corpus-assisted* translation studies (CATs).

4.3. Thick descriptions

A data-based empirical study often requires thick descriptions of the research design so as to ensure accurate and contextualized interpretation of the results and if desired, proper replication of the study (Lincoln and Guba 1985). This requirement also applies to corpus-assisted research. That is, detailed descriptions should be given about the design and construction process of the corpus used in the project. However, little description has been given to issues such as orientation, sampling, criteria, and composition as listed by Sinclair (2005). This can be dangerous for other users of the corpus or for readers / examiners of the reports. As Sinclair (2005: 8) argues:

> A corpus that sets out to represent a language or a variety of a language cannot predict what queries will be made of it, so users must be able to refer to its make-up in order to interpret results accurately.

Toury (1995: 3) also stresses the importance of explicitation of the research methodology and research techniques in order to “ensure that the findings of individual studies will be intersubjectively testable and comparable, and the studies themselves replicable.” Therefore, as one pointer for future research, the design and composition of a corpus should be documented fully with details about the contents and arguments in justification of the decisions taken.

When providing thick descriptions, the following flowchart of corpus-making can be used as a guide regarding the aspects to be included (see Figure 11-4).

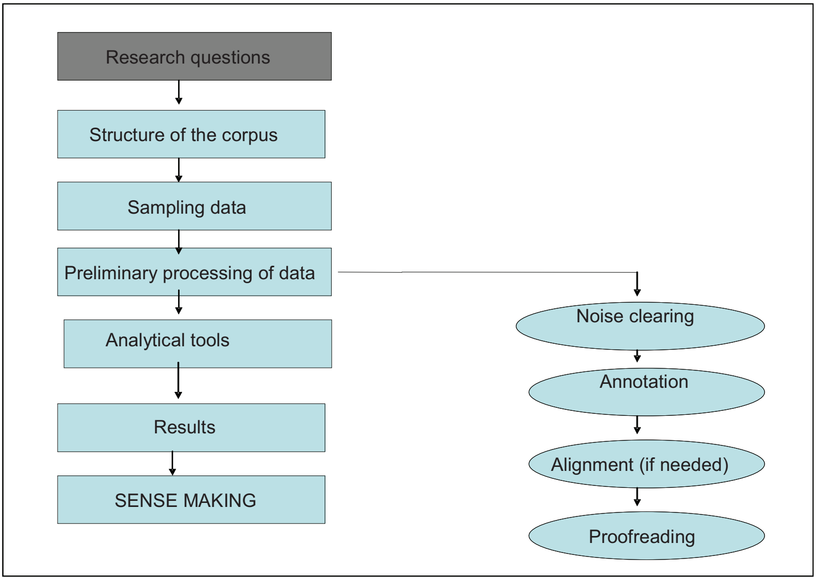


Figure 11-4. The flowchart of aspects to be covered in thick descriptions

5. Concluding remarks

Corpus-assisted translation research has attracted much attention among translation teachers, students and researchers in the Chinese mainland over the past 15 years and the interest is growing rapidly. A good number of publications in the form of research articles and monographs have been produced on this topic. However, as the present chapter shows, much has been devoted to repeated introductions of how corpora can be used in translation research and teaching, or what research has been carried out on this topic outside China. Serious data-based original studies of translational issues are rare to find.

Two aspects to be addressed immediately in future CAT research are: The misconception of obtaining statistical results as the final goal needs to be rectified. A corpus as well as a statistical presentation of translation or language facts is not the ultimate goal of our research, but rather the beginning and foundation for real research on whatever research questions the project is addressing. Attempts to answer questions such as "what do the statistical figures mean?", "what do they tell about the translation process and act?", "what factors were at play in the production of the translation(s) in question?", and "how did they work together in the translation process?", can and will bring out the real juicy findings of the research. Originality and novelty must be held as the soul of real serious research.

On a more specific level, for corpus-assisted translation research as data-based empirical research, detailed information must be provided in the research reports regarding the design of the corpus and the research methods to ensure informed interpretation and any desirable replication of the study.

Notes

1. The CJN database contains over 8,200 journals and it is continuously updated. The data for this research were collected on the 14th of September 2008.
2. The articles are published in Chinese with English titles and abstracts supplied in the original journals. Therefore the English titles and abstracts printed here are taken from the original journals without editing.
3. Taken from the Central European University Self-access Website, accessed in August 2008.

References

Anderman, G. and Rogers, M. (eds.) (2007), *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Multilingual Matters.

Baker, M. (1993), “Corpus linguistics and Translation Studies: Implications and applications”, in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology. In Honour of John Sinclair*, 233-250. Amsterdam: John Benjamins.

—. (2000), “Towards a methodology for investigating the style of a literary translator”. *Target* 12(2): 241-266.

Bosseaux, C. (2006), “Who’s afraid of Virginia’s you: A corpus-based study of the French translations of *The Waves*”. *Meta*: 51(3): 599-610.

—. (2001), “A study of the translator’s voice and style in the French translations of Virginia Woolf’s *The Waves*”, in Maeve Olohan (ed.) *CTIS Occasional Papers* (Volume 1), 55-75. Manchester: Centre for Translation and Intercultural Studies, UMIST.

Chen, W. (2007), “TEC and TEC-based descriptive translation studies”. *Journal of Foreign Languages* (1):67-73.

Chen, Y. and Shi, Y. (2005), “Corpus linguistics and C-E Spoken Corpus of Traditional Chinese Medicine”. *Journal of Jiangxi University of Traditional Chinese Medicine* (5): 67-69.

Ding, S. (2001), “A study of western Translational English Corpus”. *Journal of Foreign Languages* (5): 61-66.

Duan, J. and Li, Y. (2007), “A style study basing on corpus - Taking *Fortress Besieged* as an example”. *Journal of Huangshi Institute of Technology* (Humanities and Social Sciences) (3): 100-103.

Feng, Y. and Chen, W. (1999), “A corpus-based study on translation of subordinate ranks from Chinese to English”. *Journal of Foreign Languages* (2): 43-49

Genette, G. (1997), *Paratexts: Thresholds of Interpretation* (translated by Jane E. Lewin). Cambridge: Cambridge University Press.

Granger S., Lerot, J., and Petch-Tyson, S. (eds.) (2003), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi.

Hansen, G., Malmkjær, K. and Gile, D. (2004), *Claims, Changes and Challenges in Translation Studies*. Amsterdam: John Benjamins.

He, W. (2008), “Translation of metaphors with ‘Xin’ (heart) in Chinese: Based on instances from a Chinese-English parallel corpus”. *Journal of Sichuan International Studies University* (2): 129-135.

Hu, X. (2004), “Corpus-based studies and translation universals”. *Shanghai Journal of Translators for Science and Technology* (4): 47-49.

—. (2005), “Corpus-based studies on universals of translation”. *Journal of PLA University of Foreign Languages* (3): 45-48.

—. (2007), “A corpus-based study on the lexical features of Chinese translated fiction”. *Foreign Language Teaching and Research* (3): 214-221.

Huang, J., Huang, P., and Fan, Y. (2004), “A survey of specialized parallel corpus-based translation studies”. *Journal of Chongqing University* (*Social Sciences Edition*) (6): 91-94.

Jiang, L. and Jin, B. (2007), “On the issue of representation in corpus translation studies”. *Chinese Science and Technology Translators Journal* (1): 29-32.

Jiang, T. (2005), “On construction of forensic parallel corpus”. *Journal of Chongqing University* (Social Sciences Edition) (4): 94-97.

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

Laviosa, S. (1998), “The corpus-based approach: A new paradigm in translation studies”. *Meta* 43(4): 474-479.

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

Li, D. (2007), “The ‘representativeness’ of corpora and its enlightenment to the creation of an English-Chinese translation corpus”. *Foreign Languages Research* (5): 66-70.

Li, D. and Liu, K. (2007), “Translator’s style: A corpus-assisted study of English translations of Hongloumeng”. Paper presented at the Workshop on Corpora and Translation Studies, Shanghai Jiaotong University, March 29, 2007.

Li, T. (2007), “Corpus and translation teaching”. *Chinese Science and Technology Translators Journal* (3): 47-49.

Liao, Q. (2000), “Corpus and translation studies”. *Foreign Language Teaching and Research* (5): 380-384.

Lin, Z., Jia, J., and Guo, W. (2008), “Building Chinese-English bilingual corpus of news field and research on sentence alignment”. *Computer and Information Technology* (1): 5-7.

Liu, S. (2008), “Using parallel corpora in translation teaching”. *China Water Transport* (1): 235-236.

Lincoln, Y. and Guba, E. (1985), *Naturalistic Inquiry*. Newbury Park, CA: Sage Publications.

Lu, X. (2005), "A new way to translation teaching: Application of parallel corpora". *Foreign Language Education* (00): 65-69.

Mao, C. and Qiu, T. (2007), "Application of WordSmith Tools in translation criticism: A case study of two translations of Zhu Ziqing's prose *Cong Cong*". *Science* (Academic Research) (34): 14-16.

Mauranen, A. and Kujamäki, P. (2004), *Translation Universals: Do They Exist?* Amsterdam: John Benjamins.

Munday, J. (2001), *Introduction to Translation Studies, Theories and Applications*. London: Routledge

Qin, H. and Wang, K. (2007), "Parallel corpus in translation teaching: Theory and application". *Chinese Translators Journal* (5): 49-52.

Olohan, M. (2004), *Introducing Corpora in Translation Studies*. London: Routledge.

Shu, Q. and Nashunwuritu. (2006), "EBMT system-oriented construction of Chinese-Mongolian Parallel Corpus". *Inner Mongolia Social Sciences* (1): 140-144.

Sinclair, J. (2005), "Corpus and text: Basic principles", in M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*, 1-16. Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2009-08-05].

Song, L. (2006), "Corpus and practical translation teaching". *Journal of Guangxi University of Technology* (S2): 67-69.

Sun G., Song, J., Yuan, Q., Xiao J. and Shan, Y. (2007), "Research on extraction of translation equivalents from Chinese-English comparable corpus". *Computer Engineering and Applications* (32): 44-47.

Toury, G. (1995), *Descriptive Translation Studies and Beyond*. Amsterdam and Philadelphia: John Benjamins.

Tymoczko, M. (1998), "Computerized corpora and the future of translation studies". *Meta* 43(4): 652-659.

Wang, B. (2000), "Extraction of translation equivalents in non-aligned C-E bilingual corpus". *Journal of Chinese Information Processing* (6): 41-45.

Wang, K. (2004), "The use of parallel corpora in translator training". *Media in Foreign Language Instruction* (6): 27-32.

—. (2006), "Corpus-based translation studies: A new paradigm". *Foreign Languages in China* (3): 8-9.

Wei, Y. and Lei, L. (2007), "Translation of *Jingshen Wenming* and *Wuzhi Wenming* revisited: A corpus-based approach". *Journal of Language and Literature Studies* (19): 103-105.

Williams, I. (2007), "A corpus-based study of the verb observer in English-Spanish translations of biomedical research articles". *Target* 19(1): 85–103.

Winters, M. (2004), "F. Scott Fitzgerald's *Die Schönen und Verdammten*: A corpus-based study of loan words and code switches as features of translators' style". *Language Matters* 35(1): 248-258.

—. (2007), "F. Scott Fitzgerald's *Die Schönen und Verdammten*: A corpus-based study of speech-act report verbs as a feature of translators' style". *Meta* 52(3): 412-425.

Wu, A. and Huang, L. (2006), "On corpus-based studies of translation universals". *Foreign Language Teaching and Research* (5): 296-302.

Wu, C. (2007), "The application of English corpus in translation teaching". *Journal of Shanxi Radio and TV University* (3): 71-72.

Xu, W. and Zhang, B. (2006), "Corpus-based contrastive studies on the causal conjunctions in English Chinese classics". *Foreign Language Teaching and Research* (4): 292-297.

Zhang, M. (2002), "Using corpus for investigating the style of a literary translator: Introducing and commenting on Baker's new research method". *Journal of PLA University of Foreign Languages* (3): 54-57.

Zhao, Y. and Shi, J. (2007), "The use of parallel corpora in teaching translation". *China Electric Power Education* (10): 106-107.

Zhao, Q. (2008), "Brief survey of corpus-based translation studies abroad". *Journal of Chongqing Jiaotong University* (Social Sciences Edition) (3): 100-104.

PART II

PARALLEL CORPUS DEVELOPMENT AND BILINGUAL LEXICOGRAPHY

Chapter Twelve

Poverty Driven Bilingual Alignment

Kim Gerdes

1. Introduction

The sheer unlimited usefulness of aligned bilingual corpora in all areas of translation sciences from theoretical corpus work to dictionary development or machine translation cannot be overstated. Unfortunately, however, many translation researchers end up aligning large parts of their parallel corpora manually, lacking tools with basic heuristics to simplify this task.

It is well known that cognate alignment paired with bilingual dictionaries can give astonishingly good results and constitute the state of the art of current bilingual alignment algorithms. However, most of these systems are out of reach for a common translation researcher, because the parameterization requires insight in the underlying statistics and, even more obstructive, the systems require adaptation of the dictionaries – dictionaries that are often expensive or inexistent for the (sub)language pair.

In order to explore the possibilities of aligning parallel corpora without any such linguistic resources, we have to ask the following questions:

Is there any "visible" feature shared by any piece of written text and its translation? The common meaning of the two texts is not easily accessible, and the great variety of syntactic structures and writing systems makes any further affirmation very difficult. However, the discrete and linear nature of all languages gives us a basic access point to alignment.

Words exist. This means that language has chunks of indissociable segments while words have a unique written representation or a finite set of representations.[1] Moreover, these forms that correspond to one word (allomorphs) are often graphically similar.

Of course, every pair of a text with its translation (a bitext) has some untranslated words or words that are translated by complex constructions, distributed over different words. However, we assume that even in very distant pairs of languages most words are translated by words or contiguous sets of words. But among these words with a linearly constraint translation, even non-ambiguous words often constitute translation ambiguities (because the target language forces us to be more or less specific). We can postulate, however, that in every sufficiently long text we find words (or groups of graphically similar words) that have an "easy" translation in the sense that they correspond to a unique word (or a group of graphically similar words) in the translation (see Figure 12-1). The central hypothesis of the present study is that these words occur at similar linear positions in both texts.

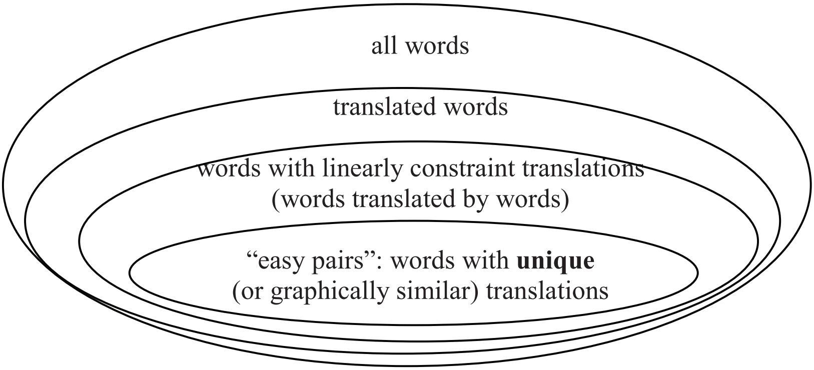

all words
translated words
words with linearly constraint translations
(words translated by words)
"easy pairs": words with **unique**
(or graphically similar) translations

Figure 12-1. Underlying hypothesis: "easy pairs"

We can consider the positions of occurrences of forms in a text as a signal. And "easy pairs" of words will have similar signals.

We will thus attempt to detect these "easy pairs" by similarity measures on all (reasonable) candidates. This chapter presents how this can be done, how to improve some known algorithms of word distance computation in order to include grouped signals of allomorphs and finally how to use these couples as anchor points for the paragraph alignment. The system presented here allows aligning any bitext on the paragraph level without any linguistic parameterization and in particular without any linguistic resources. The final version of the system will be accessible online and thus needs no installation on the user's machine, which gives all users of bitexts easy access to alignment, even without any knowledge in computer programming.

2. Approaches to alignment

2.1. Other approaches

Most alignment systems are based on some kind of graphic similarity between the source and target texts. The most common approach to alignment is based on cognates (lexical or punctuational) (see Simard *et al*. 1992). The basic idea is the exploitation of graphic similarities between a word and its translation. Proper nouns but also many words of Greco-Latin origin have similar graphic forms in many European languages. As an example, the English word *chair* corresponds to the French word *chaise*, a pair which has sufficiently similar forms to be recognized as cognates. The English-Chinese translation pair *chair* – 椅, however, cannot be detected in this way. Cognate-based systems have achieved a quite high reliability rate for most studied (European) languages although most works are highly specific to an application and a language pair, because cognate distances differ among European language pairs and the best definition of the underlying metric remains a subject of debate (see, for example, Ribeiro *et al*. 2001).

It is clearly more difficult to extend this idea to cognates in language pairs with different writing systems. But even for languages like Russian and Japanese, this approach remains interesting as has been shown by Knight and Graehl (1998), because the transcription rules (for example of katakanas in Japanese) are quite simple, although specific metrics for the computation of word distances are needed.[2] The Chinese language, on the contrary, does not have a simple phonetic transcription system. For each word of foreign origin the translator has the choice of a multitude of homophone characters that transcribe the foreign sounds in a satisfying manner. The choice is then often based on "beauty" or semantic appropriateness of the characters. In order to obtain a certain degree of coherence among different translations, the Chinese translator uses enormous specialized transcription dictionaries (for example Zhou 2003).

Thus, finding "similar" words in most language pairs requires considerable linguistic resources, while simple cognate alignment remains a privilege of the European languages with their closeness of vocabulary and their uniformity of the writing system.

First attempts to align bilingual corpora without recourse to linguistic resources have been made by Brown *et al*. (1991), Gale and Church (1991), and Kay and Röscheisen (1993). All three aim at an alignment on a sentence level and work on technical texts or particularly literal translations (e.g. Hansards). The first two are based on the closeness of the length of forms (words and sentences), while the latter, closer to our approach, describes a dynamic programming algorithm that makes hypotheses based on the overall frequency of words and enhances dynamically these hypotheses by taking into account the possible alignments of the sentences containing these words.

The hypothesis of word length similarity, confirmed for example by the English-French pair, is dubious already for pairs like German-French, because German compound nouns usually have a "noun *de* noun" translation in French.

Sentence length, too, depends on the syntactic structure of the languages; and for distant languages, we can expect to find greater difference in sentence length. Moreover, punctuation symbols vary across languages. For example the full stop indicating the end of a sentence in European languages is often represented by a small circle in Asian languages; and even if the symbols are graphically identical, they are often listed in the Unicode tables with the language they are used in, creating completely different objects from the computer's point of view.

The segmentation of texts into paragraphs seems to be the only common point between practically all modern texts. The new line character is thus the only "universal" cognate.

In this chapter we will attempt an alignment only at the paragraph level, and our approach is thus less ambitious than most approaches to alignment. Note, however, that practically all current approaches are "tweaked" for a specific language pair and they do not aspire to any universality. Moreover, the set of language pairs is very limited (mainly English and a major European language, Chinese, or Japanese).

Paragraphs constitute the next step after the alignment of chapters or sections. It seems reasonable to assume that paragraph boundaries are more often respected in the translation process than sentence boundaries, because paragraphs correspond to semantic units, whereas sentences constitute syntactic units. However, the sentence alignment can be done in a subsequent step, the task being considerably easier once paragraphs are aligned. Indeed, some sentence alignment approaches are even based on a previous paragraph alignment, which is often achieved by hand or semi-automatically (e.g. Lebart and Salem 1994, Zimina 2000).

We put one further limitation on our goal: We just try to find the best alignment of paragraph boundaries, i.e. no paragraph will remain orphan. We can thus obtain any combination of paragraph numbers being aligned. Again, finding an untranslated paragraph or inversely, the translator's insertion can be done once the best paragraph alignment is achieved.

2.2. Paragraph alignment by length

Even though paragraphs constitute semantic units, a naïve algorithm that simply aligns the first paragraph of the source language with the first paragraph of the target language and so on will not work well as long as the paragraph correspondence is not one-to-one and it is natural to want to take into account the length of paragraphs.

The first approach to paragraph alignment of a text and its translation, which will be the basis of our method, consists in finding the best alignment of the paragraph marks based simply on the length of each paragraph. The underlying idea is that aligned paragraphs should have approximately the same length. This length can neither be taken to be the number of words as the segmentation of the text into words is not always readily available, nor the number of characters, as the length in characters varies markedly between languages (see Figure 12-2). We have to take into account the relative position of paragraph marks as a fraction of the whole text. In other words, we must normalize the text length. Each paragraph position is thus taken to be a percentage of the whole text. We will now show how to find the best pairings of these percentage points.

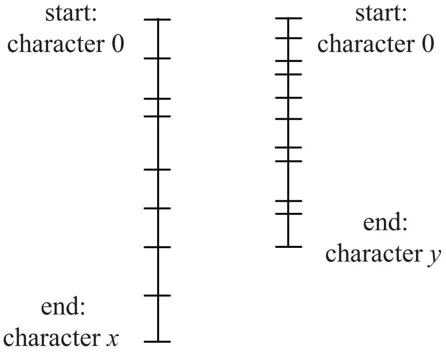


Figure 12-2. Paragraphs as marks relative to the number of characters of the text

In Figure 12-3, we indicate the proceeding graphically. An arrow goes from each paragraph mark in the source language to its closest correspondent in the target language and vice versa. We only take into account the bidirectional arrows, i.e. those arrows that correspond to a pairing of paragraph marks that are mutually their closest homologue. It is possible to obtain non-trivial pairings in this way as the multi-correspondence 2-3 indicated with curly brackets in the Figure 12-3.

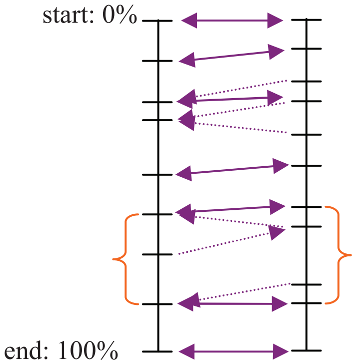


Figure 12-3. Positional alignment

From a computational point of view, this is a standard dynamic algorithm searching for the shortest path in a lattice diagram. We look for the closest path to the diagonal (i.e. the thin line in Figure 12-4) that passes through all paragraph marks of both sides if the corresponding point is a local maximum in the sense that we cannot find a horizontal or vertical neighbour point that is closer to the diagonal.

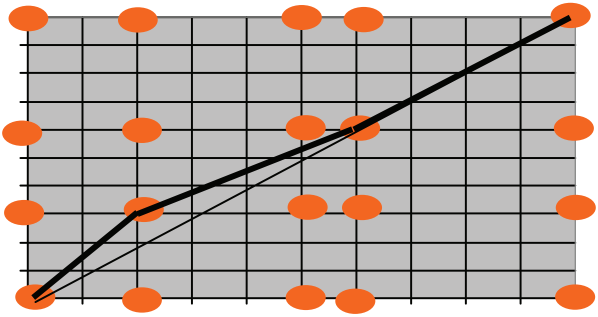


Figure 12-4. Trellis for the alignment of (0, 2, 5, 6, 10) and (0, 3, 6, 10)

This shortest path is shown as the thick line in Figure 12-4, aligning the marks (0, 2, 5, 6, 10) (horizontal) and (0, 3, 6, 10) (vertical).[3] The path starts with the alignment of the beginning of the two texts [0-0].[4] Then we obtain a two-to-one correspondence: [2, 5-3] and finally we obtain the [6-6] alignment (and the obligatory final alignment [10-10] that corresponds to no paragraph). In Figure 12-5, showing the alignment of (0, 1, 9, 10) and (0, 6, 10), no points but the start and end points are local maxima on our lattice and we obtain the grouping of three paragraphs with two paragraphs from Figure 12-3: [0, 1, 9-0, 6]

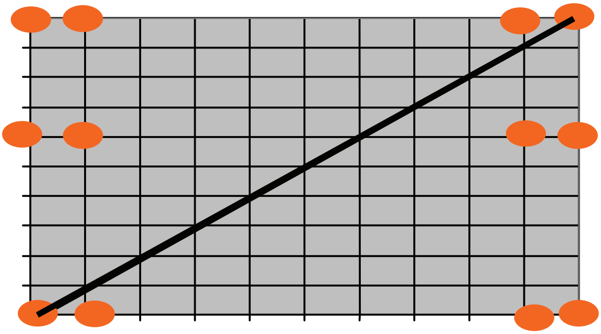


Figure 12-5. Trellis for the alignment of (0, 1, 9, 10) and (0, 6, 10)

The results obtained with this algorithm are better than those of the naïve algorithm counting paragraphs, but this approach is very sensitive to noise and will work well only on texts that are translated very precisely, homogeneously, and without omissions or insertions. If for example, the translation of a journalistic article contains an introductory paragraph that the original did not contain, all paragraph alignments will be shifted down one step too far and the alignment will thus be mostly wrong. It is clear that it is necessary to add other hints in the bitext that will make the alignment more robust.

3. Time warp

Dynamic time warping algorithms are also based on the distribution of a word in the whole text, but contrary to the paragraph marks, we first have to establish the pairings. The intuition behind the time warping approach is that a word signal resembles the signal of its translation, even if the latter is "deformed" by the translation. The signal may be reduced, occur earlier or later or even miss certain points, but it still remains "recognizable" as being the translation of the original signal.

3.1. Illustrating the intuition behind dynamic time warping

To illustrate this intuition, let us consider a French text with its Chinese translation. We use the first volume *Aube* of the epic *Jean-Christophe* by Romain Rolland (1904-1912), amounting to 226,981 characters, and its Chinese translation by Fu Lei (1957) totalling 68,062 characters.[5]

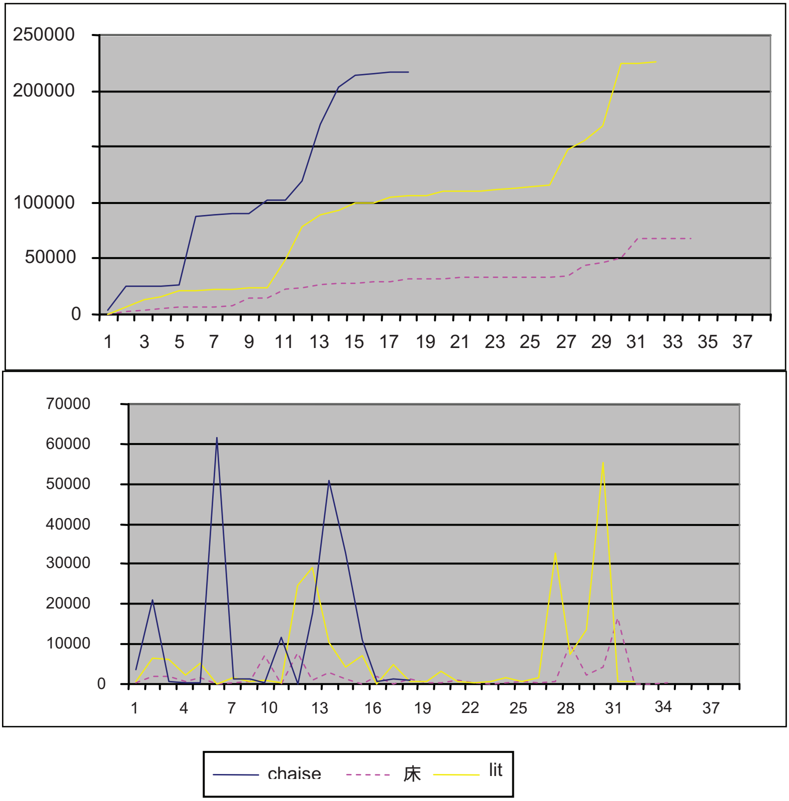


Figure 12-6. Occurrence vectors (top) and recency vectors (bottom) for three words

We consider three words: *lit* ‘bed’, its Chinese translation 床 and the word *chaise* ‘chair’. If we represent the points where these words occur simply as the number of characters from the start of the text, we obtain the graph at the top of Figure 12-6. The simple fact that *lit* and 床 occur a similar number of times (respectively 32 and 34 times) causes their curves (the light and the dotted curve) to be more similar but this similarity seems difficult to discern. It is preferable to use a recency vector. That is, instead of representing distances of occurrences of the word from the beginning of the text, we take into account the distance (in number of characters) between each apparition of the word. The representation of this vector makes the similarity of the lines of *lit* and 床 stand out much more clearly (the graph at the bottom of Figure 12-6). However, the fact that French uses many more characters than Chinese still appears in the graph as higher amplitude of the French curves.

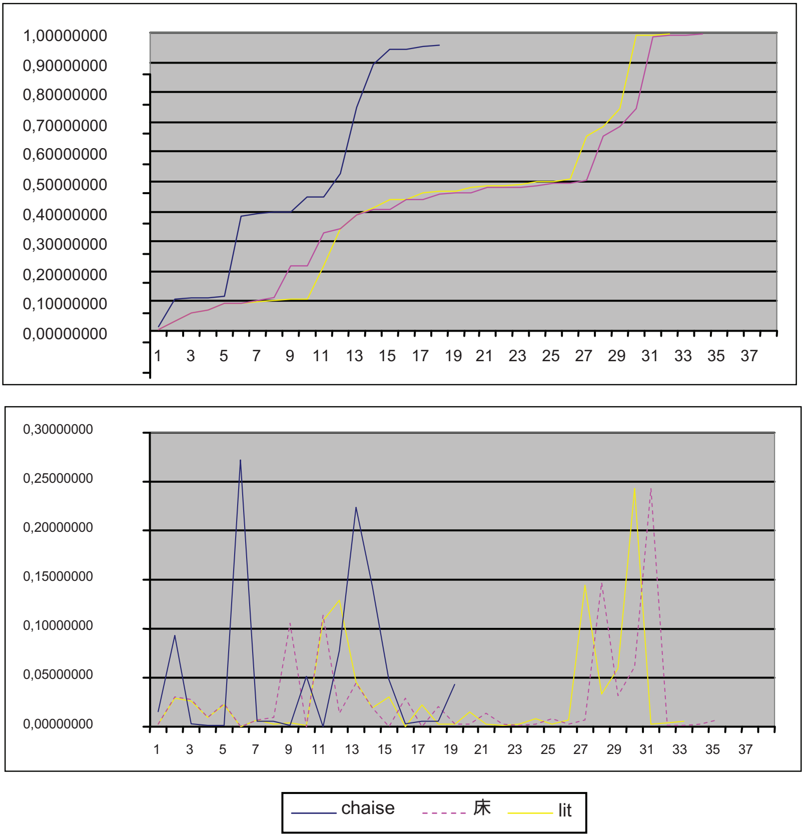


Figure 12-7. Occurrence and recency of three words in a normalized bitext

Using fractions of the whole texts instead of absolute values allows for a normalization of the curves. Now the link between the two words *lit* and 床 stands out clearly in comparison with *chaise* just as well in the fractional diagram (the top graph of Figure 12-7) as with the normalized recency vectors (using the distances between each occurrence of the words expressed in fractions of the text (the graph at the bottom of Figure 12-7). In this latter diagram one recognizes easily the slight movement to the right of the 床 curve, which is caused by two supplementary apparitions of 床 around its 9th and 10th apparitions. The time warping algorithm will allow us to establish a distance measure between two words that counts only once this right movement of the *lit* curve in relation to the 床 curve. Intuitively, the time warping distance will only count the "stretching" needed around positions 9 and 10 to superpose the two curves rather than the constant offset of the two curves. Clearly, normalization can bring the similarity of curves of 床 and *lit* to the fore.

After a short summary of works using dynamic time warping approaches, we will determine the metrics to measure the distance between curves of this type. Then we will expose the algorithm used to find word couples based on the similarity of their signals.

3.2. The use of time warping

Dynamic time warping (DTW) algorithms attempt to find optimal monotone (non-crossing) alignments of two sequences of varied length. The optimal alignment minimizes the distortion between signals. DTW is used in a wide range of domains for the recognition of forms that can be extended or contracted while preserving the information for easier recognition. Its "classic" use is in speech recognition, today usually combined with Hidden Markov Models (Jelinek 1997); moreover, it is used in image or form recognition (where the deformation can be multidimensional) as for example in signature or face recognition or even in data mining (Ratanamahatana 2004).

Concerning the use of DTW for bilingual corpus alignments, the first attempts in this direction have been made by Fung and McKeown (1994), who work on English-Chinese alignment (see also Somers 1998 for a comparison of similar approaches). Their work shows that the DTW algorithm can find pairs of words that are mutual translations.

Note that Fung and McKeown's (1994) algorithm starts with a Chinese text that is already segmented into words. They do not indicate, however, how this segmentation has been obtained, nor do they state the linguistic premises for this segmentation. However, this is important for two reasons. First, the use of pre-segmentation makes their algorithm dependent of linguistic resources because all segmentation systems of Chinese rely heavily on usually large-scale dictionaries to accomplish this task.[6] Second, the notion of "word" has an important influence on the results, because their algorithm uses the words directly as aligned units. If we wanted to obtain a Chinese-German alignment, for example, a "German-style" segmentation of the Chinese texts (i.e. a system where compound nouns constitute single words) would certainly give much better results than an "English-style" segmentation, another Germanic language, where compound nouns are written with spaces between the nouns.[7]

Thus, without any explication of this preliminary step of segmentation, Fung and McKeown's (1994) results are neither verifiable nor reproducible. We have to add here that the task of alignment, as indicated in the title of their paper, is never accomplished. They find good pairs of words that could be used as anchor points for the alignment but two points remain obscure: 1) the type of alignment (on paragraph, sentence, phrase, or word level) they want to achieve with the pairs; and 2) the actual alignment method that uses the pairs.[8]

3.3. The good distance between signals

The computation of the global distance between two sequences is based on the sum of local distances (between two elements of the two sequences). It is primordial to find a good metric of local distances, because errors will multiply up in the computation of the global distance and we have to see to it that long sequences will not have a greater distortion just because of their length (as it is the case in Fung and McKeown's (1994) metric that uses word numbers).

Let us develop this point in greater detail. In Figure 12-8, graph A shows two texts of identical length (language 1 on the left, and language 2 on the right) with a word pair that has an identical distribution (three occurrences in both languages at identical positions). These words are of course very good candidates for being mutual translations and we want to attribute 0 as the distance between these signals. Graph B shows the same pair of words in a bitext where the second text is shorter. It is clear that in this case, the word pair is a less good candidate for being a translation than in the A case. Graph C shows another bitext where a word pairing looks just as good as pairing A, because the second text is shorter. In order to obtain a distance measure that corresponds to this intuition where distance in B is greater than distance in A which equals that in C, we again have to normalize and use fractional instead of absolute positions.

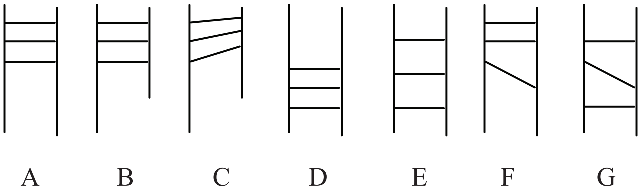


Figure 12-8. Word pairing vs. text length

A second point to take into account for the design of the distance measure is the recency vector. From the position vector of a given word (m_1, m_2, m_3, … m_n), Fung and McKeown (1994) compute the recency vector (m_1, m_2- m_1, m_3- m_2, … m_n- m_{n-1}). This recency vector is not symmetrical in the sense that it counts the distance between the beginning of the text and the first occurrence of the word (the value m_1) but it ignores the distance between the last word and the end of the text. This asymmetric recency vector does not always give bad results. For example, the couples shown in graphs D and E in Figure 12-8, which are clearly as good candidates as A or C, will all get 0 as a distance although the words occur at different positions in the text. However, our metric has to give the same value (>0) to the pairs F and G. Without taking into account the distance of the last occurrence of the word to the end of the text, the distortion of F will only be counted once (as the distance between the second and the last couple). In G, Fung and McKeown's (1994) metric will count it twice: once between the first and the second couple and another time between the second and the last couple of words. The F pairing will have a smaller distance than G, contrary to our intuition about the structure of occurrences of translations in bilingual texts.

In order to compute a correct recency vector, we use a position vector expressing fractions of the text (p_1, p_2, p_3, … p_n). The recency vector includes the distance to the end point (p_1, p_2- p_1, p_3- p_2, … p_n- p_{n-1}, 1- p_n).

Fung and McKeown's (1994) metric is not normalized length and is skewed by leaving out the final recency distance. But even if our metric seems more intuitive, we cannot compare our results directly. They start with a text segmented into words by a non-specified algorithm, as noted earlier, and moreover they only show some examples of anchor points they discover on the basis of heuristics that are not justified in the text (restriction to word frequencies of 10 to 300 words for an English-Chinese text of 700kb).

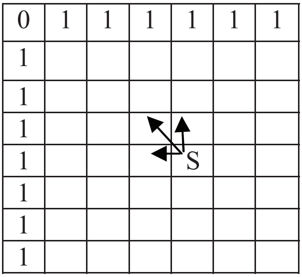
3.4. The algorithm for the computation of the time warped distance

The distance computation we use is a simple dynamic algorithm. Figure 12-9 gives the algorithm in pseudo-code: Given the position vectors of two words to be compared. After computing the recency vectors for each position vector as stated above, we construct a table crossing the two recency vectors and an additional line and colon filled with 1s (maximal distance) with the exception of the slot (0,0) containing 0 (see Figure 12-10).

```
timewarp(list1,list2):
    # takes two lists of numbers between 0 and 1
    # and computes a time warp distance
    rec1, rec2 = recency(list1), recency(list2)
    warp[(0,0)] = 0              # table initiation: corner
    for i=0 to length(rec1) do:
          warp[(i+1,0)] =  1 # table initiation: first line
    for j=0 to length(rec2) do:
          warp[(0,j+1)] =  1 # table initiation: first colon
    for i=0 to length(rec1) do:
          for j=0 to length(rec2) do:
              warp[(i+1,j+1)] = abs($rec1_i$-$rec2_j$) + min(warp[(i,j+1)],
          warp[(i+1,j)], warp[(i,j)])
    return warp[(i+1,j+1)]
```

Figure 12-9. Pseudo-code for time warping

Then the rest of the table is filled line by line as illustrated in Figure 12-10. In each slot S we enter the distance between the corresponding recency vector values, to which we add the minimal value of the following 3 slots: left of S, above S, or diagonally left above S. These three possibilities correspond to a table traversal linking slot S to one of its neighbours on its left, above, or diagonally above. The restriction to these three directions reflects the monotonicity of time warping: we can distort the signal but not tear it apart. When the table is filled, the distance between the words appears in the lowest most right slot (length of recency 1, length of recency 2), symbolizing the less costly alignment between the two words, in the same way as shown in section 2.2 for paragraph marks.



0	1	1	1	1	1	1
1						
1						
1						
1				S		
1						
1						
1						

Figure 12-10. Filling the time warping matrix

The distance computation presented here gives an advantage to rare words, because in all texts the total number of rare words is very high compared to frequent words (Zipf's law). The chances of finding two hapaxes (words with frequency 1) that are not mutual translations at some arbitrary identical fraction of the text (for example 47.6%) are very high and these pairs will thus obtain a distance close to zero, corresponding to their distance in the text. Inversely, frequent words have a very low chance of occurring all the time at exactly identical positions on both sides and they will always have a distance greater than zero. Their high frequency, however, keeps this number quite small. In heuristic tests, we wanted to give an advantage to groupings of frequent words, e.g. by dividing the distance by the number of created couples, but this will favour too boldly frequent words and exclude all rare words from the list of best couples.

The couples we want to retain depend on the use we have for them. When aligning paragraphs, we are interested in words that allow us to find as many interesting paragraph alignments as possible. In other words, we want couples that appear in a few but not in too many paragraphs, the most discriminating distribution being close to half of the number of paragraphs. These maximum and minimum values remain parameterizable by the user. Words that appear between 5% and 50% of all the paragraphs are found to be good value, but further tests will have to determine the optimal values and whether these values differ considerably between languages.[9]

In our implementation of the algorithm, we use another heuristic that does not change the results, but speeds up computation considerably. We only compute the distance between pairs of words that have a similar frequency. We take 50% to 200%; in other words, for a given word W, we do not compute distances between the word and a putative translation T, if T appears more than double or less than half the number of occurrences of W.

4. Language internal cognates and example results

We assume in the introduction section that any sufficiently long text between any two languages will contain some “easy” pairs of word-to-word translation that can be discovered by time warping signal distance comparison. This may be true, but in order to enhance our chances of finding enough “good” pairs even for languages that refuse to take most proper nouns (like Slavic languages) as usually natural candidates for “good” pairs, we propose to include groups of “graphically similar” words. How can this be done?

The answer is simply to apply a cognate search internal to the text in one language. Instead of a simple Levenshtein distance (that equals the number of changes needed to pass from one word to the other, which is used, for example, in any spell checker’s replacement options), we will go for a slightly more complex distance, i.e. the Jaro-Winkler distance, a measure that counts variations at the end of the word less than variations in the beginning of the word. This privileges the detection of word final inflection, and if languages should exist where the beginning of words is inflected more heavily than the end, this algorithm is not a good choice and could be seen as a linguistic parameterization, contrary to the stated goal.

For both texts in a bilingual pair, we compute the groups of words that are graphically similar (again using a heuristic minimum that speeds up computation) and add the discovered word groups to our list of words as if they were words with a unique form. Of course, many of the proposed word groups are not different forms of a common morpheme and have nothing in common but a similar form; but theoretically, this should not matter as a group of forms that have no common morpheme should have no counterpart in the other language with a sufficiently similar signal. To our surprise, this holds not completely true, and some of the discovered groups are slightly polluted, though the slight error does not destroy the overall advantage of using these groups.

Table 12-1 shows the 20 best pairings found for the French-German bitext *The Sorrows of Young Werther*. Word groups are based on the language internal Jaro-Winkler distance. Note that all but the second pairing are correct (sometimes partial) translation. The word ‘Daura’ is grouped with ‘Armar’ because they only appear together in a short specific section of the text, and thus this pair also helps to adjust the alignment. It is important to see this extraction of pairs not as a final goal of extraction of a translation vocabulary but exclusively as an extraction of useful anchors for the subsequent alignment process. Bilingual vocabulary extraction should be done on the basis of the final aligned corpus. Note also, that the proper nouns and the numbers in this list were not discovered by cognate matching but exclusively by their signal similarity.

Table 12-1. The 20 best pairings in *The Sorrows of Young Werther*

Distance	German words (or word groups)	French words (or word groups)	Gloss
0:0.00767	arindal	arindal	Arindal (name)
1:0.00773	daura	armar	Daura/Armar (names)
2:0.00863	daura dauras	daura	Daura (name)
3:0.01015	morars morar	morar	Morar (name)
4:0.01043	heide	bruyère bruyères	heath
5:0.01069	armins armin	armin	Armin (name)
6:0.01076	linden lindenbäume linde	tilleul tilleuls	linden (trees)
7:0.01090	bücher	livres	books
8:0.01118	paradiesisch paradies paradiese	paradis	paradise, paradisiac
9:0.01140	mai	mai	May
10:0.01144	gesandten gesandter gesandtschaft gesandte	ambassade ambassades ambassadeur	embassy, embassador
11:0.01145	schnee schneeglänzenden	neige	(sparkling) snow
12:0.01145	dezember	décembre	December
13:0.01179	krankheit	maladie	illness
14:0.01244	august	août	August
15:0.01282	buches buche buch	livre	book
16:0.01307	klaviere klavier	clavecin	piano
17:0.01324	september	septembre	September
18:0.01342	8	8	8
19:0.0141449775087	30	30	30

Table 12-2 gives the 20 best results for the Chinese-French bitext *Aube* from *Jean Christophe*. Note that the grouping of the misspelled ‘Gotttfried’ (three times the letter *t*) gives a smaller time warping distance than the correctly spelled ‘Gottfrieds’ alone. The system thus “discovered” the spelling error. Note also the greater distances than in the first examples. As expected, the second bitext offers fewer "easy" pairs than the first. Work is in progress to determine useful heuristics on the maximum distance values of useful word pairs, depending probably on the overall size of the corpus.

Table 12-2. The 20 best pairings in *Aube*

Distance	French words (or word groups)	Gloss	Chinese charac-ters	Gloss
0:0.06186	gotttried gottfried	Gottfried (name)	舅	part of the name *Gottfried*
1:0.14329	chairs chaises caisse chaise	chair, box	椅	chair
2:0.15069	table tablier tables	table, apron	桌	table
3:0.20146	lit	bed	床	bed
4:0.20734	melchior	Melchior (name)	沃	part of the name *Melchior*
5:0.21150	nuit	night	夜	night
6:0.21265	piano	piano	钢	part of the word *piano*
7:0.23058	louisa	Lousia (name)	莎	part of the name *Louisa*
8:0.26149	père	father	父	father
9:0.28664	oie voie joie	goose, voice, joy	忘	forget
10:0.29866	grands grains graisse grasses gras ras	big, grain, fat, crop	内	in(side)
11:0.29993	conseil consentait conservait conversation servait conserve	council, consented, kept, conversation, served, can	书	book, letter, write
12:0.30906	regarder	look	熟	cooked
13:0.30921	craignait crainte criant craintif craintes	fear, fearful, shouting	德	virtue, Germany
14:0.30960	petits	small	久	long (time)

15:0.31001	rêves rêveries rêver rêve	dream	梦	dream
16:0.31376	réveille réelle réveiller réveilla réveillé éveille veiller réveillait	wake, real	醒	wake
17:0.31705	soupir assoupir soupirail soupira soir souper	sigh, evening supper	晚	evening
18:0.31767	vieilles vieillots vieille vieillissait vieil vieillards vieillard	(growing) old	闷	melancholy, bored
19:0.32201	connais connaisseur connaissance connaissent connaissait	know, expert, knowledge	立	stand, establish

5. The alignment algorithm

To conclude the description of the overall alignment process, we present in this section a simple algorithm of using anchors to align paragraphs (see Figure 12-11). These anchors can be other cognates than the new line character. If we find any, of course, we have to take them into account. This step integrates well in the overall process, although the goal of this chapter is precisely the description of a solution for aligning bitext where no cognates or insufficient numbers of cognates are available. To this list of cognates we can add a search for all Unicode names of all punctuation and numeral symbols. If their names are similar, they can also be added to the cognate list. This is particularly interesting for numeral symbols or “rare” punctuations (as colons or semicolons) as the more frequent symbols like commas will not help paragraph alignment because they appear in nearly every paragraph. All cognates must have a distance value compatible with the distances of the time warping measures. The easiest step is just to give “real” cognates the value of the lowest discovered time warping distance.

```
getAlignmentMatrix(goodCoupleList):
# takes a list of good couples
    alignmatrix = numberParagraphsText1 x numberParagraphsText2
    set all alignmatrix values to 1
    for each (word1, word2, distance) from goodCoupleList do:
                parInd1 = getParagraphIndices(word1)
                parInd2 = getParagraphIndices(word2)
                for i=0 to length(parInd1) do:
                    for j=0 to length(parInd2) do:
                        alignmatrix[i,j]=alignmatrix[i,j]*distance
    return alignmatrix
```

Figure 12-11. Pseudo-code for the construction of the alignment matrix

We call the combined list of “real” cognates, Unicode cognates and time warping couples the “list of good couples”. We create an alignment matrix crossing all paragraph positions of the two texts and we initialize the matrix with ones in all slots. For each couple (*word1, word2, distance*) in the list of good couples, we obtain the two lists of paragraph indices in which *word1* and *word2* appear respectively. Each value of the slot that corresponds to a pair of paragraphs (in which *word1* and *word2* appear respectively) will be multiplied by *distance*. In this way, pairs of paragraphs that occur for various couples will receive a particularly small value.[10]

Now we only have to compute the “cheapest” path crossing this alignment matrix. For this we can practically use the same algorithm that we use for time warping (see Figure 12-12); we only have to record at each step which of the three choices (left, top, diagonal) has the minimal value, in order to be able to trace back the way through the matrix. Once we are through, we have to follow these indications from the lower right corner back to the top left corner of the matrix. Each diagonal step will correspond to a separation of two paragraph blocks, while each vertical or horizontal step adds a new paragraph to the existing block.

Note that we do not have to apply any preference of the diagonal again, as this preference is already contained in the choice of good pairs (they are declared good because they have similar signals, i.e. the pair is close to the diagonal). In other words, this algorithm will stay on the diagonal unless a detour is “cheaper” for very good reasons, i.e. lots of good couples asking for it.

```
getAlignment(alignmentMatrix):
    # takes an alignment matrix
    # and computes the diagonal path through the matrix with the lowest overall values
    # the output is a matrix that contains ones at the aligned paragraphs
    lines = number of lines of alignmentMatrix
    colons = number of colons of alignmentMatrix
    warp  =  lines +1  x colons +1
    directions = lines x colons
    finalAlignment  = lines x colons
    f,g=lines-1,colons-1
    set all warp values to ∞
    warp[0,0]=0
    for i=0 to lines do:
            for j=0 to colons do:
                        mini = min(warp[i,j+1], warp[i+1,j], warp[i,j])
                        warp[i+1,j+1] = matrix[i,j] + mini
                        if mini == warp[i,j]: directions[i,j] = 0
                        elif mini == warp[i,j+1]: directions[i,j] = 1
                        else : directions[i,j] = -1
    while f>=0 or g>=0:
            finalAlignment[f,g]=1
            if directions[f,g]==0:
                        f-=1
                        g-=1
            elif directions[f,g]==1: f-=1
            else: g-=1
    return finalAlignment
```

Figure 12-12. Pseudo-code for computing the final alignment from the alignment matrix

This algorithm gives satisfying results for insertions and deletions if sufficient good pairs have been found. At least for all concrete examples we have tested the system on, the results are always notably better than the simple paragraph length alignments. Further work will have to test systematically the advantages and disadvantages of this system in comparison with other approaches and we will explore the usage of other cognate algorithms that allow for quality values of the cognates to be taken into account.

The system is implemented on a private web server (http://elizia.net/alignator/). Although the main system is programmed in Python, the computation of the time warping distance as well as the Jaro-Winkler distance between all possible couples of words remain very heavy on long texts, even with the "tricks" of restricting the analysis to words with interesting frequencies for the paragraph alignment and to couples that have a chance of being translation based only on their frequency. This part had to be written in C (thus enhancing the speed by a factor of nearly 50) to make the system usable in a few minutes even on long texts. The user interface uses Javascript. The use of a web server allows for a direct access on all computer systems without prior installation. The complete code will be distributed as free software under the GNU licence.

6. Conclusion

To sum up, here is a brief list of the steps of the algorithm of this alignment system:

1. Word detection – if *scriptua continua*, work on the character level;
2. Cognate detection, including punctuation cognates using Unicode names (it is not necessary to find any);
3. On languages with word spacing, add "intra-language cognates" to the word list, i.e. groups of words with similar forms using the Jaro-Winkler distance;
4. Apply DTW distance measures with the normalized text length on potentially useful word pairs (or word group pairs) and extract potential translation pairs;
5. Add distance of all potential translation pairs (including cognates, if any) to the paragraph matrix of both languages and compute a minimal diagonal matrix path, corresponding to the best paragraph alignment;
6. This alignment can be corrected manually, directly on the web, and exported in different formats for further examination;
7. This approach is considerably better than naïve approaches to paragraph alignment like a purely length based alignment, but it is difficult to evaluate and compare in greater detail for two reasons:
    - on the one hand, other work is often language specific, focuses on sentence alignment or vocabulary extraction, and is often unavailable for testing;
    - on the other hand, while it is easy to construct artificial bitexts that will fool the system, the lack of large manually aligned bitexts for various non-European language pairs makes it impossible to give numbers on the reliability of the system on real texts in those languages.

In conclusion, we believe that the alignment system presented in this chapter can be of great help for researchers of translation studies working on rare language pairs when they create aligned parallel corpora; and if the automatically aligned corpora are eventually corrected manually, they can serve as control data for further systematic enhancement of the algorithm and its associated heuristic parameters. Moreover, the results obtained by this resourceless system as well as the problems encountered in its development have shed light on some universals of translation.

Notes

1. See also the classic debate on the existence of discontinuous morphemes (Harris 1945).
2. The most serious methodological problem concerning Japanese is that only texts using many foreign words can be aligned. A "pure" Japanese text, for example with its English translation, cannot be aligned in this way. In this latter case we would need a complex pronunciation lexicon, just as for Chinese texts.
3. In tenths of the whole text for simplification. We refer to the paragraphs by the fraction of the text that indicate the starting points of the paragraphs in the text.
4. The hyphen indicates here the association of two groups of paragraphs.
5. The electronic versions of these texts were graciously provided by Jun Miao from the ESIT, Sorbonne Nouvelle.
6. As the Chinese writing system does not give easy indications on the beginning or ending of words (contrary to Japanese, for example, where certain simple heuristics on the changes of the types of characters can go a long way), it is natural to use extensive lists of words. The only alternative could be a search for repeated sequences in very large corpora. This, however, will not easily give linguistically relevant results (because the definition of "word" is much more semantic than statistical – one would consider as words, in English deprived of spaces for example, nouns that are always followed by a specific preposition).
7. Fung and McKeown (1994) give an astonishing example: 一氧化碳 'carbon monoxide' is listed twice as a word, once translated as 'carbon' and a second time translated as 'monoxide'. We can thus conclude that in their corpus, the segmentation does not separate compound nouns.
8. Knowing the word pairs does not imply knowing how to align the occurrences of the words. See the extensive literature on cognate alignment and section 2.2 of this chapter where we show a possible alignment procedure for the known "pair" of the new line character.
9. It is possible to enhance the algorithm further by also taking account of high frequency word couples (or symbols like punctuations), for which we believe that they are mutual translations. However, they will have to be taken into account differently in the subsequent alignment computation (where for the moment we only count a binary absent/present feature).
10. Note that we enter all possible alignments of the couple into the matrix, not just the "best" alignment (i.e. the closest one to the diagonal). This makes it possible to

get longer distances from the diagonal, as soon as a large number of pairs point to the same paragraph pairing.

References

Fung P. and McKeown, K. (1994), “Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping”, in *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-94)*, 81-88, Columbia, Maryland.

Gale W. and Church, K. W. (1991), “A program for aligning sentences in bilingual corpora”, in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Berkeley, CA.

Harris, Z. S. (1945), “Discontinuous morphemes”. *Language* 21(3): 121-127.

Haruno M. and Yamazaki, T. (1997), “High performance bilingual text alignment using statistical and dictionary information”. *Natural Language Engineering* 3(1): 1-14.

Jelinek F. (1997), *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.

Kay, M. and Roscheisen, M. (1993), “Text-translation alignment”. *Computational Linguistics* 19(1): 121-142.

Knight K. and Graehl J. (1998), “Machine transliteration”. *Computational Linguistics* 24(4): 599-612.

Lebart L. and Salem A. (1994), *Statistique Textuelle*. Paris: Dunod.

Meng H., Lo, W. K., Chen, B. and Tang, K. (2001), “Generating phonetic cognates to handle named entities in English-Chinese cross-language spoken document retrieval”, in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*. Trento, Italy.

Ratanamahatana, C. A. and Keogh, E. (2004), “Everything you know about Dynamic Time Warping is wrong”, in *Third Workshop on Mining Temporal and Sequential Data*. Seattle, WA.

Ribeiro, A., Dias, G., Lopes, G., and Mexia, J. (2001), “Cognates alignment”, in B. Maegaard (ed.) *Proceedings of the Machine Translation Summit VIII*. Santiago de Compostela, Spain.

Simard, M., Foster, G. and Isabelle, P. (1992), “Using cognates to align sentences in bilingual corpora”, in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation TMI-92*, 67-81. Montréal, Canada.

Somers, H. (1998), “Further experiments in bilingual text alignment”. *International Journal of Corpus Linguistics* 3: 115-150.

Wagner, R. A. and Fischer, M. J. (1974), “The string-to-string correction problem”. *Journal of the ACM* 21(1): 168-173.

Yamada, K., and Knight, K. (2001), “A syntax-based statistical translation model”, in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 523-529. Toulouse, France.

Yarowsky, D., Nag, G. and Wicentowski, R. (2001), “Inducing multilingual text analysis tools via robust projection across aligned corpora”, in *First International Conference on Human Language Technologies*. San Diego.

Zhou, D. (2003), *Waiguo Diming Yiming Shouce* (Translation Dictionary of Proper Nouns and Foreign Places [modified edition]). Beijing: Commercial Press.

CHAPTER THIRTEEN

CHINESE-UYGHUR PARALLEL CORPUS CONSTRUCTION AND ITS APPLICATION

SAMAT MAMITIMIN, UMAR DAWUT

1. Introduction

It is well recognized that multilingual resources are very important to both theory-oriented linguistic research and application-oriented cross-language information processing. Therefore, many multilingual corpora have been created for a range of purposes. One of the best-known and most frequently used parallel corpora is Europarl, which is a collection of materials including 11 European languages taken from the proceedings of the European Parliament. Another parallel corpus is the JRC-Acquis Multilingual Parallel Corpus (Steinberger *et al*. 2006). It is the largest existing parallel corpus of today in terms of both size and the number of languages covered. The OPUS corpus (Tiedemann and Nygaard 2004) and the English-Norwegian Parallel Corpus (Johansson 1994) are also very famous parallel language resources.

A Chinese-Uyghur parallel corpus also has very important applications in cross-language information processing, Chinese-Uyghur bilingual lexicography, Chinese-Uyghur comparative study, translation research and language teaching. However, so far large-scale and balanced Chinese-Uyghur corpus resources are still unavailable given the difficulties of collecting bilingual translated texts and the intensive labour required in corpus processing. Recently, the project “Construction and Application of Chinese-Uyghur Bilingual Corpus”, which is funded by the China National Foundation of Social Sciences, has achieved some preliminary results in this field.

This chapter presents the work we have undertaken on the project in the building of the Chinese-Uyghur parallel corpus. We will first introduce the aim, source data collection and the workflow of constructing the corpus, and then discuss the annotation process in detail, including preprocessing, markup, sentence alignment and development of corpus tools. Finally, we will present some preliminary results and several applications based on the Chinese-Uyghur parallel corpus.

2. The Chinese-Uyghur parallel corpus

The aim of the project "Construction and Application of Chinese-Uyghur Parallel Corpus" is to create a representative language resource for Chinese and Uyghur in order to facilitate the study of the relations between the two languages. The project started at the end of 2007 and is expected to complete in early 2010. More specifically, the project aims to build and annotate a medium sized and balanced Chinese-Uyghur parallel corpus aligned at sentence level using a set of tools. The Chinese-Uyghur parallel corpus is expected to be used in linguistic research, teaching, translation studies as well as in applications such as machine translation.

Before we present the corpus data, we give a short overview of the languages involved as they belong to different language families.

2.1. Chinese and Uyghur

Chinese belongs to the Han-Tibetan language family. It is the most commonly used language in China, and one of the most commonly used languages in the world. Modern Chinese is an analytic language, and functions such as number in nouns or tense in verbs are expressed through syntax (word order and sentence structure) rather than morphology. One key feature is that all words in Chinese have only one grammatical form, as the language lacks declension, or any other inflection (there are minor exceptions). Chinese features subject-verb-object (SVO) word order similar to English.

Uyghur is a Turkic language of the Altaic family, spoken by about 10 million Uyghur people in Xinjiang Uyghur Autonomous Region of China and around the world. Uyghur is a suffixing and agglutinative language; in most cases, there is a one-to-one relationship between morpheme and function. The verbal system is rich and verbs have markers for tense, mood, aspect, and voice, as well as agreement markers in terms of the features like person and number. Considering the syntactic characteristics, Uyghur is a left-branching type of language, where the dependents precede their head (i.e. adjective or genitive modifier precedes the modified head and the object precedes the verb). Uyghur is rather free in its word order, which is based on the morphological structure. Uyghur has subject-object-verb (SOV) word order but other orders are possible depending on which element is put into the focus in the discourse. Modern Uyghur uses a modified Arabic script as its writing system.

As the official languages in Xinjiang, Chinese and Uyghur are widely used in many fields such as education, communication and publication. Bilingualism in Xinjiang requires many translations from Chinese to Uyghur or in the opposite direction. Therefore, it is possible and essential to build a Chinese-Uyghur parallel corpus for multiple purposes. So, we began to build a sentence aligned general corpus of contemporary Chinese and Uyghur.

2.2. Corpus data collection

It is now a well-known fact that a corpus is more than just a collection of electronic texts. Corpus data have to be selected with care with respect to the intended applications, which means that a corpus should contain texts of different domains and different genres in reasonable proportions so that the resulting corpus can be a reasonable reflection of language use.

In this project, we emphasize quality with regard to content and translation. We focus on a collection of written texts translated from Chinese into Uyghur to build a balanced corpus of source and target languages. However, when we decided to construct the corpus, we found that it was not easy to construct a perfectly balanced Chinese-Uyghur corpus. That is because there are not many electronic bilingual texts of Chinese and Uyghur available.

Therefore, in the first step we collected as many good quality bilingual texts as possible. Bilingual texts in electronic format were collected from several resources such as books, newspapers, journals and the Internet. Some texts that could not be obtained in electronic format were scanned, OCRed and proofread. After one year's effort, we have collected a total of three million Uyghur words and Chinese characters of untagged Chinese-Uyghur parallel texts.

In the second step, the texts have been normalized in their form (text-only), size and field in order to keep the balance in the corpus collection. Here, balance means the weighting of different fields, styles and genres in the corpus. Obviously, it does not mean to have equal amounts of texts from different domains that are covered in the corpus. In addition, to make it possible to include as many different writers and translators as possible, text extracts of 10,000 - 15,000 words will be used instead of complete books. Each text extract will start at the beginning of the book and, if possible, end at a chapter boundary. This will ensure coherent if not complete texts. Where texts are very short, they will, however, be included in their entirety.

So far, over two million Uyghur words and Chinese characters of bilingual texts, totalling 250 texts, have been included into the raw corpus. After sampling and normalization, the corpus texts cover a variety of genres, such as fiction, news stories, popular science, government documents, legal texts and daily conversation. Presently, the size of corpus is smaller than we expected because it is not easy to obtain such digital text data that we have to process before inclusion in the corpus. Proportions of different genres in the corpus are shown as Table 13-1.

Table 13-1. Genres and their proportions in the corpus

Genre	Number of samples	Chinese characters	Uyghur words	Total	Proportion
Fiction	55	386,880	195,394	582,274	29%
News	24	120,900	66,430	187,330	9.3 %
Science	85	298,383	155,407	453,790	22.6%
Government documents	34	145,080	82,000	227,080	11.3 %
Legal texts	40	277,294	149,890	427,184	21.3%
Daily conversation	12	89,078	41,432	130,510	6.5%
Total	250	1,317,615	690,553	2,008,168	100%

Table 13-1 compares the actual numbers of tokens in different genres as well as their corresponding proportions in both languages. As can be seen, fiction is the most important and frequent genre in the corpus, which accounts for 29%. News, popular science, government documents and legal texts account for 9.3%, 22.6%, 11.3% and 21.3% of the corpus texts respectively. As a special part of the corpus, we also collected some daily conversation texts from some coursebooks that recorded everyday conversations in Chinese and Uyghur. However, conversational texts occupy a smaller proportion (6.5%) than other genres because of the difficulty of obtaining such data. The Chinese-Uyghur parallel corpus is a medium-sized corpus which is roughly balanced in terms of the proportions for different genres.

2.3. The workflow of the parallel corpus construction

To facilitate the construction of the parallel corpus, we also developed a systematic workflow based on our examination of the whole process of the corpus construction.

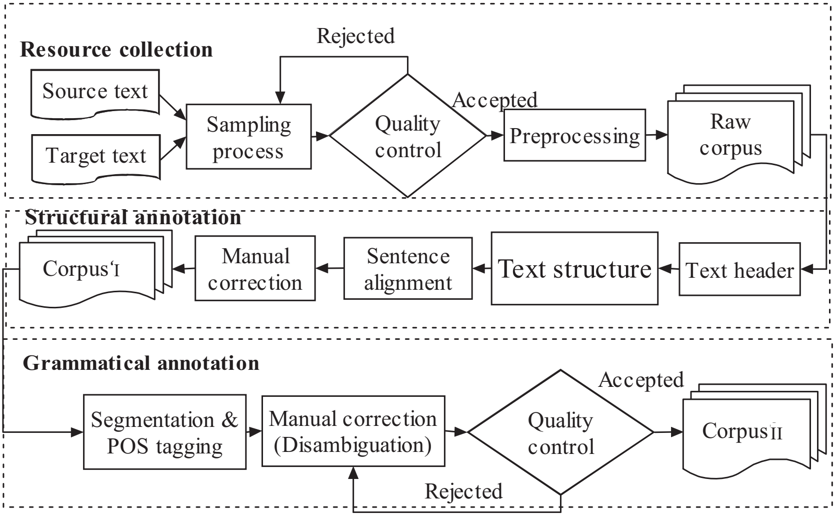


Figure 13-1. The workflow of the parallel corpus construction

According to the workflow, the construction process of the Chinese-Uyghur parallel corpus includes three steps: resource collection, structural annotation and grammatical annotation. In the first step, bilingual texts are collected, preprocessed and included into the raw corpus. In the second step, Chinese and Uyghur texts are marked up separately with the textual attributes (text header) and textual structural information. Parallel alignment at paragraph and sentence level is also undertaken in this step with the alignment tools. After this step, we can obtain a sentence aligned parallel corpus (Corpus I). However, this corpus has not any grammatical annotations such as part-of-speech (POS) tagging and parsing. As Chinese is written as running strings of characters without white spaces as word boundaries, all Chinese texts must be tokenized (segmented). Therefore, in the third step, all Chinese texts are tokenized and annotated with POS tags, while all Uyghur texts are lemmatized and annotated with POS tags. In Figure 13-1, Corpus II is the resulting parallel corpus that includes POS tags. As we can see from Figure 13-1, processing results in every step are checked for errors and corrected manually.

3. Annotation of the parallel corpus

The following steps give an overview of the annotation procedure and the tools used.

3.1. Preprocessing

The original materials that we obtained from publishers and the Internet are noisy and in various formats. In this step, text noises, i.e. irrelevant HTML tags, figures, tables etc., are removed from the texts before the texts are included in the corpus. Then, the various formats (e.g. rtf, doc, pdf) are converted to plain text format. All texts are encoded according to international standards by using UTF-8 (Unicode) and resaved using unique filenames to indicate a pair of parallel files. In some cases, we scanned and proofread the materials and, where necessary, corrected OCR errors to ensure that the plain text file is complete and correct. After the first step, all texts are pure texts in the same format.

3.2. The markup of the parallel corpus

The parallel corpus can only be useful after it is marked up. In our study, in order to make the corpus application-independent and easier for data exchange via the Internet, all parts of the corpus are clearly marked up using a uniform markup scheme. For this reason, Extensible Markup Language (XML) for corpus encoding is adopted. The XML-based markup scheme (see Table 13-2 for XML tags), very similar to that of Chang (2004), has been designed and all texts are marked up in this XML format, which is compliant with Corpus Encoding Standard (CES). The occurrence of tags and attributes of XML markup is defined formally by a Document Type Definition (DTD) file.

According to this markup scheme, all Chinese texts and Uyghur texts are encoded separately. Each text, no matter what language it is, is composed of a text head and a text body. All the global textual attributes are placed in the text head; the monolingual structural tags, grammatical tags and the text itself are put within the text body. Alignments are indicated by an alignment attribute in the text body of both languages. The structure of each text is illustrated as follows:

```
<?xml version="1.0" encoding= "Unicode" ?>
<TEXT>
<TEXT_HEAD>
Text header
</TEXT_HEAD>
<TEXT_BODY>
Text body
</TEXT_BODY>
</TEXT>
```

At present, we have only inserted some basic annotation. Considering the reliability of the corpus tools and possible applications of the corpus, we carried out the following four types of encoding.

1) Global textual attributes (text head). Global textual attributes are attributes applied to every text in the corpus. They are features that specify the domain, style, mode of the text (i.e. whether a text is written or spoken), the author of a text, the translator of a text, the time when a text was authored and translated, the title and subtitle of a text and so on. The global textual attributes will facilitate special research based on the corpus; for example, language researchers might be interested only with texts belonging to a particular domain, and they can easily extract all texts belonging to that domain.

Table 13-2. XML markup scheme for the Chinese-Uyghur Parallel Corpus

		XML element	Attribute
Text		<TEXT>	
Text Head	Text head	<TEXT_HEAD>	
	Chinese title	<CN_TITLE>	
	Uyghur title	<UY_TITLE>	
	Chinese subtitle	<CN_ SUBTITLE>	
	Uyghur Subtitle	<UY_ SUBTITLE>	
	Direction	<DIRECTION>	
	Author	<AUTHOR>	
	Translator	<TRANSLATOR>	
	Style	<STYLE>	
	Field	<FIELD>	
	Mode	<MODE>	
	Time (CN)	<CN_TIME>	

	Time (UY)	<UY_TIME>	
	Chinese source	<CN_SOURCE>	
	Uyghur source	<UY_SOURCE>	
	Text size (CN)	<CN_SIZE>	
	Text size (UY)	<UY_SIZE>	
	Collector	<COLLECTOR>	
Text Body	Text body	<TEXT_BODY>	
	Paragraph	<p>	id
	Sentence	<s>	id
	Sentence alignment unit	<a>	id, no
	Word	<w>	id, pos, lemma

2) Monolingual text structure markup. Monolingual textual structural annotation deals with text units of different levels. At present, the boundaries of paragraphs and sentences have been annotated in the corpus.

3) Parallel alignment annotation. Parallel alignment annotation establishes the correspondence between the language units of the original texts and their translations. So far, the corpus is aligned only at the sentence level. Word alignment of the corpus seems still unpractical for the massive labour required and lacking of reliable tools.

4) Grammatical annotation. After the structural annotation, the corpus is annotated grammatically for multiple purposes. Grammatical annotation actually covers any descriptive or analytic notations applied to raw language data. Although the added notations may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tagging, syntactic analysis, Named Entity identification, co-reference annotation, and so on, grammatical annotation in our corpus just includes word boundary detection and POS (part-of-speech) tagging for Chinese texts and lemmatization for Uyghur texts. Detection of word boundaries in Chinese texts, also known as word segmentation, is a very basic process in Chinese corpus construction. For the Uyghur words, the detection of word boundaries is the word tokenization or word lemmatization that is a process for reducing inflected (or sometimes derived) words to their stem, base or root form. External morphological analyzer and part-of-speech tagger are used for the specific language to the grammatical annotation.

3.3. Sentence alignment

For a parallel corpus, the most important annotation is alignment, especially sentence alignment, which is a minimal and essential requirement for a parallel corpus. Aligning Chinese-Uyghur parallel texts, however, is very difficult because of the great differences in the syntactic structures and writing systems of the two languages.

A number of alignment techniques have been proposed for other language pairs, varying from statistical methods to lexical methods. There are basically three kinds of approaches to sentence alignment: the length-based approach (Gale and Church 1991), the lexical approach (Kay and Röscheisen 1993), and the combination of the two (Chen 1993, Wu 1994).

In our project, we developed a Chinese-Uyghur sentence aligner on the basis of other sentence alignment methods. In our case, we introduced an anchor sentence based multilevel sentence alignment method, in which some sentences are used as ‘anchors’ and two steps are applied. In the first step, some types of lexical information such as proper names, technical terms, numbers, punctuation marks, location information, and length information are used to generate anchor sentences that satisfy some conditions. In the second step, texts are divided into several segments by using anchor sentences as boundaries, and then sentences in each segment are aligned by using the length-based approach. This method avoids complex computing and error spreading because of its subsection technique. Experiment results show that the method is robust, and fast enough to be practical and more accurate than previous methods for Chinese-Uyghur sentence alignment in multi-domain texts.

All sentences in parallel texts are aligned automatically by using the Chinese-Uyghur sentence aligner, and then checked semi-automatically with the help of a sentence checker for alignment errors.

3.4. The parallel corpus tool set

The corpus texts are processed semi-automatically by using various tools for markup, annotation, and alignment, and to make manual correction easier and more straightforward for users with limited computer skills. To facilitate the construction of the Chinese-Uyghur parallel corpus, we have used a set of corpus tools developed by our team or others. So far, we have the following tools in use: (1) The corpus builder: it is a text file editor developed by us for the parallel corpus construction and for future application. It is similar to UltraEdit but has very strong application in corpus building such as text editing, XML encoding, and text indexing; (2)

The Chinese-Uyghur sentence aligner: This is a sentence alignment tool developed specifically for the Chinese-Uyghur parallel corpus; (3) The Chinese segmentation and POS tagging program: This software tool, which is developed at Peking University, is used to segment and POS-tag all Chinese texts. (4) Uyghur morphological analyzer: The analyzer, developed by the researchers of Xinjiang University, has the function of Uyghur word tokenization (lemmatization) and POS tagging. The four tools have been heavily used in the construction of the Chinese-Uyghur parallel corpus.

4. Preliminary results of the corpus application

At the time of writing, only one year after its beginning, our project has not progressed far enough for us to carry out any major corpus-based studies. We can only present some very preliminary findings and some potential applications of the corpus at this point. These should, however, be of some interest in showing the sorts of analyses which can be carried out using the parallel corpus.

4.1. Some basic statistics

This section compares the number of paragraphs, sentences and words (characters) in source and target texts to find out whether different genres have an influence on translation style.

As noted earlier, the Chinese-Uyghur parallel corpus includes the texts of several genres such as fiction, newspaper stories, popular science, government documents, legal texts and daily conversation. The statistics given in Table 13-3 are based on all parallel texts in the corpus. The table shows the ratios between the number of paragraphs, S-units (orthographic sentences) and words (characters) in Chinese and Uyghur texts of different genres.

As we can see, the last three columns of Table 13-3 are ratios between the number of paragraphs, sentence, and words in Chinese texts and Uyghur texts respectively. It is not difficult to find that Uyghur translations of news and fiction use more paragraphs than original Chinese texts. The ratio of the number of sentences in Chinese originals to the number of sentences in Uyghur translations is 0.93 in fictions while it is 0.91 in popular science. This means that Uyghur translations of Chinese fiction and scientific articles have more sentences than Chinese original texts. There is an overall tendency for the Uyghur translated texts to contain more sentences than original Chinese texts. This is also true for government documents and news texts. In other words, every 95 Chinese sentences are translated as 100 Uyghur sentences on average. Another interesting result revealed by the statistics is that translators use one Uyghur word to translate 1.91 Chinese characters on average. However, the ratio changes slightly in different genres, for example, 2.15 Chinese characters for one Uyghur word in the conversational data while 1.77 Chinese characters are translated as one Uyghur word in government documents. This means that Uyghur translation of government documents tend to be more literal than that of daily conversations while translators tend to use fewer words in Uyghur to express the same ideas in daily conversations.

Table 13-3. Some basic statistics of the Chinese-Uyghur parallel corpus

	Genre	Paragraph (Cn/Uy)	Sentence (Cn/Uy)	Character/ word
1	Fiction	0.96	0.93	1.98
2	News	0.96	0.97	1.82
3	Science	1	0.91	1.92
4	Government documents	0.99	0.96	1.77
5	Law texts	1.01	0.99	1.85
6	Daily Conversation	1	1	2.15
Mean		0.99	0.95	1.91

As far as sentence length is concerned, different genres have very different mean sentence lengths in the source (Chinese) and target (Uyghur) languages. Figure 13-2 compares the mean sentence length in Chinese and Uyghur.

From Figure 13-2 we can see that sentences in news and government documents of original or translated texts are very long (47.5 and 46.5 characters for Chinese versus 25.3 and 25.2 words for Uyghur respectively), which is 2-4 times of daily conversation and fiction (10.63 and 24.86 characters for Chinese versus 4.95 and 10.63 words for Uyghur respectively). According to the statistic results of sentence length, the genre of a text has a great influence on its mean sentence length. Moreover, this influence is almost the same in original text and translated text. From this, we can also find that longer sentences tend to have longer translations, and that shorter sentences tend to have shorter translations. This proves that the correlation between the length of the sentence (or a paragraph) and the length of its translation is extremely high. The high correlation suggests that sentence length might be a strong clue for sentence alignment and other translation studies.

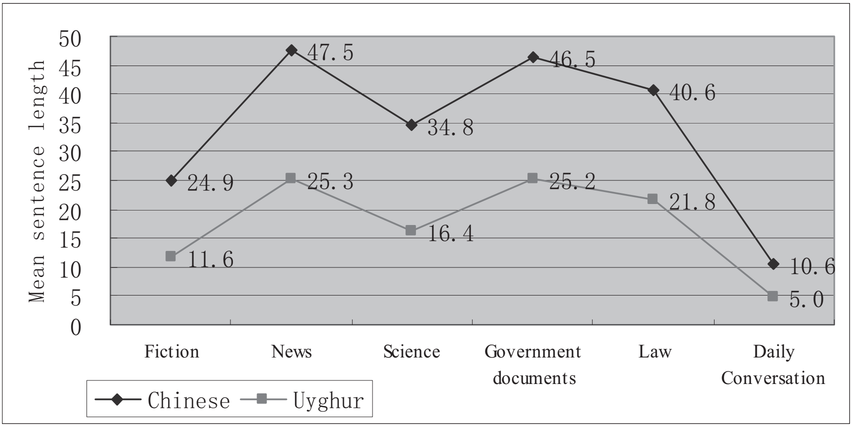


Figure 13-2. Mean sentence length in the source and target languages

4.2. Future applications of the parallel corpus

Parallel corpora are very useful for all types of cross-linguistic research. The value of a parallel corpus grows with its size and with the number of languages for which translations exist. The creation of the Chinese-Uyghur parallel corpus is great progress in corpus-based study of Uyghur language. At the current stage, however, we are chiefly focusing on developing the corpus application tools and have not been able to carry out large-scale investigations. The examples given in the previous section should be sufficient to show the possibilities of using the Chinese-Uyghur parallel corpus.

(1) The parallel corpus offers specific uses and possibilities for contrastive and translation studies. It gives new insights into the languages that are not likely to be noticed in studies of monolingual corpus; it can be used for a range of comparative purposes and increase our knowledge of language-specific, typological and cultural differences, as well as of universal features; it illuminates differences between source texts and translations, and between native and non-native texts.

(2) The parallel corpus can be used for bilingual lexicography. A parallel corpus is necessary to clarify some terminological issues and acquisition of bilingual translation patterns; researchers can use it in bilingual dictionary making with the help of the concordance tools.

(3) The parallel corpus has applications in building statistical machine translation systems and translation memories. Useful data or knowledge could also be extracted from the bilingual corpus based on statistical model to provide translation examples for MT systems.

5. Conclusions and future work

In this chapter, we have presented the process of constructing the Chinese-Uyghur parallel corpus, including data collection, preprocessing, markup and annotation, and sentence alignment. Some preliminary results and potential values of the corpus have also been discussed.

The size of the corpus is relatively small at present. We would like to expand the corpus by including other texts, both fiction and non-fiction, and to develop a Chinese-Uyghur word alignment tool to do automatic word alignment in the near future. We hope that the corpus will provide ample material for text-based contrastive studies as well as for more specialized translation studies in the future.

Notes

We are grateful to the China National Foundation of Social Sciences for supporting our project "Construction and Application of Chinese-Uyghur Bilingual Corpus" (Grant Reference 07xyy019). We would like to thank Prof. Hou Min for her guidance and great help during the study.

References

Areta N., Gurrutxaga A., Leturia I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Sologaistoa A. (2007), "ZT Corpus: Annotation and tools for Basque Corpora", in *Proceedings of Corpus Linguistics 2007*. University of Birmingham, Birmingham, UK, 2007.

Bai, X., Chang, B. and Zhan, W. (2002), "The construction of a large-scale of Chinese-English parallel Corpus", in *Proceedings of National Machine Translation Conference 2002*, 124-131. Beijing: Electronic Industrial Publisher.

Baker, M. (1993), “Corpus linguistics and Translation Studies: Implications and applications”, in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology. In Honor of John Sinclair*, 233-250. Amsterdam: John Benjamins.

—. (1995), “Corpora in translation studies: An overview and some suggestions for future research”. *Target* 7(2): 223-243.

Beata B. Megyesi, Anna Sågvall Hein, and Eva Csato Johanson. (2006), “Building a Swedish-Turkish Parallel Corpus”, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2130-2133. Genoa, Italy, 2006.

Chang, B. (2004), “Chinese-English Parallel Corpus construction and its application”, In *PACLIC 18*, 283-290. Waseda University, Tokyo, 8-10 December, 2004.

Chen, S. F. (1993), “Aligning sentences in bilingual corpora using lexical information”, in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 9-16. Columbus, Ohio, 1993.

Gale, W. and Church, K. (1993), “A program for aligning sentences in bilingual corpora”. *Computational Linguistics* (19) 1:75-102.

Johansson, S and Ebeling, J. (1994), “The English-Norwegian Parallel Corpus: Introduction and applications”, in *Proceedings of International Conference on Cross-Language Studies and Contrastive Linguistics*. Rydzyna, Poland, 15-17 December, 1994.

Kay, M., Röscheisen, M. (1993), “Text-translation alignment”. *Computational Linguistics* 19(1): 121–142.

Kennedy, G. (2000), *An Introduction to Corpus Linguistics*. Beijing: Foreign Language Teaching and Research Press.

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

McEnery, T. and Xiao, R. (2002), “Domains, text types, aspect marking and English-Chinese translation”. *Languages in Contrast* 2(2): 211-229.

McEnery, T. and Xiao, R. (2007), “Parallel and comparable corpora: What is happening?”, in M. Rogers and G. Anderman (eds.) *Incorporating Corpora. The Linguist and the Translator*, 18-31. Clevedon: Multilingual Matters.

Simard, M., Foster, G. and Isabelle, P. (1992), “Using cognates to align sentences in bilingual corpora”, in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine translation (TMI92)*, 67-81. Montreal, Canada, 1992.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006), "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages", in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, May 24-26, 2006.

Tiedemann, J. and Nygaard, L. (2004), "The OPUS corpus − parallel & free", in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'2004)*. Lisbon, Portugal, 26-28 May, 2004.

Wu, D. (1994), "Aligning a parallel English-Chinese corpus statistically with lexical criteria", in *Proceedings of ACL-94: 32nd Annual Meeting of the Assoc. for Computational Linguistics*, 80-87. LasCruces, NM, 1994.

Xiao, R., McEnery, T., Baker, P. and Hardie, A. (2004), "Developing Asian language corpora: Standards and practice", in *Proceedings of the Fourth Workshop on Asian Language Resources*, 1-8. Sanya, Hainan Island, 25 March, 2004.

CHAPTER FOURTEEN

DESIGNING AND DEVELOPING A PARALLEL CORPUS BASED ON MEDIA SUBTITLES

CHONG ZHU

1. Introduction

Corpora have become one of the most important tools in various linguistic research and parallel corpora are especially useful in translation studies and language teaching, etc. (Sinclair 1991; Baker 1995, 1998, 1999; Barlow 2000; Bowker 2001; Hunston 2002; Kennedy 1998; Laviosa 1998; Wang *et al*. 2005). However, due to the difficulties involved in their creation, especially text alignment, parallel corpora are still few in number and small in size compared to other types (Wang *et al*. 2005). In the meantime, a fast growing and easily available type of sources, namely, film and television subtitles, has been overlooked by many corpus builders.

Most audiovisual programs, except the mute or wordless ones, which are few in number compared to vocal ones, when transmitted into a foreign country, often need to be transferred linguistically. Two ways of achieving this are most common: dubbing and subtitling (Baker 1998: 74, 244; Ma 2005: 6-7).

Compared to dubbing, subtitling is a method that costs much less time and money. In subtitling, the translated lines, referred to as subtitles or captions, are “presented simultaneously on the screen” while the original sound and voice tracks are intact and played (Baker 1998: 245). Dialogues of a program are first translated and then each line is given a time code, or in Baker’s words, the lines are all “time-cued” (Baker 1998: 245).

Baker (1998: 248) pointed out that with the advances of teletext technology, subtitling and especially personal subtitling would become “standards of language transfer” in the future. Since she made that claim, a decade has passed, and it is not only the improvement of teletext technology but also the prevalence of DVDs and other media that have promoted subtitling as an important standard of audiovisual translation.

Nowadays, in addition to the professionals, an increasing number of amateur translators are also practicing screen translation and their products are abundant on the Internet. These amateur works differ greatly in quality, but their quantity increases constantly and the sheer volume of these translation products is stunning. It would be a great waste for corpus and especially parallel corpus builders to overlook this large source of multilingual materials.

2. Overall design and structure

The design of the Multi-Media Subtitle Corpus (MMSC), like the creation of any corpus, involves overall design and planning, data collection, data encoding and storing, and data processing, etc. (Sinclair 1991: 14-22, Yu 2003: 83-91). In the case of creating a parallel multi-media subtitle corpus, it is necessary to single out one particular step, namely, text alignment in data processing (Yu 2003: 93-99).

2.1. Aims and guiding principles

The overall design of a corpus centres on its building targets which establish the controlling guidelines of the whole project (Sinclair 1991: 15, Yu 2003: 83). Therefore, it is necessary to introduce these concepts before the detailed steps of construction are presented.

1. Aims: to build a database for
    a) the theoretical and practical studies on subtitle translation;
    b) facilitating translator training;
    c) facilitating language teaching and learning;.
2. Guiding principles: the subject corpus should:
    a) contain as many texts as possible;
    b) be open and extensible in a way that:
        i. it should supply researchers with transparent and friendly interfaces and standards so that they can submit new data and correct the errors included or input by mistake;
        ii. it should be able to be easily embedded into a platform that can be accessed through the World Wide Web;
        iii. it should have interfaces for multi-media data insertion and linking, which enables further efforts to make the corpus a multi-media tool for translation studies and language teaching.

c) only include suitable or eligible texts in terms of their translation quality.

2.2. Text categories and sub-categories

As mentioned earlier, the subject corpus mainly contains subtitles from films and television programs. Thus the author ventures to categorize the texts in a way consistent with the classification of these programs as shown by Table 14-1:

Table 14-1. Categories of programs included in MMSC

Categories	Film	Television
	adventure	documentary
	kids & family	situation comedy
	musicals	news
	westerns	science fiction
	thrillers/suspense	classics
	animated film	comedy
	biographical	animated series
	historical	war
	science fiction	
	documentary	
	war	
	comedy	
	romance	

2.3. Database structure and organization

The adoption of XML as the way of storing data has become more and more popular in corpus creation. Like the Chinese-English parallel corpus designed by researchers at Peking University (Chang and Bo 2003), the texts in the MMSC are also stored in XML files. Each subtitle is stored as one XML file. The query and concordance interface works as a webpage, which retrieves and displays texts in a desired format in the user’s web browser.

3. Corpus development

3.1. Text selection and collection

Olohan (2004: 47) argues that the "subjectivity of decisions" concerning which texts to be included into or excluded from a corpus causes a lot of trouble in the creation of a corpus and the translation studies based on it. So is the situation in building the MMSC corpus.

Subtitles, as mentioned before, are extremely abundant in number and easily available through different ways, including the Internet, DVDs, etc. Careful selection should be done before the subtitles are taken into the corpus in order to ensure the quality of the corpus itself. The author assumes three ways to decide whether a subtitle should be included into the corpus:

1. whether the subtitle comes from a credible source
2. whether the subtitle enjoys high recommendation on the Internet
3. whether the subtitle passes the examination of the corpus builders

When a subtitle is found to be qualified and desirable to be included, it then must be retrieved from its source medium. The author obtains all the subtitles needed from two main sources: DVDs and the Internet.

All the raw materials collected are stored in a folder named 'Draft' and need to be further treated before they can be imported into the corpus.

3.2. Text pre-processing

The texts collected and stored must be further processed before they are ready for alignment. This process includes file type conversion and text formatting. The subtitle files are first converted into plain text files which are stored in a folder named 'lexical version'. Then these plain text files are treated according to the following rules:

1. In each subtitle, all the lines with the same time code must be arranged into one line.
2. A time code must be in the same line and followed by the texts under it.
3. All time codes contained in the subtitles should be retained.

After being treated through these steps, all the texts are in a format as shown in the examples given in Table 14-2.

Table 14-2. Example lines taken from subtitles of *Big Shot's Funeral*[1]

Examples of subtitle lines
#1|000037.203>000039.671^Do you know who Tyler is?
#2|000041.174>000044.166^Yeah. He's a da waar director.
#3|000044.377>000046.777^What's a da waar?

The texts with such an internal structure are stored in a folder named 'Pre-processed Version' and ready to be aligned.

3.3. Text alignment

The alignment in the MMSC corpus is not strictly on sentence level but on 'tod' level. Tod is a word coined by the present author to represent a single unit of texts in a subtitle, which is included in one single time-stamp. In effect, it refers to each caption line or set of lines shown on the screen during one particular time span as one unit.

In Table 14-2, there are 3 separate tods, each with a particular time stamp in the format like ######.###>######.### which indicates the reference beginning time and end time when the subtitle is printed on the screen.

The whole alignment process is centred on the time-cue information contained in each tod.

If a pair of tods from two subtitles of the same program share the same time code, it would be easy to align the two tods. All that has to be done is to extract the time code and the texts associated with it, and then store them in the appropriate format. This situation between the two tods is called exact match situation.

Unfortunately, the exact match situation is extremely rare in subtitles. The time cues in most subtitles are much more complicated. Different time code match relations result in different aligning algorithms, thouth a unified algorithm to govern all the patterns, if possible, is preferred.

3.3.1. Tod match relations

The different tod match relations are described in this section with examples and graphs:

1. Exact match relation:

Exactly matched subtitles are quite rare, where parallel lines share exactly the same time codes. For example:

Table 14-3. Lines taken from subtitles of BBC's *Blue Planet*[2]

English Lines	Chinese Lines
#1|000033.566>000036.694^Dwar fed by the vast expanse of the open ocean	#1|000033.566>000036.694^因为海洋的辽阔而显得渺小的
#2|000036.870>000040.397^the biggest animal that has ever lived on our planet.	#2|000036.870>000040.397^是这个星球上最巨大的生物 舌头有一头大象那么重

2. Random match relations:

a) Self-contained

A self-contained line is a line that does not share any part of its time span with any line from the subtitle of the other language, as exemplified in Table 14-4 and Figure 14-1.

Table 14-4. Lines taken from subtitles of Episode I of *Star Wars* movies[3]

English Lines	Chinese Lines
#1|000159.886>000200.944^-Captain. - Yes, sir?	#1|000021.228>000025.927^很久以前在遥远的银河系…
#2|000202.521>000204.648^Tell them we wish to board at once.	#2|000030.571>000034.803^星际大战
#3|000204.757>000207.385^-[Machinery Beeping] - With all due respect	#3|000042.917>000049.413^首部曲：威胁潜伏

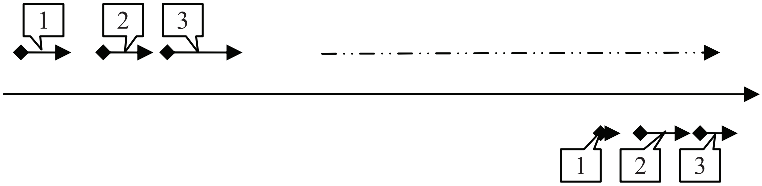


Figure 14-1. Graphic representation of self-contained lines

The corresponding relations between these subtitle lines can be roughly reflected in Figure 14-1, where the solid line in the middle stands for the time line and the arrows above it are Chinese subtitles and those below are English ones. In the figure, the first 3 lines from the Chinese subtitle do not share any time span with any English line. These solitary lines are defined as self-contained. Self-contained lines do not have parallel counterparts. Therefore they should be separately treated and marked.

b) Contain and Contained

'Contain' and 'contained' are a pair of complementary relations. When the time span a Chinese line occupies is larger than and totally covers that of an English line, the Chinese line is said to contain the English line and vice versa. In Figure 14-2, the Chinese line (above the time line in the middle) numbered 1 is contained by the English line (below the time line) numbered 4; the English line numbered 5 contains the Chinese line numbered 2.

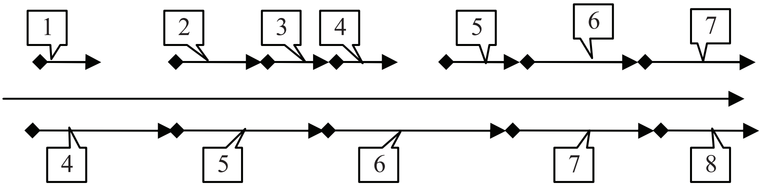


Figure 14-2. Graphic representation of contained and overlapped lines

c) Overlap

Overlap is the situation where one line does not contain and is not contained by any other lines (of the other language) while it is not self-contained either. In Figure 14-2, Chinese line 3 overlaps with English line 5; Chinese line 5 overlaps with English line 6; Chinese line 7 overlaps with English lines 7 and 8.

These three basic match relations can result in a multiplicity of tod match situations, which must be dealt with in the alignment process.

3.3.2. Tod match situations in alignment

At the beginning of the aligning process, the first Chinese line (the leftmost one above the time line) is taken out and examined against the first English line (it does not matter which language is taken first in this discussion). There may be five possible situations which are discussed below.

1. The Chinese line is self-contained as shown in Figure 14-3:

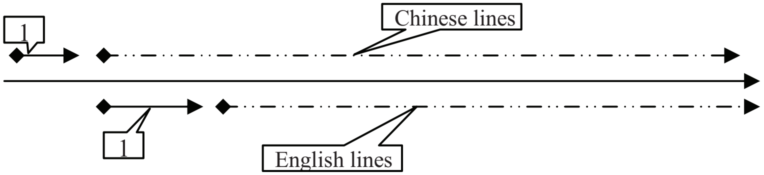


Figure 14-3. Graphic representation of self-contained Chinese line

2. The English line is self-contained as shown in Figure 14-4:

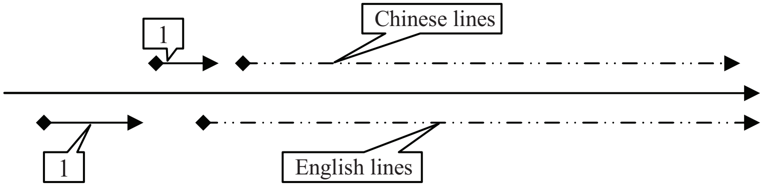


Figure 14-4. Graphic representation of self-contained English line

3. The Chinese line contains the English line as shown in Figure 14-5:

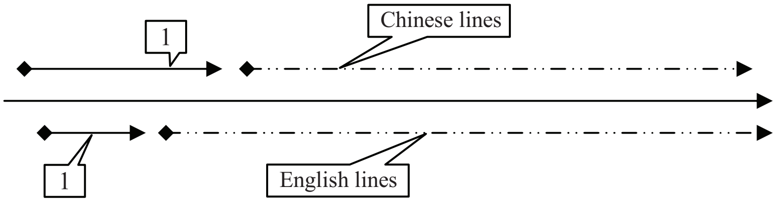


Figure 14-5. Graphic representation of Chinese line containing English line

4. The Chinese line is contained by the English line as shown in Figure 14-6:

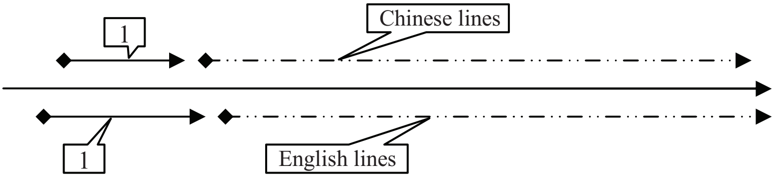


Figure 14-6. Graphic representation of Chinese line contained in English line

5. The Chinese line overlaps with the English line as shown in Figures 14-7 and 14-8:

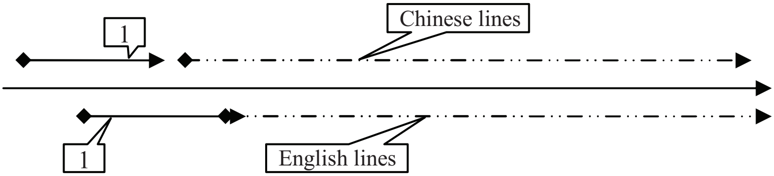


Figure 14-7. Graphic representation of overlapped lines

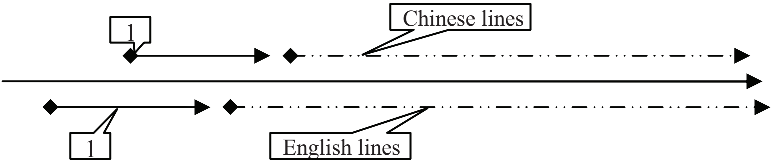


Figure 14-8. Graphic representation of overlapped lines

These are actually the only possible situations that can happen from the perspective of the three basic match relations. A governing strategy to deal with all these situations is fundamental to text alignment.

3.3.3. The aligning strategies

Several points need to be made clear before the explanation of the strategies:

In the description of the strategies, those words in capitals like ‘MARKED WITH A FLAG’ and ‘DROPPED’ are macro-processes that are to be realized in a certain computer language.

‘The first line’ in this context refers to the current Chinese line or English line which is being examined or processed; ‘the second line’ refers to the line immediately to the right of ‘the first line’ on the time axis and so on.

The strategies are described as follows:

The Governing Strategy: DECIDE the relation between the first Chinese line and the first English line for further treatments.

Strategy 1. When the first Chinese line is self-contained:

The Chinese line is MARKED WITH A FLAG ‘self-contained’ and DROPPED from the aligning process; the second Chinese line is SET as the first (current) Chinese line and examined with the first English line; repeat the Governing Strategy.

Strategy 2. When the first English line is self-contained:

The English line is MARKED WITH A FLAG ‘self-contained’ and DROPPED from the aligning process; the second English line is SET as the first (current) English line and examined with the first Chinese line; repeat the Governing Strategy.

Strategy 3. When the first Chinese line contains the first English line:

DECIDE whether the first Chinese line contains the second English line and:

- ◆ If yes, COMBINE the first and second English lines and SET the combination as the first English line and repeat this strategy (Strategy 3).
- ◆ If not, DECIDE whether the first Chinese line overlaps with the second English line:
    - ▫ If yes, COMBINE the first and second English lines and SET the combination as the first English line; repeat the Governing Strategy or go directly to Strategy 5. (The situation has virtually turned into Situation 5 where the Chinese line is ahead of the English line, and thus can be handled by Strategy 5.)
    - ▫ If not, consider the first Chinese line and the first English line as parallel and MARK them as ALIGNED; SET the second Chinese line and the second English line as the first

Chinese line and the first English line respectively; repeat the Governing Strategy.

Strategy 4. When the Chinese line is contained by the English line:

DECIDE whether the second Chinese line is contained by the first English line and:

- If yes, COMBINE the first and second Chinese lines and SET the combination as the first Chinese line and repeat this strategy (Strategy 4).
- If not, DECIDE whether the second Chinese line overlaps with the first English line:
  - If yes, COMBINE the first and second Chinese lines and SET the combination as the first Chinese line; repeat the Governing Strategy or go directly to Strategy 5. (The situation has virtually turned into Situation 5 where the Chinese line is behind the English line, and thus can be handled by Strategy 5.)
  - If not, consider the first English line and the first Chinese line as parallel and MARK them as ALIGNED; SET the second English line and the second Chinese line as the first English line and the first Chinese line respectively; repeat the Governing Strategy.

Strategy 5. When the Chinese line overlaps with the English line:

There are two kinds of overlap situations: the Chinese line is ahead of the English line or the Chinese line is behind the English line. To say one line is behind the other means that the end of one line is behind the end of the other one. This is shown by Figures 14-9 and 14-10:

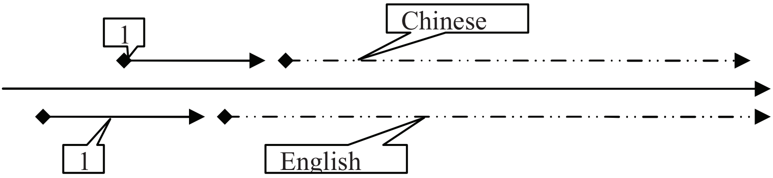


Figure 14-9. Graphic representation of English line ahead of Chinese line

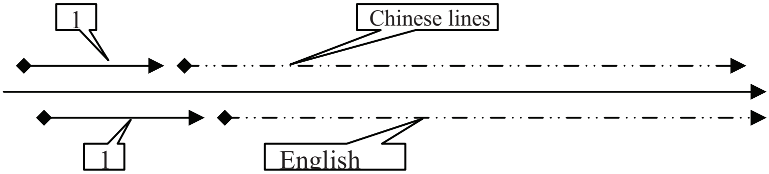


Figure 14-10. Graphic representation of English line behind Chinese line

Hence it is necessary to deal with them respectively:

For the Chinese-line-ahead situation, namely, when the first Chinese line is ahead of the first English line:

DECIDE whether the first Chinese line contains the second English line and:

- If yes, COMBINE the first and second English lines and SET the combination as the first English line; repeat this strategy (Strategy 5).
- If not, DECIDE whether the first Chinese line overlaps the second English line and:
    - If yes, COMBINE the first and second English lines and SET the combination as the first English line; repeat the Governing Strategy or go directly to Strategy 4. (The situation has virtually turned into Situation 4 and thus can be handled by Strategy 4.)
    - If not, consider the first Chinese line and the first English line as parallel and MARK them as ALIGNED; SET the second Chinese line and the second English line as the first Chinese line and the first English line respectively; repeat the Governing Strategy.

For the Chinese-line-behind situation, namely, when the first Chinese line is behind the first English line:

DECIDE whether the second Chinese line is contained by the first English line and:

- If yes, COMBINE the first and second Chinese lines and SET the combination as the first Chinese line; repeat this strategy (Strategy 5).
- If not, DECIDE whether the second Chinese line overlaps the first English line and:
    - If yes, COMBINE the first and second Chinese lines and SET the combination as the first Chinese line; repeat the

Governing Strategy. (The situation has virtually turned into Situation 3 and thus can be handled by Strategy 3.)

- If not, consider the first Chinese line and the first English line as parallel and MARK them as ALIGNED; SET the second Chinese line and the second English line as the first Chinese line and the first English line respectively; repeat the Governing Strategy.

3.3.4. The aligning algorithm

The whole procedure of text alignment is shown in Figure 14-11. In the graph, all squares refer to certain functions or macro-functions defined in the aligning program; the lozenges refer to certain judgment processes which lead to processes in different directions; ENL stands for English line and CNL for Chinese line. When the program reaches the end of the texts, the whole process comes to an end.

3.4. Text annotation

Corpus annotation means adding informative or explanatory linguistic information to a corpus as Leech (1997: 2) states. He also claims that annotation can greatly add value to a corpus and widen the range of studies the corpus can benefit. The author intends to annotate the MMSC corpus with part-of-speech (POS) tagging and several software tools are employed. However, this has not yet been completed by the time this chapter is finished.

3.5. Concordancer

The concordancer is one of the most important tools in corpus-based studies and usually is, as Kennedy (1998: 251) says, “easily available”. There are software packages like Wordsmith and AntConc, which can perform well-formed concordances with many types of corpora, including the MMSC corpus. However, as in the case of the MMSC, which is designed to be accessed through the Internet, a user friendly concordance interface is of great importance to the project. Therefore, the author has designed a simple web-based concordancer that allows users to perform KWIC (Key Word in Context) concordances with the MMSC.

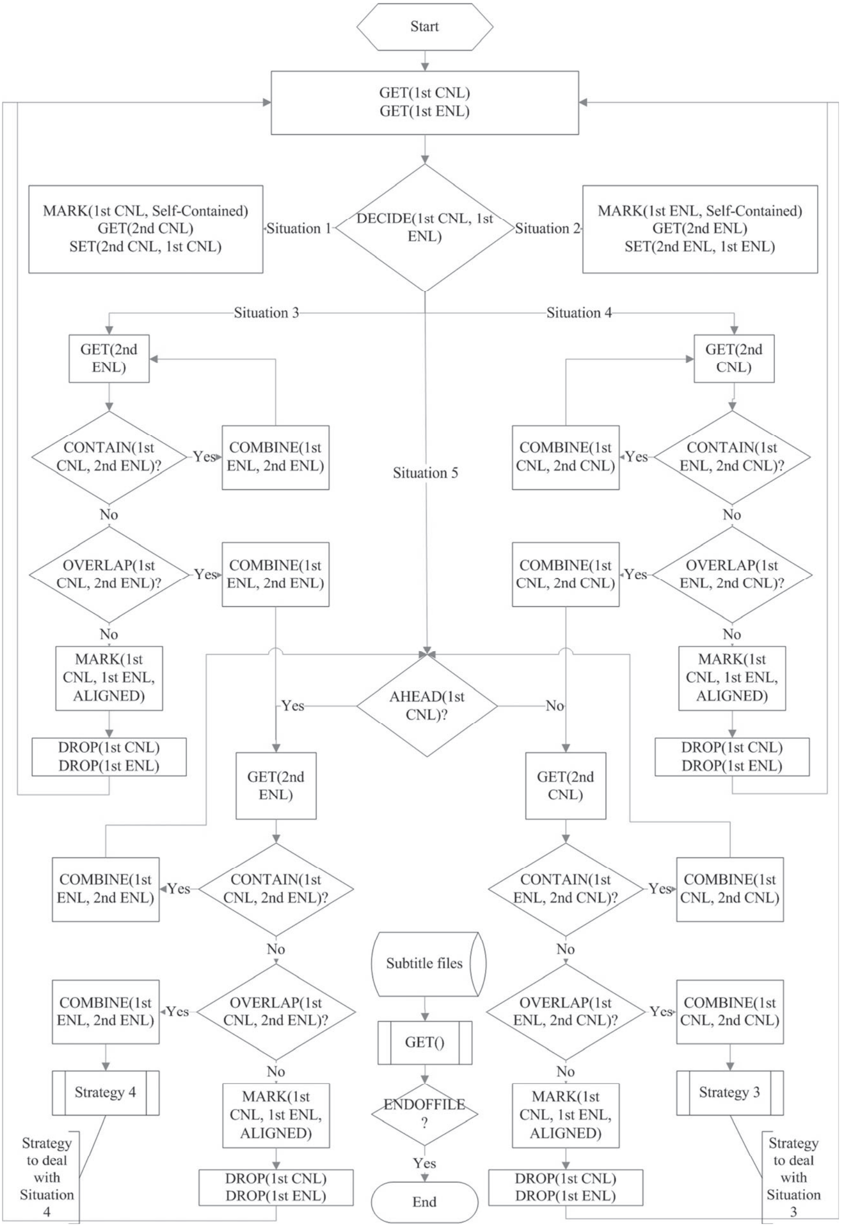

Start
GET(1st CNL)
GET(1st ENL)
DECIDE(1st CNL, 1st ENL)
Situation 1
MARK(1st CNL, Self-Contained)
GET(2nd CNL)
SET(2nd CNL, 1st CNL)
Situation 2
MARK(1st ENL, Self-Contained)
GET(2nd ENL)
SET(2nd ENL, 1st ENL)
Situation 3
Situation 4
Situation 5
GET(2nd ENL)
CONTAIN(1st CNL, 2nd ENL)?
Yes
COMBINE(1st ENL, 2nd ENL)
No
OVERLAP(1st CNL, 2nd ENL)?
Yes
COMBINE(1st ENL, 2nd ENL)
No
MARK(1st CNL, 1st ENL, ALIGNED)
DROP(1st CNL)
DROP(1st ENL)
GET(2nd CNL)
CONTAIN(1st ENL, 2nd CNL)?
Yes
COMBINE(1st CNL, 2nd CNL)
No
OVERLAP(1st ENL, 2nd CNL)?
Yes
COMBINE(1st CNL, 2nd CNL)
No
MARK(1st CNL, 1st ENL, ALIGNED)
DROP(1st CNL)
DROP(1st ENL)
AHEAD(1st CNL)?
Yes
No
GET(2nd ENL)
CONTAIN(1st CNL, 2nd ENL)?
Yes
COMBINE(1st ENL, 2nd ENL)
No
OVERLAP(1st CNL, 2nd ENL)?
Yes
COMBINE(1st ENL, 2nd ENL)
Strategy 4
Strategy to deal with Situation 4
No
MARK(1st CNL, 1st ENL, ALIGNED)
DROP(1st CNL)
DROP(1st ENL)
GET(2nd CNL)
CONTAIN(1st ENL, 2nd CNL)?
Yes
COMBINE(1st CNL, 2nd CNL)
No
OVERLAP(1st ENL, 2nd CNL)?
Yes
COMBINE(1st CNL, 2nd CNL)
Strategy 3
Strategy to deal with Situation 3
No
MARK(1st CNL, 1st ENL, ALIGNED)
DROP(1st CNL)
DROP(1st ENL)
Subtitle files
GET()
ENDOFFILE ?
Yes
End

Figure 14-11. Aligning algorithm flowchart

3.6. Online subtitle processor

The MMSC corpus is designed as a dynamic corpus whose size increases as more subtitles are processed and fed into it. Upon the completion of this chapter, the MMSC corpus contains a simple version of online subtitle processor (beta version 1.3) in order to automatically align and include subtitles donated by users.

4. Potential applications

The MMSC corpus can serve as a data bank for media translation studies, especially subtitling studies. It is also supposed to be of use to English learning and teaching practices. Translator trainees and trainers may also find the corpus helpful.

If properly annotated with POS information, the MMSC can greatly enhance its usage in the above mentioned areas, and may be expected to be of use to the practice in more areas such as example-based machine translation (EBMT), bilingual lexicography, subtitle quality control, automatic subtitling, etc. These applications, of course, may also require the MMSC corpus to be upgraded and improved in some aspects to better suit their needs.

It is also worth mentioning that the MMSC corpus keeps all the time-cue information intact and thus can be extended and added with new functions to combine audio-visual materials from movies and television programs with the subtitles. In this way, the MMSC corpus shall become a multi-media corpus.

5. Conclusion

This chapter has proposed the idea of using film and television subtitles as the sources for creating a parallel corpus and has discussed the building process based on the author’s ongoing work. This pilot study and the simple corpus product, however, as the author hopes, might be of some use to scholars and researchers interested in the area of parallel corpus building and subtitling studies.

Notes

1. *Big Shot’s Funeral* (2001) is a film directed by the Chinese director Feng Xiaogang.
2. *Blue Planet* (1990) is a documentary program produced and published by BBC.

3. Episode I of *Star Wars* movies, *The Phantom Menace* (1999), is a film directed by George Lucas.

References

Baker, M. (1995), "Corpora in translation studies: An overview and some suggestions for future research". *Target* 7(2): 223-243.

—. (1998), *Routledge Encyclopaedia of Translation Studies*. London: Routledge.

—. (1999), "The role of corpora in investigating the linguistic behaviour of professional translators". *International Journal of Corpus Linguistics* 4(2): 281-298.

Barlow, M. (2000), "Parallel texts and language teaching", in S. Botley, T. McEnery and A. Wilson (eds.) *Multilingual Corpora in Teaching and Researchin*g, 106-115. Amsterdam: Rodopi.

Bowker, L. (2001), "Towards a methodology for a corpus-based approach to translation evaluation". *Meta* 46(2): 345-364.

Chang, B. and Bai, X. (2003), "The markup guidelines for the Chinese-English parallel corpus of Peking University". *Journal of Chinese Language and Computing* 2:195-214.

Hunston, S. (2002), *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kennedy, G. (1998), *An Introduction to Corpus Linguistics*. New York: Addison Wesley Longman Ltd.

Laviosa, S. (1998), "The corpus-based approach: A new paradigm in translation studies". *Meta* 43(4): 474-479.

Leech, G. (1997), "Introducing corpus annotation", in R. Garside, G. Leech and T. McEnery (eds.) *Corpus Annotation*, 1-16. London: Longman.

Ma, Z. (2005), *An Introduction to Film and Television Translation*. Beijing: Communication University of China Press.

Olohan, M. (2004), *Introducing Corpora in Translation Studies*. London: Routledge.

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

Wang, K (2004), *Bilingual Parallel Corpus: Design and Applications*. Beijing: Foreign Language Teaching and Research Press.

Yu, S. (2003), *An Introduction to Computational Linguistics*. Beijing: Commercial Press.

CHAPTER FIFTEEN

FINDING THE PARALLEL: AUTOMATIC DICTIONARY CONSTRUCTION AND IDENTIFICATION OF PARALLEL TEXT PAIRS

SUMITHRA VELUPILLAI, MARTIN HASSEL, HERCULES DALIANIS

1. Introduction

Dictionaries are an important part of natural language processing tasks and linguistic work. Domain-specific dictionaries can, for example, be used in cross-language web and intranet search engines. Creating dictionaries manually is labour intensive and time consuming, and as a result, many methods have been proposed to automate this process. Word alignment tools are often used for the creation of bilingual word lists. Many assumptions about the characteristics of words and their translations for extracting bilingual vocabulary underlie the algorithms in such tools, which need parallel or comparable corpora as input. However, finding such corpora is often a difficult and arduous task, especially for small languages. The Internet is a useful resource for finding corpora in different languages, and many large corporations and organizations have abundant information on their multilingual websites. However, these text sets are often noisy, containing a lot of non-parallel parts which need to be removed in order to create useful parallel corpora.

In this chapter, three experiments are described. The first, described in section 3, is an experiment on creating parallel corpora and bilingual dictionaries from the website Hallå Norden (Hello Scandinavia).[1] After extracting text pairs covering all the Nordic language pairs by treating the entire set of texts on the website as one multilingual parallel corpus, ten parallel corpora were created. These were further used as input to the word alignment tool Uplug (Tiedemann 2003) for the automatic creation of dictionaries covering the Nordic languages.

However, in these corpora, we discovered that all text pairs were not completely parallel. Therefore, we have developed and evaluated methods for identifying parallel and non-parallel texts in corpora covering different language pairs. In section 3, an initial experiment on deleting non-parallel texts from the ten Nordic corpora is described. This method did not prove very successful, and two more thorough experiments on alternate methods for automatically identifying non-parallel texts in bilingual corpora have been performed.

The first experiment, described in section 4, exploits the frequency distribution of word initial letters in order to map a text in one language to a corresponding text in another. In this experiment, the JRC-Acquis corpus (European Council legal texts) was used,[2] with English and Swedish as language pair. In the second experiment, described in section 5, a memory-based machine learning technique was used with simple frequency features such as word, sentence and paragraph frequencies. The method was evaluated on a subset of the JRC-Acquis corpus as well as the entire set of Hallå Norden texts described above, and used on Swedish-Danish, Swedish-Finnish and Finnish-Danish pairs respectively.

The experiments described in this chapter show very promising results. However, further development and evaluation is needed. Language-independent methods for creating language resources, especially for small languages, are still scarce but important. Some concluding remarks and thoughts on future work are described in the final section, with the intent of exploring some directions for further studies in this intriguing and important research area.

2. Related work

Bilingual parallel corpora are useful for many natural language processing tasks, such as machine translation systems. For the automatic creation of dictionaries, word alignment systems are often used. Such systems need to make some assumptions regarding translated texts (Somers 2001):

- Words have one sense per corpus
- Words have a single translation per corpus
- There are no missing translations in the target document
- The frequencies of words and their translations are comparable

- The positions of words and their translations are comparable

These assumptions affect word alignment algorithms and, as can be seen, for the systems to work optimally, parallel or comparable corpora are needed.

The distinction between a parallel and a comparable corpus has been discussed in several research articles. In Somers (2001), it is pointed out that a “comparable” corpus has been used both interchangeably with “parallel” corpus, and as a term describing a corpus with similar but not necessarily equivalent texts. A more detailed discussion of the distinctions between how the terms ‘parallel’, ‘comparable’ and ‘non-parallel’ corpora are used can be found in Fung and Cheung (2004), for instance.

Ma and Liberman (1999) and Chen and Nie (2000) describe different heuristics for downloading and identifying parallel text. However, these methods are insufficient since the downloaded parallel text still can be very noisy. Such freely available multilingual resources are often noisy and non-parallel sections need to be removed. Many methods for identifying such sections automatically have been proposed. Maximum entropy (ME) classification is used in Munteanu and Marcu (2005) in order to improve machine translation performance. From large Chinese, Arabic and English non-parallel newspaper corpora, parallel data was extracted. For this method, a bilingual dictionary and a small amount of parallel data for the ME classifier is needed. By selecting pairs of similar documents from two monolingual corpora, all possible sentence pairs are passed through a word-overlap based filter and then sent to the ME classifier. The results were evaluated in different ways. One evaluation was made by testing the system on the news test corpus used for the NIST 2003 MT evaluation,[3] using the BLEU score, reporting significant improvements over the baseline (the highest score for Arabic-English was 47.97 and for Chinese-English 30.03).

In Fung and Cheung (2004), a method for extracting parallel sentences through bootstrapping and Expectation Maximization (EM) learning methods is presented. An iterative bootstrapping framework is presented, based on the idea that documents, even those with a low similarity score, containing one pair of parallel sentences must contain others. In particular, the proposed method works well for corpora with very disparate contents. The approach achieves 65.7 percent accuracy and a 50 percent relative improvement over their baseline.

Latent Semantic Indexing (LSI) has been experimented with in Katsnelson and Nicholas (2001) in order to identify parallel sequences in corpora. In this work, the hypothesis that LSI reveals similarities between parallel texts not apparent in non-parallel texts is presented and evaluated. Corpora from digital libraries were used with the language combinations English-French, English-Russian, French-Russian and English-Russian-Italian. Applying correlation coefficient analysis, a threshold of 0.75 was reported to successfully hold as a lower bound for identifying parallel text pairs. Non-parallel text pairs did not, in these experiments, exceed a correlation coefficient value of 0.70.

Unfortunately, most work has been performed on different types of corpora and on different language pairs. Moreover, they have been evaluated differently depending on available resources and the nature of the experiments, which makes them difficult to compare. However, the different approaches show the need for these types of methods.

3. Automatic construction of domain-specific dictionaries based on sparse corpora of the Nordic languages

In an experiment described in Velupillai and Dalianis (2008), dictionaries covering the Nordic languages using corpora obtained from the website Hallå Norden (Hello Scandinavia) were automatically created. Hallå Norden contains information regarding mobility between the Nordic countries in five languages: Swedish, Danish, Norwegian, Icelandic and Finnish. Treating the entire set of texts on the website as one multilingual parallel corpus, ten parallel corpora for each Nordic language pair were extracted and used for the creation of ten different dictionaries. The creation of the corpora was semi-automatic. The texts on the website were structured in a site map which was exploited to automatically find parallel text pair candidates. However, after manual inspection of these candidates, we discovered that only around 45 percent of the initial corpora from the website contained parallel text pairs. The remaining texts were either single texts with no matching translations, texts in the wrong language, or just empty pages. We removed almost all such texts manually.

Creating parallel corpora from multilingual websites often involves analyzing the contents and structures, as well as removing a lot of noise. For instance, on a Scandinavian bank corporation website with information in Swedish, Danish and Finnish, more than 50 percent of the texts were non-parallel. However, although a lot of texts may be removed, the final size of the resulting parallel corpora will naturally depend on the types of texts. The Hallå Norden texts, for example, are in general very short, while other types of texts available on other websites, annual reports for instance, may be much longer.

The final version of the resulting Hallå Norden corpora contained on average less than 80,000 words per language pair, which are considered as sparse corpora. For the creation of the dictionaries we used the word alignment system Uplug, since it is a non-commercial system which does not need a pre-trained model and is easy to use. It is also updated continuously and incorporates other alignment models, such as GIZA++ (Och and Ney 2003).

The produced dictionaries gave on average 213 dictionary entries (frequency > 3). Combinations with Finnish, which belongs to a different language family, had a higher error rate, 33 percent, whereas the combinations of the Scandinavian languages only yielded on average 9 percent errors. Despite the corpus sparseness the results were surprisingly good compared to other experiments with larger corpora.

However, we discovered that the created corpora were to some extent non-parallel containing some extra non-aligned paragraphs. We believed that these text pairs affected the results negatively, and made a small experiment on trying to automatically delete text pairs that were not parallel.

Table 15-1. Produced dictionary words and error rate for the initial and the refined corpora, from Velupillai and Dalianis (2008)

	Initial		Deleting non-parallel	
Language pair	No. dictionary words	Erroneous translations, %	No. dictionary words	Erroneous translations, %
sw-da	322	7.1	305	7.2
sw-no	269	6.3	235	9.4
sw-fi	138	29.0	133	34.6
sw-ice	151	18.5	173	16.2
da-no	322	3.7	304	4.3
da-fi	169	34.3	244	33.2
da-ice	206	6.8	226	10.2
no-fi	185	27.6	174	30.0
no-ice	159	14.5	181	14.4
Average	213	16.4	219	16.1

We used an algorithm simpler than that in for instance Munteanu and Marcu (2006). The total number of paragraphs and sentences in each parallel text pair was counted. If the total number for each language in some language pair differed by more than 20 percent either in the total number of paragraphs, sentences, or both, these texts were automatically deleted. On average 5 percent of the manually processed corpora were detected as being non-parallel using this algorithm. The refined corpora were re-aligned with Uplug and evaluated, but unfortunately about the same error rate as before deleting the non-parallel texts was obtained, although with some differences in the produced word pairs (see Table 15-1). Perhaps our simple algorithm was too coarse for these corpora, especially since they were so sparse. The texts were in general very short and simple frequency information on paragraph and sentence numbers might not have captured non-parallel fragments on such texts. A more detailed discussion of the results of this experiment can be found in Velupillai and Dalianis (2008). More elaborate and efficient methods for identifying parallel and non-parallel texts in bilingual corpora are described in the following sections.

4. Identifying parallel and non-parallel texts in bilingual corpora using fingerprints

When comparing documents for content similarity it is a common practice to produce some form of document signatures, or ‘fingerprints’. These fingerprints represent the content in some way, often as a vector of features, which are used as the basis for such comparison. One common method when comparing the likeness of two documents is to utilize the so-called Vector Space model (Salton 1971, 1983). In this model the documents’ fingerprints are represented as feature vectors consisting of the words that occur within the documents, with weights attached to each word denoting its importance for the document. We can, for example, for each feature (in this example, a word) record the number of times it occurs within each document. This gives us what is commonly called a document-by-term matrix where the rows represent the documents in the document collection and the columns each represent a specific term existing in any of the documents (a weight can thus be zero). We can now, in a somewhat simplified way, compare the documents’ fingerprints by looking at how many times each feature occurs in each document, taking the cosine angle between the vectors, and pair the two most similar together. One obvious drawback of the basic use of this model is that when comparing texts written in different languages we do not necessarily know which feature in one language corresponds to which feature in another.

Another drawback when building a word vector space representing more than one language is that the vocabulary, i.e. the number of features in the feature vectors, grows alarmingly (this is in many cases already a problem representing just one language) (Sahlgren 2005). Ways of limiting the vocabulary include using stop-word lists to remove "information poor" features, frequency thresholding and conflation into feature classes (for example lemmatization). In word vector spaces the latter is often accomplished by bringing semantically related words to a common lemma or stem. In the experiments described below conflation was attempted by moving from term frequency classes towards prefix frequency classes, i.e. the leading characters of each token. In this way, a document's fingerprint is effectively represented by a feature vector containing the frequency of each prefix of a set length *n* occurring in the corpus. This has, for example, been used in information retrieval for filtering of similar documents written in the same language (Stein 2005). Here we attempt to utilize this notion in cross-language text alignment.

4.1. Data sets and experimental setup

In this set of experiments we have used the JRC-Acquis corpus (Steinberger *et al*. 2006). This corpus consists of European Union law texts, which are domain specific and also very specific in their structure. Many texts are listings of regulations with numerical references to other law texts and named entities (such as countries).[4] The corpus is very large, containing a different amount of texts depending on the language. Here we have investigated the language pair Swedish-English, i.e. we used Swedish as a source language attempting to find the corresponding parallel text in English. We have also used only those documents that have a counterpart in both languages, resulting in a total of 20,145 document pairs. In Appendix A, a Swedish example file along with its corresponding, parallel, English translation from the JRC-Acquis corpus is given.

In order to delimit the search space for the practicality of this experiment we have not compared each Swedish source text with each and every English text. Instead we compared, in one experiment, the similarity between a true positive (the corresponding, parallel, English text) and one true negative (a randomly chosen non-parallel English text), letting the algorithm choose the closest match (as defined by the cosine angle between the feature vectors for each text). In another experiment we repeated the setup, but instead of only using one true negative we used nine. This setup gave us a random chance of picking the true positive of 50 percent in the case of one true positive and one true negative (k=1), and 10 percent in the case of one true positive and nine true negatives (k=10, see Table 15-2). In order to rule out any random fluke in the choice of true negative(s) for each true positive, both experiments were carried out 10 times, making new random pairings each time. An average was then taken, calculated over these ten runs.

As in Stein (2005) we have extracted *a priori* probabilities of prefix classes from reference corpora. Since we are dealing with the language pair Swedish-English, we have used a Swedish reference corpus, the Swedish Parole corpus (Gellerstam *et al*. 2000), and an English ditto, the British National Corpus (Aston and Burnard 1998). The Swedish reference corpus is composed of roughly 20 million words. In order to have a comparable English reference corpus we have only used the first 20 million words of BNC (out of roughly 100 million).[5] These two corpora can be seen as the expected distribution of the prefix classes for each language, while each text's feature vector is then the deviation to the expected distribution. We would like to find out whether a deviation from the expected frequency distribution pattern in one language in the pair could possibly reflect a similar deviation in the other. In this set of experiments the feature vector for each text was pre-processed in two ways:

1. Using Parole as the reference corpus for Swedish texts and the BNC as the reference corpus for English texts, by calculating the difference in frequency between the occurrences of a prefix in the reference corpus and in each text. The prefixes in these vectors were then sorted by the frequency in each respective reference corpus. The most common feature in the source language corresponds to the most frequent feature in the target language, and so on. The comparison of the text's feature vectors is then based on the deviation from the expected and normalized distribution for each language.
2. No normalization using reference corpora. Instead the raw frequencies are compared directly. However, matching of features is still based on the frequency in each language's respective reference corpus.

As mentioned above, feature vectors were created using the leading *n* characters of each word occurring in each reference corpus, as well as in any of the 20,145 documents used in the tests. A fingerprint was constructed for each reference corpus and each document, in both languages, for *n*=1..3, both using all lower case, (lc), prefixes as well as prefixes maintaining their original capitalization (see Table 15-2). To be noted here is the fact that the vocabulary size grows at an explosive rate as *n* grows, especially when the original capitalization is preserved.

4.2. Results

Table 15-2 shows the following information: Swedish source, one true positive and one true negative English target (k=2); one true positive and nine true negatives (k=10). Lower case is abbreviated as lc. The precision is calculated over 10 random selections of the non-parallel text(s). Also given are the lowest and the highest results of the ten runs. At k=2 baseline-random is 50 percent and our results indicate up to 87 percent precision; at k=10 baseline-random is 10 percent and our results indicate up to 68 percent precision. As can be seen, it is far more favourable to compare the raw frequencies of the features in the source and target vectors, rather than comparing the deviation based on the frequency distribution in the reference corpus of the respective languages. This is further supported by the fact that model two stands even stronger, relatively speaking, when pin-pointing the right match out of ten possible target texts.

Table 15-2. Results, fingerprints

Model	1. Parole / BNC		2. No normalization	
Prefix size	mean precision %	lowest – highest	mean precision %	lowest – highest
k=2, *n*=1	50	0.496 - 0.503	**87**	0.865 - 0.872
k=2, *n*=1, lc	50	0.497 - 0.502	86	0.852 - 0.858
k=2, *n*=2	50	0.497 - 0.502	80	0.794 - 0.799
k=2, *n*=2, lc	50	0.498 - 0.502	76	0.756 - 0.762
k=2, *n*=3	50	0.496 - 0.502	76	0.759 - 0.769
k=2, *n*=3, lc	50	0.495 - 0.505	75	0.747 - 0.753
k=10, *n=1*	10	0.097 - 0.102	**68**	0.674 - 0.678
k=10, *n=1, lc*	10	0.098 - 0.102	65	0.646 - 0.655
k=10, *n=2*	10	0.099 - 0.104	54	0.534 - 0.543
k=10,*n=2, lc*	10	0.098 - 0.103	45	0.450 - 0.455
k=10, *n=3*	10	0.100 - 0.102	50	0.497 - 0.504
k=10,*n=3, lc*	10	0.097 - 0.102	44	0.438 - 0.442

We can also see that the results are very stable – there is only a slight difference in the precision between the best and the least good run – even though there is little overlap between the 10 randomly generated lists of pairs. The highest number of pairs that one of the lists has in common with any of the other lists is 12 (out of 20,145). When it comes to the lists containing 10 target words this number is nearly non-existent.

One possible answer for the success of the second model could of course be that the source and target texts are always lexically very similar. This could be the case if they to a high degree share the same vocabulary, for instance named entities. However, this does not seem to be the case if we take a look at Table 15-3. Here, results for three different baselines using only basic features are given. Each baseline tracks the number of occurrences of the following: baseline1={bytes, tokens, dot, comma, percent, digit, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9}, baseline2={bytes, tokens, dot, comma, percent} and baseline3={tokens, dot, comma}.

Table 15-3. Results using baselines with basic features

	k=1		*k*=10	
Baseline	mean precision %	lowest – highest	mean precision %	lowest – highest
1	50	0.496 - 0.503	10	0.097 - 0.102
2	50	0.497 - 0.503	10	0.099 - 0.102
3	50	0.497 - 0.504	10	0.098 - 0.102

The degree of precision and the stability of the results are encouraging. However, for the sake of a fairer comparison one might want to reconsider the baselines used in this experiment as being too naïve. In the next section, a different set of roughly language-independent features, as well as some language-dependent features (relying on the use of a part-of-speech tagger), is presented, evaluated on some of the Nordic language pairs.

5. Identifying parallel and non-parallel texts in bilingual corpora using simple frequency features and memory-based learning

In the final experiment on trying to identify whether two texts in different languages in a bilingual corpus are parallel or not, a memory-based machine learning technique was used. The identification problem can be viewed as a classification problem where the possible classes are *Parallel* and *Non-parallel*. We put forward the hypothesis that simple frequency counts of, for instance, paragraphs, sentences and words, as well as part-of-speech information, could be valuable features for detecting whether a text pair in two different languages is parallel or not.

The following language pairs were used: Swedish-Danish, Swedish-Finnish and Danish-Finnish (treating the leftmost language in each language pair as the source language, and the rightmost language as the target language). Using language pairs from both related and non-related language families is important in order to investigate if such issues influence the results. Two bilingual corpora for each language pair were created, consisting of an equal amount of *Parallel* and *Non-parallel* instances (only one true positive and negative instance, thus giving a 50 percent random chance of picking the true positive), amounting to a total of six corpora. The corpora were extracted from the JRC-Acquis corpus (described in section 4) and the Hallå Norden corpus (described in section 3).

As stated in section 4, many texts in the JRC-Acquis corpus contain listings of regulations and numerical references to other law texts, thus containing very short sentences. The Swedish, Danish and Finnish text sets contain around 20,000 texts, where most of the texts also exist in a parallel version in the other two languages.

The Hallå Norden corpus consists of short texts regarding mobility information in the Nordic region (see section 3). The corpus is very small (around 200 texts per language pair), but provides a different type of text from a different domain that reflects another type of language use than the texts in the JRC-Acquis corpus. Although the texts are short and may also contain a lot of listed information, they are not as fragmented as the texts in the JRC-Acquis corpus. In Appendix B, a Swedish and a Danish example file from the Hallå Norden corpus are given. These examples illustrate the type of texts this corpus contains, and how they contain sequences that are parallel translations but also sequences that may be missing. Moreover, they exemplify how differently the texts can be formatted, especially with regards to paragraphs. This text pair was recognized as non-parallel using the simple algorithm for detecting non-parallel files described in section 3.

5.1. Machine learning algorithm

For this experiment the machine learning algorithm used was memory-based learning, using the TiMBL software (see Daelemans *et al*. 2007 for a reference guide). It was used with the classification algorithm IB1, applying default settings with regards to algorithmical settings. This means that the distance metric used was *Overlap* and the feature weighting used was *Gain Ratio*. A feature selection experiment was performed on these default values, testing different combinations of features. The tests were performed through 10-fold cross-validation, splitting the entire data sets into 10 parts, equal in size, containing the same amount of *Parallel* and *Non-parallel* classified text pairs, using nine parts for training and one part for testing in turn for each part.

5.1.1. Features

For each text in the bilingual corpora, the following features were extracted:

- Total number of words
- Total number of sentences
- Total number of paragraphs
- Average length of words
- Average (word) length of sentences
- Average (word) length of paragraphs
- The five most frequent part-of-speech bi- and tri-grams

Moreover, the difference (in percent) in the total number of words, sentences and paragraphs between a text pair as well as the difference in the average number of words, sentences and paragraphs between a text pair was calculated and used as features. Here, difference is calculated the following way: $(\max(s\text{-}t))/(s+t)\times 100$, where s is the value of the total number or average length of words, sentences or paragraphs for the source language text and t is the value of the total number or average length of words, sentences or paragraphs in the target language text. In total, each instance in the data set consisted of 39 features (including an instance id, which was never included in the feature selection).

5.1.2. Definitions

A simple approach was used in order to identify words, sentences and paragraphs. Words are defined as a sequence of characters separated by space. No punctuation characters are included as words (a word such as “EG/EEG” is replaced with “EGEEG”), and digits are not counted as words. When calculating the average length of a word, the number of characters in each word is used.

Sequences of characters ending with a full stop "." and / or newline are defined as sentences. When calculating the average length of a sentence the number of words in each sentence is used. Sequences of characters ending with newline are defined as paragraphs. When calculating the average length of a paragraph the number of words in each paragraph is used. More sophisticated identification of words, sentences and paragraphs could of course be employed.

5.1.3. Part-of-speech tagging

Before extracting words, sentences and paragraphs, all texts were part-of-speech tagged. For Swedish Granska was used,[6] for Danish CST's Part-of-Speech Tagger,[7] and for Finnish Fintwol.[8] The taggers use different sets of tags, and have, naturally, been evaluated on different corpora. However, they are state-of-the-art tools for the respective languages. Fintwol, for instance, is the only available tool for tagging Finnish and was used for creating gold data in the Morpho Challenge 2007.[9] For this experiment, the different tagsets were not mapped to a uniform tagset. The idea was that the distribution patterns of part-of-speech bigrams and trigrams for each language would reflect the relationship between the texts.

5.2. Data set

For each corpus, all features for each text in one language chosen as the source language was paired with the corresponding (true positive) text in the target language, creating an instance with the classification *Parallel*. The source language text was also paired with a randomly selected target text (true negative), creating an instance with the classification *Non-Parallel*.

The Hallå Norden corpus consists of the following corpora:

- Swedish-Danish, 191 text pairs
- Swedish-Finnish, 196 text pairs
- Danish-Finnish, 239 text pairs

The JRC-Acquis corpus consists of the following corpora:

- Swedish-Danish, 14 231 text pairs
- Swedish-Finnish, 14 226 text pairs
- Finnish-Danish, 23 238 text pairs

The Swedish-Danish and Swedish-Finnish data sets from the JRC-Acquis corpus were smaller than the Finnish-Danish pair due to part-of-speech tagging problems on the Swedish texts. Each data set was divided into 10 subsets for the 10-fold cross-validation process, containing an equal amount of *Parallel* and *Non-parallel* instances.

5.3. Results

In Table 15-4 the performed feature tests are described. In total, eleven feature tests were performed on each data set. The extracted features were divided into the following sub-groups: total numbers and average lengths of words, sentences and paragraphs, part-of-speech tag information, and differences between each text with respect to total numbers and average lengths of words, sentences and paragraphs. These groups of features were tested independently. Also, the sub-groups were further divided into smaller subsets of features, in order to test which feature(s) produced the best results. Test 1, which includes all features except the instance id, was used as the baseline. The groups of features and the baseline were chosen based on intuition, and should of course be scrutinized and tested further in future developments.

Table 15-4. Feature test descriptions

Test	Descriptions
1	Default, all features except first feature (instance id), used as baseline
2	Total number and average length of words, sentences and paragraphs
3	All part-of-speech features
4	Part-of-speech bigrams
5	Part-of-speech trigrams
6	Difference in total number and average length of words, sentences and paragraphs
7	Difference in total number of words, sentences and paragraphs
8	Difference in average length of words, sentences and paragraphs
9	Difference in total and average number of words
10	Difference in total number and average length of sentences
11	Difference in total number and average length of paragraphs

The results for the Hallå Norden data sets were surprisingly good**,** despite the small size of the corpora. In Table 15-5 the results on average accuracy (in percent) of the 10-fold cross-validation tests are given. One can see that all tests from 6 to 11 yield good results. It is interesting to note that the part-of-speech information yielded very poor results. Perhaps this could be improved by mapping the different tagsets into a uniform tagset. Moreover, choosing the five most frequent part-of-speech bi- and trigrams may not distinguish parallel and non-parallel text pairs very well, as they may be common in all texts. Extracting discriminative part-of-speech patterns would be desirable. However, the features containing information about the differences between the number of, or the average length of, words, sentences and paragraphs in the text pairs yielded promising results. In particular, the feature test where all information about differences between the texts (Test 7) produced good results for all language pairs.

Table 15-5. Results, Hallå Norden

Test	Swedish-Danish %	Swedish-Finnish %	Danish-Finnish %
1	74.7	52.0	69.8
2	7.9	9.6	13.8
3	9.5	14.5	16.9
4	20.1	33.4	30.1
5	8.7	13.2	16.7
6	79.9	65.9	73.7
7	82.4	68.1	73.7
8	76.9	60.1	67.8
9	85.3	63.0	68.5
10	72.3	68.3	77.7
11	59.0	55.2	76.3

The results for the JRC-Acquis data sets are given in Table 15-6. The results, reported as the average accuracy (in percent) of the 10-fold cross-validation tests, are very encouraging, especially all tests from 6 to 11. As in the tests on the Hallå Norden corpora, the method of using the features that reflect the differences in the total number and average length of words, sentences and paragraphs produced good results for all language pairs. Using the information about the total number and average length of words for each text separately did not yield good results for any data set. Perhaps normalizing them in some way would be advantageous.

Overall the result patterns are similar for the two different corpora, even though the results for the JRC-Acquis corpora are better than those for the Hallå Norden corpora. It is interesting to note that the patterns are so similar despite the different characteristics of the text sets (in size, domain type and text type for instance).

The results are very promising. Even for a small data set such as the Hallå Norden corpora, it is possible to detect parallel and non-parallel text pairs on simple frequency features. However, more tests would need to be performed in order to verify the results properly. In particular, both text sets are very homogeneous, which might affect the results. The texts are similar in both their content and structure. The method should also be evaluated on more diversified text sets.

Even though the Swedish-Danish and Swedish-Finnish JRC-Acquis corpora were smaller than the Finnish-Danish corpus, the results were similar. It would be interesting to investigate at which point in the size of the data set results seem to decrease. Perhaps fairly small corpora are sufficient in order to obtain good results.

Table 15-6. Results, JRC-Acquis

Test	Swedish-Danish %	Swedish-Finnish %	Finnish-Danish %
1	92.2	90.1	88.1
2	25.0	24.9	22.7
3	37.0	46.8	50.5
4	59.4	65.1	66.5
5	52.6	54.7	54.2
6	92.7	90.3	88.6
7	93.2	90.7	89.2
8	93.1	90.5	88.5
9	93.3	89.7	88.5
10	89.3	89.7	85.9
11	93.1	89.2	89.0

Experiments with other language pairs should also be performed. For instance, part-of-speech information might prove more valuable to other language pairs. Moreover, as stated above, other approaches to using the part-of-speech information should be investigated. Also, the length measures for paragraphs and sentences used here are not normalized in any way. An interesting experiment would be to use language normalized number of characters instead of measuring the raw word lengths. Furthermore, other settings in the chosen machine learning algorithm should be tested. Parameter optimization tests using other distance metrics or weighting schemes might yield improved results. Given the features used, perhaps a different machine learning algorithm such as SVM (support vector machine) might produce better results.

6. Conclusions and future work

In the experiments described above we have demonstrated that our methods for identifying and deleting non-parallel texts from different corpora covering different language pairs show great potential. However, the results are, unfortunately, currently not comparable. In future experiments, we will apply the methods on the same corpora and language pairs, and evaluate the results in a comparable manner.

Methods for identifying parallel texts or sequences in texts can be used for many natural language processing tasks, including machine translation systems and dictionary construction. Evaluating and comparing such methods can be difficult, as they are developed on the basis of different types of corpora and languages. Moreover, there are many evaluation metrics that can be used, depending on both the availability of gold standard corpora and the purpose of the studies.

We have developed methods with the intention of keeping them as language-independent as possible. For the fingerprint method (described in section 4), the only language-dependent feature is the use of a reference corpus for each language. Such corpora may, unfortunately, still be difficult to obtain for very small languages with scarce resources. The use of language-dependent part-of-speech information for the simple frequency method (described in section 5) did not improve results. However, this information should probably be used differently. It is interesting to note that the best results in this experiment were obtained through the purely language-independent frequency features.

Moreover, in further work all our experiments on the identification of parallel text pairs should be run on more language pairs, preferably such that contain languages belonging to different language groups (as has, for instance, been carried out with the combinations with Finnish in the memory-based learning experiments). An obvious observation here is that the language pairs should also be tested reversely; that is, if one is to investigate the performance on, for instance, the language pair Swedish-English, it should also be evaluated on the corresponding pair English-Swedish. Also, the experiments should be re-run on other corpora than the JRC-Acquis and Hallå Norden in order to ensure that we are not just investigating peculiarities of these specific corpora.

In a real-world setting, attempting to identify whether a text in one language is parallel with a text in another means that it needs to be compared with many texts in the target language. For instance, the method described in section 5 should be tested against several true negatives, as was with the fingerprint-method as described in section 4. We also intend to investigate and develop methods for reducing the search space for candidate translations.

An important aspect of developing methods for cross-linguistic tools or resources is the possible need for preprocessing tools, such as part-of-speech taggers, covering all languages. This may be difficult to obtain, and different tools use different formatting and tagging schemes. Moreover, they might differ in robustness, which also affects the end results. Evaluating the performance of such preprocessing steps might be desirable.

Creating parallel corpora from Internet resources is both practical and convenient, as many texts are freely available. It is, however, not always trivial to extract the reqired sequences of web texts. Methods for utilizing the structure(s) of different site maps and removing tags and other web-specific formatting details are needed in order to minimize manual work. Moreover, many alternative sources for finding parallel corpora exist, e.g. digital libraries.

Parallel corpora covering different language pairs and text types are still very scarce, especially for small languages. Such corpora are important for many aspects of translation studies and need to be compiled. Moreover, the access to freely available parallel corpora provides the possibility of creating gold standard corpora that could be used for evaluating and comparing different methods. However, the difficulty of evaluating methods that are needed and used for different purposes still remains.

Notes

1. www.hallonorden.org
2. http://wt.jrc.it/lt/Acquis/
3. http://www.nist.gov/speech/tests/mt
4. Referencing systems do however differ between languages. For example, while some use Hindu-Arabic numerals, others use Roman.
5. In hindsight one should perhaps not simply use the first *n* words in the larger corpus, but instead take a random sample of the desired amount of words, when taking care so that reference corpora are of equal size.
6. www.nada.kth.se/theory/projects/granska/
7. http://www.cst.dk/online/pos_tagger/uk/index.html
8. http://www2.lingsoft.fi/doc/fintwol/
9. http://www.cis.hut.fi/morphochallenge2007/

References

Aston, G. and Burnard, L. (1998), *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Chen, J. and Nie, J-Y. (2000), "Parallel Web text mining for cross-language IR", in *Proceedings of RIAO-2000: Content-Based Multimedia Information Access*, 62-77. College de France, Paris, 12-14 April 2000.

Daelemans, W., Zavrel, J., Van der Sloot, K. and Van den Bosch, A. (2007), *TiMBL: Tilburg Memory Based Learner, version 6.1, Reference Guide*. Technical Report, ILK Research Group Technical Report Series No. 07-07.

Fung, P. and Cheung, B. (2004), "Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM", in *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*. Barcelona, 25 – 26 July 2004.

Gellerstam, M., Cederholm, Y. and Rasmark, T. (2000), "The Bank of Swedish", in *Proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000*, 329–333, Athens, Greece, 2000.

Katsnelson, Y. and Nicholas, C. (2001), "Identifying parallel corpora using latent semantic indexing", in *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, 30 March – 2 April 2001.

Ma, X. and Liberman, M. Y. (1999), "BITS: A method for bilingual text search over the Web", in *Proceedings of MT Summit VII*, 538-542. Singapore, September 1999.

Munteanu, D. S. and Marcu, D. (2006), "Extracting parallel sub-sentential fragments from non-parallel corpora", in *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics*, 81-88. Sydney, Australia, 17 – 21 July 2006.

Munteanu, D. S. and Marcu, D. (2005), "Improving machine translation performance by exploiting non-parallel corpora". *Computational Linguistics* 31(4): 477-504.

Och, F. J. and Ney, H. (2003), "A systematic comparison of various statistical alignment models". *Computational Linguistics* 29(1): 19-51.

Sahlgren, M. (2005), "An introduction to random indexing", in *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*. Copenhagen, Denmark, 16 August 2005.

Salton, G. (ed.) (1971), *The Smart Retrieval System – Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. and McGill, M. (1983), *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill.

Somers, H. (2001), "Bilingual parallel corpora and language engineering", in *Anglo-Indian Workshop "Language Engineering for South-Asian Languages" (LESAL)*. Mumbai, India, April 2001.

Stein, B. (2005), "Fuzzy-fingerprints for text-based information retrieval", in K. Tochtermann and H. Maurer (eds.) *Proceedings of the I-KNOW '05, Graz 5th International Conference on Knowledge Management Journal of Universal Computer Science*, 572-579. Graz, Austria: Know-Center.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. and Varga, D. (2006), "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages", in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06).* Genoa, Italy, 24 – 26 May 2006.

Tiedemann, J. (2003). *Recycling Translations: Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. PhD Thesis, Acta Universitatis Upsaliensis: Studia linguistica upsaliensia.

Velupillai, S. and Dalianis, H. (2008), "Automatic construction of domain-specific dictionaries on sparse parallel corpora in the Nordic languages", in *Proceedings of the 2nd MMIES Workshop: Multi-source, Multilingual Information Extraction and Summarization*. Manchester, UK, 23 August 2008.

Appendix A: Example files from the JRC-Acquis Corpus (Swedish and English)

(Apart from some minor differences we see that the files are very parallel translations. Also, we see the specificity of the text type: short sentences, named entities and many listings.)

Swedish:

2006/796/EG: Rådets beslut av den 13 november 2006 om evenemanget Europeisk kulturhuvudstad år 2010

Rådets beslut
av den 13 november 2006
om evenemanget Europeisk kulturhuvudstad år 2010
(2006/796/EG)
EUROPEISKA UNIONENS RÅD HAR BESLUTAT FÖLJANDE
med beaktande av fördraget om upprättandet av Europeiska gemenskapen,
med beaktande av Europarlamentets och rådets beslut nr 1419/1999/EG av den 25 maj 1999 om att inrätta en gemenskapsåtgärd för evenemanget Europeisk kulturhuvudstad för åren 2005 till 2019 [1], särskilt artikel 2.3 och 2.4,
med beaktande av den rapport från juryn från april 2006 som lagts fram för kommissionen, Europaparlamentet och rådet i enlighet med artikel 2.2 i beslut nr 1419/1999/EG,
med beaktande av att kriterierna i artikel 3 och bilaga II i beslut nr 1419/1999/EG,
med beaktande av kommissionens rekommendation av den 23 oktober 2006.
HÄRIGENOM FÖRESKRIVS FÖLJANDE.
Artikel 1
Essen och Pécs skall utses till europeiska kulturhuvudstäder 2010 i enlighet med artikel 2.1 i beslut nr 1419/1999/EG.
Artikel 2
Istanbul skall utses till europeisk kulturhuvudstad 2010 i enlighet med artikel 4 i beslut nr 1419/1999/EG.
Artikel 3
De tre städerna skall vidta alla åtgärder som krävs för att säkerställa att artiklarna 1 och 5 i beslut nr 1419/1999/EG genomförs på ett effektivt sätt.

Utfärdat i Bryssel den 13 november 2006.
På rådets vägnar
S. Huovinen
Ordförande

[1] EGT L 166, 1.7.1999, s. 1. Beslutet ändrat genom beslut nr 649/2005/EG (EUT L 117, 4.5.2005, s. 20).

English:

2006/796/EC: Council Decision of 13 November 2006 on the European Capital of Culture event for the year 2010

Council Decision
of 13 November 2006
on the European Capital of Culture event for the year 2010
(2006/796/EC)
THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty establishing the European Community,

Having regard to Decision No 1419/1999/EC of 25 May 1999 of the European Parliament and the Council establishing a Community action for the European Capital of Culture event for the years 2005 to 2019 [1], and in particular Articles 2 paragraph 3 and 4, thereof,

Having regard to the Selection Panel report of April 2006 submitted to the Commission, the European Parliament and the Council in accordance with Article 2 paragraph 2 of Decision 1419/1999/EC,

Considering that the criteria laid down in Article 3 and Annex II of Decision No 1419/1999/EC are entirely fulfilled,

Having regard to the recommendation from the Commission of 23 October 2006,

HAS DECIDED AS FOLLOWS:

Article 1

Essen and Pécs are designated as %quot%European Capital of Culture 2010%quot% in accordance with Article 2 paragraph 1 of Decision No 1419/1999/EC as amended by Decision No 649/2005/EC.

Article 2

Istanbul is designated as a %quot%European Capital of Culture 2010%quot% in accordance with Article 4 of Decision No 1419/1999/EC as amended by Decision No 649/2005/EC.

Article 3

All cities designated shall take the necessary measures in order to ensure the effective implementation of Articles 1 and 5 of Decision 1419/1999/EC as amended by Decision No 649/2005/EC.

Done at Brussels, 13 November 2006.
For the Council
The President
S. Huovinen

[1] OJ L 166, 1.7.1999, p. 1. As amended by Decision No 649/2005/EC (OJ L 117, 4.5.2005, p. 20).

Appendix B: Non-parallel example files from the Hallå Nården Corpus (Danish and Swedish)

(The underlined parts of Danish text are missing in the Swedish translation, and the two first sentences are juxtaposed. Also, the second last sentence in the Swedish file is missing in the Danish translation)

Danish:

Stemmeret i Danmark
Kun danske statsborgere med fast bopæl i Danmark som er myndige og fyldt 18 år har stemmeret til folketingsvalg.
Du har stemmeret til kommunalvalg, hvis du er over 18 år, har fast bopæl, *er dansk statsborger eller har boet i landet uafbrudt de seneste tre år. Det betyder, at indvandrere og flygtninge kan stemme ved kommunal- og amtsrådsvalg, selv om de ikke har dansk statsborgerskab. Ophold regnes fra den dag man registreres i folkeregistret.*
Statsborgere fra EU-lande, Island og Norge kan stemme ved kommunal- og amtsrådsvalg, hvis de har fast bopæl i Danmark. Det samme gælder personer, der arbejder for staten i udlandet eksempelvis diplomater og soldater, samt i enkelte tilfælde deres ægtefælle eller samlever.
Borgere fra andre EU-lande har stemmeret til EU-parlamentet, hvis de har fast bopæl i Danmark og er fyldt 18 år.

Senest opdateret: 16-11-2006

Swedish:

Rösträtt i Danmark

Alla myndiga personer över 18 år som är fast bosatta i Danmark har rösträtt i kommunala val.

Endast danska medborgare har rösträtt i valet till det danska folketinget.

Medborgare i andra EU-länder har rösträtt i EU-parlamentsvalet om de är fast bosatta i Danmark och har fyllt 18 år.

För mer information, se lag 730 av den 9 oktober 1998 på www.retsinfo.dk.

Senast uppdaterad: 24-11-2006

CHAPTER SIXTEEN

WEB CORPORA FOR BILINGUAL LEXICOGRAPHY: A PILOT STUDY OF ENGLISH / FRENCH COLLOCATION EXTRACTION AND TRANSLATION

ADRIANO FERRARESI, SILVIA BERNARDINI, GIOVANNI PICCI, MARCO BARONI

1. Introduction

This chapter describes two very large (> 1 billion words) Web-derived "reference" corpora of English and French, called *ukWaC* and *frWaC*, and reports on a pilot study in which these resources are applied to a bilingual lexicography task focusing on collocation extraction and translation.

The two corpora were assembled through automated procedures, and little is known of their actual contents. The study aimed therefore at providing mainly qualitative evaluation of the corpora by applying them to a practical task, i.e. ascertaining whether resources built automatically from the Web can be profitably applied to lexicographic work, on a par with more costly and carefully-built resources such as the British National Corpus (BNC) for English.

The lexicographic task itself was set up simulating part of the revision of an English / French bilingual dictionary. Focusing on the direction from English to French, it first of all compared the coverage of ukWaC vs. the widely used BNC in terms of collocational information of a sample of English SL nodewords. The evidence thus assembled was submitted to a professional lexicographer who evaluated relevance. The validated collocational complexes selected for inclusion in the revised version were then translated into French drawing evidence from frWaC, and the translations were validated by a professional translator (native speaker of French). The results suggest that the two Web corpora provide relevant and comparable linguistic evidence for lexicographic purposes.

The chapter is structured as follows: section 2 sets the framework for the study, reviewing current approaches to the use of the Web for cross-linguistic tasks, describing the Web corpora used, and the applications of corpora in lexicography work. Section 3 presents the objectives of the pilot investigation, the method followed and its results. In section 4, we draw conclusions and suggest directions for further work.

2. Corpora, lexicography and the Web

2.1. Web corpora for cross-linguistic tasks

In many fields, ranging from corpus linguistics to Natural Language Processing (NLP) and machine translation (MT), the Web is being increasingly used as a source of linguistic data. This is the case, for instance, when traditional corpus resources prove inadequate to answer certain research questions, either because they are too small and do not contain sufficient evidence for analysis (Kilgarriff and Grefenstette 2003), or because they are not up-to-date enough to document relatively new linguistic phenomena (Brekke 2000). In other cases, e.g. the study of specialized linguistic sub-domains or minority languages, no resource exists (Baroni and Bernardini 2004, Scannel 2007).

The lack of adequate corpus resources is particularly felt in cross-language studies and NLP, where parallel corpora (originals in language A and their translations into language B) are often needed but are not available due to the scarcity of relevant (easily and freely accessible) textual materials. In these cases, too, attempts have been made to use the Web as a data source. Resnik and Smith (2003) and Chen and Nie (2000), for example, propose two distinct algorithms to automatically build bilingual corpora from the Web for a variety of language pairs. Their corpora, however, suffer from a number of problems, such as their relatively small size (Resnik and Smith 2003 report that their largest corpus, for the language pair English-Chinese, contains fewer than 3,500 document pairs), and the impossibility to distinguish with certainty which document in a pair is the original and which is the translation.

A more promising approach to using the Web for mining cross-linguistic data is to exploit Web texts to build ‘comparable’ – rather than ‘parallel’ – corpora, and design algorithms that do not require input texts to be one the translation of the other. Drawing on early work by Rapp

(1995) and Fung (1995), there is by now a large and growing literature on using "unrelated" (non-parallel) corpora for tasks such as MT and automatic construction of bilingual lexicons (see also section 4). Witness to this is a workshop organized at the 2008 LREC conference, whose aim was to explore the potential of comparable corpora in tasks for which parallel corpora are traditionally considered the mainstays (Zweigenbaum *et al*. 2008). The Web was used extensively by the workshop contributors to retrieve (monolingual) corpora for multiple languages sharing similar topics or genres, such as corpora composed of science news texts (Saralegi *et al*. 2008) or online newspaper texts (Otero 2008).

In the pilot study described in this chapter we used two very large, Web-derived corpora of British English (ukWaC) and French (frWaC). Our aim in building them was to set up resources that would be similar, in terms of the variety of text types and topics, to more traditional general language corpora (in particular, the BNC, a well-established standard for British English; to the best of our knowledge, no similar resource exists for French). UkWaC and frWaC thus aim at providing similar "reference" resources for the languages of interest, rather than being comparable to each other by design, as was the case for the corpora used in the experiments discussed above.[1] However, given their large dimensions (>1 billion words), and since they were built following the same procedure, which is described in greater detail in section 2.2 below, we hypothesize that they could perform comparably in a task whose aim is to extract lexicographically relevant information for the languages in question.

2.2. Introducing the WaCky pipeline: frWaC

2.2.1. Introduction

This section briefly describes the procedure that was followed to construct the corpora used in the experiment. It should be noted that the construction of ukWaC and frWaC follows the procedures of two similar corpora of German (deWaC) and Italian (itWaC) – these resources are among the achievements of an international research initiative called *WaCky* (***W****eb* ***a****s* ***C****orpus* ***k****ool* ***y****nitiative*).[2] Since the procedure developed within this project (described in detail in Baroni *et al*. 2009, and Ferraresi *et al*. 2008, the latter focusing on ukWaC) is largely language-independent, in this section attention will be paid especially to those aspects specific to the construction of frWaC. We will focus in particular on the initial steps of the procedure, i.e. "seed" URLs selection and crawling, during which critical language-specific decisions regarding the document sampling strategy are made.

2.2.2. Seed selection and crawling

Our aim was to set up resources comparable to more traditional general language corpora, containing a wide range of text types and topics. These should include both "pre-Web" texts of a varied nature that can also be found in electronic format on the Web (ranging from sermons to recipes, from technical manuals to short stories, and ideally including transcripts of spoken language as well), and texts representing Web-based genres (Mehler *et al*. forthcoming), like personal pages, blogs, or postings in forums. It should be noted that the goal here was for the corpora to be representative of the languages of interest, i.e. (for frWaC) contemporary French in general, rather than representing the French Web.

The first step consisted in identifying sets of seed URLs which would ensure variety in terms of content and genre. In order to find these, around 1,800 random pairs of randomly selected content words were submitted to Google. Previous research on the effects of seed selection upon the resulting Web corpus (Ueyama 2006) suggested that automatic queries to Google which include words sampled from traditional written sources such as newspapers and reference corpus materials (which typically privilege formal written language) tend to yield "public sphere" documents, such as academic and journalistic texts addressing socio-political issues and the like. Submitting queries with words sampled from a basic vocabulary list, on the contrary, tends to produce corpora featuring "personal interest" pages, like blogs or bulletin boards. Since it is desirable for both kinds of documents to be included in the corpus, different seed sources were sampled. Two sets of queries were generated: the first set (1,000 word pairs) was obtained by combining mid-frequency content words from a collection of texts published between 1980 and 2000 in the *Le Monde Diplomatique* newspaper. In order to obtain more basic, informal words, the second list of queries (769 word pairs) was generated from a vocabulary list for children from eight to ten years old.[3] The URLs obtained from Google were submitted to the Heritrix crawler in random order,[4] and the crawl was limited to pages in the .fr Web domain whose URLs do not end in a suffix cuing non-html data (.wav, .jpg, etc.).

2.2.3. Post-crawl cleaning and annotation

The crawled documents underwent various cleaning steps, which were meant to drastically reduce noise in the data. First, only documents that were of mime type text/html and between 5 and 200 KB in size were kept for further processing. As observed by Fletcher (2004), very small documents tend to contain little genuine text (5KB counting as "very small" because of the html code overhead) and very large documents tend to be lists of various sorts, such as library indices, shop catalogues, etc. We also identified and removed all documents that had perfect duplicates in the collection, since these turned out to be mainly repeated instances of warning messages, copyright statements and the like. While in this way we might also have wasted relevant content, the guiding principle in our Web-as-corpus construction approach is that of privileging precision over recall, given the vastness of the data source.

All the documents that passed this pre-filtering stage underwent further cleaning based on their contents. First, code (html and Javascript) was removed, together with the so-called "boilerplate", i.e., following Fletcher (2004), all those parts of Web documents which tend to be the same across many pages (for instance disclaimers, navigation bars, etc.), and which are poor in human-produced connected text. From the point of view of our target users, boilerplate identification is critical, since too much boilerplate will invalidate statistics collected from the corpus and impair attempts to analyze the text by looking at KWiC concordances.

Table 16-1. Size data for frWaC and ukWaC

	frWaC	ukWaC
n of seed word pairs	1,769	2,000
n of seed URLs	6,166	6,528
raw crawl size	470 GB	351 GB
size after document filtering and near-duplicate cleaning	9 GB	12 GB
n of documents after near-duplicate cleaning	2.2 M	2.69 M
size with annotation	27 GB[6]	30 GB
n of tokens	1,027,246,563[6]	1,914,150,197
n of types	3,987,891[6]	3,798,106

Relatively simple language filters were then applied to the remaining documents, so as to discard documents in foreign languages and machine-generated text, such as that used in pornographic pages to "trick" search engines. Finally, near-duplicate documents, i.e. documents sharing considerable portions of text, were identified and discarded through a re-implementation of the "shingling" algorithm proposed by Broder *et al.* (1997).

At this point, the surviving text was enriched with part-of-speech and lemma information, using the TreeTagger.[5] Table 16-1 gives size data about each stage of the construction of frWaC; the same kind of information is also provided for ukWaC.

2.3. Corpus use and dictionary making

The lexicographic industry has always been one of the driving forces behind corpus development, as well as being one of its main beneficiaries. Two of the major corpus building projects of the 1990s, leading to well-known and widely used resources like the Bank of English and the BNC, were carried out by academic-industrial consortia in which publishing houses featured prominently, and which saw "reference book publishing" as the primary application area for the corpora (Burnard 1995). Sinclair's work on the Cobuild dictionaries (described e.g. in Sinclair and Kirby 1990) shows how corpus-informed methods could profitably be applied to obtain information about word and word sense frequency (thus guiding selection from a pre-compiled – e.g. dictionary-derived – headword list), collocation, syntactic patterning, typical usage and typical form (e.g. of a verb). But the corpus also made it to the published Cobuild dictionaries in a more noticeable way, providing not only examples but also the raw material for the well-known Cobuild definitions, for example, of *immune*: "if you are immune to a disease you cannot be affected by it" as opposed to "Protected from or resistant to some noxious agent or process" (OED online). These definitions sometimes also included subtle meaning generalizations unlikely to be obtainable from sources other than the corpus, e.g. typical 'semantic prosodies', for example, of *set in*: "if something *unpleasant* sets in it begins and seems likely to continue or develop" (emphasis added).

Within English lexicography, corpus resources are nowadays generally recognized as indispensable tools of the lexicographer's trade, even by professionals stemming from a non-corpus tradition (see e.g. Landau 2001). Caveats and limitations do remain, of course, both with regard to corpus construction and processing. However large and carefully built, no corpus will ever represent the whole of a language, including its potential for creativity; furthermore, corpora soon become obsolete for the purposes of lexicography, requiring constant updates and enlargements (Landau 2001). In terms of corpus processing, reliance on automation (of corpus annotation and querying) is becoming indispensable as corpora become larger and larger; yet NLP tools (taggers, lemmatizers, parsers) might hide evidence about uncommon or novel usages, while "smart" query tools (Kilgarriff *et al*. 2008), welcome as they are for speeding up the lexicographer's work, inevitably reduce her control over data selection. Nonetheless there seems to be general consensus that, as claimed by de Schryver (2003: 167), "no serious compiler would undertake a large dictionary project nowadays without having one (and preferably several) [corpora] at hand." Interestingly, availability of texts in electronic format through search engines such as Google has not made corpora obsolete, quite the contrary. At the moment, these tools are not sophisticated enough to cope with the needs of linguists (Lüdeling *et al*. 2007), and chances are slim that they will ever be, thus making the provision of very large and up-to-date corpora still a priority for linguists and for the language industry. This is especially true of languages like French, for which large and easily accessible corpus resources are still scarce.

3. Evaluating Web corpora for lexicography: Our pilot investigation

3.1. Objectives and method

In the pilot study described in this chapter our aim was that of using our automatically-constructed Web corpora for a practical application, namely to derive information about language use for dictionary making / revision. This task can only provide us with indirect evidence about corpus contents and cross-linguistic comparability, yet our take on such issues as quality and representativeness in corpus construction, especially when it gets to large and automatically-constructed corpora, is that the proof of the corpus is in the using. The number of users and usages to which a corpus is put is the ultimate testimony to its scientific as well as practical value, and this applies to automatically and manually constructed corpora alike – cf. also the position taken by Atkins *et al*. (1992: 5) in a seminal paper on the representativeness of (manually constructed) corpora:

> [a]ll samples are biased in some way. Indeed, the sampling problem is precisely that a corpus is inevitably biased in some respects. *The corpus*

> *users must continually evaluate the results drawn from their studies and should be encouraged to report them*. (emphasis added)

The task described in this chapter simulates corpus-based lexicographic practice and combines / contrasts corpus insights with translator / lexicographer input. Collocational information about three English lexical headwords is collected from ukWaC and submitted to a lexicographer for validation. The validated collocations are then translated with the help of frWaC and with the assistance of a professional translator from English into French. A detailed description of the method follows.

The extraction of potentially interesting English collocational complexes was done in three steps. First, we needed to select words which a lexicographer may want to analyse when dealing with a dictionary revision task. We therefore asked a lexicographer (a native speaker of British English) to provide us with a list of English words (one for each of the three main lexical word classes, i.e. one adjective, one noun and one verb), whose entries might in his opinion be in need of revision in a French-English bilingual dictionary. The words selected by the lexicographer were *hard* (adjective), *point* (noun) and *charge* (verb).

The second step consisted in extracting potentially interesting collocational complexes these headwords may take part in. To do this, we wrote simple rules for the extraction of candidate pairs according to syntactic criteria. While this method has potentially lower recall than one based on simple co-occurrence (i.e., one that disregards syntactic patterning), and is vulnerable to tagging errors, we reckoned that precision should be favoured over recall: since professional lexicographers are typically hard-pressed for time, limiting the amount of noise in the lists was crucial. The idea of overcoming the limitations of "grammatically blind" collocation lists relying on syntactic patterning is also at the basis of the Sketch Engine,[7] a widely used (commercial) corpus query tool especially designed for lexicographic needs. The patterns we chose were:

- for *hard*: all the nouns that occur in a span of one-three words on the right of the adjective;
- for *point*: all the nouns that occur in a span of one-three words on the left of the noun;
- for *charge*: all the nouns that occur in a span of one-three words on the right of the verb.

Notice that these grammatical patterns are also used in the Sketch Engine (Kilgarriff *et al*. 2004) and are among what Atkins *et al*. (2003: 278) describe as “lexicographically relevant sentence constituents”. In all the three cases we extracted lemmas, and did not take into account the words intervening between node and collocate (i.e. we do not distinguish between, e.g. *access point* and *access to this point*).

The extracted pairs were ranked according to the log-likelihood scores (Dunning 1993).[8] The top 30 collocational complexes extracted from ukWaC and the BNC were merged into a single list and sorted in alphabetical order. The lexicographer was then asked to look at the three lists and flag the sequences he reckoned might be considered for inclusion in the English part of an English/French bilingual dictionary (whether as a usage example, or a collocation, or anywhere else in the entry), and to provide at his discretion additional comments and observations. Table 16-2 reports data about the number of word pairs which were sent out to him for evaluation, split according to the corpus they were extracted from. The lexicographer analyzed the three lists, evaluated their relevance to the specified task, added his comments and returned the files.

Table 16-2. The extracted English collocations

Source corpus	N.	%
ukWaC and BNC (shared pairs)	51	39.6
Only ukWaC	39	30.2
Only BNC	39	30.2
Total	129	100

The second part of the study consisted in finding likely translation equivalents in frWaC for some of the collocational complexes previously validated by the lexicographer. For this task, which was substantially more labour-intensive than the previous one, we focused on the two major senses / uses of the verb *charge* identified by the English lexicographer, roughly corresponding to the following collocate sets:

1. charge -- assault, burglary, connection, conspiracy, crime, fraud, kidnapping, manslaughter, misconduct, murder, offence, possession, rape, sedition, theft, treason
2. charge -- amount, commission, fee, interest, penalty, pound, premium, price, rate, rent, tax, VAT

For the first sense (“bring an accusation against”, OED online), two translation equivalents of the node word *charge*, namely *inculper de* and *accuser de* were looked up in frWaC, and the 60 most frequent noun collocates in a span of 1-3 words to the right were selected. Out of this list, the most likely potential equivalents of the English noun collocates were selected and submitted for evaluation to a professional translator from English into French (a native speaker of French).

For the second sense, i.e. "to impose, claim, demand, or state as the price or sum due for anything" (OED online), the method was reversed. The translator was asked to provide equivalents for the collocate nouns (*somme / montant*, *commission*, *frais*, *intérêt*, *pénalité*, *prime*, *price*, *taux*, *loyer*, *taxe / impôt*", *TVA*).[9] The verb collocates in a span of 1-3 words to the left of these nouns were searched for in frWaC and the 30 most frequent ones were extracted. Potential translation equivalents found in these lists (KWIC concordances were consulted when in doubt) were then compared with those suggested intuitively by the translator.

3.2. Results

3.2.1. The validated English collocations

The lexicographer analyzed the 129 submitted word pairs, put a tick (✓) next to those that he found to have lexicographic relevance, and provided comments about the different ways in which these expressions might be treated in a dictionary. For instance, with reference to the "hard + [noun]" bigrams, he commented that "Almost every item […] would be an essential inclusion in a (bilingual) dictionary. They are what I consider to be lexicalized, 'hard' collocations with independent meanings." In other cases (e.g. *charge*), he pointed out that most of the submitted pairs would only be included as "example collocates given under productive sense categories", and provided labels for such potential sense categories, roughly corresponding to Sinclair's (1996) 'semantic preferences' (e.g. "charge + [offence: murder, assault, theft…]"), or commented that a given sequence would probably only be included in larger dictionaries (e.g. *acupuncture point*). In a few cases (about 8% of the total submitted pairs) he was unsure about the lexicographic relevance of the pair, or his intention was unclear (these cases were marked by a question mark **?**). Given that a corpus can play several roles in the making and revision of a dictionary, including signalling semantic prosodies and preferences and providing examples, and that evaluation of relevance is conditional on the specific task at hand, we consider as validated all word pairs for which we had a definite tick ✓ or an uncertain question mark **?**, regardless of accompanying comments (though we do take comments and uncertainties into account in the more qualitative part of the analysis of results).

The results of the expert validation (Table 16-3) suggest that more than 70% of the word pairs automatically extracted from both the BNC and ukWaC would be potentially relevant for lexicographic purposes, with both corpora contributing very similar numbers of valid collocations. In fact, a slightly higher overall number of valid collocations come from ukWaC than from the BNC (76 vs. 72), even though ukWaC also has a higher number of uncertain cases, which, if factored out, tip the balance in favour of the BNC (69 vs. 67). The similarity in the numeric results obtained from the two corpora is confirmed by the substantial overlap in terms of the actual sequences found. While each corpus contains between 25% and 30% of collocations not found in the other, as many as 45% are present in both lists. These are likely to be the stronger, more time-resistant "core" collocations in the language, e.g.:

- power point, vantage point, melting point
- hard cash, hard hat, hard shoulder
- charge battery, charge offence, charge fee

Table 16-3. Results of expert validation

Source corpus	Yes	Maybe	Selected	%
ukWaC and BNC (shared pairs)	45	1	46	45.0
BNC (not in ukWaC)	24	2	26	25.4
ukWaC (not in BNC)	22	8	30	29.4
ukWaC total	67	9	76	74.5
BNC total	69	3	72	70.5
Total	91	11	102	100
Out of total submitted	129			

Moving on to an analysis of results broken down by pattern (Table 16-4), more than 50% of the validated sequences for both the "hard + [noun]" and "[noun] + point" sequences are found in both corpora. Yet the two patterns differ in terms of percentages of valid collocations found only in one or the other corpus. While ukWaC and the BNC have similar numbers of "hard + [noun]" collocations (26 and 27), this is not the case with "[noun] + point" collocations. As many as 25 out of 28 sequences following this pattern taken from ukWaC are judged to be valid, in relation to only 18 from the BNC. This seems mainly due to several occurrences in the BNC list of "[number] point" (e.g. *eleven point*, *fourteen point*, *nought* *point*, *O point*, *twelve point*; notice that this pattern is not attested in the ukWaC list).

Table 16-4. The validated English collocations broken down by pattern

Source corpus	Yes	Maybe	Selected	%
hard				
ukWaC and BNC (shared pairs)	18	1	19	55.8
BNC (not in ukWaC)	7	1	8	23.5
ukWaC (not in BNC)	6	1	7	20.5
ukWaC total	24	2	26	76.4
BNC total	25	2	27	79.4
Total selected	31	3	34	100
Out of (submitted)	41			
point				
ukWaC and BNC (shared pairs)	15	0	15	53.5
BNC (not in ukWaC)	2	1	3	10.7
ukWaC (not in BNC)	7	3	10	35.7
ukWaC total	22	3	25	89.2
BNC total	17	1	18	64.2
Total selected	24	4	28	100
Out of (submitted)	41			
charge				
ukWaC and BNC (shared pairs)	13	0	13	31.7
BNC (not in ukWaC)	15	0	15	36.5
ukWaC (not in BNC)	9	4	13	31.7
ukWaC total	22	4	26	63.4
BNC total	28	0	28	68.2
Total selected	37	4	41	100
Out of (submitted)	47			

The "charge + [noun]" pattern is even more interesting in terms of qualitative differences between the two corpora. While similar numbers of valid collocations are found in ukWaC and the BNC, with the BNC performing slightly better than ukWaC (28 vs. 26 collocations), an analysis of the actual patterns found in the two corpora and of the lexicographer's comments suggests that the BNC output may in fact be less relevant for lexicographic purposes than that from ukWaC. This is because as many as 15 (out of 30) pairs exemplify the pattern “charge (someone) with [offence]”. The lexicographer rightly commented that these would only be relevant as examples of the general pattern, but each actual sequence would contribute little to an understanding of word usage, and would certainly not be included as strong collocations. The BNC output thus provides fewer instances of collocations featuring the *charge* sense (i.e. *take as payment*) of the verb (e.g. “charge + fee, price, VAT, penalty, rent”), and no instance of the pattern “charge + [person]”, e.g. *charge customer* found in ukWaC and validated by the lexicographer.

```
to despatch . We will not <charge your card> until we have confirmed
hed in the UK . We do not <charge credit cards> until goods are avail
 . I do n't mind manually <charging credit cards> at all and if I too
 and Switch . We will not <charge your card> until your order is disp
booking if less ) will be <charged to your card> by the Rowcroft Hote
 price you pay . Will you <charge my credit card> when I book ? No ,
 N.B. Boys STUFF will not <charge your card> until we are ready to di
to despatch . We will not <charge your card> until we have confirmed
 over the phone . We will <charge your credit card> manually . Pre-pa
ment and clothing . Goods <charged by credit card> are normally dispa
```

Figure 16-1. Ten occurrences of “charge + card” from ukWaC

```
ess of the central London <charging zone> has shown that tolls on
. But unlike the existing <charging zone> , there would be no flat
Traffic delays inside the <charging zone> remain 30 % lower than b
ondon or other congestion <charging zones> . Overspeed warning , th
ing the eight square mile <charging zone> . Anyone who enters the
 further extension of the <charging zone> should only be considere
i of synagogue within the <charging zone> , quoted in the Observer
ing within the congestion <charging zone> . This will involve a su
 ride into the congestion <charging zone> . This may be on-street
to pay ? Residents in the <charging zone> can register their vehic
```

Figure 16-2. Ten occurrences of “charge + zone” from ukWaC

More importantly, the only two collocations to get two ticks out of the total submitted (signalling high relevance according to the lexicographer) were found in the ukWaC output for “charge + [noun]”, namely: “charge + card” and “charge + zone” (see Figures 16-1 and 16-2). While a few occurrences of “charge + card” do occur in the BNC (7, the numbers are too small to make it to our list), the collocation *(congestion) charging zone* is completely absent from the corpus. As can be seen in Figure 16-2, the expression refers to a traffic regulation scheme first implemented in London in 2003, and nowadays operating in many other cities within and outside the UK. It is not surprising that the expression is absent from the

BNC, created in the early 1990s, and that the lexicographer found it particularly relevant for purposes of dictionary revision.

3.2.2. Translation equivalents from frWaC

With reference to the "bring an accusation against" sense of *charge*, a quick browsing of the top 60 noun collocates of *inculper de* and *accuser de* in frWaC shows that 12 out of 16 collocation equivalents of the noun collocates found in ukWaC are present in the output, namely:

- charge burglary ~ inculper vol; accuser vol
- charge connection ~ inculper complicité; accuser complicité
- charge conspiracy ~ inculper conspiration
- charge crime ~ inculper crime; accuser crime
- charge fraud ~ inculper fraude; accuser fraude
- charge manslaughter ~ inculper homicide
- charge murder ~ inculper homicide; inculper meurtre; accuser meurtre
- charge offence ~ inculper délit; accuser délit
- charge possession ~ inculper détention
- charge rape ~ inculper viol
- charge theft ~ inculper vol; accuser vol
- charge treason ~ accuser trahison; inculper trahison

All these translation equivalences were validated by the French translator.

With reference to the second sense, i.e. "to impose, claim, demand, or state as the price or sum due for anything" (OED online), Table 16-5 shows the potential equivalents for *charge* in these collocations, which were found in browsing the top 30 verbs co-occurring with each noun in frequency order.

In Table 16-5, a tick means that a given verb was found in frWaC among the top collocates for the corresponding noun, while a star means that the verb was given (intuitively) by the French translator as an equivalent of *charge* in that collocation. While a star not accompanied by a tick does not imply that the translator's intuition is faulty,[10] the partial overlap between corpus evidence and the translator's intuition does suggest that browsing a large corpus such as frWaC is crucial for several reasons.

Table 16-5. French equivalents of *charge* (“demand a sum”)

	commission (commission)	droit (duty)	frais (fee)	intérêt (interest)	loyer (rent)	pénalité (penalty)	prime (premium)	prix (price)	somme/ montant (amount)	taux (rate)	taxe/ impôt (tax)	TVA (VAT)
appliquer	*	*				✓	*	*		✓*	✓	✓*
demander	*						✓					
facturer	*		*		*			*	*			✓
faire									✓			
imposer						✓				✓		
infliger						✓						
lever											✓	
payer			✓	✓			✓		✓		✓	✓
percevoir	✓	*			✓						✓	
pratiquer					✓			✓				
prendre	✓											
prélever				*			*				✓*	
recevoir	✓											
réclamer					✓							

```
Les transportés <paieraient une somme> qui serait un peu supérieure aux tarifs des transports en commun
Les emprunteurs <paient une commission> de risque, versent souvent une contribution restituable
ceux qui ont une activité nécessairement polluante <payent une taxe> et on droit à un allègement ensuite
```

Figure 16-3. Examples of “payer + [noun]” from frWaC

First, a translator might want to enlarge the pool of possible translation equivalents to be evaluated, rather than relying on her intuition only. Second, certain equivalents of *charge* (e.g. *appliquer* and *percevoir*) seem to collocate more widely than others, and would therefore provide safer bets, e.g. for compact dictionaries with limited space. Finally, the presence of *payer* in several lists would seem to suggest that the action of "imposing a price / sum / tax etc. for something" might preferably be encoded in French from the perspective of the *imposee*. Both the verbs *faire* and *payer* appear among the collocates as part of the phrase *faire payer*, and several occurrences of "payer + [noun]" would indeed appear to be translatable into English as "be charged + [noun]" (see Figure 16-3).

3.3. Discussion

The results discussed in section 3.2 suggest that the *WaCky* pipeline, in which texts are collected opportunistically from the Web through automatic routines requiring no manual intervention, produces corpora that compare favourably with high-quality benchmark resources. With reference to the English part of the study, previous attempts at evaluating ukWaC have shown that the corpus is indeed reasonably similar to the BNC in terms of topic coverage (Ferraresi *et al*. 2008). The current, more functionally-oriented study shows that the two corpora perform comparably in the collocation extraction task, with a slight edge for ukWaC. Besides (obviously) providing more up-to-date results, the latter corpus has a better coverage of different word senses, and does not give undue prominence to uninteresting collocates (e.g. "[number] point"). While one cannot rule out the possibility that this is an effect of the statistical measure used to extract the collocates, the co-occurrence statistic used, log-likelihood, is standard in corpus work, and has been shown to provide stable results from corpora of the size of the BNC and smaller (Dunning 1993). The edge in favour of ukWaC is likely to be the result of its substantially larger size, which makes up for the (probably) more reduced variety of texts sampled. Given that constantly updated and very large corpora are required for lexicographic purposes (Landau 2001), and that building a carefully designed corpus like the BNC is very costly and time-consuming, Web corpora would appear to be a viable alternative.

The results of the second part of our study, using the newly-built French corpus for finding translation equivalents, could not rely on a benchmark for comparison, given that no corpus comparable to the BNC is available in the public domain for the French language. Nonetheless, our study has shown that frWaC provides a very large number of plausible potential collocations that a lexicographer / translator could draw from when translating collocations or examples from English into French. For one of the senses of the verb *charge* (i.e. “bring an accusation against”), an automatic search for two central equivalents returns most of the French noun collocates corresponding to the English noun collocates found in ukWaC. For a second sense of *charge* (“demand a sum”), the opposite method (searching for the verbal collocates of specific nouns) provides a list of verbs that a lexicographer could choose from when translating *charge* collocations, and gives an idea of which verbs have a wider or more restricted range of collocating nouns, and what these are. Comparison between the collocations intuitively suggested by a native speaker translator and those found in the frWaC lists (limited to the top 30 pairs in frequency order) shows that all the verbs the translator came up with are also present in frWaC, though not necessarily in combination with the same nouns. More interestingly perhaps, frWaC suggested that *(faire) payer* is a favourite option that the translator did not come up with, possibly because of its less lexicalized status.

4. Conclusion and further work

This article has introduced two very large corpora of English and French built from the Web following similar procedures, namely ukWaC and frWaC. Previous studies have shown that evaluating Web corpus contents is an extremely arduous task. This becomes daunting when one attempts to also compare these contents cross-linguistically; therefore an empirical approach was favoured. A bilingual lexicographic task was set up simulating part of a dictionary revision project. English source language collocations were extracted from ukWaC and from a benchmark corpus (the BNC), and validated by a lexicographer. This phase of the task suggested that an automatically built corpus like ukWaC provides results that are both quantitatively and qualitatively similar to those provided by a smaller and older but much more carefully constructed corpus, and (as could be expected) that these results are more useful for a lexicographer because they provide a more up-to-date snapshot of language in use, and because a larger corpus provides a better coverage of certain word senses. In the bilingual task, the French corpus was used to seek likely translations for the English collocations. This second part of the task was meant to ascertain that the two Web corpora are similarly adequate for practical purposes. Since we are not aware of any publicly available benchmark corpus for French, we tried to establish the validity of frWaC by implication, first by showing that ukWaC performs slightly better than the BNC in this task, and then establishing that comparable linguistic information can be obtained from ukWaC and frWaC, thus suggesting that frWaC is a valid reference resource for French lexicography, and arguably for the French language in general.

While the results obtained in this study are very encouraging (one should not forget that Web corpora such as those described here are built fully automatically, with no control over corpus contents, and that therefore their validity cannot be assumed *a priori*), a lot remains to be done. We see two main areas in which further work is needed, first to improve on the Web corpora themselves with respect to the requirements of lexicographers, and secondly to investigate the extent to which these new and freely-available resources can be exploited for lexicographic purposes.

With regard to the first area, in the immediate future we intend to make frWaC available through the Sketch Engine. This software provides users with so-called "word sketches", i.e. summaries of a word's collocational behaviour, generated automatically using rather sophisticated pre-defined syntactic rules. In this chapter very simple rules were used for extracting collocations, which only specified the distance between the node word and its collocates, and allowed for virtually no flexibility in the searched patterns, thus probably discarding several interesting collocations not matching the search pattern exactly, and including some noise. While leaving the tasks of devising search rules and the need of browsing large amounts of data with the individual user places her in control, we should not forget that "[t]ime pressures too often push the lexicographer to cut corners to avoid time-consuming analyses", and that "no one will have the time to sort through thousands of hits […] in order to find a particular usage that has to be included" (Landau 2001: 322-323).

The trade-off between ease of consultation and control is certainly not news to corpus users, but with Web corpora being constructed opportunistically and reaching sizes of one billion words or more, it is likely to become a major issue in the future, with practical and theoretical implications that need to be explored. In the longer run, we hope to be able to include genre / topic information with the texts in the two corpora. In this sense, one aspect that seems particularly worthy of attention in Web-as-corpus linguistics at large is the development of classificatory schemes that can be applied automatically to Web corpora. Often a lexicographer or translator will need specialized (sub-)corpora rather than huge undifferentiated masses of text, e.g. when seeking evidence about specialized senses of a certain word, or when compiling thematic sections. While traditional corpora often contain extra-linguistic information annotated with the texts by the corpus compilers, the creation method used for Web corpora and their size makes manual annotation impossible. Work is therefore underway (see e.g. Biber and Kurjan 2007, Sharoff forthcoming) to come up with genre / topic typologies adapted to Web genres that can then be used to classify documents within Web corpora using probabilistic classifiers.

Moving on to the second area, i.e. potential applications of multilingual Web corpora to lexicography, we see two main ways in which our corpora can be of immediate relevance. First, in the production of headword lists of currently used single words and phrases, providing suggestions for new inclusions in revised editions of existing dictionaries. Even relatively straightforward methods, e.g. filtering a lemma list from ukWaC using an older corpus as a stoplist, work very well. For instance, the top ten most frequent nouns in ukWaC after filtering out nouns also found in the BNC are: *website*, *dvd*, *websites*, *sudoku*, *linux*, *html*, *Google*, *url*, *blog* and *homepage*. An even simpler procedure, applying no filtering whatsoever and simply listing the most frequent noun-noun sequences in the corpus, provides the following list (Table 16-6) of high usage potential multiword expressions in English and in French (in the case of French an optional empty slot is allowed between the two nouns).

Table 16-6. Frequent noun-noun sequences in ukWaC and frWaC

French phrase	fq.	English phrase	fq.
mot de passe	30,379	Web site	175,642
chiffre d’affaires	31,831	case study	81,127
projet de loi	42,517	search engine	70,514
site Internet	44,578	application form	66,693
millions d’euros	44,901	credit card	65,198
prise en charge	48,657	Web page	60,626
base de donnée	50,725	car park	56,721
site Web	55,954	health care	48,833
point de vue	69,419	climate change	47,655
mise en place	73,216	email address	46,643

A lexicographer can quickly browse through these lists to pick expressions that might be included in a revised edition of a dictionary, or whose entries are in need of revision due to their having become key in a given culture (note e.g. the high frequency of *climate change* in ukWaC). Secondly, and more challengingly, we intend to investigate the potential of our Web corpora for the automatic extraction of bilingual collocation pairs. In previous work (see section 2.1), attempts have been made at developing algorithms that find likely translations of single words from relatively small comparable corpora composed of homogeneous classes of textual materials – mainly newspaper texts. Collocational complexes such as those described in this chapter, i.e. noun-noun, verb-noun and adjective-noun word pairs, constitute much rarer events in a language than words taken in isolation. For this reason, larger data sets are needed to address the problem of data sparseness in tasks involving automatic extraction of collocations. We would therefore like to test the suitability of corpora like ukWaC and frWaC for such tasks, on the hypothesis that size could compensate for reduced comparability by design. The results presented in section 3.2 are very encouraging in terms of the comparability of the linguistic evidence obtainable from the two corpora, especially since the likely translations were extracted through computationally simple methods. Applying a more fine-tuned algorithm on the *WaCky* corpora, we hope to be able to assist lexicographers in the complex task of establishing translation equivalents above the word level.

Notes

* We wish to thank Martyn Back and Christine Gallois for their help with the English collocations and French translations; Federico Gaspari and Sara Piccioni for input on the collocation extraction method, and Sara Castagnoli and Eros Zanchetta for this and for their contribution to the development of frWaC.

[1] Corpus comparability is far from being a clear-cut notion, especially when corpora contain non-homogeneous classes of textual materials (Kilgarriff 2001), or pertain to different languages (Bernardini and Zanettin 2004). Moreover, ukWaC and frWaC were built with semi-automated procedures (see section 2.2), thus reducing the possibility to control for the materials that ended up in their final set up. Given these difficulties, in this chapter we do not attempt to provide an evaluation of the similarity between the two corpora in terms of their contents, but rather try to establish whether they can provide comparable resources in the framework of a practical task, chosen among those most central to corpus linguistics (see section 2.3).

[2] http://wacky.sslmit.unibo.it/

[3] http://o.bacquet.free.fr/index.html

[4] http://crawler.archive.org/

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] These data refer to the beta version of frWaC which is available at the time of writing. We expect that after further processing token and type counts will stabilize to numbers similar to those pertaining to ukWaC.

[7] http://www.sketchengine.co.uk

[8] In order to decide on the best measure to use, a distinct mini-pilot study was conducted. Three lists displaying the top 100 collocational complexes of the noun *course* in ukWaC were first sorted according to raw frequency (FQ), log-likelihood (LL) and mutual information (MI). Seven expert linguist informants were then

asked to judge what list best fit their intuitions about the collocations of *course* (the word was picked opening a monolingual dictionary at a random page). Four people favoured the LL list, two the FQ list and only one the MI list. Based on these results, and on previous work on collocation extraction (cf. Evert 2008), the LL measure was adopted.

9. In this search we ignore one of the collocates validated by the English lexicographer, namely *pound*, since this has no obvious equivalent in French (*livre sterling*? *euro*? *franc*?).

10. Note that raw frequency was considered here, and only the top most frequently co-occurring verbs were analyzed; it is quite likely that other relevant verbs would turn up if one browsed a longer collocate list, and / or if more sophisticated co-occurrence statistics were used.

References

Atkins, S., Clear, J. and Ostler, N. (1992), “Corpus design criteria”. *Literary and Linguistic Computing* 7(2): 1-16.

Atkins, S., Fillmore, C. J. and Johnson, C. R. (2003), “Lexicographic relevance: Selecting information from corpus evidence”. *International Journal of Lexicography* 16(3): 251-280.

Baroni, M. and Bernardini, S. (2004), “BootCaT: Bootstrapping corpora and terms from the web”, in *Proceedings of LREC*, 1313-1316. Lisbon.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009), “The *WaCky* Wide Web: A collection of very large linguistically processed web-crawled corpora”. *Language Resources and Evaluation* 43(3): 209-226.

Bernardini, S. and Zanettin, F. (2004), “When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals”, in A. Mauranen and P. Kuyamäki (eds.) *Translation Universals: Do they Exist?*, 51-62. Amsterdam: Benjamins.

Biber, D. and Kurjan, J. (2007), “Towards a taxonomy of Web registers and text types: A multidimensional analysis”, in M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*, 109-131. Amsterdam: Rodopi.

Brekke, M. (2000), “From the BNC towards the cybercorpus: A quantum leap into chaos?”, in J. M. Kirk (ed.) *Corpora Galore: Analyses and Techniques in Describing English*, 227-247. Amsterdam: Rodopi.

Broder, A., Glassman, S., Manasse, M. and Zweig, G. (1997), “Syntactic clustering of the Web”, in *Proceedings of the Sixth International World Wide Web Conference*, 391–404. Santa Clara (CA), 7-11 April 1997.

Burnard, L. (1995), *Users Reference Guide for the British National Corpus*. Oxford: OUCS.

de Schryver, G. M. (2003), “Lexicographers’ dreams in the electronic-dictionary age”. *International Journal of Lexicography* 16(2): 143-199.

Chen, J. and Nie, J. Y. (2000), “Parallel web text mining for cross-language information retrieval”, in *Recherche d’Informations Assistée par Ordinateur (RIAO)*, 62–77. Paris, 12-14 April 2000.

Dunning, T. (1993), “Accurate methods for the statistics of surprise and coincidence”. *Computational Linguistics* 19(1): 61-74.

Evert, S. (2008), “A lexicographic evaluation of German adjective-noun collocations”, in *Proceedings of the Workshop on ‘Towards a Shared Task for Multiword Expressions’ at LREC*, 3-6. Marrakech, 1 June 2008.

Ferraresi, A., Zanchetta, E., Baroni, M. and Bernardini, S. (2008), “Introducing and evaluating ukWaC, a very large web-derived corpus of English”, in *Proceedings of the WAC4 Workshop at* LREC, 45-54. Marrakech, 1 June 2008.

Fletcher, W. (2004), “Making the Web more useful as a source for linguistic corpora”, in U. Connor and T. Upton (eds.) *Corpus Linguistics in North America 2002*, 191-205. Amsterdam: Rodopi.

Fung, P. (1995), “Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus”, in *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, 173-83. Boston (MA), 30 June 1995.

Kilgarriff, A. (2001), “Comparing corpora”. *International Journal of Corpus Linguistics* 6(1): 97–133.

Kilgarriff, A. and Grefenstette, G. (2003), “Introduction to the special issue on the Web as corpus”. *Computational Linguistics* 29(3): 1-15.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. and Rychlý, P. (2008), “GDEX: Automatically finding good dictionary examples in a corpus”, in *Proceedings of Euralex*, 425-432. Barcelona, 15-19 July 2008.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004), “The Sketch Engine”, in *Proceedings of Euralex,* 105-116. Lorient, 6-10 July 2004.

Landau, S. (2001), *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Lüdeling, A., Evert, S. and Baroni, M. (2007), “Using web data for linguistic purposes”, in M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus Linguistics and the Web*, 7-24. Amsterdam: Rodopi.

Manning, C., and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*. Cambridge (MA): MIT Press.

Mehler, A., Sharoff, S., Rehm, G. and Santini, M. (eds.) (forthcoming), *Genres on the Web: Computational Models and Empirical Studies.*

Otero, P. G. (2008), "Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora", in P. Zweigenbaum, E. Gaussier and P. Fung (eds.) *Proceedings of the Workshop on Comparable Corpora at LREC*, 19-26. Marrakech, 1 June 2008.

Rapp, R. (1995), "Identifying word translations in non-parallel texts", in *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, 320-322. Cambridge (MA), 26-30 June 1995.

Resnik, P. and Smith, N. (2003), "The web as a parallel corpus". *Computational Linguistics* 29(3): 349-380.

Saralegi, X., San Vicente, I., and Gurrutxaga, A. (2008), "Automatic extraction of bilingual terms from comparable corpora in a popular science domain", in P. Zweigenbaum, E. Gaussier and P. Fung (eds.) *Proceedings of the Workshop on Comparable Corpora at LREC*, 27-32. Marrakech, 31 May 2008.

Scannel, K.P. (2007), "The Crúbadán Project: Corpus building for under-resourced languages", in C. Fairon, H. Naets, A. Kilgarriff and G. M. de Schryver (eds.) *Building and Exploring Web corpora. Proceedings of the WAC3 Conference*, 5-15. Louvain, 15-16 September 2007.

Sharoff, S. (forthcoming), "In the garden and in the jungle: Comparing genres in the BNC and Internet", in A. Mehler, S. Sharoff, G. Rehm and M. Santini (eds.) *Genres on the Web: Computational Models and Empirical Studies.*

Sinclair, J. McH. (1996), "The search for units of meaning". *Textus* 9(1): 71-106.

Sinclair, J. McH. and Kirby, D. (1990), "Progress in English computational lexicography". *World Englishes* 9(1): 21-36.

PART III

CORPUS-BASED CONTRASTIVE STUDIES

CHAPTER SEVENTEEN

CONTRASTIVE CORPUS ANALYSIS OF CROSS-LINGUISTIC ASYMMETRIES IN CONCESSIVE CONDITIONALS

BART DEFRANCQ

1. Introduction

This chapter aims at describing the use of *wh*-items in universal concessive conditionals in English and French. It will do so on the basis of several corpora, including a parallel corpus of translated text.

Wh-items are nonspecific indefinite items which are used in interrogatives, free relatives, concessive conditionals and other free choice contexts. These items typically tend to form morphological paradigms, as in English where most items are written with initial '*wh*', hence the name *wh*-items. In French most items are written with initial *qu*-. Consequently, these items are referred in French as '*qu*-'. The following examples illustrate some of the various uses of *wh*-items in English (1-4) and in French (5-8):

(1) **Who** is responsible for purchase of equipment, etc? (BNC: B2M)

(2) **Whoever**'s running the course needs to fill in this particular form. (BNC: G4X)

(3) He won't have a go if you have a bad game, but he expects everyone to give their all. If he feels someone is not applying themselves 100 per cent, he won't spare reputations or ego --; **whoever** you are. (BNC: HTY)

(4) How did Andropulos or **whoever** know when, and from where, that bomber was leaving? (BNC: CKC)

(5) **Qui** viendra donc se plaindre, depuis la terrasse ensoleillée du huitième étage de l'Hôtel Bernini [...] de passer trois jours à Rome, même avec son pire ennemi ? (Le Monde, 20 octobre 2006)
[Who will complain on the sunny terrace on the eighth floor of Hotel Bernini that he has to spend three days in Rome, even with his worst enemy?]

(6) **Qui** dort dîne. (Le Monde, 27 septembre 2007, p. 36)
[Who sleeps, dines]

(7) "Il a été reconnu coupable. Cela prouve que le système fonctionne, **qui que** vous soyez", a commenté pour Reuters Pat McQuaid, président de l'Union cycliste internationale. (Le Monde, 21 septembre 2007, p. 19)
["He's been found guilty. That proves that the system works, whoever you are", Pat McQuaid, president of the Union cycliste internationale, said to Reuters.]

(8) "Je n'ai jamais cherché à nuire à **qui que ce soit**, j'invente rien, je ne peux rien ajouter."
[I have never sought to harm anyone, that is the truth, I've got nothing to add.]

As examples (1) and (5) show, *wh*-items appear as independent morphemes in interrogative clauses. In free relatives they can be supplemented with an additional *ever* in English, as in (2). In the other two uses, additional items are required in both languages. English resorts to the same item as the one used in free relatives (3 and 4). French makes use of different items depending on the nature of the *wh*- (6 and 7).

Examples (3) and (7) are both illustrations of the subordinate clauses that will be at the heart of this chapter. Following a well established tradition since Haspelmath and König (1998) (henceforth H&K), I will call them universal concessive conditionals (henceforth UCC), even though I have reasons to believe that the name is based on two misunderstandings (see below for further discussion). The semantics of these clauses will be discussed in detail in section 2. At this stage, it suffices to say that UCCs declare a specific category of items irrelevant for the realization of the content of the main clause. In (3), for instance, 'who you are' is irrelevant to being a possible victim of the person referred to by *he*.

UCCs are an interesting context to study the semantic properties of *wh*-items, because not all *wh*-items appear to be equally suited to be used in

UCCs. Several restrictions are reported in the literature (Quirk *et al*. 1985, Declerck 1991, Huddleston and Pullum 2002, Morel 1996, Grevisse and Goosse 2008), but hardly any credible explanation is offered. In addition, there are substantial differences between languages with respect to these restrictions. In English, UCCs can host all *wh*-items with the exception of *why* (Quirk *et al*. 1985, Declerck 1991, Huddleston and Pullum 2002). In French, on the other hand, only a small set of items is reported to be allowed, including the equivalents of *who* 'qui', *what* 'quoi', *which* 'quel' and *where* 'où'. German seems to have one of the most liberal UCCs, since even the equivalent of *why* 'warum' is marginally used in UCCs.

Since no actual corpus research has been performed on the compatibilities between *wh*-items and UCCs, they allow a corpus linguist to delve eagerly into virtually unexplored terrain. The contrasts that exist between languages are *gefundenes fressen* for contrastive linguists, as they may inform them of unsuspected and deep-rooted differences in the semantic structure of languages (*wh*-items are among the most fundamental and diachronically stable items in any particular language). Finally, they offer translation theorists a different perspective on how translators deal with lexical gaps. As the use of *wh*-items is diversely restricted in different languages, translators have to come up with creative solutions to compensate for the absence of particular items. Their choices can reveal basic semantic properties of structures and items.

The following corpora are used in the present study:

- the BYU-BNC, the online available version of the British National Corpus (100 million words). For some of the searches only the 'Newspaper' section was consulted;
- Le Monde via LexisNexis. Searches were restricted to 2005, 2006 and 2007 (approximately 63 million words);
- the JRC Acquis Communautaire Corpus. A parallel corpus of original texts with their translations in 21 languages. The JRC AC is essentially a corpus of legal and administrative texts produced by the institutions of the European Union. Its total size is approximately 636 million words. The size of the English part is approximately 34.5 million words. Searches were restricted to the most recent part of the corpus (2000-2006), totalling around 26.5 million words. Most of the recent texts are originally written in English and translated into the other languages.

This chapter is structured as follows. Section 2 discusses the semantic properties of UCCs. Section 3 gives an overview of the semantic properties of *wh*-items in English and in French, discussing similarities and differences on the basis of a number of independently defined categories. Section 4 presents more detailed information on the use of *wh*-items in UCCs on the basis of grammars and linguistic research on the one hand, and of monolingual corpus data on the other. Section 5 presents the translation data and section 6 makes room for some conclusions.

2. Semantics of UCCs

H&K point out that UCCs have conditional properties. They share the typical tense / mood combinations of conditionals and are analyzable in terms of protasis – apodosis, i.e. clauses bound by a conditional relationship. The main difference between prototypical conditionals and UCCs is that the protasis of the latter denotes several alternatives. Example (10) represents the conditional scheme of (9) with a, b, c and d representing the alternatives:

(9) Whatever medication you take, it won't help you

(10) If {*a* or *b* or *c* or *d* ...} then *q*

The conditional relationship underpinning UCCs is subject to caution: in UCCs, the apodosis is often presupposed, which is normally not the case of prototypical conditional clauses. In addition, the relationship cannot be described in terms of a logical entailment, as is the case of conditional clauses in general. As the apodosis is claimed to hold irrespective of the particular value instantiating the variable in the protasis, there is in fact no connection between both propositions in reality. As Gawron (2001) states, conditionals are used to express dependence, while UCCs are used to assert independence. The conditionality of UCCs resides in the fact that the protasis is presented as containing a parameter that is potentially (but not actually) influential for the apodosis. In (9), for instance, taking medication of some sort can help when feeling ill.

For most scholars, conditionality does not suffice to describe the semantics of UCCs. Indeed, other clause types involving *wh*ever*-items can express the same kind of conditionality as found in UCCs. The following is an example of a free relative clause:

(11) Both sides want to be the last to put forward a proposal, reckoning that if no rate is set, a court will pin the blame **on whoever** was last to reject one. (BNC: HSF)

It yields a semantic description in the following terms:

(12) if (a or b or c or d...) is last to reject a rate, the court will blame (a or b or c or d...)

which, apart from the repetition of the variable, is identical to H&K's proposal for UCCs: the parameter identity of the rejecting individual is potentially (but not actually) influential for the court's decision. To distinguish UCCs from free relatives, it is therefore necessary to integrate an additional, concessive component, by virtue of which the apodosis can be interpreted as running counter general intuitions about the result of what is denoted in the protasis.

As repeatedly stressed (König and Traugott 1982, Anscombre and Ducrot 1983), concessivity arises from the fact that a particular clause denotes a proposition that runs counter basic inferences drawn by individuals on the basis of what they consider to be plausible outcomes of a particular premise.

In the case of UCCs, concessivity adds up with conditionality. As pointed out before, the protasis of a UCC is presented as containing a parameter that is potentially (but not actually) influential for the apodosis. Concessivity implies that the relationship between the apodosis and the protasis can be understood in such a way that the apodosis is contrary to inferences about the plausible outcomes of what is denoted by the protasis. For instance, in (9) a plausible inference of the hearer taking medication is that it helps him / her recover. This inference is contradicted by the apodosis.

As far as quantification is concerned, H&K analyze UCCs as involving "some kind of universal quantification" (Haspelmath and König 1998: 566). To take this quantification into account, H&K propose the following formula for UCCs:

(13) $(\forall x)$ (if px then q)

where x represents the *wh*-variable (***who****-ever,* ***what****-ever*, etc) and p and q the protasis and apodosis respectively. As the protasis is quantified over, it denotes a set of propositions. The universal operator proposed in the formula seems to be meant as a first attempt at describing the semantics of UCCs. H&K specify that there is evidence that the exact nature of the quantification is free choice quantification and not universal quantification. This view is supported by Giannakidou (2001) and Vlachou (2003).

There is also a general intuition that UCCs involve scalarity. *Wh*ever*-items are believed to be associated in some way with extreme points on scales of values. However, the relationship is seldom clearly described and interpretations of scalarity vary considerably from one author to another. H&K, for instance, claim that the *wh*ever*-item denotes the two extreme points of a scale and, therefore, ranges over the whole set of values. Gawron (2001), on the other hand, states that *wh*ever*-items denote the minimal value of a scale, a point of view shared by Vlachou (2003), who specifies that the value concerned is a likelihood value: *wh*ever*-items denote values that are extremely unlikely to "cause" the event in the apodosis. It is also far from clear how a *wh*ever*-item can both denote a free-choice variable and an extreme point on a scale, as most authors seem to suggest. Free-choice quantification and scalarity are contradictory: the former implies an unordered set all members of which are representative of the set as a whole, whereas the latter implies an ordered set (scale) from which one value is singled out.

In sum, conditionality, concessiveness, free choice and scalarity are the main ingredients of UCCs. In Section 5 it will become clear that when facing translation problems, translators do opt to drop one of these meaning components and that their choice depends on the nature of the *wh*-item.

3. Semantics of *wh*-items

All known languages have paradigms of indefinite items that can be used in interrogative clauses. As pointed out, in most Indo-European languages these items are also used in (free) relatives, UCCs and other free-choice contexts. In quite a number of non-Indo-European languages, interrogative indefinites coincide with ordinary indefinites. This is for instance the case in Korean and Japanese (see for instance Cheng 1997).

Wh-paradigms are usually composed of a limited number of items which denote basic semantic categories such as referents: 'human' (*who*), 'non-human' (*what*); 'predicates' (*what* + V); features of referents or events: 'identification' (*which, what*), 'nature' (*what (kind of)*), 'place' (*where*), 'time' (*when*), 'manner' (*how*), 'reason-cause' (*why, how come*), 'purpose' (*why, what for*), 'amount-number' (*how much, how many*) and 'extent' (*how*). Obviously, most of these meanings can also be expressed by a composite form, called periphrastic here, involving the 'identification' item and a specific noun referring to time, place, etc., as in *what time, what place, what way,* and *what reason* etc. In this chapter, periphrastic forms will be considered to be instances of the 'identification' category and not of the different other semantic categories they would belong to according to the meaning expressed by the head.

In many languages the paradigm of *wh*-elements tends to be organized roughly along the same lines. Usually there are between four and seven elements covering the major semantic categories. As the number of semantic categories exceeds the number of available items, some of the items are used for more than one category. In English this is particularly the case of *how* and *what*. In a small number of cases, semantic categories are expressed by more than one element. This is the case of the *why-how come* pair in English, although there are admittedly subtle differences in use (cf. Collins 1991).

Languages differ only moderately with regard to the precise form-function correspondences. The general picture is one of great symmetry: in English and French, for instance (see Table17-1), there is a fairly straightforward relationship between the individual items of both languages in most cases.

Table 17-1. Semantic categories and *wh*-items in English and French

Categories	EN	FR
human	who	qui
non-human	what	que, quoi
predicate	what + V	que, quoi + V
identification	which (one) / what	(le)quel
nature	what (kind of), who	que, quoi ... comme, qui
place	where	où
time	when	quand
manner	how	comment
reason	why, how come	pourquoi, comment se fait-il
purpose	why, what for	pourquoi, pour quoi
amount, number	how much/many	combien
extent	how (much)	combien

Some asymmetry can be observed in the categories of ‘identification’, ‘amount, number’ and ‘extent’: in French, the ‘amount, number’ and ‘extent’ categories make use of a separate item *combien,* whereas in English the ‘manner’ item *how* is used for ‘extent’, sometimes in combination with *much* (*how bad is his condition?*[1] *how much do you* *really love her?*), while *how much* and *how many* are used for number and amount. On the other hand, *what* can be used as an interrogative determiner in English indicating identification, but its French non-human equivalents *que* and *quoi* cannot. Instead, French consistently uses the equivalent of *which* (*quel*).[2]

The similarity of both paradigms precludes any explanation of cross-linguistic differences in the context of UCCs in terms of paradigmatic differences between the languages involved.

In semantic theory, with the possible exception of *why*, *wh*-items are believed to be variables, which means that their interpretation partly depends on the kind of quantifier they appear with, much like indefinite expressions. This is certainly the case with adverbial quantifiers, such as *usually* and *mostly*, as shown by Berman (1994). However, when it comes to combining them with ordinary quantifiers, all logic seems to disappear. The universal quantifier *all*, for instance, can only be combined with *who* and *what* (the latter only in American English). The free-choice quantifier *any* can only be combined with *where* and *how*. The French free-choice quantifier *n'importe*, on the other hand, combines with all *wh*-items except *pourquoi*. This is hardly the kind of distribution one would expect when dealing with variables. *Wh*-items should therefore not be considered on the same footing as the variables used in logical formulae. They have semantic properties of their own, which make them more or less suitable for specific contexts.

4. *Wh*-items in UCCs

4.1. English

Grammars of English are consistent in terms of what they consider to be the *wh*-items that can be found in UCCs. Quirk *et al*. (1985), Declerck (1991), Tsai (1999), and Huddleston and Pullum (2002) all contend that all *wh*-items are used in UCCs, except the causal item *why*. The examples quoted above usually illustrate only the most typical meanings of the items involved: 'manner' for *how,* 'non-human referent' for *what*, etc. For a number of categories mentioned in Table 17-1, there is no information. Some corpus research is therefore needed to fill in the gaps.

Table 17-2 provides an overview of the corpus examples provided by the BNC:

Table 17-2. Wh-items used in UCCs in English

Categories	Examples
human	(14) Anti-fascists argued that **whoever** was to blame for the violence, the police and courts treated them more harshly, and the National Council for Civil Liberties certainly produced reliable testimony to back up those claims. (BNC: CS6)
non-human	(15) But **whatever** the papers think, and **whatever** the English management says, there has never been any trouble between us and the English players. (BNC: CH7)
predicate	(16) But **whatever** happens we're providing a platform for the season that is exciting. (BNC: CH3)
identification	(17) Now that Lamb has blown the whole affair into the open, Sir Colin must see the cheats are exposed and the door slammed forever on the ball doctors, **whichever** country they belong to. (BNC: CH3)
nature	(18) **Whatever kind** of music we are writing, we must move forward with the most essential factor (usually melody) for at least an adequate distance before turning back to consider the rest. (BNC: GVJ) (19) He won't have a go if you have a bad game, but he expects everyone to give their all. If he feels someone is not applying themselves 100 per cent, he won't spare reputations or ego --; **whoever** you are. (BNC: HTY)
place	(20) "**Wherever** Steve went Sarah would find him. It was embarrassing" (BNC: CH1)
time	(21) Under this head belongs every form of words by which, in speaking of a proposed measure of relief, an intimation is given that the time at which the proposal is made, **whenever** it may be, is too early for the purpose. (BNC: EEC)
manner	(22) It is excellent **however** you use it, but rather than fiddling with small mince pies, my great aunts from Norfolk Island made huge double-crusted pies, the pineapple layered between home-made mince meat, rich with rum and spices. (BNC: AHK)
reason	*whyever, *however come
purpose	*whyever, *whatever for
amount, number	(23) **However many** women he took to himself, they were not Beth. (BNC: FPK)
extent	(24) That means shifting the patients around **however** sick they are. (BNC: CH1) (25) I was so downhearted and at such a low ebb because he made it painfully clear I didn't figure in that great club's future, **however much** I loved the place. (BNC: CH3)

All categories are represented, except ‘reason’ and ‘purpose’. As pointed out in the literature, there is no instance of *whyever* in a UCC[3], nor of *whatever for* or *however come* in the intended meanings.

On the Web, there is no first-hand example of *whyever*. However, there is a Wiktionary article on *whyever* which lists three occurrences from different written sources:

(26) Whatever we do, and **whyever** we do it, does not every motive originate in self, and does not every act proceed out of the individual's instinct for self-fulfilment ? (Wilson Follett 1918. *The Modern Novel: A Study of the Purpose and the Meaning of Fiction*, p. 79)

(27) **Whyever** they began, there was no perceptible wolf at their door. (Steven Polgar 1975**.** *Population, Ecology, and Social Evolution*, p. 74)

(28) “And **whyever** they were doing it, they were the ones responsible for what happened to her and all of the rest of my friends in the first place.” (David Weber, Linda Evans 2006**.** *Hell's Gate*)

Despite these three examples, it is fair to say that the views expressed in the literature are confirmed by the observed facts: *whyever* is practically banned from UCCs. This is not only the case in English, but can be observed in French and Spanish as well. In German and Dutch, causal items are occasionally found in UCCs, but their use is marginal in comparison with other items (cf. Defrancq and Leuschner in preparation). In Defrancq and Leuschner (in preparation), it is also suggested that the incompatibility between causal items and UCCs is due to the scalar component of the concessive semantics.

This is not to say that the concepts of cause and purpose cannot be expressed in the context of a UCC. The BNC provides quite a number of examples where they are, but in all these cases a periphrastic form is used involving nouns such as *reason*, *cause* and *purpose*, as in the following examples:

(29) In taking this position, the bishops were also following the lead of Pope John Paul II who, on his visit to Ireland in 1979 had argued: “Divorce, for **whatever reason** it is introduced, inevitably becomes easier and easier to obtain and it gradually comes to be accepted as a normal part of life” (1979). (BNC: A07)

(30) But **whatever purpose** your music has been commissioned for, it is typical that the company who

commissions the piece obtains the copyright to it as well. (BNC: C9J)

Speakers can also resort to expressions of irrelevance, which are natural producers of new concessive clauses (cf. Thompson and Longacre 1985, Leuschner 2006). In English, *matter* can be used with a negation to form a clause very similar to a UCC:

(31) I was a gambler on a winning streak: it didn't **matter what** number I placed my bet on, it always came up a winner. (BNC: ASV)

In this case, *why* seems to be allowed, as the following example shows:

(32) Article 2 (h) of the Vienna Convention on the Law of Treaties provides that “A third State means a State not a party to a treaty.” It does not **matter why** a State has failed to become a party to a treaty, or whether it is eligible to become a party and intends at some time to do so. (BNC: EF3)

Admittedly, the punctuation of both examples is different. In (31), the use of a comma between the clause with the expression of irrelevance and the clause with which it is associated suggests that they both belong to the same utterance and that their relationship can be seen as one in which the *matter*-clause is a dependent clause, as UCCs normally are. In (32), on the other hand, the punctuation clearly marks the *matter*-clause as an independent clause. The difference could be significant: if causal items (such as *why*) are really incompatible with the context of a UCC, as suggested by cross-linguistic evidence, then it should not come as a surprise that causal items are not admitted in grammaticalized instances of *matter*-clauses (with soft punctuation), as these resemble UCCs most.

4.2. French

In French, the UCCs described in the literature take the form of a *wh*-item followed by the complementizer *que* and a verb in the subjunctive, as in example (7), repeated here as (33) for easy reference:

(33) "Il a été reconnu coupable. Cela prouve que le système fonctionne, **qui que** vous soyez", a commenté pour Reuters Pat McQuaid, président de l'Union cycliste internationale. (Le Monde, 21 septembre 2007, p. 19)

Their use is much more restricted in French than in English. Various sources point out that UCCs tend to appear in fixed or semi-fixed expressions, such as *quoi qu'il en soit* 'anyway', *quoi qu'il dise* 'whatever he says', etc. (cf. Morel 1996). Their frequency in French is also much lower than in English: French UCCs introduced by *qui que*, for instance, have a frequency of 0.1 per million words, whereas their English *whoever*-counterparts occur 2.9 times per million words. Both figures are based on similar genres: the *Le Monde* newspaper 2005-2007 and the 'Newspaper' section of the BNC. The only really productive UCC seems to be the one with *quel que*, partly because it has to compensate for other *wh*-items that cannot be used in UCCs.

According to various sources, the paradigm of *wh*-items that can be used in UCCs is indeed rather limited (Hadermann 1993, Morel 1996, Grevisse and Goosse 2008). Only *qui* 'who', *quoi* 'what', *où* 'where' and *quel* 'which' are reported to lend themselves to such a use. In the latter case, the conjunction *que* is used twice: once attached to *quel* and once after the noun which is determined by *quelque*, the correct form thus being: "*quelque* N *que*". However, in most cases this complex form is avoided and replaced by a cleft form based on the ordinary form of the *wh*-item, followed by "*que ce soit* N *que*" or "*que ce soit* N *qui*". In other words, *quelque groupe qu'il rejoigne* as an equivalent of *whatever group he joins* is much less frequent than *quel que soit le groupe qu'il rejoint*.

Grevisse and Goosse (2008) quote various examples of *comment que* 'however', but acknowledge that these are rare. They also report that *comme que* 'however' can be found in Swiss French. Morel (1996) contends that even though some grammars quote examples of *quand* 'when', *comment* 'how' or *combien* 'how much / many' used in UCCs, these are either "archaic or very colloquial" (Morel 1996: 127). Benzitoun (2006) quotes some Web examples of *quand que ce soit*, but used as a free choice indefinite. Finally, Hadermann (1993) suggests that *pourquoi* is also one of the items that cannot be used in UCCs.

These claims are proven true by corpus research. Distributed over the previously identified semantic categories, the attested forms are shown in Table 17-3:

Table 17-3. *Wh*-items used in UCCs in French

Categories	Examples
human	(34) La nazification de l'ennemi, **qui que** soit cet ennemi, semble avoir caractérisé, à très peu d'exceptions près, les modalités du discours des élités d'Israel. (Le Monde, 1 février 2005) [The discourse modalities of Israel's elites seem to have been characterized, with few exceptions, by nazification of the enemy, whoever that enemy is.]
non-human	(35) Ils ont découvert, avec l'hitlérisme, que, **quoi qu**'ils disent, fassent ou rêvent, ils étaient rivés à leur judéité. (Le Monde, 11 novembre 2007, p. 14) [They found out under Hitler's regime that whatever they said, did or dreamt, they were riveted to their jewness.]
predicate	(36) Allant plus loin encore, il se disait soulagé de savoir que, **quoi qu**'il arrive, cet enfant avait un avenir assuré. (Le Monde, 29 mars 2007, p. 33) [Going even further, he said he was relieved to see that, whatever happened, this child's future was safe.]
identification	(37) On peut dire que partout en Afrique noire, à la différence de l'Algérie [sic], les Européens ont "pris le virage" et ont admis - **quelque** préjugé **que** certains puissent conserver au fond d'eux-mêmes - la collaboration avec les Noirs, voire une éventuelle subordination à ceux-ci. (Le Monde, 11 mai 2007, p. 32) [It is not false to say that everywhere in Subsaharian Africa, except in Algeria, Europeans have "made the twist" and have accepted cooperation with black people or even to be at their orders, whatever prejudice some may still have deep down.] (38) L'étude note que "**quel que** soit le groupe", les résultats des élèves aux tests sont moins bons dès lors qu'ils doivent "mettre en jeu des repères temporels et spatiaux". (Le Monde, 29 décembre 2007, p. 10) [The study points out that "whatever group is concerned", the test results obtained by pupils worsen when they have to use reference points in time and space.]
nature	(39) "Il a été reconnu coupable. Cela prouve que le système fonctionne, **qui que** vous soyez", a commenté pour Reuters Pat McQuaid, président de l'Union cycliste internationale. (Le Monde, 21 septembre 2007, p. 19) ["He's been found guilty. That proves that the system works, whoever you are", Pat McQuaid, president of the Union cycliste internationale said to Reuters.]

place	(40) Hasard ou maladresse - c'est une question devenue rituelle avec lui -, il commençait souvent l'entretien en néerlandais, d'**où que** vint son intervieweur. (Le Monde, 26 septembre 2007, p. 18) [Coincidence or clumsiness – an almost ritual question in his case – he often used to begin the interview in Dutch, wherever the interviewer came from.]
time	*quand que
manner	*comment que
reason	*pourquoi que, *comment que se fasse-t-il
purpose	*pourquoi que, *pour quoi que
amount, number	*combien que
extent	*combien que

It should be noted that the 'extent' meaning can be expressed by means of *quelque ... que*, which derives from the 'identification' item. Examples are rare and usually quoted from other, older sources, as in the following case:

(41) Ce que signifiait peut-être à sa manière La Rochefoucauld, qui affirmait que "**quelque** rare **que** soit le véritable amour, il l'est encore moins que la véritable amitié". (*Le Monde*, 22 juin 2005)
[That is perhaps what La Rochefoucauld meant to say in his own personal way when he stated that however rare real love is, it is less so than real friendship.]

Other adverbs such as *aussi* and *si* and even the preposition *pour* can be used instead of *quelque* with the same concessive meaning.

Examples of some of the missing categories can be found on the Web. Occurrences of *comment que* usually are either uses found in old documents made available through the Web or modern uses made possible by analogy, as in (42):

(42) Pour ceux qui n'arrivent pas encore à se dire que voler, **quoi** ou **comment que ce soit**, est "mal", passez du temps sur qqchose que vous mettez en vente (www.cuk.ch/articles/2675)
[For those who can still not understand that stealing, whatever it is or however it is done, is "bad", spend some time putting something on sale]

Quand que is surprisingly frequent on the Web. In many cases it is used as a free choice indefinite, but there are some genuine UCCs as well, even in contexts without analogy:

(43) Au cours des quelques derniers mois, il n'avait rien tenté; mais elle était morte de peur qu'il puisse, et il le ferait probablement, retourner à ses vieilles habitudes une fois qu'il pouvait être sûr d'être tranquille. Severus avait purement et simplement refusé de venir chez les Rosiers, mais il lui offrait l'hospitalité-ou plutôt un asile-**quand que** ce soit qu'elle en ait besoin. (http://www.fanfiction.net/s/1227698/16/LOracleDeLaSybilleLivre2_Le_Cr_ne_et_les_Serpents)
[During the last couple of months, he hadn't made any attempt; but she was terrified that he would fall back into old habits, and he probably would, once he was sure to be left alone. Severus had simply refused to go at the Rosiers' place, but he offered her hospitality – or was it an asylum – whenever it was she needed one.]

Finally, there are some examples of *combien que*, but they all come from old documents made available through the Web. No examples of other disallowed *wh*-items were found.

French speakers have two alternatives for the disallowed combinations. Morel (1996) reports that French possesses periphrastic forms, such as *à quelque moment que*, *de quelque façon que*, and *pour quelque cause que*. However, these have not been found in the corpus. What the corpus did provide was a number of periphrastic forms with the item *quel* in a clefted structure:

(44) Selon ce dispositif, **quel que** soit le **moment** où l'assuré sortira de son contrat, les frais ne devront pas être supérieurs à 5 % du montant qu'il percevra. (Le Monde, 4 octobre 2005)
[According to this provision, whenever the insurance taker comes to be released from his or her contract, the costs should not exceed 5% of the sum he or she will receive.]

(45) «**Quelle que** soit la **façon** dont on la présente, la pause décidée par M. Barroso est rassurante, car elle montre que la libéralisation à tout crin n'est plus possible », dit un responsable bruxellois. (Le Monde 3 février 2005)

[However it is presented, the break decided on by M. Barroso is reassuring, because it shows that ruthless liberalization is no longer possible, a Bruxelles based official says.]

(46) **Quelles que** soient les **raisons**, ces mesures donnent une idée de l'absurdité qui régit la vie économique [...] (Le Monde, 9 octobre 2007, p. 2)
[Whatever may be the reasons behind them, these measures give an idea of the absurdity of the economic system.]

(47) Kokopelli milite pour la création d'un fichier de variétés que chacun pourrait enrichir et utiliser à sa guise, **quel que** soit son **objectif**. (Le Monde, 3 janvier 2007, p. 7)
[Kokopelli promotes the creation of a file of species everyone can use and contribute to, whatever their objectives.]

(48) En France par exemple, la différence de prix entre le générique et le médicament premier est de 40 % **quel que** soit le **nombre** de produits en compétition. (Le Monde, 25 octobre 2007, p. 18)
[In France, for instance, the price difference between generic brand medication and name brand medication amounts to 40%, however many products compete.]

On the other hand, French also has expressions of irrelevance that can be used as concessive-like clauses. In previous work (Defrancq 2005), I have pointed out that *peu importe* can introduce clauses with a concessive meaning and the punctuation of a dependent clause,[4] as in the following example:

(49) C'est la seule famille que je connaissais. **Peu importe ce qu**'elle pouvait me faire, c'est la seule famille que je connaissais", répète-t-elle. (Le Monde, 22 octobre 2006, p. 4)
[It's the only family I knew. No matter what they did to me, it's the only family I knew.]

Some of the *wh*-items that are disallowed in UCCs can be used in combination with an expression of irrelevance. However, the punctuation is always that of an independent clause, suggesting a less grammaticalized kind of relationship:

(50) En revanche, il faut bloquer sur le «oui» tous ceux qui sont contre la Turquie en Europe. **Peu importe comment** l'on obtient ces «oui». (Le Monde, 11 février 2005)
[On the other hand, we have to make sure that those who are opposed to Turkey being part of the European Union stick with the "yes". No matter how we persuade them to vote "yes".]

The concessive relationship between the two clauses may not be directly observable, but it becomes clear when the preceding context is taken into account: just before this utterance, the author of the example, Nicolas Sarkozy, is reported to have been criticized for his half-hearted campaign in favour of the "yes" in the French referendum on the European Constitutional Treaty. Obviously, a half-hearted campaign makes it likely that voters abandon the yes-camp, and that Sarkozy is held responsible for it. Therefore, *peu importe comment* has to be interpreted as referring to the end point of a scale, i.e. the campaign most likely to lose the yes-camp voters; Sarkozy's campaign in other words. To this Sarkozy objects that voters will stay in the yes-camp if they are sure that this does not imply that Turkey becomes a member of the European Union.

(51) Nicolas Sarkozy est la démonstration vivante que la notion de congé est essentiellement psychologique. Etre en vacances ne suppose pas de vivre vraiment des vacances, mais simplement de partir et de revenir. Peu importe où, **peu importe combien** de temps... (Le Monde, 15 mai 2007, p. 2)
[Nicolas Sarkozy is living proof of the fact that holidays are essentially a psychological concept. To be on holiday does not imply that one really experiences a holiday, but merely that one leaves and comes back. No matter where, no matter for how long...]

There are no examples of *quand* or *pourquoi* in the range of newspapers that constitute the corpus I used. Extending the search to all *Le Monde* newspapers from 2001 to 2007 provided one example of *peu importe quand*:

(52) **Peu importe quand** la cassette fut enregistrée, et remise au correspondant à Kaboul d'Al-Jazira. Elle était à l'évidence programmée pour être diffusée quelques heures après le début des bombardements, alors que les télévisions du monde passeraient en boucle de la

> neige verte qui ne montrerait rien. (Le Monde, 15 octobre 2001)
> [It does not matter when the tape was recorded and handed over to the Al-Jazeera correspondent in Kaboul. It was obviously programmed to be broadcast few hours after the bombings started at a time when broadcast companies all over the world would show over and over again this green snow that would not tell anything.]

The absence of *peu importe pourquoi* could suggest that this kind of clause is more grammaticalized as a UCC in French than in English,[5] even in cases with strong punctuations. Genuine UCCs indeed resist causal items. If French does not allow causal items to be combined with a particular expression of irrelevance, but English does, it is possible that the French cases are closer to the genuine UCCs than the English ones.

4.3. Summary

Putting the two languages side by side, the following picture emerges: *wh*-items representing the categories ‘human referent’, ‘non-human referent’, ‘predicate’, ‘identification’, ‘nature’ and ‘place’ can be used in UCCs in both languages. Neither English nor French admits *wh*-items representing ‘cause’ or ‘purpose’. As for the other categories, English allows them to be used, but French does not. Both languages offer alternative structures for the disallowed items. Periphrastic forms of UCCs with generic head nouns are available, but in French they only appear in clefted structures. Expressions of irrelevance constitute a less grammaticalized alternative in both languages. Many of the items that are disallowed in UCCs do combine with those expressions, with the notable exception of the French causal item *pourquoi*. Different possibilities are represented schematically in Table 17-4.

It is a basic assumption in contrastive linguistics that contrasts – paradigmatic asymmetries we may call them – of this kind can and will have effects on the process of second language acquisition and translation. Languages with paradigmatic deficits typically suffer from overgeneralization: language learners fill in the gaps by generalizing the rules that apply in the existing cases. Translators, on the other hand, are forced to find viable alternatives for the absent items, respecting at the same time as much as possible the semantic integrity of the source text.

Table 17-4. Concessive structures using *wh*-items in English and French

Categories	EN	FR
human	UCC	UCC
non-human	UCC	UCC
predicate	UCC	UCC
identification	UCC	UCC
nature	UCC	UCC
place	UCC	UCC
time	UCC	?UCC (exc. generic noun) > irrelevance
manner	UCC	*UCC (exc. generic noun) > irrelevance
reason	*UCC (exc. generic noun) > irrelevance	*UCC (exc. generic noun) *irrelevance
purpose	*UCC (exc. generic noun) > irrelevance	*UCC (exc. generic noun) *irrelevance
amount, number	UCC	*UCC (exc. generic noun) > irrelevance
extent	UCC	UCC (other *wh*-)

In section 5, I will examine how translators deal with the differences between English and French with respect to the use of the *wh*-paradigm in UCCs. As the paradigmatic gaps concern French, I will mainly focus on translations from English into French. Obviously, the cases that will be of most interest are those where a contrast exists between the two languages and where, consequently, a problem needs to be solved by the translator. Some solutions to ungrammatical UCCs have been suggested in this section on the basis of monolingual corpora. Parallel data will show whether or not these solutions are taken up by translators.

5. Parallel data

As pointed out before, the parallel data are extracted from the JRC Acquis Communautaire Corpus. The JRC AC is a corpus of administrative and legal texts from the EU institutions, translated into all the official languages of the European Union (minus 1). The text genre is of course very specific, which could raise problems with respect to the representativeness of the results, but the huge amount of texts it contains makes it one of the largest parallel corpora in the world, especially when taking into account the number of languages concerned. In addition, legal and administrative texts are likely to present the kind of complicated reasoning and universal claims that underpin the use of concessive clauses. The quality of the texts is unequal, as quite some original texts in the EU are drafted by non-native speakers of the language involved. Even native British drafters are quite often accused of producing a kind of Eurospeak that is considered awkward in the UK.

5.1. English originals

The part of the corpus that was consulted covers the most recent material. It contains texts from 2000 to 2006 and comprises 26.5 million words. Most of the texts are originally drafted in English and then translated into the 21 other languages. For the purposes of this chapter, every occurrence of a UCC in a text translated into English was excluded, as the main concern of the research is to check how translators deal with items that exist in English, but not in French. An substantial number of duplicates had to be removed as well: legal and administrative texts have a strong propensity for repeating whole stretches of text. Some occurrences, mainly of *whenever* and *wherever*, have been ignored when the concessive meaning component was absent, the meaning of the clause being purely free-choice.

A total of 418 occurrences of UCCs were found. Their distribution over the different cases is shown in Table 17-5 (as in the previous tables, the examples of *whatever* and *whichever* used as determiners of head nouns such as *moment, reason,* etc. are grouped and not assigned to the semantic categories their head noun could belong to).

The best represented category is the ‘identification’ category. Categories corresponding to predicate or argument expressions are rare: there are only 6 examples of *whoever*, 2 of which are in fact introduced by the archaic sounding *whomsoever*:

(53) The AMM, its property and assets, wherever located and by **whomsoever** held, shall enjoy immunity from every form of legal process. (jrc22005A1029_01-en)

Table 17-5. Distribution of *wh*-items over semantic categories in UCCs in JRC-AC English

Semantic category	Item	in UCC
human	whoever	6
non-human	whatever	8
predicate	whatever	5
identification	whichever whatever	11 229
nature	whatever whoever	55 0
place	wherever	44
time	whenever	2
manner	however	19
reason	whyever however come	0 0
purpose	whyever whatever for	0 0
amount, number	however much/many	1
extent	however (much)	38
Total		418

The frequency of *whoever* in the corpus is 0.2 occurrences per million words (6 examples on a corpus of 26.5 million words). This is about fifteen times lower than the frequency of *whoever* in UCCs in the Newspaper section of the BNC, which was 2.9 occurrences per million words (see section 4.1).

Only 8 examples present *whatever* in an argument position of a verb, 7 of which are nearly identical: they illustrate a combination of *whatever* with the verb *call*, as in (54).

(54) [...] "university" means any type of higher education institution, according to national legislation or practice, which offers qualifications or diplomas at that level, **whatever** such establishments may be called in the Member States [...]. (jrc32000D0253-en)

Of the other categories, only 'place', 'manner' and 'extent' are well represented. As expected, no occurrences were found of the items expressing 'reason' and 'purpose', these concepts being expressed by means of a periphrastic form involving *whatever*, as in the following examples:

(55) The activities comprise in particular:
- organising, offering for sale and selling, outright or on commission, single or collective items (transport, board, lodging, excursions, etc.) for a journey or stay, **whatever the reasons** for travelling (Article 2(B)(a)) (jrc32005L0036-en)

(56) **Whatever the purpose** of the measure may be, state aid is determined on the basis of effects and not objectives. (jrc32006D0748-en)

Only very few examples were found of the items expressing 'time' and 'number / amount'. For most of the categories, there is a transfer towards other expressions, in particular expressions involving *whatever*, such as *whatever means* (instead of *however*), *whatever date* (instead of *whenever*), *whatever number / amount* (instead of *however much*):

(57) Each Member State shall take the necessary measures to ensure that the following conduct is punishable:
(a) any fraudulent making or altering of currency, **whatever means** are employed; (jrc32000F0383-en)

(58) The obligation laid down in (a) shall apply to all relevant acts in force at any given moment, **whatever** their **date** of adoption. (jrc22005D0092-en)

(59) **Whatever** their nitrogen **content**, all solutions of UAN are considered to have the same basic physical and chemical characteristics and therefore constitute a single product for the purpose of this investigation. (jrc32000R1995)

5.2. French translations

The French translation data are presented in Table 17-6. As can be seen, the total number of examples in French data is lower than that in the English data because 66 occurrences of *wh*-items in UCCs were translated using a structure without a UCC. The most striking aspect of these data is the fact that, except for the 'identification' category, all the frequencies are considerably lower than the corresponding frequencies in Table 17-5 for English. The frequency of the 'identification' category, on the other hand, is considerably higher in French than in English. This means that in the translation process, a transfer has taken place from the other categories towards ‘identification’, because translators used a periphrastic form instead of a simple *wh*-item. As this does not account for all the discrepancies, a number of occurrences have also been translated using other structures. A closer look on the different translation strategies reveals a number of noticeable facts.

Table 17-6. Distribution of *wh*-items over semantic categories in UCCs in JRC-AC French translations

Semantic category	Item	in UCC
human	qui que	0
non-human	quoi que	0
predicate	quoi que	0
identification	quelque N que quel que soit N qu	9 280
nature	quelque N que qui que quel que soit N qu	0 48 0
place	où que	15
time	quand que	0
manner	comment que	0
reason	pourquoi que comment qu’il se fasse	0 0
purpose	pourquoi que pour quoi que	0 0
amount, number	combien que	0
extent	combien que quelque A que	0 0
Total		352

1. *Wh*-items denoting referential entities are mostly translated by means of a periphrastic form in French: *whoever* is never translated as *qui que*. Alternative translations include: *quelle que soit la partie*, *quel que soit l’auteui,* and *quell que soit le détenteur* (five cases); one occurrence is translated by means of a simple relative clause, abandoning the concessive meaning of the example:

(60a) [...] the Commission shall make arrangements so as to ensure that in the event referred to in paragraph 2 the costs for the following actions are borne in appropriate proportions by the competent authorities of Australia or New Zealand, **whoever** has requested the formulation into vaccines of antigens stored in the Community reserves: [...]. (jrc32004D0288-en)

(60b) [...] la Commission prend des dispositions afin de s'assurer que, dans le cas visé au paragraphe 2, le coût des mesures énumérées ci-après soit supporté selon des proportions appropriées par les autorités australiennes ou néo-zélandaises compétentes **qui** ont demandé la formulation de vaccins à partir d'antigènes stockés dans les réserves communautaires: [...] (jrc32004D0288-fr)

There are no instances of the concessive-like irrelevance expression *peu importe*.

Whatever, in its autonomous referential use, is always translated by means of a periphrastic form:

(61a) [...] “university” means any type of higher education institution, according to national legislation or practice, which offers qualifications or diplomas at that level, **whatever** such establishments may be called in the Member States; [...]. (jrc32000D0253-en)

(61b) [...] “université”: tout type d'établissement d'enseignement supérieur, au sens de la réglementation ou de la pratique nationale, qui confère des titres ou des diplômes de ce niveau, **quelle que** soit son **appellation** dans les États membres ; [...] (jrc32000D0253-fr)

When *whatever* is used in a periphrastic form in combination with a noun (with or without a copular verb), the preferred translation is a periphrastic cleft form in French (229 cases):

(62a) A variety shall be regarded as distinct if, **whatever** the origin, artificial or natural, of the initial variation from which it has resulted, it is clearly distinguishable on one or more important characteristics from any other variety known in the Community. (jrc32002L0053-en)

(62b) Une variété est distincte si, **quelle que soit** l'origine, artificielle ou naturelle, de la variation initiale qui lui a donné naissance, elle se distingue nettement par un ou plusieurs caractères importants de toute autre variété connue dans la Communauté. (jrc32002L0053-fr)

The same is true of *whichever*, which is always translated by means of a periphrastic cleft (11 cases). Five translations of periphrastic *whatever* lack the cleft, as in the following example:

(63a) [...] accidental marine pollution risks include releases of harmful substances into the marine environment, **whatever** their origin, both from ships and from the shoreline or estuaries, including those linked to the presence of dumped materials, such as munitions, but excluding authorised discharges and continuous streams of pollution originating from land-based sources; [...] (jrc32000D2850-en)

(63b) [...] les risques de pollution marine accidentelle incluent les rejets de substances nocives dans l'environnement marin de **quelque origine qu**'ils soient, tant en provenance des navires que du littoral ou des estuaires, y compris ceux liés à la présence de matériaux immergés, comme les munitions, à l'exclusion des déversements autorisés et des flux continus de pollution d'origine tellurique; [...] (jrc32000D2850-fr)

Other translations include the universal quantifier *tout* 'all' (five cases) and the irrelevance adverb *independamment* 'independently / regardless' (four cases), as illustrated by the following examples:

(64a) It shall be possible to leave the wheelhouse safely **whatever** its position. (jrc52006AG0008-en)

(64b) Il doit être possible de quitter sans danger la timonerie dans **toutes** ses positions. (jrc52006AG0008-fr)

(65a) **Whatever** the approach adopted, the study would help in the setting of concrete objectives to be given priority status in the CI and which give the CI itself added value. (jrc52004SA0004-en)

(65b) **Indépendamment** de l'approche suivie, l'analyse servirait à fixer des objectifs concrets considérés comme prioritaires pour l'IC et lui donnant une valeur ajoutée. (jrc52004SA0004-fr)

In these cases the concessive value is lost. Wherever the universal quantifier is used, the translator inferred the universal quantification from the free choice quantification that was intended. Another occurrence was translated by means of the complex preposition *en dépit de* 'in spite of', focusing on the concessive relationship. In four cases, *whatever* remained untranslated. There are no instances of the concessive-like irrelevance expression *peu importe*.

Wherever is translated by *où que* in 15 out of 44 cases only:

> (66a) [...] (a) an accident occurring within its territory involving any of the following installations or in connection with any of the following fields of activity:
> - any nuclear reactor, **wherever** located, [...] (jrc22003A0429_01-en)
>
> (66b) [...] a) d'un accident, survenu sur son territoire ou en dehors de celui-ci, dans les installations ou dans le cadre des activités suivantes:
> - tout réacteur nucléaire, **où qu**'il soit implanté, [...] (jrc22003A0429_01-fr)

Most of the examples are translated by means of a clefted periphrastic form (21 cases):

> (67a) These people should be guaranteed appropriate assistance, **wherever** they go and whatever the form of transport used, so that they can travel with confidence throughout the European Union. (jrc52005DC0046-en)
>
> (67b) Ces personnes devraient avoir la garantie d'une assistance appropriée, **quel que soit l'endroit** où elles se rendent et le mode de transport utilisé, afin de pouvoir voyager en confiance dans toute l'Union européenne. (jrc52005DC0046-fr)

The cleft is not used in two cases, as illustrated in example (68):

> (68a) The AMM's archives and documents, including multimedia support, either in conventional or in digital form, shall be inviolable at any time, **wherever** they may be. (jrc22005A1029_01-en)
>
> (68b) Les archives et les documents, y compris les supports multimédias, qu'ils se présentent sous forme conventionnelle ou numérique, de la MSA sont inviolables à tout moment et en **quelque lieu qu**'ils se trouvent. (jrc22005A1029_01-fr)

Other translations include the universal place quantifier *partout* 'everywhere' (three cases) and the irrelevance adverb *indépendamment* 'independently / regardless' (two cases). One case is translated by means of the free choice indefinite *n'importe où* 'anywhere'. No instances have been found of the concessive-like irrelevance expression *peu importe*.

2. As far as predicates and properties are concerned, there is much more variation among the translation strategies. *Whatever*, in its non-referential autonomous use, only occurs in combination with the verb *happen*. Four out of the five cases of *whatever happens* are translated by means of the fixed expression *en tout état de cause* 'in any case', including the universal quantifier *tout*. In the remaining case, a clefted periphrastic form is used: *quel que soit l'avenir* 'whatever the future'. In its non-referential use in combination with a noun, some of the previously commented translations are used: clefted periphrastic forms (48 cases), the universal quantifier *tout* (two cases) and the irrelevance adverb *indépendamment* (five cases). No examples have been found of the concessive-like irrelevance expression *peu importe*.

Whenever, which does not has a *wh*-equivalent in French, is twice translated by means of a clefted periphrastic form (with the nouns *date* 'date' and *moment* 'moment'), as in the following example:

(69a) However, Article 13 would apply to all batteries that become waste after transposition of the Directive, **whenever** they were placed on the market. (jrc52005AG0030-en)

(69b) Toutefois, l'article 13 s'appliquera à toutes les piles qui deviendront des déchets après la transposition de la directive, **quel que soit le moment** où elles ont été mises sur le marché. (jrc52005AG0030-fr)

However used as a manner item, which does not have a *wh*-equivalent in French, is usually translated with a clefted periphrastic form involving a head noun which is semantically similar to the verb used in combination with *how* (17 cases):

(70a) [...] bonuses to which policy holders are already either collectively or individually entitled, **however** those bonuses are described - vested, declared or allotted [...]. (jrc32002L0083-en)

(70b) [...] des participations aux bénéfices auxquels les assurés ont déjà collectivement ou individuellement droit, **quelle que soit la qualification** de ces participations, acquises, déclarées, ou allouées [...]. (jrc32002L0083-fr)

Examples of head nouns referring to ‘manner’ are found in two cases only: *modalités*, as illustrated in (71), and *titre*:

(71a) Any compensation, **however** it is assigned, must conform with these provisions. (jrc52006AE0734-en)
(71b) Toute compensation, **quelles qu**'en soient les **modalités** d'attribution, doit être conforme à ces dispositions. (jrc52006AE0734-fr)

In two cases the UCC appears non-clefted. In two more cases the English UCC is not translated. No instances have been found of the concessive-like irrelevance expression *peu importe*. In the one case where *however* is used as denoting ‘number / amount’, a clefted periphrastic form is used in French:

(72a) The principle of the common system of VAT entails the application to goods and services of a general tax on consumption exactly proportional to the price of the goods and services, **however many** transactions take place in the production and distribution process before the stage at which the tax is (jrc32006L0112-en)
(72b) Le principe du système commun de TVA est d'appliquer aux biens et aux services un impôt général sur la consommation exactement proportionnel au prix des biens et des services, **quel que soit le nombre** des opérations intervenues dans le processus de production et de distribution antérieur au stade d'imposition (jrc32006L0112-fr)

When *however* denotes ‘extent’, translations vary widely. Expressing a free choice concessive meaning with respect to ‘extent’ appears to be quite a challenge for translators. Not a single instance has been found of the only *wh*-item French admitted in this case, i.e. “*quelque* A *que*”. The clefted periphrastic form, which is most frequently used as an alternative for ungrammatical combinations in French, appears in only six examples out of a total of 38, as illustrated in the following example:

(73a) **However** useful biometrics may be for certain purposes, their widespread use will have a major impact on society, and should be subject to a wide and open debate. (jrc52005XX0723_01-en)
(73b) **Quel que soit** l'intérêt de la biométrie à certains égards, son utilisation généralisée aura un impact majeur sur la société et devrait faire l'objet d'un débat

large et ouvert. (jrc52005XX0723_01-fr)

There are seven cases in which translators used an adverb (*aussi* 'as'*, tout* 'all'*, si,* 'so') or a preposition (*pour* 'for') in combination with the adjective. This is an accurate translation, even though no *wh*-item is involved:

(74a) All accidents to staff members, whether incurred at work or outside the Institute, **however** trifling they may appear at the time, must be reported immediately by the staff member to the Head of Administration and Personnel, together with the names and addresses of any witnesses. (jrc32005Q0912_01-en)

(74b) Tout accident dont pourrait être victime un agent, soit sur le lieu de son travail, soit en dehors, **aussi** bénin **qu**'il puisse paraître sur le moment, doit être signalé dans les plus brefs délais au chef de l'administration et du personnel par l'intéressé, avec les noms et adresses des témoins éventuels. (jrc32005Q0912_01-fr)

In nearly all other cases, the free-choice meaning is abandoned and only the concessive relationship is maintained. Various conjunctions, adverbs and prepositions serve this purpose: *même* 'even', *même si* 'even if', *bien que* / *quoique* 'although', *en dépit de* / *malgré* 'in spite of', *cependant* 'however', and *néanmoins* 'nevertheless'. This seems to indicate that translators are more likely to treat the UCC as a concessive than as a conditional clause. In four of these cases, the free-choice component is replaced by an intensifying item:

(75a) Moreover, as regards agents of OHIM having contracts of less than three years, it would be difficult for them to gain access to an invalidity allowance **however** incapacitated they were, because they could never satisfy the criteri ... (jrcC2006#096#53-en)

(75b) En outre, il serait difficile pour les agents de l'OHMI ayant des contrats de moins de trois ans d'accéder à la pension d'invalidité, car, **même** **<u>très</u>** malades, ils n'atteindraient jamais la limite prévue par l'article 59, paragraphe 4, du statut. (jrcC2006#096#53-fr)

As intensifying items refer to a scale, this clearly supports the idea that the concessive relationship found in UCCs involves a scalar component (see section 2). Finally, one surprising translation was found, i.e. *y compris* 'including' and two cases remain untranslated. No examples have been found of the concessive-like irrelevance expression *peu importe*.

5.3. Summary

The data yield both expected and unexpected results. Obviously, the observed differences described in Table 17-2 show up in the parallel data: *wh*-items that are disallowed or strongly marked do not appear in the French translated data. As expected, translators resort very frequently to periphrastic alternatives. However, they do not resort to using the structure that was shown to be grammaticalizing into a new type of French UCC (*peu importe...*). Periphrastic alternatives turn out to be very frequent, even in cases where there is no need for them. This is possibly a case of normalization (Baker 1996): translators may be aware that in many cases the simple form is not allowed in French and generalize this awareness to cases where a simple form would be admitted.

When translators neither opt for the simple *wh*-item nor for a periphrastic alternative, they usually omit one of the meaning components of the UCC. In cases where the *wh*-item represents a referential category (humans, objects, places), the concessive component tends to be absent, with most of the translators opting for universal quantification. In cases where non-referential categories are concerned (manner, extent), the concessive component is maintained.

6. Conclusion

What this chapter has intended to show is how a contrastive analysis based on monolingual corpora can provide a framework for the analysis of parallel data. The case analyzed here involves universal concessive conditionals. These have been shown to offer interesting contrasts between English and French, as some of the French equivalents of *wh*-items that can be used in English cannot be used in French UCCs. The analysis of monolingual corpora has also uncovered what alternatives speakers of French and English have in cases where specific *wh*-items cannot be used, while the analysis of parallel data has confirmed that French translators indeed use one of the alternatives, but that they seem to avoid the other and prefer to resort to unexpected strategies which result in the loss of specific meaning components.

Notes

[1] It should be noted, however, that French would not even use *combien* in this particular case. For some yet unexplained reason, there is no way in French to ask for the extent of some property. A tentative translation of this example would be: *dans quel état il est?* 'in what state is he?'.

[2] There are other minor differences which have not been listed here, because they concern specific individual uses. English *how* used in *how far*, for instance, has *où* as its French counterpart as part of *jusqu'où* (*how far will he go* 'jusqu'où ira-t-il').

[3] A number of misspelled occurrences of interrogative *why ever* can be found, as shown in example (i). And there is one occurrence of *whyever* used as a free-choice indefinite, as exemplified in (ii).

> (i) "He was such a charming fellow," said Martin. "And I thought some people might like them." "But nobody liked them," said Clelia. "Nobody at all." "**Whyever** did you buy it then?" said Clara. "My mother bought it," said Clelia. (BNC: EFP)
>
> (ii) SAGITTARIUS (Nov 23 --; Dec 21) On the soccer field of life, Sagittarians are (of course) the centaur forwards. Not only do you have to be up there, in the thick of the action, the whole time, but you also specialize in following your balls! This month, though, the sky suggests mental agility is your best asset. Sex: Whoever … Income: Whenever … Expenditure: However … Creativity: Whatever … Travel: Wherever … Work: **Whyever** … Opportunity: Forever … Adventure: Whichever … Success: As ever. (BNC: ECU)

[4] It should be noted that *ce que* is considered an allomorph of *quoi* 'what' in the context of an embedded interrogative.

[5] *Peu importe pourquoi* can be found on the Web, but it is either used as a free choice indefinite or in a series of different *wh*-items, in which case its appearance is probably motivated by analogy (cf. Defrancq 2006).

References

Anscombre, J. C. and Ducrot, O. (1983), *L'argumentation dans la langue*. Bruxelles: Pierre Mardaga.

Baker, M. (1996), "Corpus-based translation studies. The challenges that lie ahead", in H. Somers (ed.) *Terminology, LSP and Translation*, 175-186. Amsterdam: John Benjamins.

Berman, S. (1994), *On the Semantics of WH-clauses*. New York/London: Garland.

Benzitoun, C. (2006), *Description morphosyntaxique du mot* quand *en français contemporain*. PhD thesis, Université de Provence.

Cheng, L. (1997), *On the Typology of Wh-Questions.* New York/London: Garland.

Collins, C. (1991), "Why and how come", in L. Cheng and H. Demirdache (eds.) *More Papers on Wh-Movement*, 31–45. Cambridge: MIT Press.

Declerck, R. (1991), *A Comprehensive Descriptive Grammar of English*. Tokyo: Kaitakusha.

Defrancq, B. (2005), *L'interrogative enchâssée. Structure et interprétation*. Gembloux: Duculot.

—. (2006), "Etudier une évolution linguistique en ligne: *n'importe* et *peu importe*". *Le Français Moderne 74*: 159-182.

Defrancq, B. and Leuschner, T. (in preparation), "Scalar reason. Why ever *whyever* is so rare". ms.

Gawron, J. M. (2001), "Universal concessive conditionals and alternative NPs in English", in C. Condoravdi and G. Renandel de Lavalette (eds.) *Logical Perspectives on Language and Information*, 73-105. Stanford: CSLI.

Giannakidou, A. (2001), "The meaning of free choice". *Linguistics and Philosophy* 24: 659-735.

Grevisse, M. and Goosse, A. (2008), *Le Bon Usage* (14th edition). Bruxelles: De Boeck & Larcier.

Hadermann, P. (1993), *Etude morphosyntaxique du mot* où. Paris / Louvain-la-Neuve: Duculot.

Haspelmath, M. and König, E. (1998), "Concessive conditionals in the languages of Europe", in J. van der Auwera (ed.) *Adverbial Constructions in the Languages of Europe,* 563-640. Berlin: Mouton de Gruyter.

Huddleston, R. and Pullum, G. (2002), *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

König, E. and Traugott, E. (1982), "Divergence and apparent convergence in the development of *yet* and *still*", in *Proceedings of the Eighth Annual Meeting of the Berkeley Linguistics Society*, 170-179. Berkeley: Berkeley Linguistics Society.

Leuschner, T. (2006), *Hypotaxis as Building-Site: The Emergence and Grammaticalization of Concessive Conditionals in English, German and Dutch.* München: Lincom Europa.

Morel, M. A. (1996), *La Concession en Français*. Gap: Ophrys.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985), *A Comprehensive Grammar of the English Language.* London: Longman.

Thompson, S. and Longacre, R. (1985) “Adverbial clauses”, in T. Shopen (ed.) *Language Typology and Syntactic Description: Complex Constructions* (Volume 2), 171-234. Cambridge: Cambridge University Press.

Tsai, W. (1999), “Lexical courtesy revisited: Evidence from Tsou and Seediq *wh*-constructions”. *Journal of East Asian Linguistics* 8: 39-73.

Vlachou, E. (2004), “Greek bare *wh*-items: Evidence from *otan* and *opote*”, in *Studies in Greek Linguistics 24* (*Proceedings of the 24th Annual Meeting of the Department of Linguistics*, 728-738. University of Thessaloniki.

CHAPTER EIGHTEEN

CONTRASTIVE CONNECTORS IN ENGLISH AND CHINESE: A CASE STUDY OF *HOWEVER* IN TWO TRANSLATION CORPORA

JIANXIN WANG

1. Theoretical framework and research objective

There are two broad traditions in language study in the English speaking countries. The first is the American structuralist tradition, represented by Chomsky's generative grammar, which regards language as in-born, biologically pre-formed and consisting of abstract rules. It uses introspection as the main data source. The second is the British tradition, represented by Firth, Halliday and Sinclair, which regards language as a social process. It has authentic text in context as the target and source of language study. This tradition has been carried forward since the 1970s, when computer corpora and software tools became increasingly available and powerful, especially by Sinclair (2004) who regards corpus linguistics and discourse analysis as the twin pillars of language study, which nicely sums up one balanced and fruitful way for language research.

This case study follows the British tradition and adopts usage-based models which maintain, among others, the following important assumptions. First, "language use shapes grammar" by "frequent repetition" (Bybee 2007: 269); second, usage events in context "are crucial" to the forming of linguistic system and its operation; third, "linguistic representations" are neural connections of "recurrent patterns of mental activation" which are dynamic and emergent, not "fixed entities"; fourth, frequency is "of fundamental importance" and "indispensible" in forming such representations; and finally, text corpora contain usage "sampling" and are "ideal" for such study, if "used sensibly" with their limitations borne in mind (Kemmer and Barlow 2000: vii-xxviii).

The purpose of this parallel-corpus-based translation comparison is to examine the usage of contrastive connectors in Chinese-to-English (C/E) and English-to-Chinese (E/C) translations, as exemplified by the concessive connector *however* in the Hóng Lóu Mèng C/E corpus (HLM) and the Babel E/C corpus. Specifically I will compare *however* (and related contrastive connectors) and its Chinese counterparts (overt or implied) in these two corpora regarding the following three aspects (and explore the possible reasons):

- Coverage: what contrastive Chinese connectors are translated into *however* and related English connectors and vice versa
- Feature: what characteristics the contrastive connectors in these two languages have
- Position: where in the sentence these connectors are used

2. *However* in the C/E HLM Corpus

2.1. Sample description

I retrieved all the aligned samples containing *however* and its Chinese counterparts from the free online HLM Parallel Corpus (http://score.crpp.nie.edu.sg/hlm/index.htm), which consists of 红楼梦 (Hóng Lóu Mèng) in Chinese and its two complete English translation versions. HLM is a classic Chinese novel by Cao Xueqin, which is widely acknowledged as the "zenith of Chinese classical fiction"; it is written in Vernacular Chinese and Beijing Mandarin dialect which later became the standard written and spoken Chinese (http://en.wikipedia.org/wiki/Dream_of_the_Red_Chamber). One of the two English translations of this classic work is by David Hawkes and John Minford, entitled *The Story of the Stone;* the other translated version is by Gladys Yang and Yang Hsien-yi, entitled *A Dream of Red Mansions.* Both versions are used for the comparison of *however* and its Chinese counterparts, as the translators of both versions are highly professional in both languages. (The Hawkes translation seems more accurate and literal, while the Yangs translation more explanatory.)

Twenty complete pairs of aligned text segments containing *however* were retrieved from the HLM Parallel Corpus, after omitting 9 repetitions or wrong matches and two incomplete matches, as only part of the HLM Parallel Corpus was accessible on the Internet. Each pair consists of one Chinese original and two versions of English translation.

2.2. Data analysis and discussion

The results of data analysis are summarized and presented and discussed one by one as follows. The Chinese words of first appearance are accompanied with Romanized Pinyin glosses to facilitate reading. The contrastive and concessive connectors in these 20 alignments are listed in Table 18-1, showing the different types of connectors found. This table shows that in the 20 Chinese originals and 40 English translations from HLM, most of the Chinese contrastive connectors, implied or overt, are translated into *however* (22 cases), some into *but* (7 cases), some into other connectors (5 cases), and some into 0 (6 cases), i.e., expressed by other means. Further observation reveals that 18 cases of Chinese originals are translated into *however* etc. to signal concession between sentences (summarized in rows 1-18 of Table 18-1), and two cases to signal concession within a sentence, which will be discussed separately.

One remarkable feature in these 18 cases, as the first 18 rows of Table 18-1 indicate, is that the majority of the original Chinese counterparts of *however* (or other contrastive connectors such as *but*) are zero: there are simply no contrastive or concessive Chinese connectors in the corresponding Chinese text at all. Instead, in most of the cases, such Chinese connectors are implied. Out of the 18 cases there are 14 such zero or implied Chinese connectors, accounting for 78% of the total instances. These zero Chinese counterparts for the English *however* (and other contrastive connectors) can be divided into two types. One type, as shown in the first 7 rows of the table, consists of 7 cases with an initial implied 卻 (*què* 'but, however'), and another type, as shown in rows 8-14, consists of 7 cases with an initial implied 不过 (*bùguò* 'nevertheless') or 可是 (*kěshì* 'but, however') or 可 (*kě* 'but'), all of which are contrastive or concessive Chinese connectors. The ratio between these two types is half and half. So in most cases the implied Chinese connectors for *however* are 卻 (què), 不过 (bùguò), or 可是 (kěshì).

Interestingly, in the first group of 7 zero Chinese connectors (see the first 7 rows of Table 18-1) with an implied initial 卻 (què), 6 of them are immediately followed by a Chinese word or phrase, made up of one to two characters, which indicate some sense of contrast or concession; the seventh is followed by a causal linker. These words seem to have made it unnecessary to use any overt contrastive or concessive Chinese connectors before them.

Table 18-1. Chinese contrastive connectors and English translations in 20 aligned samples from HLM

Row	Sentence ID	HLM in Chinese	Position in sentence	Hawkes translation	Position in sentence	Yangs translation	Position in sentence
1	5	0 (卻 què)誰知 shuízhī	Sentence initial (SI)	however	SI	0	
2	14	0 (卻)誰知	SI	however	After initial subject	but	SI
3	12	0 (卻)只管 zhǐguǎn	After initial subject	however	After initial subject	0	
4	24	0 (卻) 只	SI	however	After initial adverbial	however	After initial adverbial
5	7	0 (卻)只可惜 zhǐkěxī	SI	however	After main clause	although	SI
6	10	0 (卻)無奈 wúnài	SI	but	SI	however	After main clause
7	30	0 (卻)因 yīn	Clause initial	and, as	SI	however	After initial adv. Clause
8	27	0 (不过, bùguò 可是 kěshì)	SI	however	SI	however	SI
9	11	0(不过可是)	SI	but	SI	however	After main clause
10	9	0 (可, 不过)	SI	but	SI	however	SI
11	13	0 (可, 不过)	SI	however	SI	but	SI
12	17	0 (不过)	SI	however	SI	0	
13	19	0(可, 不过)	SI	but	SI	however	After initial adv. clause

14	18	0 (可)	SI	however	SI	but	SI
15	8	到底 dàodǐ	After initial adverbial	however	After initial adverbial	0	
16	1	卻	Clause initial	as a matter of fact	After subject+be	however	After subject+be.
17	4	卻	Clause initial	of course	SI	However	SI
18	6	雖 suī…卻	Clause initial	though… nevertheless	Clause initial	however	After subject
19	16	任憑是什麼好的 rènpíng shì shénme hǎo de	Clause initial	however good they are	Clause initial	0	
20	31	人來客往 rén lái kè wǎng	2nd clause initial	0		however many guests	Clause initial

Here is an example:

(1a) 誰知自娶了他令夫人之後，倒上下無一人不稱頌他夫人的，璉爺倒退了一射之地。(Cao: 002)

(1b) However, ever since he married this young lady I mentioned, everyone high and low has joined in praising her, and he has been put into the shade rather. (Hawkes: 002)

(1c) Since his marriage he's been thrown into the shade by his wife, who is praised by everybody high and low. (Yangs: 002)

In the Chinese part of this aligned pair, the initial 誰知 (shuízhī) means *who knows* or *who expects that*, which implies unexpectedness. It renders 卻 (*què* 'but, however') before it unnecessary. If the contrastive 卻 (què) is used before 誰知 (shuízhī), they together still mean the same thing with slightly increased emphasis on unexpectedness, and the Chinese sentence is still correct. Without it, the sentence is natural. Therefore this implied Chinese connector 卻 (què) before 誰知 (shuízhī) is optional. It is also optional in the other 6 cases before 只管 (zhǐguǎn), which implies continuation against expectation, 只 (zhǐ), which means 'only, simply', 無奈 (wúnài), which implies helplessness and concession, and 因 (yīn), which means 'because of'. Most of these Chinese words that immediately follow the optional 卻 (què) indicate denial of expectation and helpless concession, which must be used and cannot be replaced by 卻 (què). When these words are translated into English, in most cases their implied concession, denial of expectation or helplessness is conveyed by the overt concessive English connector *however* (and in some cases by *but*).

In the second group of 7 zero Chinese connectors (rows 8-14 of Table 18-1) with an implied concessive Chinese connector 不过 (bùguò), or 可是 (kěshì) or 可 (kě), there are no other overt Chinese words in the text to show contrast or concession. The contrast or concession demonstrated by *however* or *but* in the English translation seems to be based on the overall meaning and logical relation of the Chinese text. Here is an example:

(2a) 今兒你既老遠的來了，又是頭一次見我張口，怎好叫你空回呢。(Cao: 006)

(2b) However, since you have come such a long way, and since this is the first time you have ever said a word about needing help, we obviously can’t let you go back empty-handed. (Hawkes: 006)

(2c) But since you’ve come so far today and this is the first time you’ve asked me for help, I can’t send you away empty-handed. (Yangs: 006)

In the Chinese part, there is no overt evidence to indicate contrast or concession. Instead, there is a clear causal relation between the first two clauses and the third main clause, expressed by 既 (*jì* ‘because, since’) and correlated by 又 (*yòu* ‘and (because, since)’), and the effect or result is expressed by 怎好…呢 (*zěn hǎo…ne* ‘how can I…’). This causal relation is clearly shown in both translations: in Hawkes’ by *since…and since…*, and in Yangs’ by *since…and…*. In addition to this overt causal relation, the implied mental contrast of the speaker between dismissing the guest empty-handed or with some reward, which is implied in the Chinese text, is explicated by *however* and *but* in the two English translations.

Example (3) in this group also contains an implied concessive relation in an overt causal relation. This implied concession is expressed explicitly in both English translations: in Hawkes’ by *however* and in Yangs’ by *but*. The causal relation in Chinese is expressed by 因 (*yīn* ‘because’) and correlated by 又 (*yòu* ‘and (because)’), and the effect or result is expressed by 敢不 (*gǎnbù* ‘how dare I not to…’). This causal relation is reflected in Hawkes’ translation in a cause-effect relation linked by *so*, and in Yangs’ translation by a subordinate clause followed by a main clause of result. In both examples (2 and 3) the causal relation is clearly marked in English and Chinese, while the concessive relation is implied in Chinese but clearly marked in English.

(3a) 昨因馮大爺示知，大人家第謙恭下士，又承呼喚，敢不奉命。(Cao: 010)

(3b) However, Mr Feng was telling me yesterday of the courteous and considerate patronage of scholars which is traditional in your family, so when I received your summons I felt unable to refuse. (Hawkes: 010)

(3c) But when I heard yesterday from Mr. Feng that Your Lordship's family is considerate to ordinary scholars and had condescended to send for me, how could I disobey your orders? (Yangs: 010)

Rows 15-18 of Table 18-1 contains 4 cases out of the 18 (22%) where, except in one translation, both the Chinese original and the English translations have overt concessive connectors. In Chinese the connectors are 卻 (què, 2 cases), 雖…卻 (suī… què, 1 case), and 到底 (dàodǐ, 1 case). In English they are *however* (4), *as a matter of fact* (1), *of course* (1), and *though… nevertheless* (1). The exception is Yangs' translation of 到底 (dàodǐ), in which no overt English concessive connector is used. This indicates that in the sample only a minority of Chinese contrastive relations are clearly expressed by Chinese connectors.

Rows 19-20 of Table 18-1 contain two cases where one Chinese original has an overt unconditional concessive expression (任憑是什麼好的 (*rènpíng shì shénme hǎo de* 'no matter how (good)'), and the other implied one in 人來客往 (*rén lái kè wǎng* 'no matter how (many guests come and go)'). Hawkes used *however* to translate the overt Chinese expression and a conditional clause to translate the implied one; Yangs used the superlative degree to translate the overt unconditional concession in Chinese, and *however* to translate the implied one. This shows that the unconditional concessive expression *no matter how* in Chinese can be overt or implied, while its English translation can use *however* as an equivalence, or can resort to other means to do this. Here is an example.

(4a) 衣裳任憑是什麼好的，可又值什麼！ (Cao: 010)
(4b) Never mind about the clothes, for goodness' sake, however good they are! (Hawkes: 010)
(4c) This will never do. The finest clothes are nothing compared with her health. She can wear new ones every day if it comes to that. (Yangs: 010)

In this aligned pair, Hawkes used *however good* but Yangs used the superlative degree (*the finest clothes*) to translate the Chinese original 任憑是什麼好的 (*rènpíng shì shénme hǎo de*). Both translators convey the original meaning accurately. In fact, as an unconditional concessive marker, *however* still signals concession in these two cases, although it also functions locally as an intensifier modifying its ensuing adjective. This concession occurs within the sentence, between the *however* subordinate clause and the main clause. Therefore these two cases actually demonstrate concessive relations between clauses, while the previous cases demonstrate concessive relations at discourse level between sentences.

2.3. Overt vs. implied contrastive connectors

To further compare the 20 Chinese originals with their English translations, the contrastive connectors in the Chinese originals are regrouped into two types: implied vs. overt. The 40 translations in the two versions (Hawkes and Yangs**)** are first divided to four types, based on frequency: *however* vs. *but* vs. 0 vs. other (which includes other connectors), and then reduced to two categories: overt vs. zero. The result is shown in Table 18-2.

Table 18-2. Connector types of the Chinese original and English translations in the HLM sample

Function	No.	Chinese original	English translation (Hawkes + Yangs)	No. of such translations
Sentence contrastive connector	14	Implied	*however*	16
			but	7
			0	3
			although	1
			and, as	1
	4	Overt	*however*	4
			as a matter of fact	1
			of course	1
			though…neverthe-less	1
			0	1
Clause contrastive connector	1	Overt	*however*	1
			0	1
	1	Implied	0	1
			however	1
Total	**20**			**40**
Sentence connector: Clause connector	**18:2** **=90%: 10%**	**Overt: Implied =5: 15 =25%: 75%**	***However* + *but* + *other: 0*** **Overt: Implied**	**22+ 7+ 6: 5 =34: 6** **=85%:15%**

As demonstrated in Table 18-2, 75% of the concessive relations in the original Chinese text segments are expressed by implied contrastive Chinese connectors while only 25% by overt ones. In the English translations, by contrast, 85% of these relations are expressed by overt contrastive or concessive English connectors while only 15% by zero English connectors, i.e., by other means such as using conditional relations. This is indeed a striking difference. It has been observed and suggested (cf. Cao 1994, Lin and Li 2004, Pan 2004, Wang and Zheng 2004) that Chinese is an implicit language, whereas English is an explicit language. In terms of expressing contrastive or concessive relations, this observation is certainly true. The English language tends to use clear and overt connectors such as *however* and *but* to indicate these relations, whereas in Chinese such connectors are often implied. Another finding, as witnessed in this sample, is that most of these contrastive connectors in both languages are used at discourse level. The ratio of these contrastive connectors used between sentences and within sentences is 90% vs. 10%.

2.4. Positional distribution of C/E contrastive connectors

The positional distribution of the Chinese and English contrastive connectors in this sample is summarized in Table 18-3. As indicated in the table, most of the Chinese contrastive connectors, implied or overt, occur in sentence or clause initial position, which total 85%. The positions of connectors in the English translations are more varied, where the initial position totals 52.5%, and the second initial 32.5%. A further analysis reveals that this high percentage of second initial position is mainly caused by the concessive connector *however*, whose occurrences in the sentence initial position take up 45.5% while those in the second sentence initial position account for 54.5%. This high percentage of initial Chinese contrastive connectors seems to be related to the phrase-centred characteristic of the Chinese language, particularly in spoken Chinese, where phrases combine freely to form sentences, leaving the subject implied or unexpressed. The high percentage of second initial position of *however* is partly caused by the characteristic of the English syntax, where the subject is normally required in the sentence and it often occurs at the sentence beginning, partly by the flexible positions *however* can have in the sentence, and especially by the double functions it serves in second initial position: emphasizing the immediate preceding part and signalling a contrast or concession (Altenberg 2006). Here is an example.

(5a) 誰知狗兒利名心甚重，聽如此一說，心下便有些活動起來。(Cao: 006)

(5b) Gou-er's cupidity, however, had been aroused by the words of his mother-in-law, and his reaction to them was less discouraging than his wife's. (Hawkes: 006)

In the Chinese part, 誰知 (shuízhī) implies unexpectedness and occurs in sentence initial position, which can be optionally preceded by the contrastive connector 卻 (què) or 可 (kě) but either is implicit. They all indicate what ensues is against expectation. Gou-er's strong greed is emphasized which leads to his unexpected change of mind. In Hawkes' English translation, *however* occurs after the initial Gou-er's cupidity. This second initial position both highlights cupidity by foregrounding it and signals Gou-er's unexpected change of mind.

Table 18-3. Position comparison of contrastive Chinese connectors and English connectors in the HLM sample

Position / Connectors	Sent initial	Clause initial	2nd sent initial	Other	Total	Initial vs. other
Chinese	**12**	**5**	**2**	**1**	**20**	**17: 3 =85%:15%**
English	**18**	**3**	**13**	**6**	**40**	**21:19 =52.5%:47.5%**
however	**8**	**2**	**12**		**22**	**10:12 =45.5%:54.5%**
but	7				7	
Other	3	1	1		5	
zero connector				6	6	

2.5. A summary

To sum up, (1) *however* is most of all a contrastive or concessive connector between sentences. In this sample of 20 aligned text segments and translations from the HLM corpus which contain *however*, 90% of them are used between sentences. Only 10% of them are used within the sentence as an adverbial intensifier before an adjective in the subordinate clause, which concedes to the main clause. (2) The Chinese contrastive connectors are more implicit than explicit. Among the 20 original Chinese contrastive relations, 75% of them involve an implied Chinese contrastive connector, half of which is followed by some other concessive Chinese expressions which render the overt contrastive Chinese connectors unnecessary. Only 25% of them are expressed by overt contrastive or concessive connectors. The English connectors are used more explicitly than implicitly in expressing contrastive relations: the ratio between overt and zero (implied) English connectors being 85% to 15%. (3) The positional distributions of the contrastive connectors in these two languages differ considerably. 85% of the Chinese contrastive connectors occur in the beginning of the sentence or clause, whereas in English only 52.5% do so. The second initial position of *however* is especially common: 54.5% in the sample, due to its double functions in this position: highlighting the initial element of the sentence and indicating a contrast or concession. (4) The differences between the usage of contrastive connectors in Chinese and English are likely to be related to the implicit characteristic of the Chinese language, the frequent omission of subjects in Chinese sentences, the explicit characteristic of the English language and the constraint of its syntactic structure, where the subject is normally required. It seems especially related to the rhetoric structure of the English language where the initial part of the sentence tends to be primed, fore-grounded and thus emphasized, which often makes it necessary to put contrastive connectors such as *however* in the second initial position.

3. *However* in the E/C *Babel* corpus

3.1. Sample description and result

To study the usage and behaviour of *however* and its translations from English into Chinese, I used the *Babel English-Chinese Parallel Corpus* (www.lancs.ac.uk/fass/projects/corpus/babel/babel.htm)*,* which consists of 327 English articles and their translations into Mandarin Chinese, totalling 544,095 words (253,633 English words and 287,462 Chinese tokens), about half of them taken from the *World of English* and half from *Time,* both between the years 2000 and 2001.

A total of 102 aligned sentence pairs of texts are retrieved from the Babel which contain *however* in the English sentences and its translations in the Chinese sentences. One alignment pair is a false match and is excluded from the sample, leaving 101 alignment units. I counted the sample regarding three aspects as follows (and the results are recorded in Table 18-4).

- What Chinese words *however* is translated into
- What frequencies these Chinese translations are

- What positions *however* takes in the English sentences; what positions the Chinese translations take in the Chinese sentences

Table 18-4. *However* and its Chinese translations in the Babel corpus

Eng.	No.	Sent initial	2nd sent ini-tial	Sent final	Chin.	No.	%	Sent initial	2nd sent initial
however	101	39	56	6	16 items	101		95	6
%		38.6 %	55.4 %	5.94 %				94.06 %	5.94 %
					然而 ránér	38	37.6		
					不过 bùguò	26	25.7		
					但是 dànshì	11	10.89		
					但 dàn	7	6.93		
					可是 kěshĭ	4	3.96		
					0	4	3.96		
					而 ér	2	1.98		
					还是 háishì	1	0.99		1
					仍然 réng rán	1	0.99		1
					不管怎样 bùguăn zěnyàng	1	0.99		
					不管怎么说 bùguăn zěnme shuō	1	0.99		
					不管多 bùguăn …duō	1	0.99		1

					(不论)…多么 (bùlùn …duōme	1	0.99		
					则 zè	1	0.99		1
					竟 jìng	1	0.99		
					其实 qíshí	1	0.99		1

[Notes: **Second sentence initial:** after initial subject, initial adverbial, initial subject + verb). **Sentence final** includes main clause final.]

3.2. Data analysis and discussion

As indicated by the middle column in Table 18-4, *however* is translated into 16 different Chinese connectors (including 4 zero or implied connectors). These Chinese connectors are synonyms, expressing contrast or concession. Some of them are very strong, e.g. 但是 (dànshì), 但 (dàn), 不管…多 (bùguăn …duō), (不论)…多么 (bùlùn…duōme); some of them are mild, e.g. 然而 (ránér), 不过 (bùguò); some of them are very weak, e.g. 则 (zè), 其实 (qíshí). This range of Chinese translations reflects that the meaning of *however* is interpreted slightly differently by different translators. Translation seems to be a process of understanding and re-expression. 16 different Chinese connectors are used to translate the same English connector *however,* which shows the wide range of possible Chinese translations of this English contrastive connector.

Among the 101 Chinese translations of *however*, 然而 (ránér) is used most frequently (37.6%), which is followed by 不过 (bùguò, 25.7%), 但是 (dànshì, 10.89%), 但 (dàn, 6.93%), 可是 (kěshĭ, 3.96%), zero Chinese connector (3.96%), and 而 (ér, 1.98%). Each of the rest nine connectors takes up less than 1%. This shows that *however* is widely regarded in the Chinese translation as the equivalent of the concessive Chinese connectors 然而 (ránér) and 不过 (bùguò), which total 63.3%. It is also fairly commonly interpreted as a strong contrastive connector in Chinese translations, as indicated by the two strong contrastive connectors in Chinese, namely 但是 (dànshì) and 但 (dàn), which total 17.82%. The implied Chinese connectors for *however* in the translations are far less common than in Chinese-to-English translations (as evidenced in the HLM corpus): only 4 such cases of translation occur, about 3.96%. This seems to suggest that translators tend to have the contrastive or concessive relations signalled by *however* in English clearly expressed in Chinese translations.

The most frequent position of *however* in the sample is in second sentence initial, i.e. after an initial subject, adverb, adverbial phrase, or subject plus (different forms of) a verb, which totals 55.4%. The second most frequent position of *however* is in the sentence beginning position (38.6%). The final position of *however* totals 5.94%, where *however* occurs at the end of sentence or main clause. In contrast, most Chinese contrastive or concessive connectors occur in the sentence initial position in the sample (94.06%). Only 5.94% occur in the sentence medial position. Here are three examples from the Babel corpus where the Chinese translations of *however* all occur in the sentence beginning position.

(6a) This time, **however**, it will be written by Keiko himself.
(6b) **不过**，这次剧本的编写者不是别人，而是凯科自己。

(7a) One man, **however**, is working overtime to get the American worker more vacation time.
(7b) **不过**，有一个人为他人争取更多休假时间而超时工作。

(8a) It isn’t certain, **however**, that others will follow the trend.
(8b) **不过**，其他航空公司是否加入它们的行列还很难说。

This big difference in the distributional positions of *however* and its Chinese translations reveals their different usage patterns in the two languages, which tend to have a strong mother-tongue influence on the learners in learning the other language. This is clearly evidenced by the misuse of *however* by learners of English in China, who tend to heavily overuse *however* in sentence initial position and underuse it in second sentence initial position, as demonstrated by my survey of the positional distributions of *however* in three one-million-word corpora: the Brown, the LOB, and the Chinese Learner English Corpus (CLEC).

3.3. A summary

As indicated by the above analysis of the 101 aligned E/C translation pairs, (1) the contrastive connector *however* is translated into a range of Chinese connectors, most frequently into 然而 (ránér , 37.6%) and 不过 (bùguò, 25.7%). *However* is interpreted in Chinese translations as being a formal contrastive connector, in most cases the contrast being mild to middle, and in some cases strong. (2) Most cases (96%) of *however* are translated into an overt Chinese contrastive connector. This is in sharp contrast with the findings in section 2 where most (75%) of the Chinese counterparts of *however* are implied. This seems to indicate that contrastive connectors such as *however* are more explicitly translated in E/C translation than in the Chinese original. If this is true, then the E/C translation process may have produced an inter-language which is similar to but is not natural Chinese. It also implies that the translation process is a clarifying process, which makes the translation more explicit than the original. It remains to be proved if this is the case. (3) The positional distributions of *however* and its Chinese translations are rather different. *However* is most frequently used in second sentence initial (55.4%) and initial (38.6%) positions. The Chinese counterparts are mostly in the sentence initial (94.06%) position. This confirms the findings in section 2, where the initial positions of the Chinese connectors total 85%, but those of the English translations total 52.5%, and the second initial positions of *however* total 54.5%.

4. Conclusions

The evidence of *however* and its counterparts in the samples from the Chinese-to-English translation corpus of HLM and the English-to-Chinese translation corpus of Babel brings us to the following tentative conclusions in terms of usage, position, translation process, and translated language.

In expressing contrastive and concessive relations, the Chinese language tends to be more implicit by using fewer such connectors, while the English language tends to be more explicit by using more such connectors. In the HLM sample, 75% of such relations in Chinese are implied without using overt connectors, whereas in its English translations only 15% are implicit, with 85% of such relations clearly expressed by explicit connectors.

The positional distributions of *however* and its Chinese translations are rather different. For *however,* the second initial position in the sentence is common: 54.5% in HLM and 55.4% in Babel. For Chinese contrastive connectors, the sentence initial position is the most common: 85% in HLM and 94% in Babel.

The translation process seems to be a clarifying and explicating process. In the HLM sample, the implied contrastive relations in Chinese are translated explicitly into English with *however* etc. In Babel, 96% of occurrences of *however* are explicitly translated into a Chinese contrastive connector.

The translated language seems to be somewhat different from the native language such as Chinese. In translated Chinese, the contrastive relations are often expressed by overt contrastive connectors (Babel). In native Chinese language (HLM), however, such relations are often implicit or expressed by other means than using overt contrastive connectors.

The above findings in this small-scale study are based on limited samples from the HLM and Babel corpora, but they have revealed some interesting characteristics of contrastive connectors in both languages. To reveal the overall usage patterns of these connectors in the two languages, large balanced monolingual and parallel corpora are needed for further research. This seems to be a promising field of research, the findings and results of which can be applied to the teaching and learning of these two languages among non-native speakers, and to the (automatic) translation practice between English and Chinese.

Notes

The research presented in this chapter is part of the pilot study of my PhD research on contrastive and concessive connectors in English and Chinese, funded by the IDS, University of Auckland, New Zealand. I am grateful to Professors Michael Barlow and Jim Miller, whose supervision and advice are highly appreciated. I also want to thank Huaqing Hong for allowing me to use the online HLM Parallel Corpus, and Richard Xiao for using the online Babel English-Chinese Parallel Corpus and his painstaking work in editing the whole book in which this chapter is included.. An earlier version of the manuscript of this chapter was presented at the International Symposium on Using Corpora in Contrastive and Translation Studies, 2008, Hangzhou, upon which further revisions have been made.

References

Altenberg, B. (2006), “The function of adverbial connectors in second initial position in English and Swedish”, in K. Aijmer, and A. M. Simon-Vandenbergen (eds.) *Pragmatic Markers in Contrast* (Volume 2), 11-37. Amsterdam: Elsevier.

Bybee, J. (2007), *Frequency of Use and the Organization of Language*. Oxford/New York: Oxford University Press.

Cao, H. (1994), “A quantitative comparison of English and Chinese”, in China English and Chinese Comparative Studies Association (ed.) *A Comparative Study of English and Chinese*, 220-233. Changsha: Hunan Publishing House of Science and Technology.

Firth, J. R. (1957), “A synopsis of linguistic theory, 1930 – 1955”, in J. R. Firth (ed.) *Studies in Linguistic Analysis. Special volume of the Philological Society,* 1-32. Oxford: Basil Blackwell.

Kemmer, S. and Barlow. M. (2000), *Usage-based Models of Language*. Stanford, Calif.: CSLI Publications.

Lin, R and Li, M. (2004), "Experiment report of English words and Chinese characters", in J. Wang and L. Zheng (eds.) *Language and Culture: Contrastive Studies between English and Chinese 1995-2003*, 440-454. Shanghai: Shanghai Foreign Language Education Press.

Pan, W. (2004), "A hundred years of comparative study between English and Chinese", in J. Wang and L. Zheng (eds.) *Language and Culture: Contrastive Studies between English and Chinese 1995-2003*, 102-140. Shanghai: Shanghai Foreign Language Education Press.

Sinclair, J. (2004), *Trust the Text: Language, Corpus and Discourse*. London/New York: Routledge.

CHAPTER NINETEEN

A CORPUS-BASED COMPARISON OF SATELLITES IN CHINESE AND ENGLISH

HUI YIN

In this study, two balanced corpora (the Academia Sinica Balanced Corpus of Modern Chinese and the British National Corpus) were used to compare satellites in Chinese and English. The corpus data show that English and Chinese are quite different in the nature of satellites. Satellites in English are mainly verb particles while satellites in Chinese are basically second elements in verbal compounds. The corpus findings inform us that the most frequently used satellites in English are path satellites such as *in* and *out*, *up* and *down* whereas in Chinese the most frequent ones are motion verbs such as *lai* 'come' and *qu* 'go'. The corpus evidence also indicates that there are more satellites used in Chinese than in English. This difference is largely due to their differences in verb lexicalization. Chinese regularly uses satellites to specify realization or fulfilment but that is not the case in English.

1. Verb-framed languages vs. satellite-framed languages

Talmy (1985, 2000) proposes that language can be classified into two typological categories on the basis of where a particular language characteristically expresses the core schema of the event complex: in the main verb or in a satellite to the verb. Talmy (2000: 222) defines the satellite as "the grammatical category of any constituent other than a nominal or prepositional-phrase complement that is in a sister relation to the verb root." The satellite can be a bound morpheme or a free word and it includes English verb particles, German separable and inseparable verb prefixes, Chinese verb complements among others.

Languages that characteristically express the schematic core by the verb are verb-framed languages while languages that characteristically express the schematic core by the satellite are satellite-framed languages. The following examples illustrate such a distinction.

(1) Maotouying cong shandong li fei
cat-head hawk from mountain-hole inside fly
chulai. (Chinese)
exit-come (out)
'The owl flew out from the cave.'

(2) el buho salió volando de la cueva (Spanish)
'the owl exited flying from the hole.'

In the Chinese example, the verb *fei* 'fly' indicates the flying movement. It is the job of the satellite *chulai* 'exit-come, out' to express the direction. If we take the basic message of a movement-event communication to be that an entity has moved along a path in a specified direction (Berman and Slobin 1994), we can say that Chinese is a satellite-framed language, because the core information 'path' is conveyed by the satellite.

However, in the Spanish (a verb-framed language) example, the verb *salió* 'exited' alone indicates the core information of direction. The encoding of motion is conveyed by the satellite *volando* 'flying'.

English is also a good example of satellite-framed language. In the English example *the owl flew out from the cave*, the satellite *out* conveys the core information (path), while the main verb *flew* expresses the co-event (motion).

In English and Chinese, a verb and its satellite(s) constitute a verb complex to form a macro event (e.g. motion plus path). The satellite relates to the verb root as a dependent to its head. A set of forms that can serve as satellites in a particular language often overlap with a set of forms in another grammatical category in that language, generally the category of prepositions, verbs or nouns.

2. English satellites: mainly verb particles

In English, a verb and its satellite(s) form a verb complex, which often denotes a conceptual macro event with two phases (Berman and Slobin 1994). A set of forms that can serve as satellites in English largely overlap with prepositions while in Chinese, satellites largely overlap with main verbs.

2.1. Differences between satellites and prepositions

Since English satellites overlap with prepositions, how can we distinguish satellites from prepositions in actual English language contexts? Talmy (2000) notices some important differences between satellites and prepositions. First, these two categories do not have exactly identical membership, that is, there are forms serving only one function or the other. For example, in English, *apart, away, back and forth* always serve as satellites while *of*, *from* and *toward* always act as prepositions. Moreover, items that can serve both functions have different senses in each as illustrated in examples (3-4). It is obvious from these examples that *to* as a preposition in (3) has a different sense from *to* as a satellite in (4).

(3) He came to the university.

(4) He came to.

Satellites and prepositions also differ greatly in their properties. "With regard to phrase structure and co-occurrence, a satellite is in construction with the verb, while a preposition is in construction with an object nominal" (Talmy 2000: 107). Therefore, when a nominal is omitted, the preposition that would have co-occurred with that nominal should also be omitted; in contrast, the satellite should remain because it is closely associated with the verb as in the following example:

(5) When he saw a snake in the house he ran away (from the house) as fast as possible.

In addition, satellites and prepositions are different in positional properties. A preposition should precede its nominal whereas a satellite has more complex positional properties. It could either precede or follow a full NP, but it should follow a pronominal NP that lacks a preposition.

The satellites in English are mostly involved in the expressions of path. The main verb and its satellite constitute a verb complex that conveys a macro event.

2.2. Simple type: motion + path

In English, the "motion + path" type is basic on the evidence that it is first acquired by children (Berman and Slobin 1994). It is possible that the verb simply indicates the fact of movement without specifications of manner while its satellite specifies direction or path. Satellite-framed languages like English allow for detailed descriptions of paths within a verb complex, "because the syntax makes it possible to accumulate path satellites to a single verb, along with preposition phrases that add further specification" (Berman and Slobin 1994: 118), as exemplified in (6):

(6) The man *went out* of the house into a cave.

The simple type (also basic type) of such a verb complex expresses motion and path. Verbs in such constructions are general-purpose verbs such as *come*, *go*, *arrive*, and *move*, which simply indicate movement with no co-event involved and are acquired earlier by children (Berman and Slobin 1994).

The following are the commonly used path satellites: *in*, *out*, *up*, *down*, *away*, *through*, *past*, *on*, *under*, *over*, *below*, *across*, *off*, *back*, *forth*, etc. A particular verb can be followed by a bunch of satellites to indicate different directional specifications (paths). For example, we can pair different satellites with the general-purpose verb *go* to form different verb complexes:

(7) He went in. He went out.
He went up. He went down.
He went across. He went by.
He went off. He went along.
He went through. He went past.
He went above. He went below.
He went back. He went over.

2.3. Co-event type: motion and manner (or cause) + path

In a satellite-framed language like English, since the path components are tucked away in satellites, what kind of other semantic element can be encoded in the main verb? In fact, the main verb often encodes co-events. In English, in which path is expressed by satellites, a whole series of verbs in common use could express motion occurring in various manners or by various causes (Talmy 2000). Verbs in such constructions are more specific and more complex than general-purpose verbs such as *come* and *go*.

As Berman and Slobin (1994) suggested, satellite-framed languages like English have a tendency towards greater specification of manner than verb-framed languages, probably because the lexicon provides a large collection of verbs that conflate manner with change of location. Instead of using general-purpose verbs such as *go* and *move*, the following examples give us illustrations of event conflation of motion with manner in the verb root:

(8a) The rock slid/rolled/bounced down the mountain.
(8b) I ran/limped/stumbled/hopped/rushed my way down the hill.

(9) I knocked the nail into the wall.

Besides conflating manner with motion, the main verb in a verb complex consisting of a verb root and its satellite can encode co-events of both motion and cause. Consider the example in (9), where *knocked* basically refers to what the speaker did to the nail, so it expresses the cause of the event.

3. Chinese satellites: basically verb complements (compounds)

3.1. Directional complement: motion + path

In Chinese, some verbs, typically verbs of displacement, can serve as the main verbs (V_1) in directional verbal compounds. As Li and Thompson (1981) observe, the most obvious type of displacement verb is a verb of motion such as *hui* 'return', *zou* 'walk', and *guo* 'cross'. Another common type of displacement verb is a dislocation verb "that inherently implies that the direct object undergoes a change of location" (Li and Thompson 1981: 58), e.g. *ban* 'remove', *reng* 'throw', *song* 'send', *ji* 'mail', *ju* 'lift', *fang* 'put', and *duan* 'carry'. These verbs conflate movement with some other activity. As for satellites, that is, V_2 denoting path or direction, they are highly limited lexically. The prototypical satellite verbs functioning as directional complements in VV compounds are *lai* 'come' and *qu* 'go', although there is a small set of additional verbs which serve as complements of direction.

3.1.1. Satellite verbs *lai* 'come' and *qu* 'go' as complements

The satellite verbs *lai* 'come' and *qu* 'go' are used extensively in Chinese as complements of direction (path). They occur after verbs of movement or action to indicate path or direction 'towards' and 'away from' the speaker respectively (Yip and Don 1998a). Typically, these involve events of TRANSPORTATION as in (10) or TRANSACTION (TRANSLOCATION) as in (11):

(10a) Zhangsan **zou** **lai**-le.
Zhangsan **walk** **come**-ASP
'Zhangsan came over here on feet.'
(10b) Lisi **zou** **qu**-le.
Lisi **walk** **go**- ASP
'Lisi went over there on foot.'

(11a) Zhangsan **na** **lai**-le yiben shu.
Zhangsan **carry** **come**-ASP one-CL book
'Zhangsan brought a book.'
(11b) Lisi **na** **qu**-le yiben shu.
Lisi **carry** **go**- ASP one-CL book
'Lisi took a book with him.'

The verbs in these sentences are bound together and the verb of movement or moved action is naturally accompanied by path or direction. These verb complexes actually form directional compounds in which the main verb V_1 expresses motion or co-event while the satellite V_2 conveys the core information, i.e. path.

3.1.2. Double complements (or VVV compounds) and their figurative uses

There is a small group of motion verbs in Mandarin Chinese other than *lai* and *qu* which also participate in VV compounds. These verbs have directional meanings denoting path when they occur in directional complements in addition to verbal meanings when they are used as independent verbs (Li and Thompson 1981). Two examples are given below:

(12) Ta **zou** **jin** le jiaoshi.
S/he **walk** **enter** ASP classroom.
'S/he walked into the classroom.'

(13) Ta **fang** **xia** le shubao.
S/he **put** **descend** ASP schoolbag
'S/he laid down her/his schoolbag.'

There are about eight verbs in this group (Li and Thompson 1981): *shang* 'ascend', *xia* 'descend', *jin* 'enter', *chu* 'exit', *qi* 'rise', *hui* 'return', *guo* 'cross', and *kai* 'open'. *Lai* 'come' and *qu* 'go' may be linked to this group of 8 motion verbs (Yip and Don 1998b) in Chinese to form a set of double directional complements elaborating path. Therefore, there are 16 members in this category of double complements when the 8 verbs combine with *lai* and *qu*.

A. Following verbs of movement (absolute motion)

(14) Huar **diao xia-lai**-le.
Picture **drop descend-come**-ASP
'The picture fell down.'

(15) Che **kai guo-qu**-le.
Car **drive cross-go**-ASP
'The car went past.'

B. Following verbs of action (translocation)

(16) Shu **fang hui-qu**-le.
Book **put return-go**-ASP
'The book was put back.'

(17) Cai **duan jin-lai**-le.
Dish **bring enter-come**-ASP
'The dishes were brought in.'

Sometimes these double complements can have metaphorical interpretations in appropriate contexts besides being used literally as in (18) and (19). In that case, the VV complements (satellites) in VVV compounds could be regarded as having been lexicalized.

(18) Ni yinggai ti ta **shang-lai**.
You should pick him **ascend-come**
'You should lift him up.'/'You should promote him.'

Here *shanglai* 'ascend-come' can be used figuratively: come up high in social (or administrative) positions and the metaphorical meaning is derived from the basic meaning *shanglai* 'come up'.

(19) Ta xiang huo **xia-qu**.
S/he want live **descend-go**
'He wants to live on.'

In (19), *xiaqu* 'descend-go' is also used figuratively. The directional aspect of *xiaqu* is metaphorically extended to the aspect of time (Li and Thompson 1981). Therefore, *huo xiaqu* 'live descend-go' is interpreted as 'live on'. The double satellite *xiaqu* has been lexicalized to indicate path.

In Chinese, path satellites are very lexically restricted. If given a particular verb of motion or action, we can combine it with different path satellites to make different VV compounds. Thus, VV compounds of the "motion + path" type are very productive and frequent.

3.2. Fulfilment complement

Fulfilment verb compounds are important in Chinese and they are widely used both in speech and writing (Li and Thompson 1981). In Mandarin Chinese, complements of fulfilment in VV compounds are cases in which the second verb indicates fulfilment or result of the action of the first verb. Given an appropriate verb, we can freely create new fulfilment verb compounds. Verbs used as complements of fulfilment are very restricted lexically. The commonly used ones are the following phase verbs or achievement verbs: *po* 'break', *dao* 'fall', *diao* 'drop', *kai* 'open, separate', *wan* 'finish', and *dao* 'attain, achieve'. These verbs serving as complements express the phases or achievements of the first verbs in the compounds. In (20), the result of pushing is that the person being pushed fell; in (21) the result of wiping the dirty things is that the dirty things were gone.

(20) Ta **tui dao** le wo.
S/he **push fall** ASP I
'S/he pushed me down.'

(21) Zhangsan **mo diao** le zang dongxi.
Zhangsan **wipe drop** ASP dirty thing
'Zhangsan wiped out the dirty things.'

However, in English, the fulfilment or resulting state is usually indicated by an adverb or a particle – in short, by an atemporal relational predicate (Langacker 1987) whereas in Chinese, the resulting state is often indicated by a complement verb or adjective which usually follows the first verb immediately. It is obvious in (20) that the fulfilment is indicated by the satellite verb *dao* 'fall' while in the English translation, it is expressed by a particle (satellite) *down*. Usually, the action verb and the complement verb in Chinese form a VV compound. That is one of the main reasons that explains why there are much more compounds in Chinese than in English (Nicoladis and Yin 2001).

Fulfilment verb compounds are always compounds of two parts, although each part may be a compound itself. In such a compound, the second part signals fulfilment or some result of the action or process conveyed by the first part. Fulfilment verb compounds can express the following different kinds of fulfilment or result (Li and Thompson 1981):

1. Cause

(22)	Wo	**da**	**po**	le	huaping.
	I	hit	broken	ASP	vase

‘I broke the vase.’

(23)	Ta	**la**	**kai**	le	men.
	S/he	pull	open	ASP	door

‘S/he pulled the door open.’

In this kind of VV compound, the first verb indicates the cause and the second verb signals the result. In (22), for example, the action *da* ‘hit’ produces the result of being broken of the vase while in (23), the action of *la* ‘pull’ results in *kai* ‘open’ (of the door).

2. Achievement

(24)	Ta	**zhao**	**dao**	le	na	ben	shu.
	S/he	search	arrive	ASP	that	CL	book

‘S/he succeed in searching (found) that book.’

(25)	Wo	ba	yifu	**xi**	**ganjing**	le.
	I	BA	clothes	wash	clean	ASP

‘I washed the clothes clean.’

In this kind of fulfilment verb compound, the first element denotes the action and the second element expresses the achievement of the action verb. In (24), the meaning of *dao* is derived from its independent verbal meaning ‘arrive’ and the meaning of *dao* in this example can be described as ‘succeed in or achieve the goal’ of *zhao* ‘searching’. In (25), the action of *xi* ‘wash’ achieves the result of *ganjing* ‘being clean’ of the clothes.

3. Phase

There are some fulfilment verb compounds in which the second part denotes something more like the type of action described by the first verb or the degree to which it is carried out than its result (Li and Thompson, 1981). These compounds can be called phase fulfilment verb compounds, in which the second element is highly restricted lexically. The following are the most commonly used phase verbs (the second element) in this kind of fulfilment verb compound.

(a) *wan* ‘finish’, which signals the completion of an action

(26)	xie wan	'write-finish' — finish writing
	du wan	'read-finish' — finish reading
	zuo wan	'do-finish' — finish doing

(b) *zhao* ‘be on target’

(27)	zhao zhao	'search-be on target — find
	shuo zhao	'say-be on target' — say (it) right
	cai zhao	'guess-be on target' — guess right

(c) *zhu* ‘hold on’

(28)	zhan zhu	'stand-hold on' — stand still
	ting zhu	'stop-hold on' — stop firmly
	zhua zhu	'grab-hold on' — grab onto

(d) *hao* ‘completing the task signalled by the first verb’, which is similar to but not identical with the meaning of *wan* ‘finish’.

(29)	xi hao	'wash-complete task — complete the task of washing
	zuo hao	'do-complete task — complete the task of doing
	tian hao	fill out-complete task — complete the task of filling out

4. Chinese and English satellites in the corpora

In this study, two balanced corpora were used to compare satellites in Chinese and English: the British National Corpus (BNC) for English and the Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus) for Chinese. The BNC is a 100 million word corpus collected from samples of written and spoken English through a wide range of sources. This corpus was designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The written part of the BNC takes up 90% of the entire corpus while the spoken portion occupies 10% of the corpus. The BNC is a balanced corpus which collected samples from different genres. The building of the corpus started in 1991 and was finished in 1994. No new texts have been added after the completion of the project. However, the corpus was slightly revised before the release of the second edition BNC World (2001) and the third edition BNC XML Edition (2007).

The Sinica Corpus is the first balanced Chinese corpus which was designed for analyzing modern Chinese. The preliminary version was developed on a small scale and became available to the academic community in 1994. The current corpus (Sinica 3.0), which was completed in 1997, contains 5 million words. Samples of the corpus were collected from different areas and classified according to five criteria: genre, style, mode, topic, and source. Therefore, the Sinica Corpus is a balanced corpus which is a representative sample of the modern Chinese language. Like the BNC, written texts make up approximately 90% of this corpus and the spoken portion takes up 10% of the entire corpus. Both the BNC and the Sinica Corpus are balanced corpora and they are useful for linguistic comparisons between modern English and modern Chinese.

In order to compare English satellites with Chinese satellites, 1,000 sentences were randomly selected from each corpus, which formed the basis for the comparison. Each sentence was checked for whether it contains verb complexes of my interests and both Chinese and English satellites were identified. The corpus findings indicate that there are more satellites used in Chinese than in English as Figure 19-1 illustrates.

It can be seen from Figure 19-1 that the token frequency of satellites (the total number of satellites) is higher in Chinese than in English. In Chinese about 5 out of 10 sentences contain satellites while in English nearly 3 out of 10 sentences have satellites in them. In terms of type frequency (the number of different satellites), it is higher in Chinese than in English as Figure 19-2 indicates.

Talmy (1985, 1991) claims that both English and Chinese are basically satellite-framed languages, in which the core information of path expressions is conveyed by satellites rather than by main verbs. However, the corpus evidence shows that English and Chinese are quite different in the nature of satellites. As Tables 19-1 and 19-2 show, satellites in English are mainly verb particles whereas satellites in Chinese are basically the second elements in verbal compounds – resultative (or directional) complements.

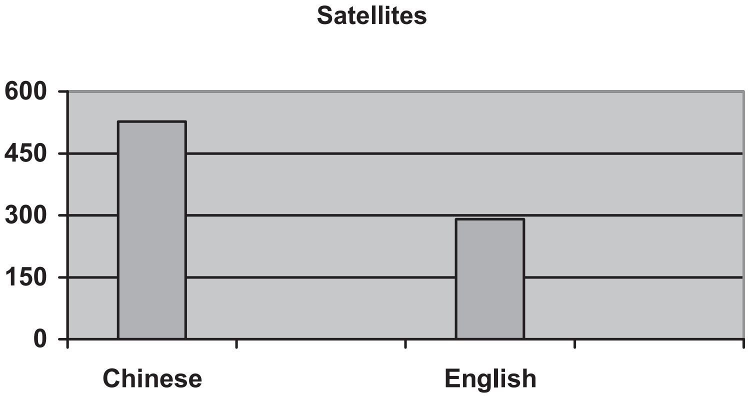


Figure 19-1. Frequency of satellites in Chinese and English

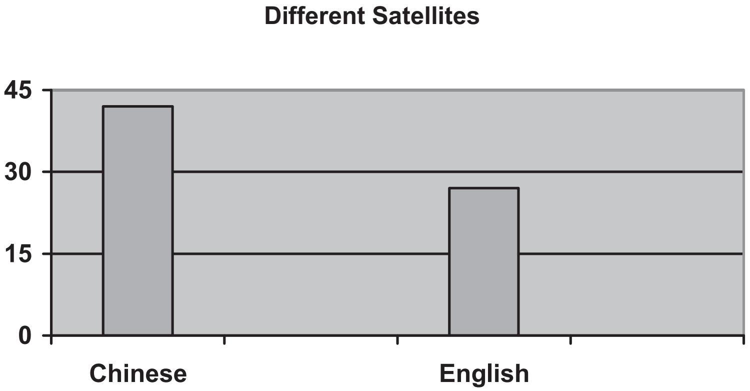


Figure 19-2. Frequency of different types of satellites in Chinese and English

Tables 19-1 and 19-2 list individual satellites and their frequencies (where the number is lager than one) in English and in Chinese. In English, satellites largely overlap with prepositions while in Chinese, satellites overlap with verbs.

Table 19-1. English satellites and their frequencies (number >1)

English satellites	N	English satellites	N
out	55	round	9
up	31	along	8
in	28	across	6
back	27	around	5
down	23	over	4
into	20	past	3
on	17	to	3
through	14	beyond	2
away	13	above	2
off	13		

Table 19-2. Chinese satellites and their frequencies (number >1)

Chinese satellites	English gloss	N	Chinese satellites	English gloss	N
lai	come	75	wan	finish	7
qu	go	55	chuqu	exit come	7
chu	exit	52	guoqu	cross go	6
chulai	exit come	36	jin	enter	6
dao	arrive, achieve	35	diao	drop, away	5
shang	ascend	32	hao	complete	4
qilai	rise come	32	huilai	return come	3
zou	walk, away	21	qing	clean	3
qi	rise	18	jinlai	enter come	3
zhu	hold on	17	zhao	be on target	3
xia	descend	16	ding	stop, hold	3
kai	open	14	jinqu	enter go	2
shangqu	ascend go	12	po	break	2
xialai	descend come	10	qingchu	clear	2
xiaqu	descend go	10	si	die, dead	2
hui	return	8	cheng	achieve, succeed	2
huiqu	return go	8	ru	enter	2
guolai	cross come	7			

The frequency data displayed in Figures 19-1 and 19-2 suggest that English and Chinese have different frequency patterns of different kinds of satellites in verb complexes. The corpus findings suggest that the most frequently used satellites in English are path satellites such as *in* and *out*, *up* and *down* whereas in Chinese the most frequent ones are motion verbs used as verb complements such as *lai* ‘come’ and *qu* ‘go’, which often indicate direction toward or away from the speaker. Figures 19-3 and 19-4 display the distribution patterns of the top eight satellites in English and Chinese. In English, the eight most frequent satellites account for 74% of all the satellites while in Chinese, the eight most frequent satellites take up 64% of all the satellites.

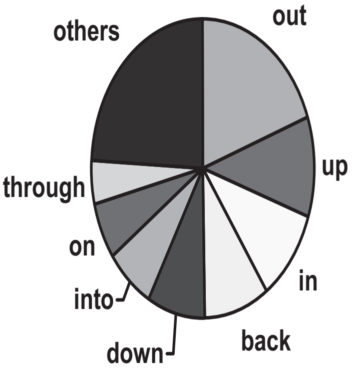


Figure 19-3. Frequency distribution of the eight most frequent satellites in English

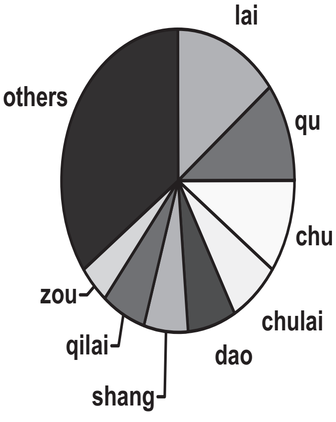


Figure 19-4. Frequency distribution of the eight most frequent satellites in Chinese

As Figure 19-4 indicates, the most frequent double complement (satellite) in Chinese is *chulai* ‘exit come’. The common use of *chulai* with another verb to form a compound is due to the fact that many Chinese verbs do not specify fulfilment or result and thus, they often need another verb or compound to perform this function. *Chulai* is the most frequent satellite used together with another verb to signal fulfilment or result. Another double complement frequently used as a satellite is *qilai* ‘rise come’. Most of the instances of *qilai* signal the aspectual meaning of inceptiveness. In this kind of use, *qilai* ‘rise come’ does not specify direction of real motion but rather indicates that the situation has started and will continue as in the case *ku qilai* ‘began to cry’. Here, this directional verb has been extended to function as an aspectual marker.

5. Differences of English and Chinese verb lexicalization

Another important factor to account for the fact that there are more verb complexes in Chinese than in English is the differences of English and Chinese verb lexicalization. Chinese is a strongly satellite language, which regularly uses satellites to specify realization or fulfilment. Perhaps most of Chinese verbs require a satellite for their realization. The following example is entirely acceptable in Chinese but sounds strange in English:

(30) Wo sha le zhu (keshi mei sha si)
I kill asp pig (but not kill die)
*‘I killed the pig but it didn't die’

(31) Wo sha si le zhu.
I kill die asp pig
‘I killed the pig.’

The semantics of the above examples can be explained as follows. In (30), the first clause means that the speaker performed the action with the intention of killing the pig and the conjoined second clause in parentheses indicates that the action did not achieve the goal, i.e. success in killing the pig. In contrast, with the confirmational satellite *si* ‘die’ in (31), the sentence is now an undeniable assertion that the speaker succeeded in killing the pig.

So the English verb *kill* used to gloss the Chinese verb *sha* does not really correspond in meaning. Therefore, a gloss like ‘I killed the pig but the pig didn’t die’ is really contradictory in English, but it accurately expresses the non-paradoxical meaning in the Chinese original, i.e., ‘I performed the action with the intention of killing the pig, but the pig didn't die.' English verbs such as *kill*, *open*, *kick* are generally construed to refer to a simplex action of the fulfilment type and they specify the attainment of a certain final state.

In Chinese, the concept covered by a typical English verb such as *kill* is divided into two parts: the final outcome, usually confirmed by a verb satellite and an action intending that outcome, which is signalled by the main verb. As a result, the unitary concept of an English verb often has a counterpart in Chinese with two-part conceptualization expressed by a verb plus another verb (satellite). Hence, quite a few fulfilment verb compounds in Chinese come into being this way.

Furthermore, the semantics of the Chinese verb-satellite system ranges more widely than in English. Some Chinese verbs can enter into constructions not only with resultative verbs (satellites) to indicate fulfilment, but also with those that express underfulfilment, overfulfilment, antifulfilment and other event (Talmy 2000).

(a) Fulfilment

(32) Wo ba kuaizi **zhe** **duan** le.
I BA chopstick break broken ASP
'I broke the chopstick.'

In (32), the first verb *zhe* means to squeeze in on an object with the intention of breaking it and the second verb *duan* expresses the fulfilment that the action achieves its goal of breaking it.

(b) Underfulfilment

(33) Wo ba kuaizi **zhe** **wan** le.
I BA chopstick break bend ASP
'I broke the chopstick bent.' (I squeezed in on the chopstick to break it, but only managed to bend it.)

In (33), the verb *zhe* 'break' takes a state-change satellite *wan* that denotes a 'bent' state. Usually in the efforts of breaking something, a bent state for the object is on the way to a broken state. Therefore, the verb *wan* 'bent' indicates an insufficient fulfilment of the full scope of intention. Thus, the resultative verb *wan* in this example sentence marks underfulfilment.

(c) Overfulfilment

(34) Wo ba kuaizi **wan** **zhe** le.
I BA chopstick bend broken ASP
'I bent the chopstick broken.' (I squeezed in on the chopstick to bend it, but wound up breaking it.)

In (34), the verb *wan* 'bend' takes a state-change satellite *zhe* that denotes a broken state. Since the concept of breaking is on a continuum with that of bending and conceived as lying beyond it, the resultative verb that marks this excess is properly termed as overfulfilment (Talmy 2000).

(d) Antifulfilment

(35) Wo ba yifu **xi** **zang** le.
I BA clothes wash dirty ASP
'I washed the clothes dirty.' (I washed the clothes [e. g., in a lake] but they turned out dirtier than before.)

In (35), the verb *xi* 'wash' takes the state-change satellite *zang* 'dirty' to express the following combined meaning: immerse and rub the clothes with the intention of making them clean, but they turned out to be dirtier than before. Talmy (2000) terms a satellite for this semantic effect on the verb as an antifulfilment satellite.

(e) Other-event

(36) Wo ba yifu **xi** **po** le.
I BA clothes wash torn ASP
'I washed the clothes torn.' (I washed the clothes and they got torn in the process.)

In verb-satellite relations, the state indicated by the satellite could lie somewhere along the conceptual axis leading to the intended goal. "Thus, the state expressed by the satellite was either before the starting point, almost at the goal, or past the goal" (Talmy 2000: 277). However, in (36), the verb *xi* 'wash' takes the satellite *po* with the meaning of 'torn'. This satellite expresses a state that results from the action of *xi* 'wash' but *po* 'torn' does not lie somewhere along the axis of the intended goal. Therefore, such a satellite like *po* 'torn' in (36) can be termed as an other-event satellite.

Unlike Chinese, English generally uses one word to express action and goal such as *pull*. However, it is very common in Chinese to use two words such as *pull open* to indicate action and goal respectively. As a result, VV fulfilment compounds are very common in Chinese to denote action and goal.

6. Conclusions

Both English and Chinese are typically satellite-framed languages. However, they are quite different in the nature of satellites. Satellites in English are mainly verb particles while satellites in Chinese are basically second elements in verbal compounds, i.e. the resultative complements. This distinction is largely due to the fact that Chinese is a verb-serialized language in which verbs in a sequence without any intervening conjunctions are quite common but English is not.

The corpus findings suggest that these two languages have different frequency and distribution patterns of different kinds of satellites in verb complexes. The results indicate that the most frequently used satellites in English are path satellites such as *in* and *out* whereas in Chinese the most frequent ones are motion verbs such as *lai* ‘come’ and *qu* ‘go’. The corpus evidence also demonstrates that there are more satellites used in Chinese than in English. This difference largely results from their differences in verb lexicalization. Chinese regularly uses satellites to specify realization or fulfilment (e.g. *la kai* ‘pull open’, *sha si* ‘kill die’) but that is not the case in English.

References

Berman, R. A. and Slobin, D. I. (1994), *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, New Jersey: Lawremce Erlbaum Associates, Inc.

Chang, C. H. (1990), “On serial verbs in Mandarin Chinese: VV compounds and co-verbal phrases”. *Ohio State University Working Papers in Linguistics* 39: 316-339.

Chao, Y. R. (1968), *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Chen, H. C. (ed.) (1997), *Cognitive Processing of Chinese and Related Asian Languages*. Hong Kong: The Chinese University Press.

Huang, S. (1998), “Chinese as a headless language in compounding morphology”, in J. L Packard (ed.) *New Approaches to Chinese Word Formation*, 261-283. Berlin: Mouton de Gruyter.

Langacker, R. (1987), *Foundations of Cognitive Grammar, I: Theoretical Prerequisites*. Stanford, CA: Stamford University Press

Langacker, R. (1991), “Cognitive Grammar”, in F. G. Droste and J. E. Joseph (eds.) *Linguistic Theory and Grammatical Description,* 275-306. Amsterdam: John Benjamins.

Li, C. N. and Thompson, S. A. (1981), *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Nicoladis, E. and Yin, H. (2001), “Acquisition of Chinese and English compounds by bilingual children”. Paper presented at Boston University Conference on Language Development. Boston, 2-4th November 2001.

Packard, J. L. (ed.) (1998), *New Approaches to Chinese Word Formation*. Berlin: Mouton de Gruyter.

Talmy, L. (1985), “Lexicalization patterns: Semantic structure in lexical forms”, in T. Shopen (ed.) *Language Typology and Syntactic Description: Vol. 3. Grammatical Categories and Lexicon*, 36-149. Cambridge: Cambridge University Press.

—. (1991), “Paths to realization: A typology of event conflation”, in *Proceedings of the Annual meeting of the Berkeley Linguistic Society*, 17. [Supplement on *Buffalo Papers in Linguistics 91-01* (182-187).]

—. (2000), *Toward a Cognitive Semantics*. Cambridge, Mass.: MIT Press.

Wardhaugh, R. (1997), *Understanding English Grammar: A Linguistic Approach*. Oxford: Blackwell Publishers Ltd.

Yip, P. and Rimmington, D. (1998a), *Basic Chinese: A Grammar and Workbook*. London and New York: Routledge.

Yip, P. and Rimmington, D. (1998b), *Intermediate Chinese*. London and New York: Routledge.

CHAPTER TWENTY

REPETITION PATTERNS OF RHETORIC FEATURES IN ENGLISH AND CHINESE ADVERTISEMENTS: A CORPUS-BASED CONTRASTIVE STUDY

GUILING NIU, HUAQING HONG

1. Introduction

There have been myriad studies concerning repetition use in advertising in a general context (McQuarrie 1996, Leigh 1994) or a monolingual context (Goddard 1998, Phillips 2002, Lagerwerf 2005, Zeng 2005) and some contrastive studies regarding repetition use of different languages in different cultures or countries (Ahmed 2000) whereas the comparative analysis of repetition use in advertising in a bilingual/multilingual context is rarely seen. There are also many studies concerning bidirectional translation of repetition, but mainly concerned with literary works (Han 2001, Ma 2004, Jiang 2007, Yang and Wu 2006) while research on repetition translation in advertising has received much less attention. Moreover, most related studies in view of the repetition use or translation are qualitative while quantitative analysis is seldom operated.

To follow up the previous studies and to address the issue of repetition in a bilingual context with a corpus linguistics approach, based on the data from the on-going project of a parallel corpus of English and Chinese advertisements in Singapore print media, a multilayered and annotated corpus, this chapter combines McQuarrie and Mick's (1996) taxonomy and Leigh's (1994) classification of rhetoric to investigate how different types of rhetorical repetition are used in the bilingual context in Singapore, to explore respective rhetorical repetition patterns (e.g. rhyme, assonance, alliteration, anaphora, etc), and to find the distributive properties of their uses in the corpus. The properties are demonstrated with corpus annotation, and quantitative analysis is employed to justify their similarities and differences with statistical tests. In so doing, this study is expected to shed light on how a corpus-based investigation can contribute to contrastive linguistic study of such a kind.

2. Literature review

The art of repetition, one important expression form of rhetoric, is universally employed in advertisements in any context and has long been the hot topic for language and advertising researchers (Cooper 1960; Cooker 1992; Cook 1992, 2001; McQuarrie and Mick 1999). Aristotle defined rhetoric as "the faculty of discovering all the available means of persuasion in any given situation" (cited in Corbett 1990: 3). The use of rhetorical repetition is universal in advertisements just because of its strongly persuasive strengths.

Traditionally, rhetorical figures are subdivided into schemes, superficial deviations such as rhyme and alliteration, and tropes, meaningful deviations such as metaphors and puns. Repetition is one of the two rhetoric operations under schemes (repetition, reversal) (McQuarrie and Mick 1996). Repetition patterns of rhetorical figures have been catalogued (McQuarrie and Mick 1996, Leigh 1994) ranging from the familiar (rhyme, alliteration) to the obscure (antimetabole, parison). Despite the frequent appearance of rhetorical repetition expressions in print advertisements, their incorporation into advertising theory and research has been minimal. This chapter develops a framework for classifying the repetition patterns of rhetorical figures that distinguishes among three repetition categories (sound repetition, word repetition and repetition of structure-phrase) and is intended to follow up and to expand prior studies in this area.

2.1. Functions of repetition

Psychological studies have shown that materials with rhythm and repetition read pleasantly and fluently, and this biological inertia leads to memorial inertia. Nearly all people have the similar experience that articles with rhythm and repetition are much easier to remember than those awkward sounding ones, and repetitive materials have become a psychological need in reading. Considering this, many advertisers apply this approach to the promotion strategy of their products or services and achieve good effects. These ads employ repetition to stress the features, brand and functions of the goods or services, and it is hard for the recipients to forget them as long as they read or hear them.

McQuarrie (2003) conducted an experiment, called Aided Recall, using a hierarchical log-linear analysis, a saturated model containing all interaction terms, finding that rhetorical device was better recalled than its controls and that participants were more likely to recall a rhyme such as *pop the top* as opposed to a pun such as *pull a fast one*. This result is consistent with an ease of processing explanation that favours schemes (the two main categories of rhetoric and the other one is trope as mentioned earlier); the excess regularity that characterizes schemes appears to be advantageous when processing conditions are degraded and resource availability is minimal.

In addition to bringing about beauty of rhyming, facilitating comprehension and recalling of the products or services, rhyming and repetition can also serve the following functions (Zeng 2005):

1. Onomatopoeia, to produce vivid acoustic effect;
2. Enabling advertisements to be precise and active, intense and striking, which helps strengthen recipients' effective recalling of them;
3. Generating greater influential and persuasive force and inducing their purchasing impetus in time;
4. Promoting cultural association;
5. Acting as the text mark of adverting semantic focus and contrast with other products, and thus leading the receivers to follow the current message schemata and inferring new and potential inducing information.

The function of repetition in advertising can never be exaggerated. It serves to highlight the content, features and strengths that the text stresses. Texts with rhyme or repetition read rhythmically, euphoniously and consonantly and it is easy to remember them. Advertising frequently employs the rhetoric of repetition to highlight the information of products or services via the repetition of the same or similar sounds, words or structures, intensifying the semantic meaning and facilitating the receivers' memory.

2.2. Objectives

Advertising has traditionally communicated messages with strong local and national identities to consumers. However, with the increasingly quickening development of globalization, advertising is also bearing more internationalized characteristics. It is advisable and helpful to understand the universal nature of advertising translation and to distinguish the differences between diverse languages by analyzing the relationship of the matching advertisements in different languages, which will facilitate the advertising translation.

As well as having different structural rules about how texts work, different cultures bring different attitudes and values to the reading of any text (Goddard 1998). Ads are highly condensed artistic discourse. In translating advertisements, translators are confronted with myriad problems in that it involves linguistic and cultural differences ranging from formal aesthetics to advertising strategy. A headache problem for translators is how to deal with rhyming. In order to address this question, the present study seeks help from the matching repetition pairs to find the patterns, which will facilitate finding of the answer.

Different from literary translation (for enjoyment and appreciation), translation of advertisements must serve the function of attracting receivers' attention at the utmost degree, leaving them a best impression, making them remember these products or services with least effort to achieve the best persuasive effect and to induce their purchasing impetus as the final result.

The present study adopts McQuarrie and Mick's (1996) taxonomy and Leigh's (1994) classification to distinguish between different types of rhetorical figures, with the focus on sub-categories of repetition (see further discussion below). In light of prior research, the current study attempts to investigate the repetition phenomenon in print advertisements found in English and Chinese magazines in Singapore context. The researchers hope to find out the respective repetition patterns used in these English and Chinese ads, the similarities and differences between the two languages in repetition use, and to explore the properties in advertising translation in terms of rhetorical repetition.

Singapore is a typical multilingual nation and a rare country in which Chinese and English are the main streams among the four national languages (English, Chinese, Malay and Tamil). That is why most producers, businesspersons and advertising agencies choose to use these two languages in their advertisements to appeal to the maximum potential consumers. In view of the factors above, Singapore is seen as a perfect nation to collect and analyze the characteristics of advertising translation between English and Chinese.

In brief, the main goal of this Singapore advertising parallel corpus is to provide a large general purpose resource for advertising and translation researchers. More specifically, we attempt to achieve the following objectives:

- To build a bilingual multilayered parallel corpus of advertisements;
- To provide empirical modelling of parallel advertising patterns;
- To inspire and facilitate corpus-based investigation of linguistic variation within or across cultures in advertising;
- To explore the features of bidirectional translation of English and Chinese advertisements in Singapore, a bilingual / multilingual context.

The focus of the present research is rhetorical repetition, one of the several layers in this parallel corpus. Comparison will be conducted with regard to the distinct repetition features in English and Chinese advertisements in Singapore.

3. Data and methodology

3.1. Data source

To achieve the goal of comparing the linguistic features and patterns of the matching ads published in the two newspapers and exploring their intrinsic relationship to facilitate bidirectional translation, this parallel corpus is compiled using data collected from the two most influential newspapers in Singapore, namely the *Straits Times* (in English) and *Lianhe Zaobao* (联合早报 'United Morning News', in Chinese). Both of the two newspapers are published daily and have a large readership. This corpus is multilayered which includes an appraisal subcorpus, a repetition subcorpus, and will be extended to several other layers at later stages. The compilation of this corpus will help to find the peculiar advertising language features and the strategies which are used to attract a large multilingual base, and from which the mechanism concerning bidirectional advertising translation from English into Chinese or vice versa, and the similarities and differences in language use in English and Chinese advertisements in Singapore will be explored and analyzed.

This corpus aims to cover the matching pairs of advertisements published in Singapore from January 2008 through July 2008; and two months of newspapers, with 125 pairs of ads altogether, have been processed up to now. The present chapter is mainly concerned with rhetorical repetition use as mirrored in the annotated data in the parallel corpus.

3.2. Tools for this study and word processing

Three software packages are used to serve different functions in the present study. ICTCLAS Chinese preprocessing software is used for word segmentation and POS-tagging of Chinese texts; Wmatrix is used for POS-tagging of English texts and the MMAX2 Toolkit is used for alignment, annotation and query.

3.3. Corpus size

Some English sentences do not match Chinese ones in language form or in meaning, and they do not form aligned pairs, so we do not get the same numbers of sentences (see Table 20-1).

Table 20-1. Corpus size

Corpus	Number of words	Number of sentences
English	20,221	1,733
Chinese	18,499 (28,437 Chinese characters)	1,598

4. Feature selection

Advertisements are decorated as rhythmical as songs and as metric as poems, not only beautiful to the eyes, but also to the ears and pleasant to the mind. Theoretically, it is best to reproduce the Sound beauty, Form beauty and Beauty in artistic conception of the original text. However, aesthetically, it is hard to embody the beauty in sound, form and artistic conception all the time, because different languages adopt different linguistic devices to convey these beauties. When it is hard to reconcile them from each other, it is advised that the last should be sacrificed first (Qiao 2006). That is, Sound beauty and Form beauty are two factors that count most in bidirectional translation.

Rhetorical repetition, the main manifesting operation of Sound beauty and Form beauty, is one of the main strategies in advertising composition, and successful rhyming and repetition promotes the attraction, charm and power of advertisements, and increases receivers' cognition of them.

Rhyming and repetition advertisements are always the concentration of the sound, rhythm, meaning and culture of the product, service or concept it disseminates, that is, one can see its spirit from its details, and that is why it is rather difficult to embody its sound, rhythm, meaning and culture in the translated text. Therefore, translation of advertisements is supposed to choose appropriate expression, based on the content and features of the original advertisement, not only taking account of the stylistic characteristic of the original text, but also conveying the most of the original information.

According to McQuarrie and Mick's (1996) taxonomy, repetition is categorized into three main types: sound, word and structural phrase, under each of which there will be one or more subcategories (see Table 20-2). McQuarrie and Mick maintain that the rhetorical operation of repetition combines multiple instances of some element of the expression without changing the meaning of that element. In advertising we find repetition applied to sounds so as to create the figures of rhyme, alliteration, and assonance or consonance. Repetition applied to words creates the figures known as anaphora (beginning words), epistrophe (ending words), epanalepsis (beginning and ending), and anadiplosis (ending and beginning). Repetition applied to phrasal structure yields the figure of parison.

This chapter develops a framework for classifying repetition types that distinguishes between the texts of Chinese advertisements and those of English ones, and among the 12 rhetorical operations of repetition that underlie individual scheme figures (sound, word and phrasal structure) (McQuarrie and Mick 1996, Leigh 1994).

4.1. Categories explained

4.1.1. Assonance and Alliteration

It is widely believed that patterns of sound repetition such as Rhyme and Alliteration are particularly noticeable and memorable (Cook 2000) and this strategy is widely used in advertising. Chinese Rhyme and English Rhyme are quite similar in features, and what needs more attention is Assonance and Alliteration, which are two very important sound repetition forms in both languages.

Chinese *Shuangsheng* 'alliteration' and Vowel Rhyme, and English Alliteration and Assonance, as effective means to enable a language to be musical and rhythmical, are both important factors to produce musical beauty. Appropriate application of them can promote linguistic infecting power, and thus enabling the text to bear acoustic beauty in rhythm. Therefore, it is advisable to transmit this musical beauty in translating. Because both Alliteration and Shuangsheng are concerned with the repetition of words or characters with the same initial consonants, and both

Assonance and Vowel Rhyme deal with words or characters with the same vowels, we define and process the repetition types in a flexible way to cater to the situation so that they are applicable and comparable. In this study, Chinese Shuangsheng and Vowel Rhyme are also categorized into Alliteration and Assonance.

English Alliteration and Assonance, however, are composed of words while Chinese Shuangsheng and Vowel Rhyme comprise characters, that is, a Chinese character is only equivalent to an English syllable. In light of the fact that each Chinese Shuangsheng covers two characters which share the same initial consonants, such as 坎坷 (kǎnkě), and these two characters are intrinsically linked not only because of word segmentation but because of their natural attributes, so it is with Vowel Rhyme, for a Vowel Rhyme generally consists of two characters with similar vowels, for example, 轻盈 (qīngyíng). One word of Shuangsheng or Vowel Rhyme is counted as one pair of Alliteration or Assonance per se in this corpus to make comparison with English Alliteration and Assonance in the English ads, and so it is with Word Repetition because there is no Link-type between each two "repeated" characters, for example, 潺潺 (chánchán) is segmented as one word but it also covers two "repeated" characters and is counted as two "Word Repetition" instances.

4.1.2. Consonance

Consonance is exclusive of English because Chinese words are composed of one vowel or one consonant plus a vowel, while an English word can end with, besides vowels, one or more consonants which results in the repetition of consonance.

4.2. Taxonomy of rhetorical repetition

Based on McQuarrie's (1993, 1996) and Leigh's (1994) taxonomy and definitions of figures of speech, the present study respectively classifies and defines the specific items of rhetorical repetition. Detailed descriptions of the repetition types are shown in Table 20-2.

Table 20-2. Taxonomy of rhetorical repetition

Repetition elements		Brief descriptions	Text		Ad source
			English	Chinese	
Sounds	Rhyme	Repetition of identical or similar sounds in two or more different words in the same phrase or between a word in one phrase and one in the next	Our engineering has sent vehicles soaring since the very beginning.	一个人嗜赌，身边人受苦。	Ads_00046E[1] Ads_00082C
			1. THEY BREED, YOU BLEED. 2. FlexiMortgage is a competitively-priced home loan for stability. It's also a low-interest overdraft for financial flexibility.	1. 鼠年添色彩，喜气自然来！ 2. 好运庆丰年　健康财富跟着来	Ads_00084E Ads_00089E Ads_00030C Ads_00010C
	Consonance	Similarity between consonants, but not between vowels, as between the *s* and *t* sounds in *sweet silent thought*.	My style varies from the east to the west.		Ads_00076E
			Aerospace-Defense-Security		Ads_00072E
	Assonance	Repetition of similar vowel sounds preceded or followed by different consonants	Health & Wellness TODAY	观赏专业大厨烹制美味风暴。	Ads_00103E Ads_00103C
			Because the good life shouldn't cost a lot.	长12公里的球道蜿蜒曲折，环绕着整个市	Ads_00037E Ads_00112C

				镇，让住户每天在回家的路上都能享受风光美景。	
	Alliteration	Repetition of initial or medial consonants in two or more adjacent words (Leigh 1994)	this sedan is offered with enticing installment subsidies and substantial financial savings.	展望未来，何不让崭新优雅的宝马520i 豪华轿车与您同步驰骋，一路长虹？	Ads_00013E Ads_00013C
			Gifts of Prosperity Greetings of Wealth	地势绵延起伏，风光明媚，宛如一幅群山环抱，流水潺潺的风景画。	Ads_00027E Ads_00112C
Words	Anaphora	Repetition of words at the beginning of phrases (McQuarrie and Mick 1996); Repetition of a word or phrase at the start of successive phrases (usually for emphasis)	Share the moments, Share the minutes. (Anaphora + Rhyme + Consonance)	1.让爱升温，让爱起飞!——Valentine’s Day 2. 新年新春新气象。	Ads_00066E Ads_00025C Ads_00013C
			I'M IN A GOOD HEALTH, GOOD FORTUNE, GOOD LOOKS, KIND OF MOOD.	我报给你更多，现增添50%英文版，给你不同的内容，不一样的视野。	Ads_00042E Ads_00036C
	Epistrophe	Also known as *epiphora*, the counterpart of anaphora. It is the repetition of	A place for your family to call home, any time, all the time.	吉祥鼠年，该是丰收的一年！ 源自灵感，激发灵感，甚至就是灵感本身。	Ads_00078E Ads_00030C Ads_00085C

		the same word or words at the end of successive phrases, clauses or sentences.	BUY IT. SELL IT. FIND IT	额外红利 有效年利	Ads_00025E Ads_00004C
	Epanalepsis	Repetition of a word toward the beginning and end of a phrase	Smart phone smarts. (McQuarrie and Mick 1996)	1. 财神送财喜迎新春 2. 花样年华新达城	Ads_00110C Ads_00110C
	Anadiplosis	Repetition of the last word or any prominent word in a sentence or clause, at the beginning of the next, with an adjunct idea	1. Kleenex Ultra. Ultra softness Is all you feel. (McQuarrie and Mick 1996) 2. He gave his life; life was all he could give. (Wikipedia[2])	让活力注入生命中的每一天 vivolife 充实度过每一天　每一天活得精彩，才是生活的真谛。	Ads_00090C
	Antimetabole	Repetition of words in successive clauses in reverse grammatical order (Wikipedia)	I know what I like, and like what I know. (Wikipedia)	上海自来水来自海上. 黄山落叶松叶落山黄 (Chinese antithetical couplet)	
	Word Repetition	The repeating of a word or phrase for emphasis or rhetorical effect	1. A residential haven situated on acres upon acres of pristine, undulating terrain. 2. Once you set sail on our ships, you'll understand	1. 到飞禽公园迎春接福，可爱鸟儿的璀璨色彩，随着艳阳向您洒下，象征您一整年喜气洋洋、多姿多彩！	Ads_00112E Ads_00119E Ads_00030C Ads_00103C

			why you'll want to do it again... and again.	2. 囊中满满？何不来选择多多的 John Little 新年展销会！	
	Polyptoton	Repetition of words derived from the same root	1. Inspired, inspiring and, for many, even inspiration itself. 2. Be rewarded with the attractive rates and a Maturity Bonus when your fixed deposit matures.	快来迎接财神的财运祝福吧！	Ads_00085E Ads_00034E Ads_00110C
Phrase structure	Parison	Marked parallelism between successive phrases; often involving the use of one of more embedded repeated words	The more energy efficient an appliance is, the less electricity it consumes and hence, a lower energy bill every month.	电器的能源效率越高，耗电越低，每个月的能源消耗账单就会更低。	Ads00107 E&C

5. Distribution of repetition categories

Because the two subcorpora of English and Chinese advertisements are composed of matching pairs of ads, which deal with the same content, nearly with the same pairs of sentences, it can be assumed that they are equivalent in volume and comparable to each other, and it is unnecessary to normalize the data in terms of repetition frequencies in the two corpora.

5.1. Repetition distribution across categorized ads

Table 20-3 shows the distribution pattern of repetition across categorized ads.

Table 20-3. Repetition distribution among categorized ads

Categorized items	No. of ad pairs	Repetition frequency		
		Chinese	English	Total
Investment	22	120	90	210
Shopping	13	68	59	127
Housing	13	71	51	122
Show & Entertainment	12	68	45	113
Public Affairs	12	30	30	60
Car	12	58	45	103
Media & Classified	9	59	48	107
Food	5	44	12	56
Telecom & Appliance	5	7	21	28
Medicine & Health	5	10	10	20
Education & Career	4	27	20	47
Travel	4	33	12	45
Airlines	3	12	15	27
Religion & Apology	2	12	5	17
Furniture	2	11	17	28
Animal	2	30	7	37
TOTAL	125	660	487	1147

5.2. Frequency description of all repetition types

In terms of the repetition distribution in the parallel corpus, the separate frequencies of all repetition types are counted to display how they are respectively used in the English and Chinese advertisements as illustared in Table 20-4 and Figure 20-1.

Table 20-4. Frequency description of all repetition categories

	Rhyme	Consonance	Assonance	Alliteration	Anaphora	Epistrophe	Epanalepsis	Anadiplosis	Antimetabole	Repetition	Polyptoton	Parison
Chin.	80	0	214	130	84	50	5	9	2	78	0	8
Eng.	71	14	59	239	58	11	0	0	0	3	16	16
Total	151	14	273	369	142	61	5	9	2	81	16	24

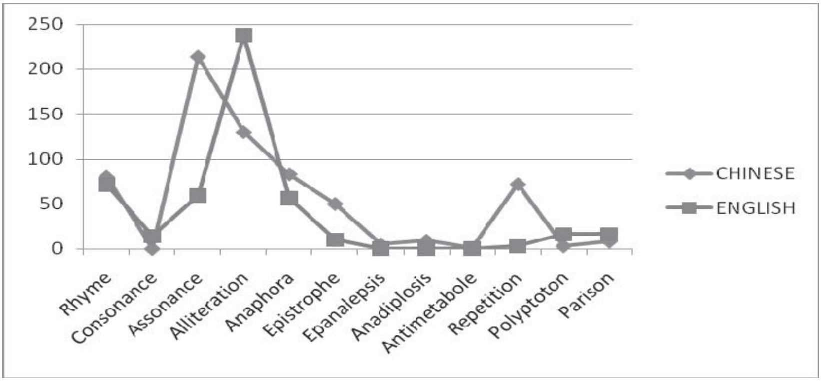


Figure 20- 1. Repetition frequency in English and Chinese advertisements

5.3. Similarities and differences

Based on the data retrieved from the corpus, the repetition use is generalized into three patterns: similarities, differences, and the exclusive type of English rhetoric, consonance.

Consonance is exclusive of English as stated earlier, so it is natural that the frequency of consonance in Chinese advertisements is zero.

5.3.1. Similarities: Rhyme, Anaphora

As can be seen from Table 20-4 and Figure 20-1, the frequency of Rhyme and Anaphora is quite similar in English and Chinese advertisements. We can make the assumption that both English and Chinese ads pay equal attention to the beginning and end (most Rhyme items are end rhyme) of a phrase or sentence in word choice.

5.3.2. Differences: Alliteration, Assonance, Epistrophe, Word repetition, Polyptoton, Parison

Based on the frequency result, we can see that there is a striking difference in the use of Alliteration, Assonance, Epistrophe, Polyptoton, Word repetition and Parison between English and Chinese ads. English ads employ Alliteration, Polyptoton and Parison more frequently to impress consumers while Chinese ads tend to use Assonance, Epistrope and Word repetition to attract the recipients' attention.

5.4. Frequency order

The linguistic analysis reveals that the top four frequently used English repetition types are Alliteration, Rhyme, Assonance and Anaphora while the top four frequently used Chinese repetition expressions are Assonance, Alliteration, Anaphora and Rhyme. Although the four items are the same, they rank differently in frequency order. In general, English ads use Alliteration and Rhyme more often than their Chinese counterparts while Chinese ads utilize Assonance and Anaphora more frequently than their English counterparts.

Tables 20-5 and 20-6 show the distribution of repetition of English and Chinese ads. As can be seen, in the English ads, alliteration accounts for half of all the English repetition expressions used, which manifests the importance of Alliteration in English advertising, whereas Chinese Shuangsheng (Alliteration) occupies 20% of all repetition occurrences in Chinese ads. Alliteration is often used to add to the colourfulness and attraction of an ad in English expressions. We can assume that the status of Chinese Alliteration and Assonance (Shuangsheng and Vowel Rhyme) is unparalleled to that of English Alliteration and Assonance. For one thing, Chinese Shuangsheng and Vowel Rhyme are stereotyped and the number is limited, while the formation of English Alliteration and Assonance is flexible and there are much more English words with the same initial consonants or medial vowels which can form countless Alliteration and Assonance expressions. For another, Alliteration originates from poetry and has a longer history than End rhyme. It used to take a more prominent position in English poetry. In spite of its decline in status, Alliteration still plays a very important role in all kinds of genres in English, and advertising in Singapore context is not an exception.

Table 20-5. Repetition frequency of English ads

Category	Frequency	Percentage
Alliteration	239	50%
Rhyme	71	15%
Assonance	59	12%
Anaphora	58	12%
Parison	16	3%
Polyptoton	16	3%
Consonance	14	3%
Epistrophe	11	2%
Repetition	3	1%
Anadiplosis	0	0%
Epanalepsis	0	0%
Antimetabole	0	0%
Total	487	100%

Table 20-6. Repetition frequency of Chinese ads

Word	Frequency	Percentage
Assonance	214	32%
Alliteration	130	20%
Anaphora	84	13%
Rhyme	80	12%
Repetition	78	12%
Epistrophe	50	8%
Anadiplosis	9	1%
Parison	8	1%
Epanalepsis	5	1%
Polyptoton	0	0%
Antimetabole	2	0%
Consonance	0	0%
Total	660	100%

5.5. Co-occurrences of several repetition types and their translations

Like poems, English advertisements often utilize many devices to be effective and successful. Three related terms referring to sound repetition, alliteration, assonance, and consonance, are often confused for one another in ads as in poems, or used in place of each other. Though they are related, the use and effect are quite different.

The ad for Model BMW320i (Ads_00013C & E) is a good case in point, in which all three of these repetition types appear in one line, "this sedan is offered with enticing installment subsidies and substantial financial savings." This line clearly contains all three, and can show the difference between assonance, consonance and alliteration. Alliteration is the repetition of the sound of /s/ in "sedan, installment, subsidies, substantial, savings"; "Substantial and financial" are also rhyme and "Subsidies and substantial" are alliteration + assonance, which can never occur in their Chinese counterparts. These repetition types are very closely related in English, though the distinction between them comes in determining vowels versus consonants, and then placement within the words. However, the bidirectional translation of ads with repetition types is demanding and challenging because it is nearly impossible to find the matching types in Chinese, and therefore the corresponding Chinese version of the above ad with complex repetition categories used only plain expression to convey the message as 我们还为您准备了超值分期付款及优惠配 (wǒmen hái wèi nín zhǔnbèi le chāozhí fēnqífùkuǎn jí yōuhuì pèi), which includes no repetition types at all.

5.6. Statistical significance

5.6.1. Significance test

To verify the validity of the hypothesis that there is significant difference in the use of rhetorical repetition between Singapore Chinese Ad corpus and Singapore English Ad Corpus, the frequencies shown in Table 20-3 are cross-tabulated (see Tables 20-7 and 20-8). Considering the factor that Consonance and Polyptoton are exclusive of English, it is better to leave out the data of these two pairs before operating the significance test. The results indicate that the calculated log-likelihood (LL) ratio is 231.850 for 9 degrees of freedom, and the 2-sided significance level (0.000) is less than 0.001.

Table 20- 5. Corpus * Type cross-tabulation

			Type										Total
			Rhyme	Repetition	Parison	Assonance	Alliteration	Anaphora	Epistrophe	Epanalepsis	Anadiplosis	Antimetabole	Rhyme
Cor-pus	English Ad Corpus	Count	71	3	16	59	239	58	11	0	0	0	457
		Expected Count	61.8	33.2	9.8	111.8	151.1	58.1	25.0	2.0	3.7	.4	457.0
		Residual	9.2	-30.2	6.2	-52.8	87.9	-.1	-14.0	-2.0	-3.7	-.4	
	Chinese Ad Corpus	Count	80	78	8	214	130	84	50	5	9	1	659
		Expected Count	89.2	47.8	14.2	161.2	217.9	83.9	36.0	3.0	5.3	.6	659.0
		Residual	-9.2	30.2	-6.2	52.8	-87.9	.1	14.0	2.0	3.7	.4	
Total		Count	151	81		273	369	142	61	5	9	1	1116
		Expected Count	151.0	81.0	.0	273.0	369.0	142.0	61.0	5.0	9.0	1.0	1116.0

Table 20- 6. Chi-square test

	Value	d.f.	Asymp. Sig. (2-sided)
Pearson Chi-Square	207.789(a)	9	.000
Likelihood Ratio	231.850	9	.000
N of Valid Cases	1116		

a. 5 cells (25.0%) have expected count less than 5. The minimum expected count is .41.

If we consult the distribution table in a reference book for statistics, we will find the critical value for statistical significance at p<0.001 is 27.88 with 9 d.f. The calculated LL score is considerably greater than this critical value (see Table 20-8). Therefore, we can be more than 99.9% confident that the difference in the frequencies of rhetorical repetition in Singapore Chinese Ad Corpus and Singapore English Ad Corpus is statistically significant.

5.6.2. Highly figurative

In order to determine the extent to which rhetorical repetition is utilized in English and Chinese advertising in Singapore, the total number of running words / characters and related rhetorically repeated words / characters are counted in both English and Chinese ad corpora.

Table 20- 7. Percentage of rhetorically repeated words / characters

Corpus	**Total no. of words / characters**	**Rhetorically repeated words / characters**	**Percentage**
English	20,221 words	602	2.98%
Chinese	28,437 characters	1,086	3.82%

The high percentages of rhetorical repetition in Table 20-9 show that both English and Chinese ads in this parallel corpus are highly figurative in terms of the use of rhetorical repetition, respectively reaching 2.98% and 3.82%. That is a good indication that advertisers, manufacturers and businesspeople attach much importance to the use of rhetorical repetition in their promotion or sale activities of their products and services.

5.7. Discussion

In spite of statistical difference in the way that the two languages produce advertising messages with rhetorical repetition, successful bidirectional translation of ads is possible between English and Chinese. One cannot simply accredit the failure to match the two languages to ‘untranslatability’. In fact, it is not necessary to translate the original text in a literal way, or to translate it using the same rhetorical device, which is nearly impossible. Qualified translators know that the essence of good advertising translation is not about simply translating the words; it is about encoding the right concepts and those concepts may well vary from culture to culture. This process is called copy ADAPTION – adapting the text to fit the culture of its targeted group (Goddard 1998).

In practical translation, the rhetoric adopted in the original text can be dealt with in a circuitous way, not word-to-word, not necessarily in the same rhetoric form, but with similar artistic form and effect. This kind of translation is also viewed as a successful one. For example, in view of the deviation in rhyming between English and Chinese, it is extremely hard to operate literal translation with the original rhetorical rhyming from Chinese Shuangsheng and Vowel Rhyme to English Alliteration and Assonance, and therefore, the bidirectional translating does not need to be restricted to the literal form, but can make use of the methods of correspondence, supplementation, transformation or ellipsis and the like, based on the context and situation, to reproduce the musical beauty of both English and Chinese rhetorical language, and to preserve the linguistic influence of the original text (Yang and Wu 2006). Because both English Alliteration and Chinese Shuangsheng are concerned with the repetition of words or characters with the same initial consonants, and both English Assonance and Chinese Vowel Rhyme deal with words or characters with the same vowels, it is possible to convey the acoustic effect in an integral way in translating English into Chinese or vice versa, for example, translating texts with Alliteration or Assonance in a flexible way by not restricting to the specific location of Alliteration or Assonance but turning to the sentence it lies in (Yang and Wu 2006). For example, 好运庆丰年 / 健康财富跟着来 (hǎoyùn qìng fēngnián / jiànkāng cáifù gēnzhe lái) can be translated as “Gifts of Prosperity / Greetings of Wealth”. In this example, the sound repetition of Chinese vowel rhyme of 年 and 来 at the end of two utterances is successfully converted into another type of sound repetition, Alliteration of ‘Gifts’ and ‘Greetings’ at the beginning of the corresponding English translations. Both expressions convey the message of warm greetings, in different repetition forms at different positions but with similar sound and rhythm beauty.

6. Conclusion and future research

Most studies of rhetoric reject the view that figurative language is simply an attractive ornament that can be added as an afterthought to a basic message that does the real work of communication (Eco 1976). Rather, the preferred view has been that a rhetorical figure both communicates differently, and communicates more, than a straightforward delivery of the message would have (Scott 1990, McQuarrie 1993).

Rhetorical repetition of sounds, words or phrasal structures makes texts read or sound like rhythmical, euphonious, consonant as music or poems, and easy to remember. Figures of repetition may facilitate both comprehension and recall (McQuarrie 1993). Such figures may facilitate comprehension by indicating to the reader that the repeated term is a key word around which the message is organized. Similarly, repetition, with the stimuli of repeated sound, word and structure, is a key factor in learning and memory, hence, the potential value of a scheme that repeats a key term in a prominent way. We might say that these are figures that seduce or invite readers into co-constructing the advertiser's message. Therefore, this rhetorical art has been universally applied to advertisements whose main target is to impress the recipients deeply, to attract their attention and persuade them to buy the advertised goods or services.

The results from this parallel ad corpus support the idea that contemporary advertisements are highly figurative, that there is a significant difference in the use of rhetorical repetition between English and Chinese ads in Singapore context, and that Chinese advertisements utilize more repetition expressions than their English counterparts.

We hope to have opened the reader's eyes to the rich resources offered by the discipline of rhetoric, with the focus on repetition. Once consumer researchers acknowledge the profusion of figures and their features and patterns in contemporary advertising in different languages, bidirectional / multidirectional translating and advertising theory building and testing can advance.

In view of the limited size of processed data in the current corpus, categorized analysis regarding rhetorical repetition used in tourism, investment, shopping, etc. is not taken as planned. A categorized and stratified corpus will be more meaningful in analyzing and identifying the patterns across strata and this will be done in the future research and more features are to be investigated if possible. Besides, further research needs exploring with regard to comparison of repetition use between this advertising parallel corpus and another one in a monolingual context, and comparison of bidirectional translation between advertising and literary works.

Notes

* The first author would like to extend her gratitude to her friends Hui Chenri, Ao Ran, Zheng Jianzhen and Wu Pengcheng who contributed help to the accomplishment of this study.

1. The pairs of ads in the corpus are respectively numbered from Ads_00001E and Ads_00001C in the advertising parallel corpus. E stands for English ads while C represents Chinese ones.

2. See the official website of Wikipedia (http://zh.wikipedia.org/wiki/Wikipedia).

References

Ahmed, N. (2000), *Cross-cultural Content Analysis of Advertising from the United States and India*. Dissertation.com.

Bolls, P. (2006), “It’s just your imagination: The effect of imagery on recognition of product- versus non-product-related information in radio advertisements. *Journal of Radio Studies* 13(2): 201-213.

Bulmer, S., and Buchanan-Oliver, M. (2006), “Visual rhetoric and global advertising imagery”. *Journal of Marketing Communications* 12(1): 49-61.

Chen, C. W. (2006), “The mixing of English in magazine advertisements in Taiwan”. *World Englishes* 25(3-4): 467- 478.

Cook, G. (2000), *Language Play, Language Learning*. Oxford: Oxford University Press.

Cooper, L. (1960), *The Rhetoric of Aristotle*. New York: Appleton-Century-Crofts, Inc.

Corbett, E. (1990), *Classical Rhetoric for the Modern Student*. New York: Oxford University Press.

Eco, U. (1976), *A Theory of Semiotics*. Bloomington: Indiana University Press.

Garfield, B. (1999), “In the crucible of racial politics, Denny’s should stick to waffles”. *Advertising Age* 70(3): 49-49.

Goddard, A. (1998), *The Language of Advertising*. London: Routledge.

Hobbs, P. (2007), “Miracles of love: The use of metaphor in egg donor ads”. *Journal of Sociolinguistics* 11(1): 24-52.

Hong, H. Q. (2005), “SCoRE: A multimodal corpus database of education discourse”, in *Proceedings of Corpus Linguistics 2005*. Birmingham, July 14-17, 2005.

Leigh, J. (1994), “The use of figures of speech in print ad headlines”. *Journal of Advertising* 23(2): 17-33.

Lindstromberg, S., and Boers, F. (2008), “Phonemic repetition and the learning of lexical chunks: The power of assonance”. *System* 36(3): 423-436.

Ma, D. (2004), “The art of alliteration and assonance in Du Fu’s poems”. *Journal of Sichuan University* (Social Science Edition) 135(6): 74-78

McQuarrie, E. F. (1993), “Special topic session: The new advertising rhetoric”. *Advances in Consumer Research* 20(1): 308-308.

—. (2004), “Integration of construct and external validity by means of proximal similarity: Implications for laboratory experiments in marketing”. *Journal of Business Research* 57: 142-153.

McQuarrie, E. F. and Mick, D. G. (1992), “On resonance: A critical pluralistic inquiry into advertising rhetoric”. *Journal of Consumer Research* 19(2): 180-197.

McQuarrie, E. F. and Mick, D. G. (1993), “Reflections on classical rhetoric and the incidence of figures of speech in contemporary magazine advertisements”. *Advances in Consumer Research* 20: 309-313.

McQuarrie, E. F. and Mick, D. G. (1996), “Figures of rhetoric in advertising language”. *Journal of Consumer Research* 22(4): 4-438.

McQuarrie, E. F. and Mick, D. G. (1999), “Visual rhetoric in advertising: Text-interpretive, experimental, and reader-response analyses”. *Journal of Consumer Research* 26(1): 37-54.

McQuarrie, E. F. and Mick, D. G. (2003), “The contribution of semiotic and rhetorical perspectives to the explanation of visual persuasion in advertising”, in L. M. Scott and R. Batra (eds.) *Persuasive Imagery: A Consumer Response Perspective*, 191–221. Mahwah, NJ: Lawrence Erlbaum Publishers.

McQuarrie, E. F. and Mick, D. G. (2003), “Visual and verbal rhetorical figures under directed processing versus incidental exposure to advertising”. *Journal of Consumer Research* 29: 579-587.

McQuarrie, E. F. and Phillips, B. J. (2005), “Indirect persuasion in advertising: How consumers process metaphors presented in pictures and words”. *Journal of Advertising* 34: 7-20.

McQueen, D. (1998), *Television: A Media Student’s Guide*. London: Arnold.

Morgan, S. and Reichert, T. (1999), “The message is in the metaphor: Assessing the comprehension of metaphors in advertisements”. *Journal of Advertising* 28(4): 1-12.

Pan, W. and Mao, R. (2006), “Experiencing the beauty of the sounds and ornamenting the spirit of the translation”. *Journal of Central South University* (Social Science Edition) 12(1): 116-120.

Pawlowski, D., Badzinski, D. and Mitchell, N. (1998), “Effects of metaphors on children’s comprehension and perception of print advertisements”. *Journal of Advertising* 27(2): 83-98.

Phillips, B. J. and McQuarrie, E. F. (2002), “The development, change, and transformation of rhetorical style in magazine advertisements 1954-1999”. *Journal of Advertising* 31(4): 1-13.

Phillips, B. J. and McQuarrie, E. F. (2004), “Beyond visual metaphor: A new typology of visual rhetoric in advertising”. *Marketing Theory* 4: 13-136.

Qiao, J. (2006), “Aesthetics that China’s ancient poem translates from Chinese to English is explored”. *Journal of Hubei Broadcasting and TV University* 23(6): 154-155.

Renov, M. (1989), “Advertising / photojournalism / cinema: The shifting rhetoric of forties female representation”. *Quarterly Review of Film & Video* 11(1): 1-21.

Scanlan, R. (1954), “Advertising, rhetoric, and public opinion”. *Today’s Speech* 2(2): 9-11.

Scott, L. (1994), “Images in advertising: The need for a theory of visual rhetoric”. *Journal of Consumer Research* 21(2): 252-273.

Stern, B. (1988), “How does an ad mean? Language in services advertising”. *Journal of Advertising* 17(2): 3-14.

Toncar, M. and Munch, J. (2001), “Consumer responses to tropes in print advertising”. *Journal of Advertising* 30(1): 55-65.

Welford, W. (1992), “Supermarket semantics: The rhetoric of food labelling and advertising”. *ETC: A Review of General Semantics* 49(1): 3-17.

Yang, Z. S. and Wu, X. M. (2006), “Musical beauty and its reproduction - Chinese Shuangsheng, vowel rhyme, and English alliteration and assonance, and their translation”. *Journal of Nanjing Institute of Technology* (Social Science Edition) 6 (2):10-13.

Zeng, F. (2005), “The key factor of marketing”. *Market Modernization* 449(11): 74-75.

CHAPTER TWENTY-ONE

A CONTRASTIVE STUDY OF COMPARATIVE CONSTRUCTIONS IN ENGLISH, JAPANESE, AND TOK PISIN: USING CORPORA IN CROSS-LINGUISTIC CONTRAST

MASAHIKO NOSE

1. Introduction

When we describe a quality, quantity, or manner of one thing (or person), we contrast it with others. Comparative constructions are introduced to express such an evaluation between the two things (or persons).[1] This chapter deals with comparative constructions in English, Japanese, and Tok Pisin using corpora in cross-linguistic contrast.

First, we need to summarize the functions of comparative constructions. A comparative construction is a kind of adverbial construction, and it needs at least two participants, i.e., the comparee and the standard. The comparee is the thing (or person) we are describing, and the standard indicates the thing (or person) that we choose to contrast with the comparee. By comparing the two entities, we can express whether the comparee is big or small, quick or slow, much or little, etc. This study is an attempt to clarify the characteristics of comparative constructions using parallel texts in English, Japanese, and Tok Pisin. Tok Pisin is a creole language spoken in Papua New Guinea. Like other creole languages, Tok Pisin has a simple grammar, and its lexicon is mixed with English and other indigenous languages in Melanesia. It is worth observing the comparative constructions in Tok Pisin by comparing them with such constructions in other languages. This study uses Japanese (a typical nominative-accusative language in Eurasia) as a contrast.

Through this contrastive study of typologically distinct languages, by focusing on Tok Pisin, this chapter will make it clear how the comparative element is introduced into grammar.

2. Purpose of this study and parallel text research

Contrasting languages is an approach to explore functional-cognitive characteristics of grammatical morphemes and constructions. There are a large number of previous studies contrasting languages, and we can contrast some languages by using reference grammars, dictionaries, and data elicited from native speakers using questionnaires (cf. special issue of *Sprachtypologie und Universalienforschung* (STUF) 60, 2007). The present contrastive study claims that it is not enough to analyse the comparative constructions only in English or another single language, and Stassen (1985, 2005) has already pointed out that there are some other types of comparative constructions, such as locational, conjoined types, etc. (cf. Nose 2007, Henkelmann 2006). However, Stassen's typological view is insufficient in that he checked mainly descriptive grammars of the languages that he investigated, and it lacks text-based research of the comparatives. Thus, in this study, in terms of typological interest and availability of parallel texts, I have chosen the following three typologically distinct languages: English, Japanese, and Tok Pisin (cf. Cysouw and Wälchli 2007). Tok Pisin is a creole language based on English, and its grammatical features are borrowed or affected from the English grammar. Nevertheless, the grammar of Tok Pisin is still primitive, and more or less reflects human conceptualization (cf. Mihalic 1986).

For contrastive purposes, this chapter uses part of the *New Testament* written in English, Japanese, and Tok Pisin (see also Vries 2007). Briefly, grammatical (word order and possession) and geographical data of each language are shown below.[2]

- English (Indo-European, Europe): SVO, e.g. *John's book*
- Japanese (Independent, East Asia): SOV, e.g. *John-no hon* (possessive marker *no*)
- Tok Pisin (Creole, Papua New Guinea): SVO, e.g. *buk bilong John* (close relationship preposition *bilong*)

2.1. Types of comparative constructions

Stassen (1985, 2005) points out that there are four types of comparative constructions: locational, exceed, conjoined, and particle. Each type is illustrated with examples in (1)-(4) below (cf. Nose 2007).[3]

□ Locational: Korean (East Asia)

(1) Thalo-nun Hanakho-pota khi-ka khu-ta.
Taro-TOP Hanako-STM tallness-FOC big-ending suffix
‘Taro is taller than Hanako’

Locational comparative constructions are dominant in Eurasian languages, and standard marker includes some meanings of locative elements.[4]

□ Exceed: Nguna (Oceania)

(2) Nasuma waia e parua liu nasuma aginau.
house this COP big exceed house my
‘This house is bigger than my house.’

The comparative meaning can be expressed by using the verb meaning ‘exceed, surpass’ in this type. This exceed type is observed in African, South East Asian, and Oceanic languages.

□ Conjoined: Wari’ (South America) (Everett and Kem 1997: 194)

(3) Amon mixem na womu cwa. Om ca mixem
little black cotton-my this not-exist black

homa ca womum
much cotton-your

‘My clothes are more black than yours; my clothes are a little black, your clothes are not so black.’

Conjoined comparative constructions consist of two sentences, one positive and the other negative. This type is primitive and less grammaticalized, and is found in South American and Australian Aboriginal languages (Dimmendaal 2001).

□ Particle: Spanish (Uritani 2002: 91)

(4) Este lapis es más largo que ése.
this pencil COP PAM long STM it
'This pencil is longer than it.'

Finally, the particle type is observed in almost all European languages. The standard marker is expressed by the particle form. English is classified as this type.

2.2. Text-based research of comparative constructions

This study contrasts Tok Pisin with English and Japanese using parallel texts (cf. Vries 2007, Stolz 2007). The material is the *New Testament (new international version)* (*Nupela Testamen* in Tok Pisin), and we study comparative constructions presented in its different versions. Preliminary studies of comparative constructions (Heine 1997; Stassen 1985, 2005; Haspelmath and Buchholz 1988; Nose 2007) claimed that there are five parameters defining the comparative, as shown in (5).

(5) Five parameters of comparative constructions (Haspelmath and Buchholz 1998: 279)

This house is 1) Comparee (COM)	more 2) Parameter marker (PAM)
beautiful 3) Parameter (PAR)	than 4) Standard marker (STM)
that one. 5) Standard (STAN)	

However, comparative constructions do not always need five parameters indicated in the English example in (5): some can be omitted or do not appear at all in some languages. Accordingly, this study will observe behaviours of two of them, COM and STAN. However, we cannot identify such parameters in exceed (6) and conjoined (7) types. The five parameters are not obligatory in the comparative, but it is still worth paying attention to those of COM and STAN.

(6) Tok Pisin (Matai 35)

Em i winim Solomon
it COP surpass Solomon
COM STM? STAN (missing PAM, PAR)

‘It is greater than Solomon.’

(7) Amele (Roberts 1987: 135)

Jo	i	ben	(qa)	jo	eu	ben	ca.
house	this	big	but	house	that	big	add
COM		PAR		STAN		PAR	PAM

(missing STM)
‘This house is big but that house is bigger.’

In this study, the relationship between COM, STM, and STAN are examined, and semantic characteristics of STAN are considered. STAN has a significant role in the comparative, and we will observe which element is chosen in the *New Testament*.

Thus, this study contrasts the comparative constructions in three languages by using the same material. In contrasting the parallel texts in these languages, there might be differences of translations and other difficulties (see also Stolz 2007). Here, we will illustrate advantages of this parallel-text study, and summarize this research method. We can make use of the translation texts in the *New Testament*, and this is better than checking the reference grammar of each language. Tok Pisin is used mainly for everyday speech, and rarely for written texts. The *New Testament* in Tok Pisin is useful for such a contrastive study. On the other hand, the *New Testament* in English is an old written text, and the expressions are rather formal and not a good discourse resource (Vries 2007 also discussed benefits and challenges of using Bible texts). The Japanese translation is also formal, but the translation is a new one in spite of its written style. Then, an examination is made of the forms used to express comparative meanings by extracting them from the multilingual parallel texts.

3. Results

In this section, some examples of comparative constructions are assembled. This study contrasts the data cross-linguistically on the basis of the parallel texts in the three languages. First, we investigate the comparative constructions in the English version. There are 78 examples, excluding the comparative sentences without the standard marker and the standard *than X*. Then, we contrast the English comparatives with the translations in Japanese. The Japanese equivalents are shown in Table 21-1.

Table 21-1. Japanese equivalents of the English comparative

Comparative type	Forms	Numbers of examples
Locational	*yori* (locational STM)	51
Exceed	*masaru* 'surpass', *otoru* 'be inferior to'	8
Lexicalized STM	*ijou* 'more than'	6
PAM only	*issou, motto* 'further, more'	3
Superlative	*mottomo* 'the most'	2
Others	---	8

In Japanese, there is a typical locational standard marker *yori*, and 51 examples of this type of locational comparative were found in our data, as in (8). In addition, some forms are translated as an exceed type, using the verbs *masaru* 'be superior' and *otoru* 'be inferior', as in (9). It is also possible to express a comparative meaning by adding some intensifier (or PAM) forms like *issou* or *motto*. There are some lexical forms and two superlative forms as well (see 10). In Japanese, the locational comparative is a typical usage, although there are some other forms.

(8) Matthew 6:
En (English): But after me will come one who is *more powerful [than] I.*[5]
Jp (Japanese): Watashi-no atokara kuru kata-wa, *watashi-[yori] sugurete* orareru.
Tp (Tok Pisin): Tasol man i kam bihain long mi, *strong bilong em i [winim] strong bilong mi.*

(9) Matthew 35:
En: For she came from the ends of the earth to listen to Solomon's wisdom, and now one *greater [than] Solomon* is here.
Jp: Kono jouou-wa Solomon-no chie-wo kikutameni, chi-no hatekara kitakara dearu. Kokoni, *Solomon-ni [masaru]* mono-ga aru.
Tp: Dispela kwin i stap long arere tru bilong graun, na em i kam harim Solomon i autim gutpela save bilong em. Tasol man i stap hia, *em i [winim] Solomon.*

(10) Corinthians 15:
En: We are to be pitied *more [than] all men.*

Jp: Watashitachi-wa *subete-no hito-no-nakade mottomo mijimena* mono desu.
Tp: Em i olsem liklik tasol *[Mobeta] ol i sori tumas long yumi.*

Next, we contrast the English comparatives with the translations in Tok Pisin. The Tok Pisin equivalents are shown in Table 21-2.

Table 21-2. Tok Pisin equivalents of the English comparatives

Comparative type	Forms	Numbers of examples
Exceed	*winim* 'surpass'	41
Conjoined	*Mobeta* 'more better', *nogut* 'no good'	13
Particle STM	*olsem* (3), *long* (9)	12
PAM only	*moa* 'more', *tumas* 'too much, too many', *inap* 'enough', *tasol* 'only, just'	10
Others	---	10

Tok Pisin prefers an exceed type with the verb *winim* 'surpass'. There are 41 examples of comparative constructions using the verb *winim* (8), although it is possible in Tok Pisin to express English-based particle type of comparatives using *olsem* or *long*, as in (11a, 11b). There are 12 examples of the particle comparative type, 9 instances of STM *long*, and 3 occurrences of STM *olsem*.

(11) Tok Pisin (Mihalic 1986: 41)
a. Em i *moa bik [olsem] mi.*
b. Em i *moa bik [long] mi.*
'He is bigger than I.'
c. Em i go *antap [long] olgeta heven*. (Ephesians 4)
'One who ascended higher than all the heavens.'

It is worth noting that there are 13 sentences of the conjoined type. In (12), the sentence uses *gutpela* in the first part and then the form *nogut* follows in the next part. As a result of the conjoining, a comparative nuance (*em i gutpela* 'to marry is good') occurs.

(12) Corinthians 7
En: For it is *better to marry [than] to burn with passion.*
Jp: *Jouyoku-ni mi-wo kogasu-[yori]-wa, kekkonshita-hou-ga* mashi dakara desu.

Tp: Sapos ol i *marit, em i gutpela. [Nogut]* bel bilong ol i kirap
nabaut olsem paia.

Moreover, there are two means of expressing comparative meaning. As already shown above, one is an English-based comparative, which uses the standard marker *long* or *olsem*, as in (11a, 11c); and the other is a simple sentence with parameter markers (PAM) *moa*, *inap*, and *tasol*, as in (13).

(13) Matthew 10
En: Anyone who loves his father or mother *more [than] me* is
not worthy of me.
Jp: *Watashi-[yori]* chichi ya haha-wo aisuru-mono-wa
watashi-ni fusawakushiku-nai
Tp: Man i laikim papa no mama bilong en na i *no laikim mi*
moa yet, em i no inap i stap pren bilong mi.

In (13), the comparative constructions in English and Japanese are expressed with STM *than* in English and *yori* in Japanese. Tok Pisin, however, does not use a canonical comparative, but only uses a parameter marker *moa*.

Finally, this study investigates which element takes the standard position (STAN) in the comparative constructions of the Bible. We observed some STAN entities from the comparative constructions, which are classified into the following four types (see 14).

(14) Classification of the status of the standard in the comparative:

a. Noun, person (including God and angels):
27
b. Noun, thing (animate 4, inanimate 16, abstract 5):
25
c. Noun phrase, infinitive:
17
d. Demonstrative:
9

Basically, STAN is used to compare with the comparee (COM), and stereotypical people or things seem to be used as STAN. Indeed, the use of a person or an animate entity is frequent, but abstract thing(s) are uncommon. The Bible texts are written and regarded as a means of education on religious topics. For this purpose, some comparatives are expressed with STAN of noun phrase or infinitive. Such grammatically complex STAN is characteristic of the Bible texts.

4. Discussion

In this section, the following three points are discussed in terms of contrastive and functional perspectives. The first point is that this study examined the *New Testament* in English, Japanese, and Tok Pisin, and used translations. We have shown some advantages and disadvantages in this cross-linguistic study. There is rarely a good written resource in Tok Pisin, and the *New Testament* is one of the good resources which is also available in English and Japanese (there is no *Harry Potter* or *Le Petit Prince* in Tok Pisin, cf. Stolz 2007). Nevertheless, the *New Testament* texts in the three languages are written texts and include formal usages. Furthermore, we could not find many examples of the comparative; only 78 sentences in more than 700 pages. In the small number of comparative constructions, the comparative forms between English and Tok Pisin turned out to be different from each other, and Tok Pisin did not copy the English grammar despite earlier contact between the two languages. Japanese, on the other hand, has provided another point of view in this study. The locational comparative type is dominant in Japanese, but we have identified the exceed type usage as well.

The second point is that we considered the grammatical status of the standard (STAN) and the standard marker (STM). According to Stassen (1985, 2005) and Heine (1997), the comparative constructions can be classified into four types: locational, conjoined, exceed, and particle. The different usages of these four types are examined in this contrastive study as well. This study has observed that English has a grammaticalized particle *than* (originated from *then*) as STM, but in Japanese and Tok Pisin, there are some variations. Japanese mainly has the locational *yori* (originated from locative postposition) and exceed (verbs *masaru* / *otoru*), while Tok Pisin has exceed, conjoined, and particle types. These findings relating to the comparatives in these different languages are considered to be based on the cognitive differences of comparison in each language, and in particular, how these languages construe the relationship between the comparee (COM) and the standard (STAN). Stassen (2005) observes that the locational type is the most frequent comparative construction in the world, and the particle type comparative is observed in English and other Indo-European languages.

Below is a summary of the relationship between STAN and STM in the three languages:

(15) Formal differences of the comparative constructions in the three languages

a. English: *than he*, or *than him* (particle)[6]

b. Japanese: 1) *kare-yori* (locational), 2) *kare-ni masaru* (exceed)

c. Tok Pisin: 1) *winim em* (exceed), 2) *A i bik, mobeta B* (conjoined), 3) *long em*, or *olsem em* (particle)

In spite of the difficulties in translating them, the comparatives in English are translated into other comparative types in Japanese and Tok Pisin. Nevertheless, we can conclude that when English uses the particle type, Japanese prefers the locational, and Tok Pisin, the exceed type. In (16), such formal and constructional distributions are construed in terms of transitivity (cf. Hopper and Thompson 1980). Adopting the transitivity hypothesis of Hopper and Thompson (1980) makes it easier to explain differences between the comparatives. Comparative constructions need at least two participants (comparee and standard), as transitive sentences need two participants (agent and patient in transitive, and COM and STAN in comparative). Differences in the comparative can be explained by the grammatical relations (subject, direct object, or oblique) of STAN. This study claims that it is possible to illustrate the grammatical status of STAN by using transitivity, as shown in (16) and Figure 21-1.[7]

(16) Grammatical status of the standard, based on transitivity (A: comparee, B: standard)
Subject: A is big and B is small. (conjoined)
Object: A surpasses B. (exceed)
Oblique/locative: A-wa B-yori chiisai. (locational); A is bigger than B. (particle)

Thus, we create a triangle of grammatical relations with subject, direct object, and oblique / locative, and can posit the comparative types of the three languages as illustrated in Figure 21-1. There are three grammatical relations of STAN: subject, direct object, and oblique / locative. STAN in the conjoined type appears in the subject position, and STAN in the exceed type as the direct object. English particle STM and Japanese locational STM are classified as oblique / locative, and STAN is regarded as occupying the oblique / locative position.

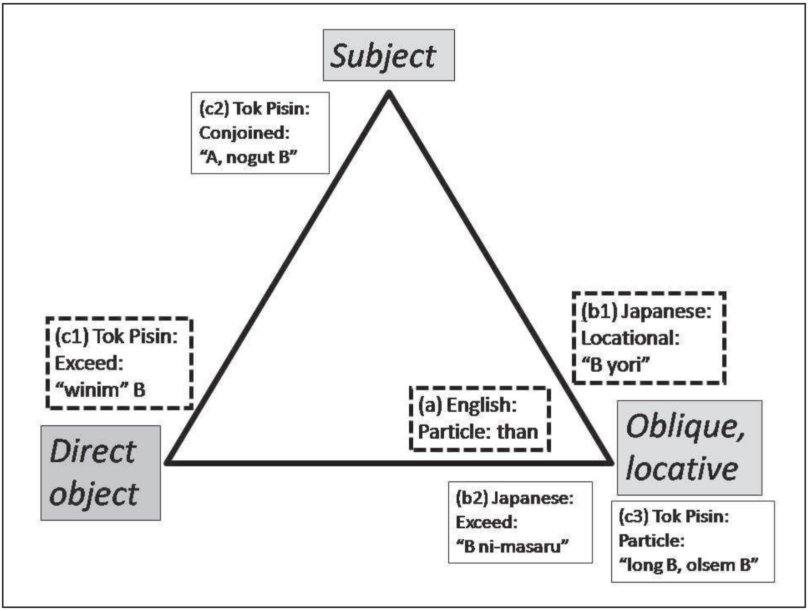


Figure 21-1. Transitivity model of standard in comparative constructions

Finally, the semantic status of STAN in comparative constructions is considered. From the explanation above, we can identify four semantic types of STAN: noun (person), noun (thing), noun phrase and infinitive, and demonstrative. The frequency of each of them is given earlier in (14), and it is probable that the person and animate entities take the position of STAN. It should be emphasized that this conclusion is based on the translated texts of the *New Testament*, and does not necessarily reflect the situation in spoken language. However, when we make a comparison between COM and STAN, a given specific person or animate one will be chosen as STAN.

5. Conclusion

This chapter examined comparative constructions using translations of texts in the *Bible* in three languages: English, Japanese, and Tok Pisin. There are clear typological implications, and this study tried to clarify them in terms of transitivity.

First, English has a fixed grammatical comparative construction that uses the particle *than*, and Japanese has another fixed locational comparative that uses *yori*. In Tok Pisin, however, there are several options for expressing comparative meanings: exceed is frequent, but there are also conjoined and particle examples. This study has illustrated differences such as these in the use of comparatives in the three languages, focusing on the grammatical status of the standard, and visualized the constructions in terms of transitivity.

Second, this study examined the semantic status of the standard position, and it turned out that a personal or animate noun frequently appears as the standard in comparative constructions. Moreover, we have observed that longer forms such as noun phrases and infinitives can appear as the standard in the *New Testament*.

Finally, this cross-linguistic study has found that the standard takes a rather higher grammatical position (subject and direct object) in primitive comparatives (conjoined and exceed), and they are observed in Tok Pisin. By contrast, through a certain grammaticalization process, the standard is demoted to the oblique / locative position in English and Japanese.

Notes

* The following abbreviations are used: COM, comparee; PAM, parameter marker; PAR, parameter; STM, standard marker; STAN, standard; COP, copula; FOC, focus; TOP, topic.

[1] Henkelmann (2006: 371) has another point of view on comparatives in terms of adjectives. He notes that there are four gradable adjectives (positive *tall*, equative *as tall as*, comparative *taller*, and superlative *tallest*), and the comparative in which the standard of comparison is implied.

[2] These grammatical and geographical descriptions of English, Japanese, and Tok Pisin are based on the information from *The World Atlas of Language Structures* (Haspelmath *et al.* 2005).

[3] Dimmendaal (2001: 70) considered a possibility of languages without a comparative construction. He observes that in Australian languages, “grammatical constructions involving comparison (‘bigger’, ‘better’, ’younger’, ‘smarter’) are unnatural in some speech communities.”

[4] Heine (1997: 112) explained the locative source meanings for the standard marker, and they are *at*, *from*, and *to*.

[5] The standard marker is indicated by the bracket [ ], and PAM, PAR, STM, and STAN are indicated by italics in the parallel text examples in (8)–(13).

[6] It is natural that there are conjoined and exceed types of the comparatives in English as well. Heine (1997) pointed out the possible constructions in English.

(i) A is small and B is big. (conjoined)
(ii) A is superior to B. (exceed)
(iii) A surpasses B in cleverness. (exceed)

Needless to say, the particle type is the most frequent, but every language has two or three options expressing comparative meaning.

7. The transitivity model suggested in this study does not imply that the other factors of transitivity (aspect, affectedness, and state of affairs) are relevant to the comparative constructions.

References

Cysouw, M. and Wälchli, B. (2007), “Parallel texts: Using translational equivalents in linguistic typology”. *Sprachtypologie und Universalienforschung* (STUF) 60: 95-99.

Dimmendaal, G. J. (2001), “Places and people: Field sites and informants”, in P. Newman and M. Ratliff (eds.) *Linguistic Fieldwork*, 55-75. Cambridge: Cambridge University Press.

Everett, D. L. and Kern, B. (1997), *Wari: The Pacaas Novos Language of Western Brazil*. London/New York: Routledge.

Haspelmath, M. and Buchholz, O. (1998), “Equative and similative constructions in the languages of Europe”, in J. van der Auwera and D. P. Ó. Baoill (eds.) *Adverbial Constructions in the Languages of Europe*, 277-334. Berlin/New York: Mouton de Gruyter.

Haspelmath, M., Dryer, M. S., Gil, D. and Comrie, B. (eds.) (2005), *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Henkelmann, P. (2006), “Constructions of equative comparison”. *Sprachtypologie und Universalienforschung* (STUF) 59: 370-398.

Heine, B. (1997), *Cognitive Foundations of Grammar*. New York/Oxford: Oxford University Press.

Hopper, P. J. and Thompson, S. (1980), “Transitivity in grammar and discourse”. *Language* 56: 251-299.

Mihalic, F. (1986), *The Jacaranda Dictionary and Grammar of Melanesian Pidgin*. Milton: The Jacaranda Press/Web Books.

Nose, M. (2007), “A typological study of Standard marker in comparative and similative constructions (in Japanese)”, in *Proceedings of the 134th Linguistic Society of Japan*, 288-293.

Roberts, J. R. (1987), *Amele*. London/New York/Sydney: Croom Helm.

Stassen, L. (1985), *Comparison and Universal Grammar*. Oxford/New York: Blackwell.

—. (2005), “Comparative constructions”, in Haspelmath *et al.* (eds.) *The World Atlas of Language Structures*, 490-493. Oxford: Oxford University Press.

Stolz, T. (2007), “Harry Potter meets Le petit prince – On the usefulness of parallel corpora in crosslinguistic investigations”. *Sprachtypologie und Universalienforschung* (STUF) 60: 100-117.

Uritani, R. (2002), *Introduction to Spanish: Revised version* (Kaitei Supein-go no nyuumon). Tokyo: Hakushisha.

Vries, L. De. (2007), "Some remarks on the use of Bible translations as parallel texts in linguistic research". *Sprachtypologie und Universalienforschung* (STUF) 60: 148-157.

CHAPTER TWENTY-TWO

A CORPUS-BASED STUDY OF RESTRICTIVE RELATIVE CLAUSES

HUI-CHUAN LU, YUN-HUI CHEN

1. Introduction

This chapter investigates the differences and similarities in Restrictive Relative Clauses (RRCs) in three languages, namely Spanish, English and Chinese, by means of comparing and contrasting parallel data extracted from a part-of-speech (POS) tagged trilingual corpus. Our study further offers experiences of corpus-based analysis and application in Second Language Acquisition (SLA) research. The investigation is comprised of three major sections, which begin with the construction of a POS-tagged trilingual parallel corpus (*Corpus Paralelo de Español, Inglés y Chino*, CPEIC) in order to search the parallel texts of the three languages. Following that, based on Keenan and Comrie's (1977) Noun Phrase Accessibility Hierarchy (NPAH) hypothesis, we study RRCs in Spanish, English and Chinese respectively and then in parallel with the assistance of two corpus exploration tools, WordSmith (Scott 2004) and ParaConc (Barlow 1995), to analyze our corpus data. Finally, we will apply the results of our analysis to the area of SLA by contrasting the results of parallel data and that of learner language.

2. Literature review

Parallel corpora were first developed as bilingual, and their texts were compiled mostly from European languages such as English in parallel to French. During the last decade, English-Chinese parallel corpora have increased in number, yet the number of corpora involving Chinese in parallel to other Indo-European languages is comparably small. For instance, on the one hand, parallel corpora involving Spanish and English include Reuters, MLCC Multilingual and Parallel Corpora, European Corpus Initiative Multilingual Corpus (ECI/MCI), Multext Official Journal of European Community (MULTEXT JOC) Corpus, CRATER Multilingual Aligned Annotated Corpus, OGI Multilanguage Corpus, and EUR-Lex among others. On the other hand, as far as parallel corpora are concerned, apart from ECI/MCI and OGI, few of them cover Spanish in parallel to Chinese, and most of the Chinese-related parallel corpora are with English: for instance, Hong Kong News Parallel Text, Hong Kong Laws Parallel Text, Hong Kong Hansards Parallel Text, the Babel English-Chinese Parallel Corpus, and the Babel Chinese-English Parallel Texts. Accordingly, due to the need for cross-linguistic research and the lack of existing parallel corpora, we set the goal of creating an annotated corpus composed of aligned parallel texts of Spanish, English and Chinese. From the process of constructing this corpus, we are able to share and provide experiences concerning the possible difficulties and solutions for the creation of a trilingual parallel corpus.

With respect to the studies on parallel corpora and translation, there are those by Aijmer and Altenberg (2002) and Santos (2004) among others, and many of them are related to English and other European languages. For instance, Aijmer and Altenberg (2002) discuss the strategies of translation, especially omission, in the English-Swedish Parallel Corpus. Santos (2004) targets the grammatical analysis of English-Portuguese translation in light of computational linguistics. In the area of corpus linguistics, corpus-based contrastive studies of Spanish and Chinese translation have rarely been reported. In relation to lexical studies, parallel corpus-based research involving syntactic or grammatical analysis is more limited, with a few exceptions. For example, Cermák and Klégr (2004) study the adverbial particles based on a parallel corpus of Czech original texts and their English translations; Uchida (2002) examines the characteristics of causal participles in English and French on the basis of a parallel corpus, the MULTEXT JOC Corpus. Accordingly, the research related to syntactic analysis needs more attention and effort.

Among the studies related to Relative Clauses (RCs), Keenan and Comrie’s (1977) NPAH hypothesis is the most widely-discussed, which proposes a hierarchical order regarding the relativization of noun phrases with respect to their relative pronouns within RCs: Subject (S) > Direct Object (DO) > Indirect Object (IO) > Prepositional Object (PO) > Genitive (G). Other research related to relative clauses includes Sheldon (1974) and Wong (1991). Proposing Parallel Function Hypothesis in her study of children’s acquisition of RCs in English, Sheldon suggests that if the antecedent noun phrases have the same or parallel grammatical function with the relative pronouns, the sentences will be easier for learners to understand. In addition to the studies of native language acquisition, these hypotheses concerning RCs have also been applied to SLA. For example, Wong (1991) analyzes the accessibility of RCs by ESL learners in Hong Kong. Still, the most widely studied hypothesis of RCs is Keenan and Comrie's NPAH.

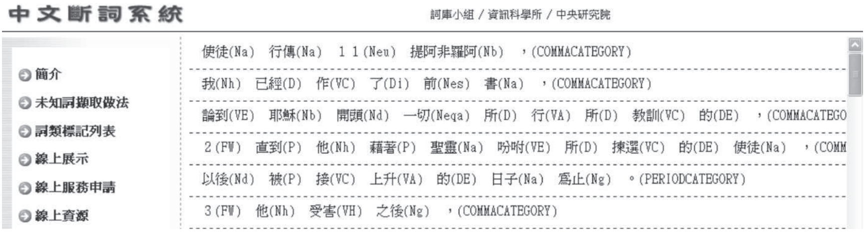
In our study of RRCs based on the trilingual parallel corpus to be presented in section 3, there are three questions we intend to answer with respect to NPAH:

(1) In terms of native language, what are the differences in expression of the languages of our study, including Spanish, English and Chinese?
(2) How is Spanish parallel to English and Chinese?
(3) As for learner language, how is the learner language of Spanish accounted for by comparing and contrasting the three native languages?

3. Corpus creation and data extraction

The corpus we have compiled for the following analysis includes three sub-corpora containing the parallel biblical texts in three languages, that is, Spanish, English and Chinese, which are extracted from the online multilingual archive BibleGateway (www.biblegateway.com/). Although it can hardly be ignored that using the Bible as our data source raises an issue since its texts were originally written in two archaic languages – old Hebrew and ancient Greek, meaning that there are several differences of language usages between the original Bible and the modern translated versions, the Bible is still regarded as a credible resource in corpus linguistics. For instance, Resnik, Olsen and Diab (1999) acknowledge the advantages of using biblical texts for cross-linguistics and parallel translation studies. Given that the Bible is one of the richest multilingual texts so far and that the available access to Spanish-Chinese parallel texts is rather limited, we consider it the most accessible data source for our research.

In order to build the Spanish-English-Chinese parallel corpus, we collected data of trilingual parallel texts from the BibleGateway.com website. Using these online and openly accessible biblical texts not only spares us the time and labour of manual inputting but also helps to facilitate the functions of the language tools, either when aligning or analyzing the texts. Furthermore, considering the various translations of the Bible, for making a quasi-parallel comparison, we used the New International Version for English, la Nueva Versión Internacional for Spanish, and the Union Version for Chinese. In dealing with the texts, we extracted from the New Testament the four gospels, Matthew, Mark, Luke, and John, as well as the Acts of the Apostles. Then Spanish and English texts were POS-tagged using Tree Tagger while the Chinese texts were tokenized and tagged using Academia Sinica's Chinese Word Segmentation System with Unknown Word Extraction and POS Tagging (Figure 22-1). The tagged results were simplified with our self-developed program (Figures 22-2 and 22-3), and after the processed texts are all saved in the text file format (*.txt), we utilized two corpus analysis tools, WordSmith and ParaConc, to search appropriate data through the three sub-corpora respectively for the analysis of RRCs. The Spanish sub-corpus contains 103,267 words, the English sub-corpus includes 105,128 words, and the Chinese sub-corpus consists of 133,078 characters in total.

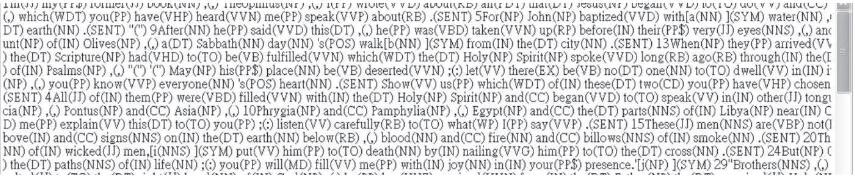


```
使徒(Na)　行傳(Na)　１１(Neu)　提阿非羅阿(Nb)　，(COMMACATEGORY)
我(Nh)　已經(D)　作(VC)　了(Di)　前(Nes)　書(Na)　，(COMMACATEGORY)
論到(VE)　耶穌(Nb)　開頭(Nd)　一切(Neqa)　所(D)　行(VA)　所(D)　教訓(VC)　的(DE)　，(COMMACATEGO
２(FW)　直到(P)　他(Nh)　藉著(P)　聖靈(Na)　吩咐(VE)　所(D)　揀選(VC)　的(DE)　使徒(Na)　，(COMM
以後(Nd)　被(P)　接(VC)　上升(VA)　的(DE)　日子(Na)　爲止(Ng)　。(PERIODCATEGORY)
３(FW)　他(Nh)　受害(VH)　之後(Ng)　，(COMMACATEGORY)
```

Figure 22-1. The interface of Chinese Word Segmentation System with Unknown Word Extraction and POS Tagging

```
Matthew       NP    Matthew
1             CD    1
The           DT    the
Genealogy     NN    genealogy
Jesus         NP    Jesus
1A            NP    <unknown>
record        NN    record
of            IN    of
the           DT    the
genealogy     NN    genealogy
of            IN    of
Jesus         NP    Jesus
Christ        NP    Christ
```

Figure 22-2. Tagged result before the procedure of simplification



```
In(JJ) my(PP$) former(JJ) book(NN) ,(,) Theophilus(NP) ,(,) I(PP) wrote(VVD) about(RB) all(PDT) that(DT) Jesus(NP) began(VVD) to(TO) do(VV) and(CC)
(,) which(WDT) you(PP) have(VHP) heard(VVN) me(PP) speak(VVP) about(RB) .(SENT) 5For(NP) John(NP) baptized(VVD) with[a(NN) ](SYM) water(NN) ,(
DT) earth(NN) .(SENT) "(") 9After(NN) he(PP) said(VVD) this(DT) ,(,) he(PP) was(VBD) taken(VVN) up(RP) before(IN) their(PP$) very(JJ) eyes(NNS) ,(,) anc
unt(NP) of(IN) Olives(NP) ,(,) a(DT) Sabbath(NN) day(NN) 's(POS) walk[b(NN) ](SYM) from(IN) the(DT) city(NN) .(SENT) 13When(NP) they(PP) arrived(VV
) the(DT) Scripture(NP) had(VHD) to(TO) be(VB) fulfilled(VVN) which(WDT) the(DT) Holy(NP) Spirit(NP) spoke(VVD) long(RB) ago(RB) through(IN) the(I
) of(IN) Psalms(NP) ,(,) "(") '(") May(NP) his(PP$) place(NN) be(VB) deserted(VVN) ;(:) let(VV) there(EX) be(VB) no(DT) one(NN) to(TO) dwell(VV) in(IN) i
(NP) ,(,) you(PP) know(VVP) everyone(NN) 's(POS) heart(NN) .(SENT) Show(VV) us(PP) which(WDT) of(IN) these(DT) two(CD) you(PP) have(VHP) chosen
(SENT) 4All(JJ) of(IN) them(PP) were(VBD) filled(VVN) with(IN) the(DT) Holy(NP) Spirit(NP) and(CC) began(VVD) to(TO) speak(VV) in(IN) other(JJ) tongu
cia(NP) ,(,) Pontus(NP) and(CC) Asia(NP) ,(,) 10Phrygia(NP) and(CC) Pamphylia(NP) ,(,) Egypt(NP) and(CC) the(DT) parts(NNS) of(IN) Libya(NP) near(IN) C
D) me(PP) explain(VV) this(DT) to(TO) you(PP) ;(:) listen(VV) carefully(RB) to(TO) what(WP) I(PP) say(VVP) .(SENT) 15These(JJ) men(NNS) are(VBP) not(I
bove(IN) and(CC) signs(NNS) on(IN) the(DT) earth(NN) below(RB) ,(,) blood(NN) and(CC) fire(NN) and(CC) billows(NNS) of(IN) smoke(NN) .(SENT) 20Th
NN) of(IN) wicked(JJ) men,[i(NNS) ](SYM) put(VV) him(PP) to(TO) death(NN) by(IN) nailing(VVG) him(PP) to(TO) the(DT) cross(NN) .(SENT) 24But(NP) (
) the(DT) paths(NNS) of(IN) life(NN) ;(:) you(PP) will(MD) fill(VV) me(PP) with(IN) joy(NN) in(IN) your(PP$) presence.'[j(NP) ](SYM) 29"Brothers(NNS) ,(,)
```

Figure 22-3. Tagged result after the procedure of simplification

We used the relative pronouns, *que* in Spanish, *that* in English, and possessive marker *DE* in Chinese, as keywords to search in the concordance tools of WordSmith. Different from the previously untagged corpus, the POS-tagged corpora enable us to filter out inappropriate sentences more effectively, by additionally including the POS-tags of these keywords. We further set up conditions for restricting the antecedent and following words or phrase as well as their POS-tags to gain more precise results. We searched for the combination of the structure 'N + *que* + V' for Spanish, 'N + *that* + V' for English, and '*DE* + N' for Chinese. After extracting the appropriate data, we annotated the relativized elements in the subordinate clauses according to their grammatical functions as illustrated in examples (1-6). Finally, we analyzed the similarities and differences in the three languages in the parallel contexts in order to answer our research questions.

- Spanish: relativized subject

(1) Entren por la puerta estrecha. Porque es ancha la
enter for the door narrow. Because is wide the

puerta y espacioso *el camino que conduce*
door and broad *the road that lead*

a la destrucción, y muchos entran por ella.
to the destruction, and many enter through it

'Enter through the narrow gate. For wide is the gate and broad is the road *that leads to destruction*, and many enter through it.' (Matthew 7:13)

- Spanish: relativized direct object

(2) Le hemos oído decir que ese Jesús de Nazaret
him have heard say that this Jesus of Nazareth

destruirá este lugar y cambiará
will-destroy this place and will-change

las tradiciones que nos dejó Moisés.
the traditions that us left Moses

‘For we have heard him say that this Jesus of Nazareth will destroy this place and change the customs Moses handed down to us.’ (Acts 6:14)

- English: relativized subject

(3) Enter through the narrow gate. For wide is the gate and broad is *the road that leads to destruction*, and many enter through it. (Matthew 7:13)

- English: relativized direct object

(4) For we have heard him say that this Jesus of Nazareth will destroy this place and change *the customs Moses handed down to us*. (Acts 6:14)

- Chinese: relativized subject

(5) 他 怎麼 進了 神 的 殿， 吃了 陳設 餅，
ta zenme jinle shen de dian chile chenshe bing
he how enter God DE palace eat display bread

這 餅 不是 他 和 *跟從 他 的 人*
zhe bing bushi ta he *gencong ta de ren*
this bread no-is he and *follow him DE person*

可以 吃 得， 惟獨 祭司 才 可以 吃。
keyi chi de weidu jisi cai keyi chi
can eat de only priest just can eat

‘He entered the house of God, and he and his companions ate the consecrated bread—which was not lawful for them to do, but only for the priests.’ (Matthew 12:4)

- Chinese: relativized direct object

(6) 我 卻 不 以 性命 為 念， 也 不 看 為
wo que bu yi xingming wei nian ye bu kan wei
I but no use life as thought too no see as

寶貴， 只要 行 完 我的 路程， 成就
baogui zhiyao xing wan wode lucheng chengjiu
precious if-only do finish my journey achieve

我 從 主 耶穌 *所 領受 的 職事，*
wo cong zhu yesu *suo lingshou de zhishi*
I from lord Jesus_*SUO receive DE duty*

證明 神 恩惠 的 福音。
zhengming shen enhui de fuyin
prove God grace DE gospel

'However, I consider my life worth nothing to me, if only I may finish the race and complete *the task the Lord Jesus has given me* – the task of testifying to the gospel of God's grace.' (Acts 20:24)

By using WordSmith, we were able to analyze and observe how the three languages are similar to and different from one another with respect to the use of RRCs in general. Going one step further, to investigate how the structures of RRCs in these three languages are parallel to one another, we used ParaConc to facilitate the parallel analysis of our sub-corpora. The texts of the three languages are aligned at sentence level with the help of ParaConc, and the analysis of the parallel materials of English and Chinese were compared with the Spanish RRCs data using WordSmith. In the section that follows, we will present the results of data analysis.

4. Analysis of RRCs

First of all, we examine the distribution of RRCs in each language. According to Table 22-1, the differences between Spanish and Chinese (7.9% vs. 1.8%) show the contrast between the two languages, which implies that there might be a higher degree of difficulty for Taiwanese learners to acquire RRCs in Spanish.

On the other hand, as can be seen in Table 22-2, the order of NPAH of *que* in Spanish RRCs and of *that* in English RRCs (PO>IO) is contrary to Keenan and Comrie's NPAH (IO>PO). The descriptions in the descriptive grammar do not appear to accord with the attested data, and our results provide further evidence to modify the proposed argument at least for the data analyzed here. Moreover, in our research the NPAH of *DE* in Chinese RRCs (DO>S>IO) shows that DO has the highest frequency (66.2%).

Table 22-1. Distribution of RRCs in three languages

Language	Sentences	RRCs	%
Spanish	5,594	443	7.9%
English	6,102	120	2.0%
Chinese	55,864	1,014	1.8%

Table 22-2. Distribution of relativized elements in three languages

Language	S	DO	IO	PO	Total
Spanish	317 (71.6%)	119 (26.7%)	1 (0.2%)	6 (1.4%)	443
English	95 (79.1%)	23 (19.1%)		2 (1.7%)	120
Chinese	325 (32%)	672 (66.2%)	17 (1.7%)		1,015

From the above results, we can observe the similarities of the orders of NPAH between Spanish and English (S>DO>PO>IO) and the discrepancies between Spanish and Chinese. The differences between Spanish and Chinese in terms of their grammatical and syntactical structures again imply that there are difficulties for Taiwanese learners to learn Spanish. In short, the results indicate that English RRCs with *that* are similar to Spanish RRCs with *que*, whereas Chinese RRCs with *DE* apparently differ from either English or Spanish.

We have discussed how RRCs behave in each of the three languages. In this section, we will focus on the parallel texts of the three languages. To begin with, we observe the distribution of Spanish RRCs in parallel with English RRCs in Table 22-3.

According to Table 22-3, compared to *que* in Spanish RRCs, *who* (22.6%) appears more frequently than *that* (8.5%) in English RRCs among others. And more than half of the RRCs of Spanish (18.9 + 33.0 = 51.9%) are paralleled with other structures of English. For instance, where there is *que* in Spanish RRCs (7a), in English the restrictive pronoun may be omitted (7b). Or, the syntactic structure of Spanish RRCs (8a) is paralleled to different sentential structure in English (8b).

Table 22-3. Distribution of Spanish RRCs paralleled in English

Eng.	who	that	which	what	where	VP	Independent sentence	∅	Others	Total
#	24	9	1	1	0	11	5	20	35	106
%	22.6	8.5	0.9	0.9	0.0	10.4	4.7	18.9	33.0	

(7a) hasta el día en que fue llevado al cielo, luego de
until the day in which was taken to-the sky after

darles instrucciones por medio del Espíritu Santo
give-him instructions by means of-the Spirit Holy

a *los apóstoles que había escogido*. (Hechos 1:12)
to the apostles that had chosen

(7b) until the day he was taken up to heaven, after giving instructions through the Holy Spirit to the apostles he had chosen. (Acts 1:12)

(8a) Qué vamos a hacer con estos sujetos? Es un
what go-1-pl to do with these persons is a

hecho que por medio de ellos *ha ocurrido* un
doing that by means of them has-happened a

milagro evidente; todos los que viven en Jerusalén
miracle obvious all the that live in Jerusalem

lo saben, y no podemos negarlo. (Hechos 4:16)
it know and no can deny-it

(8b) ‘What are we going to do with these men?’ they asked. ‘Everybody living in Jerusalem knows they have done an outstanding miracle, and we cannot deny it.’ (Acts 4:16)

Table 22-4. Distribution of Spanish RRCs paralleled in Chinese

Chinese	SUO…DE	DE	Conj	Indepen-dent sentence	Others	Total
#	20	36	22	13	15	106
%	18.9	34.0	20.8%	12.3%	14.1%	

Furthermore, in Table 22-4, we observe that there are at least two patterns in the parallel sentences between Spanish and Chinese. One of the patterns shows that *DE* functions as a connection between the subordinate clauses and the nuclear elements in Chinese RRCs. The result shows that 52.9% (18.9% + 34.0%) of Spanish RRCs with *que* are parallel to Chinese RRCs with *DE*, and 18.9% of the parallel sentences contain *SUO*. The other pattern is that the Spanish RRCs are changed into different structures in Chinese. The result shows that 33.1% (20.8% + 12.3%) of the Spanish RRCs are parallel with two separate clauses or sentences (9a-10b) or sentences with repeated pronouns in Chinese sentences.

(9a) Él cayó al suelo y oyó *una voz que le*
he fell to-the ground and heard a voice that him

decía: Saulo, Saulo, por qué me persigues?
said Saul Saul why me pursue
(Hechos 9:14)

‘He fell to the ground and heard a voice say to him, “Saul, Saul, why do you persecute me?”’ (Acts 9:14)

(9b) 他 就 仆倒 在 地， 聽見 有 *聲音 對*
ta jiu pudao zai di tingjian you *shengyin dui*
he then fell-to on ground heard have *voice toward-*

他 說： 掃羅！ 掃羅！你 爲甚麼 逼迫 我？
ta shuo Saoluo Saoluo ni weishenme bipo wo
him say Saul, Saul you why force me
(使徒行傳 9:14)

‘He fell to the ground and heard a voice say to him, “Saul, Saul, why do you persecute me?”’ (Acts 9:14)

(10a) En eso llegaron de Antioquía y de Iconio
in that arrived from Antioch and from Iconium

unos judíos que hicieron cambiar de parecer a la
some Jews who did change of think to the

multitud. Apedrearon a Pablo y lo arrastraron
crowd stoned to Paul and him dragged

fuera de la ciudad, creyendo que estaba muerto.
outside of the city thinking that was dead
(Hechos 14:19)

'Then some Jews came from Antioch and Iconium and won the crowd over. They stoned Paul and dragged him outside the city, thinking he was dead.' (Acts 14:19)

(10b) *但 有些 猶太人* 從 安提阿 和 以哥念 來，
dan youxie youtairen cong antia he yigenian lai
But some Jews from Antioch and Iconium come

挑唆 眾人， 就 用 石頭 打 保羅，以爲 他 是
tiaosuo zhongren jiu yong shitou da baoluo yiwei ta shi
incite crowd then use stone hit Paul think he is

死了，便 拖 到 城 外。
sile bian tuo dao cheng wai
dead thus drag to city outside
(使徒行傳 14:19)

'Then some Jews came from Antioch and Iconium and won the crowd over. They stoned Paul and dragged him outside the city, thinking he was dead.' (Acts 14:19)

5. Pedagogical applications

From the results of our discussion in previous sections, we would like to make a connection between the native language and the learner language. According to Lu (2007), the order of NPAH based on the analysis of the written production by Taiwanese learners of Spanish in our other corpus, CATE (*Corpus de Aprendices Taiwaneses de Españo* 'Taiwanese Learners of Spanish Corpus'), shows that S (60.45%) > DO (34.09%) > PO (5.46%) > IO (0%) (see examples 11-14).

(11) Relativized S

El criminal entregó el rifle al *policía que*
The criminal handed-over the rifle to-the *police that*

llamó al sargento.
called to-the sergeant

‘The criminal gave the rifle to the police who called the sergeant.’

(12) Relativized DO

Conozco *al estudiante que suspendió* el profesor.
Know *to-the student that flunked* the professor

‘I know the student who the professor flunked.’

(13) Relativized PO

El estudiante *con el que habló* el profesor está
The student *with the that talked* the professor was

muy triste.
very sad

‘The student to whom the professor talked was very sad.’

(14) Relativized IO

El estudiante entregó el trabajo final al professor
The student handed-in the paper final to-the professor

al que la secretaria dio el documento.
to-the that the secretary gave the document

‘The student handed in the paper to the professor whom the secretary gave the document.’

Lu’s (2007) NPAH hypothesis does not completely agree with Keenan and Comrie’s (S > DO > IO > PO > G). However, in terms of syntactical functions of RRCs, the learning order of Taiwanese learners of Spanish and the NPAH order of native Spanish speakers show a similar result, which is: S > DO > PO > IO. Yet, none of the orders is the same or similar to the NPAH of Chinese RRCs with *DE* in their native language.

Therefore, we might conclude that the RRC with *DE* of L1 (Chinese) may not play a role in language learning while the RRCs with *that* in L2 (English) can be considered as a positive transfer for the L3 Spanish learners.

6. Conclusion

Our analysis shows that, in the research of parallel texts, over half of Spanish RRCs with *que* are parallel to sentence structures other than RRCs in English, and the subordinate clauses in Spanish RRCs are parallel to independent sentences and to RRCs with *DE* in Chinese. Furthermore, in the contrastive analysis, the orders of NPAH of Spanish and of English native languages are similar (S > DO > PO > IO), but both of them are different from the order of Chinese (DO > S). However, all the orders of NPAH are different from Keenan and Comrie’s proposal.

In the study of second language acquisition of Spanish, the NPAH of the learner language in Spanish shares a similarity with native English language but both orders differ from that of NPAH of native Chinese language. As a consequence, we conclude that L1 (Chinese) may not influence the learning of a third language (Spanish), whereas L2 (English) may play a more important role in the learning of L3 in our study.

References

Aijmer, K., and Altenberg, B. (2002), “Zero translations and cross-linguistic equivalence: Evidence from the English-Swedish Parallel Corpus”, in L. E. Breivik and A. Hasselgren (eds.) *From the COLT's Mouth ... and Others*, 19-41. Amsterdam: Rodopi.

Barlow, M. (1995), *A Guide to ParaConc*. Houston: Athelstan.

Cermák, F. and Klégr, A. (2004), “Modality in Czech and English: Possibility particles and the conditional mood in a parallel corpus”. *International Journal of Corpus Linguistics* 9(1): 83-95.

Chinese Word Segmentation System with Unknown Word Extraction and Pos Tagging. Academia Sinica, and National Digital Archives Program, Taiwan. http://ckipsvr.iis.sinica.edu.tw/ (accessed July 7, 2009).

Cole, R. and Muthusamy, Y. OGI Multilanguage Corpus. Linguistic Data Consortium, University of Pennsylvania. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S17 (accessed July 7, 2009).

CRATER Multilingual Aligned Annotated Corpus. Computing Department of Lancaster University.

http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html (accessed July 7, 2009).

ELSNET. European Corpus Initiative Multilingual Corpus I (ECI/MCI). ELSNET. http://www.elsnet.org/eci.html (accessed July 7, 2009).

EUR-Lex. European Union. http://eur-lex.europa.eu/ (accessed July 7, 2009).

European Language Resources Association. MLCC Multilingual and Parallel Corpora. Evaluations and Language Resources Distribution Agency. http://www.elda.org/catalogue/en/text/W0023.html (accessed July 7, 2009).

European Language Resources Association. Multext Official Journal of European Community Corpus (MULTEXT JOC Corpus). Evaluations and Language resources Distribution Agency. http://www.elda.org/catalogue/en/text/W0017.html (accessed July 7, 2009).

Institute of Computational Linguistics, Peking University. The Babel Chinese-English Parallel Texts. Institute of Computational Linguistics, Peking University. http://icl.pku.edu.cn/icl_groups/parallel/default.htm (accessed July 7, 2009).

Keenan, E. L. and Comrie, B. (1977), "Noun phrase accessibility and universal grammar". *Linguistic Inquiry* 8: 63-99.

Lu, H-C. (2007), "A corpus-based study of Spanish teaching in Taiwan. In *Foreign Language Studies: European languages and Cultures in Taiwan*, 1-22. Taipei: National Chengchi University.

Ma, X. Hong Kong Hansards Parallel Text. Linguistic Data Consortium, University of Pennsylvania. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T50 (accessed July 7, 2009).

Ma, X. Hong Kong Laws Parallel Text. Linguistic Data Consortium, University of Pennsylvania. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T47 (accessed July 7, 2009).

Ma, X. Hong Kong News Parallel Text. Linguistic Data Consortium, University of Pennsylvania. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000T46 (accessed July 7, 2009).

Resnik, Philip, Mari Broman Olsen, and Mona Diab. (1999), "Creating a parallel corpus from the 'Book of 2000 Tongues'". *Computers and the Humanities* 33: 129-53.

Reuters.com. Thomson Reuters. http://www.reuters.com/ (accessed July 7, 2009).

Santos, D. (2004), *Translation Based Corpus Studies: Contrasting English and Portuguese Tense and Aspect System*. Amsterdam: Rodopi.

Schmid, H. and Koller, T. Tree Tagger. Institute for Natural Language Processing, University Stuttgart, and University of Nottingham. http://www.cele.nottingham.ac.uk/~ccztk/treetagger.php (accessed July 7, 2009).

Scott, M. (2004), *The WordSmith Tools* (version 4.0). Oxford: Oxford University Press. http://www.lexically.net/wordsmith/ (accessed July 7, 2009).

Sheldon, A. (1974), "The role of parallel function in the acquisition of relative clauses in English". *Journal of Verbal Learning and Verbal Behaviour* 13(3): 272-281.

Uchida, M. (2002), "From participles to conjunctions: A parallel corpus study of grammaticalization in English and French", in T. Saito, J. Nakamura and S. Yamazaki (eds.) *English Corpus Linguistics in Japan*, 131-146. Amsterdam: Rodopi.

Wong, J. (1991), "Learnability of relative clauses: A Hong Kong case". *Perspectives* 3: 108-17.

Xiao, R. The Babel English-Chinese Parallel Corpus. http://www.lancs.ac.uk/fass/projects/corpus/babel/babel.htm (accessed July 31, 2009).

Zondervan Corporation. BibleGateway.com. The Zondervan Corporation. http://www.biblegateway.com (accessed July 7, 2009).

CHAPTER TWENTY-THREE

FREQUENCY AND LEXICO-GRAMMATICAL PATTERNS OF SENSE-RELATED VERBS IN ENGLISH AND PORTUGUESE ABSTRACTS

CARMEN DAYRELL

1. Introduction

With the predominance of English as the lingua franca of research and scholarship, academics worldwide are now almost compelled to conduct and publish research in English (Hyland 2009: ix-5). However, as Hyland (Hyland 2009: ix) rightly states, academic communication poses real challenges for novice researchers, who quickly have to come to terms with the conventions and characteristics of academic discourse.

For non-native speakers of English, demands are even heavier. In addition to mastering the lexical and syntactical features of the target genre, language learners should also be able to identify the rhetorical motivations behind linguistic choices (Vold 2006). In other words, in order for language learners to write effectively, they need to recognize both the appropriate lexical and syntactical structures as well as the specialized logic of the genre (Milton and Hyland 1999). This also includes awareness of cultural differences since genre practices, expectations and values may vary across languages. As Davoodifard (2008: 40) explains, “conformity to the target language community norms, conventions and expectations can prevent writing styles from seeming inappropriate to the target language community members.” All these considerations are also valid for translators who may be required to translate academic texts and are expected to fulfil the expectations of the target academic community.

Lexical and syntactical features of English academic discourse have received increasing attention in the literature and this has been approached from different perspectives. One approach focuses on the language produced by native or expert speakers. According to this perspective, some studies have compared academic prose with other textual genres (see, for instance, Biber *et al.* 2004). Others have investigated similarities and differences across academic genres and disciplines (Charles 2006, 2007; Gledhill 2005; Groom 2005; Hyland 2008a; Peacock 2006). There are also those which have focused on specific sections of an academic text (Brett 1994, Gledhill 2000, Hyland and Tse 2005).

Of special interest to non-native writers and/or translators are cross-linguistic studies which contrast underlying characteristics of English academic prose with their counterparts in other languages (see, for instance, Cortes 2008, Davoodifard 2008, Falahati 2008, Hirano 2009, López-Arroyo and Méndez-Cendón 2007, Vold 2006). Contrastive studies are mainly motivated by the view that translators and non-native novice writers should be aware of both the discursive features of the target language as well as the similarities and differences between the two linguistic systems and cultures. Thus, learning to distinguish between what can be translated literally and what calls for adjustment can help writers achieve a more native-like performance in the target language (Cortes 2008).

Most closely related to the present research are studies which place special emphasis on the language produced by non-native novice writers and investigate what makes it different from the language in published material (for instance, Aktas and Cortes 2008, Cortes 2004, De Cock 2000, Hyland 2008b, Hyland and Tse 2005, Milton and Hyland 1999). The main rationale behind this approach is that foreign and second language "learners admittedly share a number of difficulties with novice native writers but they have also proven to have their own distinctive problems" (Gilquin *et al*. 2007: 320). Thus, when it comes to identifying the main difficulties faced by foreign or second-language learners, the data provided by native corpora will not suffice and need to be complemented with information extracted from learner corpora, that is, corpora containing texts produced by foreign or second language learners (Granger 2002, Gilquin *et al*. 2007, Nesselhauf 2004).

In comparison with published material, academic texts written by non-native novice researchers have been said to present various problems related to the overuse of certain lexical items and word combinations and the underuse or misuse of others (Aktas and Cortes 2008, Cortes 2004, De Cock 2000, Gilquin *et al*. 2007, Hyland 2008b). One plausible explanation for learners' preference for certain sequences over others which may also be possible is the fact that learners tend to have a more restricted repertoire of expressions at their disposal and, as a result, would draw more heavily on them (Gilquin *et al*. 2007). Another reason is that learners' mother tongue may have an impact on their linguistic choices (Gilquin and Paquot 2007, Gilquin *et al*. 2007). These are precisely the issues that the present study aims to investigate.

Here, the focus is on English abstracts of research papers written by Brazilian graduate students from the disciplines of physics, pharmaceutical sciences and computing. In fact, this study builds on previous work (Dayrell 2009a, 2009b, Dayrell and Aluísio 2008), which has identified various lexical and syntactical differences between English abstracts written by Brazilian graduate students vis-à-vis published abstracts from the same disciplines.

Abstracts are of special interest for their relevance in most academic contexts, even where English is not the official language. This is the case in Brazil, where they are expected to be part of most research papers written in Portuguese, be it for academic journals or conference proceedings. However, as Swales and Feak (2009: xiii) point out, constructing an efficient, clear abstract is a fairly difficult task, even for experienced and widely published writers. Abstracts are "highly polished and condensed texts" (Gledhill 2005: 41) in which authors have to capture readers' interest and convince them of the relevance and main claims of the paper (Hyland and Tse 2005). They usually function as a summary of the study in question and serve as the gatekeeper for a number of academic activities (Swales and Feak 2009: 2). Abstracts not only help readers to decide whether to read the whole paper but also allow reviewers to obtain a general picture of the paper they are about to review. Conference organizers may also be inclined to accept or decline a proposed presentation based on intial impressions gained from the abstract.

This chapter investigates the use of sense-related verbs in abstracts of research papers written by Brazilian graduate students from the disciplines of physics, pharmaceutical sciences and computing in comparison with abstracts of published papers from the same disciplines. By sense-related verbs, I refer to verbs whose core meaning is somewhat related, even though they are not necessarily interchangeable. The following sets of sense-related verbs are examined:

(1) *use/apply/employ/utili(s)ze*;
(2) *show/present/demonstrate/display/exhibit*;
(3) *obtain/collect/achieve*;
(4) *find/observe/detect*; and
(5) *study/analys(z)e/investigate/examine*.

The primary purpose of this chapter is to investigate potential differences between abstracts written by Brazilian graduates and those published by established academics in terms of the frequency and lexico-grammatical patterns of the selected verbs. The study then takes a step further and investigates whether the choices made by students can be said to have been influenced by their mother tongue, that is, Portuguese. This is done by examining the frequency and collocational behaviour of cognate translations of the selected verbs in Portuguese abstracts from the above mentioned disciplines.

2. Corpora used in this study

The data analyzed in this chapter are drawn from two independent and separate corpora of English abstracts. One corpus comprises 189 abstracts (40,278 tokens) written by Brazilian graduate students from the disciplines of physics, pharmaceutical sciences (pharmacology, chemistry and biology) and computing. These abstracts were collected in nine courses on academic English writing offered between 2004 and 2009 to graduate students (master's and PhD) from a Brazilian university. The main reason for working with these three disciplines specifically is the fact that these are the only departments that offer such courses. The number of texts from each discipline is determined by the number of abstracts submitted by students and their field of research (Table 23-1). Thus, the higher number of texts in physics is explained by the fact that this department runs these courses annually and has had the largest number of students in each. The department of pharmaceutical sciences also offers courses on an annual basis but, thus far, solely with small groups of students. Computer science has only run two courses.

All abstracts included in the corpus of English abstracts written by Brazilian graduate students (hereafter EN-STS) were written at the very beginning of the courses, before any correction or inclusion of comments and suggestions by instructors, supervisors and/or colleagues. Another point to be made here is that students' level of English varies considerably, ranging from lower intermediate to very advanced levels (further details in Genoves *et al*. 2007).

The other corpus is made up of 1,086 abstracts (187,619 words) taken from papers from the same disciplines (physics, pharmaceutical sciences and computing) published by various leading academic journals. The percentage of texts from each discipline matches the composition of the EN-STS; that is to say, both corpora contain the same percentage of texts by discipline (Table 23-1). For convenience, preference has been given to abstracts of published papers available online. Most published abstracts come from papers of multiple authorship. Given that the abstracts included in the EN-STS are, at least in principle, written by one single student, an attempt has been made to diversify the corpus of published abstracts (EN-PUB) as much as possible in terms of authors included.

Table 23-1. Composition of the EN-STS and EN-PUB corpora

	EN-STS			EN-PUB		
Disciplines	**Number of abstracts**	**%**	**Tokens**	**Number of abstracts**	**%**	**Tokens**
Physics	110	58%	25,613	630	58%	93,707
Pharmac. Sc.	53	28%	10,743	304	28%	67,552
Computing	26	14%	3,922	152	14%	26,360
Total	189	100%	40,278	1,086	100%	187,619

It is also important to stress that, although preference was given to papers from authors affiliated to universities in English speaking countries, the abstracts included in the EN-PUB are not necessarily written by native English speakers. What is valued here is that papers have been accepted by a recognized scientific body for publication and hence presumably meet the required textual quality. This is in line with Swales and Feak's (2009: xi) suggestion that, in today's research world, the number of people who do not have English as their first language has rapidly increased and "the traditional distinction between native speakers and non-native speakers (NNS) of English is collapsing."[1] For Swales and Feak (2009), the more "valid and valuable distinctions" are either between senior and junior researchers or between those who are proficient in academic English as opposed to those with a limited command of it.

In order to examine the influence of students' mother tongue on their output, I have also compiled a corpus of Brazilian Portuguese abstracts (hereafter PT-PUB). This corpus was initially intended to be of similar size and to have the same composition as the EN-PUB. However, for the disciplines under analysis, international journals vastly outnumber Brazilian journals and, to make matters worse, many Brazilian academic journals have English as their official language. Thus, in addition to abstracts of papers published in major Brazilian academic journals, the PT-PUB also includes abstracts published in conference proceedings. Even so, the PT-PUB is smaller. It comprises 610 abstracts (104,290 tokens) and the percentage of texts from each discipline is slightly different from the composition of the EN-PUB (see Table 23-2).

Table 23-2. Composition the PT-PUB corpus

Disciplines	PT-PUB		
	Number of abstracts	Tokens	%
Physics	280	50,294	46%
Pharmac. Sc.	180	33,543	30%
Computing	150	20,453	24%
Total	610	104,290	100%

The next section explains the methodological procedures adopted in this study to retrieve and manipulate the data. All procedures described below are carried out by means of the software package *WordSmith Tools*, version 5.0 (Scott 2007).

3. Methodology

Two basic criteria have been adopted in order to identify five verbs which could serve as the starting point for the analysis: (i) the frequency of the verb in the student corpus; and (ii) the frequency of the verb in academic discourse. The first criterion considers the raw frequency of verbs and selects those with the highest number of occurrences in the EN-STS. It is important to mention that the analysis takes into account lemmas (represented here in small capital letters) rather than individual forms of the verbs. For example, the label STUDY includes *study*, *studies*, *studied* and *studying*. The first criterion is adopted because the focus of this study is on the language produced by students and I am particularly interested in examining whether frequent verbs are used in a similar way by students and published writers. The second criterion selects verbs which would typically occur in academic discourse. This is established by generating a list of "key" verbs in the published corpus, that is to say, verbs whose frequency is unusually high in comparison with a reference corpus.[2] Thus, in order to be selected for analysis, the candidate verb should appear in the keyword list.

Once a given verb has been selected, the next step is to search for its near-synonyms. These have been selected on the basis of the entries in the *Collins English Dictionary and Thesaurus* (2002). In addition, the candidate verb should appear at least once per 10,000 words in either the EN-STS or the EN-PUB. The verb STUDY can serve as an example to illustrate how the near-synonyms have been selected. The following are suggested as its synonyms: *analys(z)e*, *examine*, *investigate*, *look into*, *peruse*, *research*, *scrutini(s)ze*, *survey* and *work over*. The verbs *look into, peruse*, *research*, *scrutinize* and *work over* were discarded because they did not reach the pre-defined minimum frequency in either corpus.

Table 23-3 lists the five sets of verbs selected for analysis. The verbs SHOW and PRESENT were both included among the most frequent verbs in the student corpus and were initially considered as separate entries. However, they are regarded as synonyms by *Collins English Dictionary and Thesaurus* (2002) and this is why they have been grouped together under set-2.

Table 23-3. Verbs selected for analysis

Set	Verbs	Sense-Related Verbs
1.	USE	APPLY, EMPLOY, UTILI(S)ZE
2.	SHOW	PRESENT, DEMONSTRATE, EXHIBIT, DISPLAY
3.	OBTAIN	COLLECT, ACHIEVE
4.	FIND	OBSERVE, DETECT
5.	STUDY	ANALYS(Z)E, INVESTIGATE, EXAMINE

The analysis starts by comparing the overall frequency of each verb in the EN-STS and EN-PUB corpora. This is done by applying a statistical calculation in order to determine whether the difference between the frequencies of each verb in the two corpora is statistically significant. Here, I have opted to compare two Poisson distributions, as suggested by Potthoff and Whittinghill (1966).[3] In many research fields, and the social sciences in particular, it is a standard procedure to adopt the threshold 0.05 for the level of significance.[4] Thus, in this study, the difference between the frequencies of each verb in the EN-STS and EN-PUB corpora is statistically significant if the resulting level of significance (or p-value) is less than 0.05.

The study then takes a step further and investigates whether the lexical choices made by Brazilian graduate students can be said to have been influenced by the Portuguese language. In order to do so, I examine the overall frequency of cognate translations of the selected English verbs in a corpus of Portuguese abstracts (PT-PUB). For instance, for set-1 (Table 23-3), the following Portuguese verbs are examined: USAR [USE], APLICAR [APPLY], EMPREGAR [EMPLOY] and UTILIZAR [UTILI(S)ZE].

In addition to frequency, this study also contrasts the lexico-grammatical patterns yielded by each verb in the two English corpora so as to determine whether students and published writers use them in a similar way. Lexico-grammatical patterns are identified by first retrieving all instances of the search-verb in the EN-STS and the EN-PUB and then examining the items in its surrounding context, within sentence boundaries**.** Any repeated continuous sequence occurring at least once per 10,000 words in either corpus is taken as a lexico-grammatical pattern. In case no continuous sequence stands out, I search for any other kind of regularity, be it in terms of different word-forms of the same lemma, grammatical class or semantic category. Once a given pattern has been identified, I check the remaining concordance lines and look for instances which could be regarded as variations of it. The Poisson distribution is again applied in order to determine whether the difference between the frequencies of each pattern in the EN-STS and the EN-PUB is statistically significant. The influence of the Portuguese language on students' choices of patterns is also examined, taking into consideration cognate translations of the English patterns.

As a note of caution, it is important to mention that the Poisson distribution is applied to raw frequencies. However, normalized frequencies (per 10,000 words) are used in the next section to represent the data in the graphs so that differences between the corpora are more easily spotted.

4. Results

Taking into consideration the verbs within set-1, namely USE, APPLY, EMPLOY and UTILIS(Z)E, we find that both the EN-STS and the EN-PUB show a strong preference for USE (Figure 23-1). However, this tendency is even more marked in the student corpus given that students resort to USE 49% more times per 10,000 words than published writers do. The difference between the frequencies of USE in the two corpora is highly significant ($p < 0.0001$). Interestingly, USAR [USE] is not as frequent in the Portuguese corpus, which shows a clear preference for UTILIZAR [UTILIS(Z)E]. For APPLY, EMPLOY and UTILIS(Z)E, there is no statistically significant difference between their frequencies in the EN-STS and the EN-PUB ($p > 0.05$). In other words, students' lexical choices seem to be in accordance with that of published abstracts, even though their Portuguese counterparts (APLICAR , EMPREGAR and UTILIZAR) are fairly frequent in the PT-PUB.

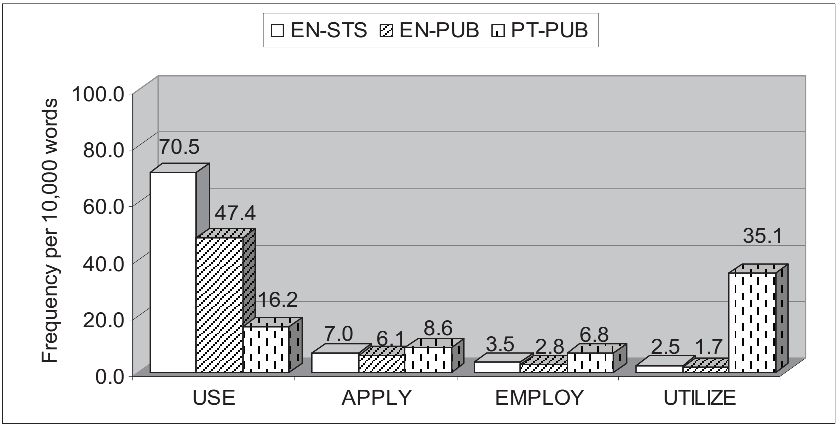


Figure 23-1. Normalized frequencies of verbs within set-1

As regards set-2 (i.e. SHOW, PRESENT, DEMONSTRATE, EXHIBIT and DISPLAY), SHOW is the preferred verb in both students' and published English corpora (Figure 23-2), and the difference between its frequencies in these two corpora is not statistically significant ($p > 0.05$). The verb PRESENT is considerably more frequent in the EN-STS than in the EN-PUB ($p < 0.0001$) and this preference seems to be driven by the influence of Portuguese. In the PT-PUB, its equivalent (APRESENTAR) is far more frequent than all other verbs within set-2. Both DEMONSTRATE and EXHIBIT are much more frequent in the EN-PUB than in the EN-STS ($p < 0.0001$ in both cases), which may be an indication that students do not feel very comfortable in using these two verbs. For EXHIBIT, students' choices seem to be related to the influence of the Portuguese language; its frequency in the EN-STS is fairly similar to the frequency of its cognate translation (EXIBIR) in the PT-PUB. However, the same cannot be said about DEMONSTRATE. The frequency of its Portuguese counterpart (DEMONSTRAR) doubles its frequency in the EN-STS. As for DISPLAY, it is the least frequent verb in both in the EN-STS and the EN-PUB and the difference between its frequencies in the two corpora is not statistically significant ($p > 0.05$).

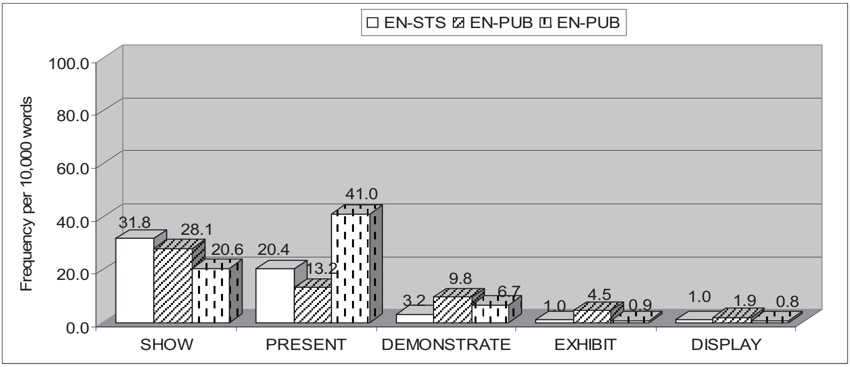


Figure 23-2. Normalized frequencies of verbs within set-2

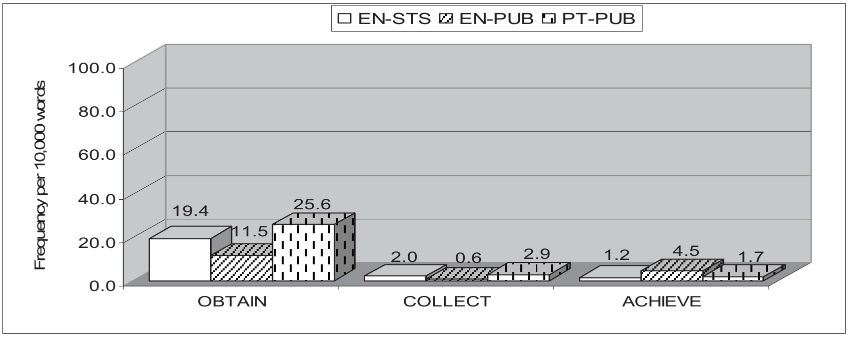


Figure 23-3. Normalized frequencies of verbs within set-3

As for the verbs within set-3 (OBTAIN, COLLECT and ACHIEVE), we find that OBTAIN is particularly salient in both student and published English corpora (Figure 23-3). However, as in the cases of USE and SHOW, the tendency is more marked in students' abstracts ($p < 0.0001$), which seems to be due to the influence of students' first language. Its Portuguese counterpart (OBTER) is highly frequent in the PT-PUB. COLLECT and ACHIEVE are rather infrequent in both English corpora. However, the differences between their frequencies in the EN-STS and the EN-PUB are both statistically significant ($p < 0.05$). COLLECT is more frequent in the student whereas ACHIEVE is more frequent in the published corpus. In both cases, the frequencies in the student corpus seem to reflect the frequencies of their cognate translations (COLETAR and ALCANÇAR).

For the verbs within set-4 (FIND, OBSERVE and DETECT), it is interesting to note that while students use both FIND and OBSERVE with similar frequencies, published abstracts exhibit a clear preference for the former (Figure 23-4). In fact, FIND is far more salient in the EN-PUB than in the EN-STS and the difference between its frequencies in the two corpora is statistically significant (p < 0.05). The lower frequency of FIND in the student corpus seems to be influenced by the lower frequencies of its translation cognates (ENCONTRAR and ACHAR) in Portuguese abstracts. Interestingly, both OBSERVE and DETECT occur with similar relative frequency in the two English corpora (p > 0.05) and they are in accordance with the frequencies of their counterparts (OBSERVAR and DETECTAR) in the PT-PUB.

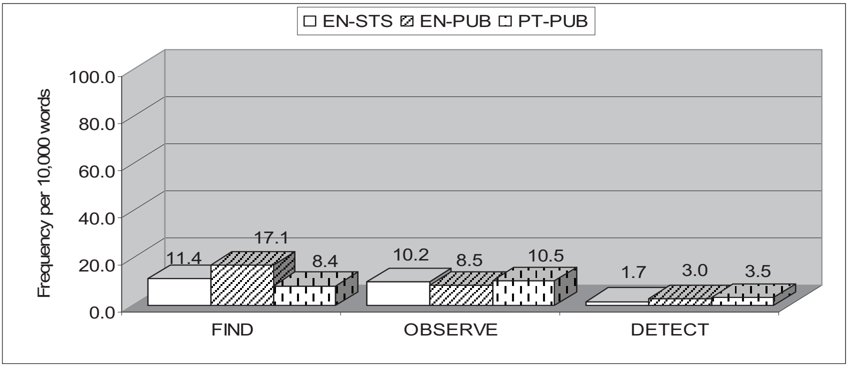


Figure 23-4. Normalized frequencies of verbs within set-4

As regards set-5 (Figure 23-5), three verbs (STUDY, ANALYS(Z)E and INVESTIGATE) occur with similar frequencies in student and published English abstracts (p > 0.05). However, the slightly higher frequency of ANALYS(Z)E and lower frequency of INVESTIGATE in the EN-STS may be related to the frequencies of their Portuguese translations (ANALIZAR and INVESTIGAR respectively) in the PT-PUB. Within set-5, the main difference between the EN-STS and the EN-PUB is related to the use of EXAMINE, which is much more frequent in the latter (p < 0.0001). The lower frequency in the EN-STS seems to be influenced by the fact that its Portuguese equivalent (EXAMINAR) is also very infrequent in the PT-PUB.

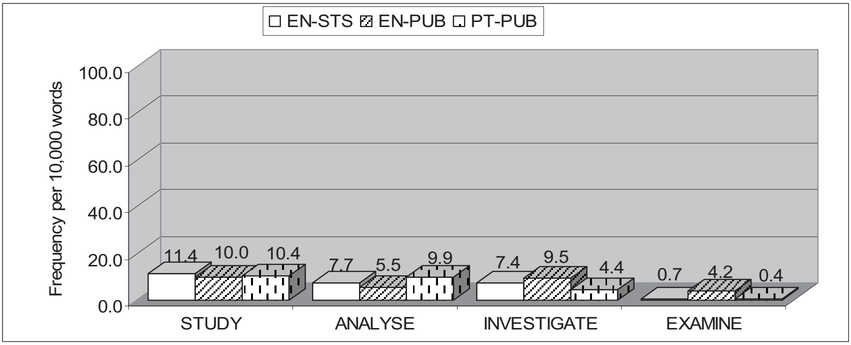


Figure 23-5. Normalized frequencies of verbs within set-5

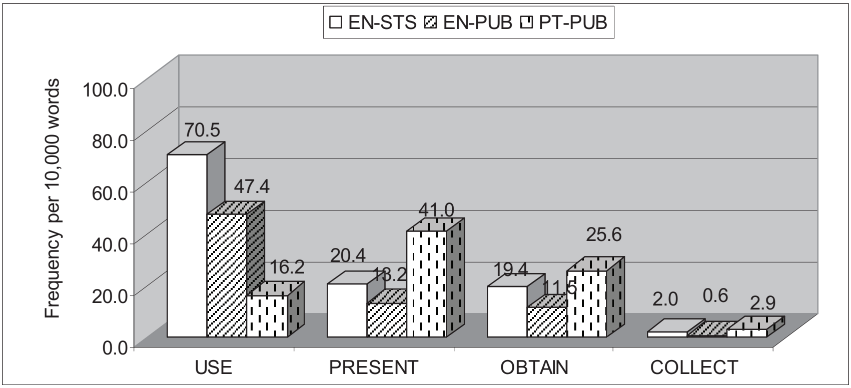


Figure 23-6. Verbs with a significantly higher frequency in students' abstracts

5. Discussion

Taking into consideration the 19 verbs analyzed in this study, the frequency counts indicate that 10 verbs (53%) are used with similar frequency in student and published English abstracts. However, nine verbs show significant differences in their frequency patterns in the two corpora. These findings can be used to draw students' attention to verbs they either over or underuse.

Four verbs (USE, PRESENT, OBTAIN and COLLECT) occur with a significantly higher frequency in students' than in published English abstracts (Figure 23-6). With the exception of USE, the higher frequency of the verbs in the EN-STS seems to reflect the frequencies of their Portuguese equivalents in the PT-PUB.

By way of contrast, five verbs (DEMONSTRATE, EXHIBIT, ACHIEVE, FIND and EXAMINE) occur with significantly lower frequency in the EN-STS when compared with the EN-PUB (Figure 23-7). With the exception of DEMONSTRATE, the lower frequencies in the EN-STS seem to be due to the interference of students' mother tongue given that their Portuguese equivalents are also infrequent in the PT-PUB.

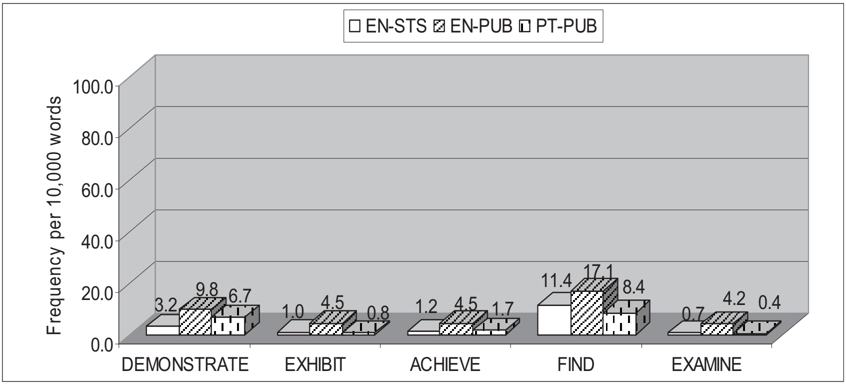


Figure 23-7. Verbs with a significantly lower frequency in students' abstracts

As regards the lexico-grammatical patterns yielded by the selected verbs, we find various similarities and differences between the two English corpora. Due to constraints of space, the present discussion will be restricted to the most dominant features.

The first aspect that stands out is the high frequency of the structure *BE* + past participle, which is overwhelmingly used in both student and published English abstracts. Portuguese abstracts follow the same tendency and also show high percentages of passive constructions.[5] These findings do not come as a surprise given that the passive structure is widely recognized as a common feature in academic discourse (Carter and McCarthy 2006: 277). What is interesting to note here is that there are some substantial differences between the EN-STS and the EN-PUB in relation to individual verbs.

Taking into consideration the verbs within set-1 (USE, APPLY, EMPLOY and UTILIS(Z)E), most instances in both corpora refer to *BE used* but its frequency in the EN-STS (26.6 occurrences per 10,000 words) is twice as high as that in the EN-PUB (13.3). The higher frequency in the

EN-STS holds true for all four verbs within set-1. Students use about twice as many passive constructions as published writers: 33.3 occurrences per 10,000 words in the EN-STS versus 17.5 in the EN-PUB ($p < 0.0001$). Overall, set-1 verbs are used in the passive form to describe methods and materials (e.g. *the method is applied to* and *p-nitrophenyl acetate was used*). Although there are various ways of doing so, the figures may be an indication of students' overuse of passive constructions to describe methodological procedures. This is possibly due to first language interference; the passive forms of the cognate translations (USAR, APLICAR, EMPREGAR and UTILIZAR) are also highly frequent in the PT-PUB.

Set-2 shows an opposite tendency as passive constructions are far more frequent in the EN-PUB than in the EN-STS: 4.2 occurrences per 10,000 words in the EN-STS and 12.6 in the EN-PUB ($p < 0.0001$). This tendency is especially evident for the verbs SHOW, PRESENT and DEMONSTRATE (2.0, 1.5 and 0.2 instances per 10,000 words respectively in the EN-STS, and 5.3, 5.1 and 2.1 in the EN-PUB). For SHOW, there is the additional fact that while published writers choose to use *BE shown*, students consistently employ *BE showed,* which is the unusual form of the past participle (Hornby 2005: 1410). Interestingly, the frequencies of *BE showed/shown* and *BE demonstrated* in the EN-STS are fairly similar to the frequencies of their equivalents in Portuguese (2.5 and 0.7). But the same cannot be said about PRESENT since the passive form of its cognate translation figures prominently in the PT-PUB (8.7).

For sets 3, 4 and 5, if we take into consideration the passive constructions yielded by all verbs within each set, we find that the overall frequencies are fairly similar in both English corpora ($p > 0.05$). What is interesting to note here is the behaviour of some individual verbs. While *BE obtained* is used with fairly similar frequencies in the two English corpora, *BE collected* is more prominent in the EN-STS and *BE achieved* is hardly used by students. This tendency seems to reflect the use of their cognate translations in Portuguese. For set-4, students' preferred pattern is *BE observed*, whose Portuguese counterpart is also highly frequent in the PT-PUB. Published writers tend to opt for *BE found*. For set-5, while *BE studied* and *BE analys(z)ed* are more frequent in the EN-STS, *BE investigated* is used with fairly similar frequencies in the two English corpora and *BE examined* is not used by students.

The most striking difference between the EN-STS and the EN-PUB is related to the use of *we* as the subject of the search-verb (e.g. *we show that*). The active voice figures prominently in the EN-PUB, especially with the verbs SHOW, PRESENT, DEMONSTRATE, FIND, STUDY and

INVESTIGATE, but it is not as recurrent in student writing (Table 23-4). The only exception is *we USE*, which occurs with similar frequency in both English corpora. This overall lower proportion of the active voice in student abstracts seems, at least in principle, to support the idea that learners tend to be uncomfortable with explicitly committing themselves to a particular proposition or evaluation (Hyland 2008a, 2008b, Hyland and Tse 2005). Students' choices may also be related to cultural practices and preferences. Unlike English, the use of first person pronouns is somewhat discouraged in academic Portuguese, which tends to rely more strongly on impersonal writing such as the passive voice. However, in the Portuguese abstracts examined in the present study, five verbs are frequently used with the first person pronoun as subject. They are MOSTRAR [SHOW], APRESENTAR [PRESENT], OBTER [OBTAIN], OBSERVAR [OBSERVE], and ESTUDAR [STUDY].

Table 23-4. Normalized frequencies of the pattern *we SEARCH-VERB* with the highest frequency in the EN-PUB

Patterns	**EN-STS**	**EN-PUB**	**Portuguese counterpart**
We USE	2.2	2.0	0.5
We SHOW	2.0	8.7	2.1
We PRESENT	2.0	5.2	3.9
We DEMONSTRATE	0.5	2.7	0.6
We FIND	1.5	6.4	0.3
We STUDY	0.2	3.2	2.3
We INVESTIGATE	0.2	3.3	1.2

Clear differences are also seen between the two English corpora when we examine the surrounding context of the passive forms of the verbs SHOW and FIND. To start with, students rarely employ a verb in the infinitive form after the passive form of these verbs (such as *was shown to have* and *is found to be*). This construction is particularly salient in the EN-PUB: 2.5 and 4.6 per 10,000 words respectively. Anticipatory *it* patterns with verb complementation do not seem to be an option for Brazilian students either. The sequences *it BE shown that* and *it BE found that* are fairly frequent in the EN-PUB (2.6 and 1.1 per 10,000 words respectively) but are rarely used by students. These lexico-grammatical patterns are examples of constructions which allow writers to conceal the source of the evaluation and present knowledge in an objective and uncontested way (Groom 2005, Hyland and Tse 2005). This discourse strategy helps “strengthen a claim as it simultaneously removes any implication of personal interest from the comment and adds rhetorical credibility” and is therefore frequently used in English academic writing (Hyland and Tse 2005: 133).

The lower frequency of these two constructions in the EN-STS may be explained by the fact that they have no direct translations into Portuguese. The closest equivalent for both is perhaps the pronominal passive (see note 5). However, with the cognate translation of SHOW [MOSTRAR], it occurs only 0.5 times per 10,000 words in the PT-PUB and no occurrences are found for the Portuguese counterparts of FIND [ENCONTRAR and ACHAR].

Interestingly, rather than using passive constructions followed by an infinitive form or anticipatory *it* patterns, Brazilian students distance themselves from their interpretation of the findings by shifting attention to the research itself and highlighting the legitimacy of the results and effectiveness of the adopted methods and techniques. The sequence *results SHOW that* is nearly twice as frequent in the EN-STS (4.7 per 10,000 words) as in the EN-PUB (2.7). Its Portuguese equivalent (*os resultados MOSTRAR que*) is also highly frequent in the PT-PUB (5.7 times per 10,000 words). Thus, in this case, students’ choice seems to have been influenced by their mother tongue. At the same time, *results DEMONSTRATE that* occurs with similar frequency in the EN-STS and the EN-PUB (1.0 and 0.9 per 10,000 words respectively) and the frequency of its Portuguese equivalent (*os resultados DEMONSTRAR que*) is also fairly similar (1.2 times per 10,000 words).

Lexical items such as *analysis*, *tests*, *experiments* and *observations*, which refer to the research procedures, are also used as the subject of SHOW and DEMONSTRATE. In the EN-STS, they are highly frequent with SHOW (3.2 per 10,000 words) but are not used with DEMONSTRATE. In both English and Portuguese published abstracts, these lexical items are not very frequently used as the subject of these two verbs, occurring not more than once per 10,000 words.

Similarly, the proportion of lexical items referring to either the study being described in the abstract or other related studies (*paper*, *work*, *study*, *studies*, *project*, *research* and *review*) as the subject of SHOW and DEMONSTRATE is also much higher in the student corpus. They occur 3.7 per 10,000 words in the EN-STS in comparison with 1.7 in the EN-PUB. Portuguese does not seem to have an impact on students’ choices, as the Portuguese counterparts of those lexico-grammatical patterns occur once per 10,000 words.

This semantic group is also used as the subject of PRESENT, STUDY, ANALYS(Z)E, INVESTIGATE and EXAMINE when the writers wish to state the purposes of their research (e.g. *this paper presents* or *the current study investigated*). Here again, they are much more frequent in student writing than in published English abstracts. In the Portuguese corpus, this lexico-grammatical pattern is very frequent with the equivalent of PRESENT (APRESENTAR), which perhaps explains its prominence in the EN-STS. However, the same cannot be said about the other verbs as their Portuguese counterparts rarely occur in the PT-PUB (no more than 0.5 times per 10,000 words with each verb).

Last but not least, USE is the only verb whose use in context seems fairly similar in both students' and published English abstracts, even though it is far more frequent in the former. With the exception of the pattern (*BE used*), which as mentioned earlier is considerably more frequent in the EN-STS, three recurrent patterns occur with very similar frequency in both English corpora ($p > 0.05$). They are (i) *we USE*, (ii) *BE used to*-infinitive (e.g. *was used to represent*) and (iii) *BE* past-participle *(by)* gerund (e.g. *were prepared using*).

6. Concluding remarks

The main purpose of this chapter has been to investigate the use of five sets of sense-related verbs in English abstracts of research papers written by Brazilian graduate students. The analysis identified some verbs that occur with a significantly higher frequency and others with a significantly lower frequency in student writing when compared with abstracts of published papers from the same disciplines. Substantial differences have also been found with respect to the lexico-grammatical patterns yielded by the verbs under consideration. For instance, students do not seem to feel very comfortable with using the active voice (e.g. *we show that*) and hence explicitly stating their commitment to a particular proposition. At the same time, although passive constructions (*BE* + past participle) were overwhelmingly used in both student and published English abstracts, students seem unaware that they can be followed by an infinitive form (e.g. *was shown to have*) or preceded by a dummy *it* (i.e. *it was found that*). By way of contrast, students remain in the background by making frequent use of impersonal subjects such as *the results*, *the analysis* and *this paper*, thereby shifting attention to the results, methodological procedures and the research itself.

Most differences between student and published English abstracts identified in this study seem to be related to the influence of students' mother tongue on their linguistic choices. This is perhaps due to lack of awareness of the main similarities and differences between English and Portuguese academic discourses, which has opened up space for first language interference.

There are, however, reasons for caution as the current study has some obvious limitations. The size of the student corpus is undeniably modest. Clearly, additional benefits could be gained from the analysis of a larger set of data so that the tendencies identified here could be further validated. In addition, disciplinary variations have not been considered. This is important to stress because several studies (Charles 2006, 2007; Cortes 2004; Groom 2005; Hyland 2008b; Peacock 2006) have shown that the frequency and behaviour of individual lexical items may vary across disciplines. This in turn suggests that phraseological characteristics need to be considered in relation to their discipline-specific use. Another limitation of this study is that further work is still needed on the discourse strategies of English abstracts. In order to write more effectively, it is crucial that learners are able to establish relationships between form, structure, and rhetorical functions.

Despite these drawbacks, the present study has offered important contributions to pedagogic practice and useful insights for teachers and materials writers. Drawing students' attention to the main differences between their own and published writing can help them identify distinctive features of their language and the preferred phraseological patterns of their academic discourse community. Learners are also most likely to benefit from consciousness raising tasks focusing on recurring linguistic features of the target genre and the rhetorical motivations behind their selection. This would undoubtedly enhance their understanding of the kind of text they are expected to write.

Notes

* I am especially grateful to Dr. Adalberto Pessoa Jr. (Faculty of Pharmaceutical Sciences), Dr. Sandra Aluísio (Institute of Computer and Mathematical Sciences) and Dr. Valtencir Zucolotto (Physics Institute of São Carlos), coordinators of the courses on academic writing offered at the University of São Paulo (Brazil), for having kindly granted me access to the data used in this study. I would also like to thank Dr. German Moreno Arenas and Dr. Júlio Singer from the Institute of Mathematics and Statistics of the University of São Paulo for their invaluable help with the statistical calculations. Thanks are also due to FAPESP (Brazil) for providing the financial support needed in this research (Grant Reference 2007/52405-3).

1. See Williams (2006) for a discussion of the various issues around the concept of native speaker.

2. For the purposes of this study, I have used the British National Corpus (BNC) as the reference corpus. The BNC is a 100-million-word corpus of texts originally produced in English. Further information is available at http://info.ox.ac.uk/bnc.
3. The Poisson formula shows the probability of *r* events occurring in *n* number of trials (Oakes 1998: 6). In the specific case of this study, the statistical calculation tests the null hypothesis that the actual frequencies of each verb (observed frequencies) in the two corpora show the same probability distributions. It uses a chi-square distribution with one degree of freedom. This test has been chosen because, as Oakes (1998: 209) explains, "in the Poisson model, the occurrence of a given word is independent of the previous occurrence of that word – it depends only on its overall rate of occurrence."
4. The level of significance is "the point at which the difference between what is observed and what is expected is too great to be due to chance or random variation" (Kurtz 1999: 151). Thus, a level of significance of 0.05 means that "the null hypothesis is not rejected unless there are fewer than five chances in 100" of obtaining that particular result (Oakes 1998: 9). In other words, the probability that the null hypothesis is wrongly rejected is no more than 5%.
5. In Portuguese, the passive voice can be of two types: (i) the analytic type corresponds to the passive structure in English, in the form of an auxiliary verb followed by the main verb in the past participle form (e.g. *é usado* 'is used'); and (ii) the pronominal passive is formed by a main verb in the third person inflection associated with the passive pronoun *se* (e.g. *usa-se*) (Cegalla 2000: 205-206). Both types are considered in the present chapter.

References

Aktas, R. N. and Cortes, V. (2008), "Shell nouns as cohesive devices in published and ESL student writing". *Journal of English for Academic Purposes* 7: 3-14.

Biber D., Conrad, S. and Cortes, V. (2004), "*If you look at* ...: Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25(3): 371–405.

Brett, P. (1994), "A genre analysis of the results section of sociology articles". *English for Specific Purposes* 13(1): 47–59.

Carter, R. and McCarthy, M. (2006), *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge: Cambridge University Press.

Cegalla, D. P. (2000), *Novíssima Gramática da Língua Portuguesa* (43rd edition). São Paulo: Companhia Editora Nacional.

Charles, M. (2006), "Phraseological pattern in reporting clauses used in citation: A corpus-based study of theses in two disciplines". *English for Specific Purposes* 25: 310-331.

—. (2007), "Argument or evidence? Disciplinary variation in the use of the noun *that* pattern in stance construction". *English for Specific Purposes* 26: 203-218.

Collins (2002), *Collins English Dictionary and Thesaurus* (Version 3.0). London: Harper-Collins Publishers Ltd.

Cortes, V. (2004), “Lexical bundles in published and student disciplinary writing: Examples from history and biology.” *English for Specific Purposes* 23: 397-423.

—. (2008), “A comparative analysis of lexical bundles in academic history writing in English and Spanish”. *Corpora* 3(1): 43-57.

Davoodifard, M. (2008), “Functions and hedges in English and Persian academic discourse: Effects of culture and the scientific discipline”. *ESP Across Cultures* 5: 23-48.

Dayrell, C. (2009a), “Sense-related verbs in English scientific abstracts: A corpus-based study of students’ writing”. *ESP Across Cultures* 6. [In press]

—. (2009b), “Lexical bundles in English abstracts: A corpus-based study of published and non-native graduate writing”, in *Proceedings of the Fifth Corpus Linguistics Conference (CL2009)*, Liverpool, 21-23 July 2009.

Dayrell, C. and Aluísio, S. (2008), “Using a comparable corpus to investigate lexical patterning in English abstracts written by non-native speakers”, in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Workshop Building and Using Comparable Corpora*, 61-72. Marrakech, 31st May 2008. Available at: http://www.lrec-conf.org/lrec2008/.

De Cock, S. (2000), “Repetitive phrasal chunkiness and advanced EFL speech and writing”, in C. Mair and M. hundt (eds.) *Corpus Linguistics and Linguistic Theory*, 51-68. Amsterdam/Atlanta: Rodopi.

Falahati, R. (2008), “A contrastive study of hedging in English and Farsi academic discourse”. *ESP Across Cultures* 5: 49-68.

Genoves Jr., L., Lizotte, R., Schuster, E., Dayrell, C. and Aluísio, S. (2007), “A two-tiered approach to detecting English article usage: An application in scientific paper writing tools”, in *Proceedings of the International Conference RANLP´2007*, 225-239. Borovetz, Bulgaria, 26th September 2007.

Gilquin, G. and Paquot, M. (2007), “Spoken features in learner academic writing: Identification, explanation and solution”, in *Proceedings of the Fourth Corpus Linguistics Conference*. University of Birmingham, 27-30 July 2007.

Gilquin, G., Granger, S. and Paquot, M. (2007), “Learner corpora: The missing link in EAP pedagogy”. *Journal of English for Specific Purposes* 6: 319-335.

Gledhill C. (2000), “The discourse function of collocation in research article introductions”. *English for Specific Purposes* 19: 115-135.

—. (2005), *Collocations in Science Writing*. Tübingen: Gunter Narr Verlag.

Granger, S. (2002), “A bird’s-eye view of learner corpus research”, in S. Granger, J. Hung and S. Petch-Tyson (eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, 3-33. Amsterdam/Philadelphia: John Benjamins Publishing.

Groom, N. (2005), “Pattern and meaning across genres and disciplines: An exploratory study”. *English for Academic Purposes* 4: 257-277.

Hirano, E. (2009), “Research article introductions in English for specific purposes: A comparison between Brazilian Portuguese and English”. *English for Specific Purposes* 28: 240-250.

Hornby, A. S. (2005), *Oxford Advanced Learner's Dictionary* (7th edition). Oxford: Oxford University Press.

Hyland, K. (2008a), “As can be seen: Lexical bundles and disciplinary variation”. *English for Specific Purposes* 27: 4-21.

—. (2008b), “Academic clusters: Text patterning in published and postgraduate writing”. *International Journal of Applied Linguistics* 18(1): 41-61.

—. (2009), *Academic Discourse*. London/New York: Continuum.

Hyland, K. and Tse, P. (2005), “Hooking the reader: A corpus study of evaluative *that* in abstracts”. *English for Specific Purposes* 24: 123-139.

Kurtz, N. R. (1999), *Statistical Analysis for the Social Sciences*. Boston: Allyn and Bacon.

López-Arroyo, B. and Méndez-Cendón, B. (2007), “Describing phraseological devices in medical abstracts: An English/Spanish contrastive analysis”. *Meta* 52 (3): 503-516.

Milton, J. and Hyland, K. (1999), “Assertions in students’ academic essays: A comparison of English NS and NNS student writers”, in R. Berry, B. Asker and K. Hyland (eds.) *Language Analysis, Description and Pedagogy*, 147-161. Hong Kong: Language Centre, HKUST.

Nesselhauf, N. (2004), “Learner corpora and their potential for language teaching”, in J. McH. Sinclair (ed.) *How to Use Corpora in Language Teaching*, 125-152. Amsterdam/Filadelphia: John Benjamin Publishing.

Oakes, M. P. (1998), *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Peacock, M. (2006), “A cross-disciplinary comparison of boosting in research articles”. *Corpora* 1(1): 61-84.

Potthoff, R. F. and Whittinghill, M. (1966), “Testing for homogeneity: II. The Poisson distribution”. *Biometrika* 53 (1/2): 183-190.

Scott, M. (2007), *WordSmith Tools* (Version 5). Oxford: Oxford University Press.

Swales, J. M. and Feak, C. B. (2000), *English in Today's Research World: A Writing Guide*. Michigan: The University of Michigan Press.

Vold, E. T. (2006), "Epistemic modality markers in research articles: A cross-linguistic and crossdisciplinary study". *International Journal of Applied Linguistics* 16(1): 61-87.

Williams, G. (2006), "Challenging the native-speaker norm: A corpus-driven analysis of scientific usage", in G. Barnbrook, P. Danielsson and M. Mahlberg (eds.) *Meaning Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, 115-127. London/New York: Continuum.

CHAPTER TWENTY-FOUR

A CORPUS-BASED CONTRASTIVE STUDY OF REPORTING IN ENGLISH MA THESES

YUECHUN JIANG, ZHIQING HU

1. Introduction

Reporting, in a broad sense, means the account of what people say or do or think with the evaluation of the reporter or writer; narrowly, according to Thomas and Hawes (1994), it refers to making references to previous research embedded in a continuous academic research article. Whether the writers have used reporting appropriately or not has great impact on the quality of academic articles. However, it seems that reporting is also a source of considerable difficulty for most EFL writers (Cadman 1997, Thompson 2000), which may undermine the theoretical basis and credibility of their research articles.

In the past the research on reporting usually fell into the traditional "indirect or direct speech" categories (Thompson 1996, Peng 2003). Different from the traditional approach, recent studies tend to provide a panorama of reporting from different perspectives.

Many studies of reporting have been published recently (Thomas and Hawes 1994; Thompson 1994; Hyland 1999, 2000, 2001, 2002, 2005; He 2000; Thompson 2000; Huang 2001; Charles 2003; Okamura 2003; Chen 2006). Roughly speaking, these studies of reporting take the following approaches, namely the stylistic approach (Leech and Short 1981, Shen 1991), the general linguistic approach, i.e. the lexical, syntactic and semantic approach (Swales 1986, 1990; Thompson 1996), the pragmatic-functional approach (Thompson 1994, 2001; Xin 1998; Jia 2000; Tang 2004) and the cognitive approach (Peng 2001, 2003).

However, to a large extent, previous studies of reporting did not pay enough attention to the reporting in learner English, especially the features of reporting in academic discourse.

This chapter investigates reporting in English MA theses, written by Chinese learners of English (CLE) and native speakers of English (NSE). We will seek to address the following research questions:

1) What features are there in MA theses in terms of reporting forms, and are there any differences in MA theses by CLE and NSE respectively?
2) What features are there in MA theses in terms of reporting structures, and are there any differences in MA theses by CLE and NSE respectively?
3) Have CLE used reporting appropriately in terms of function?

To answer these questions, an analysis of MA theses was conducted mainly from a functional and general linguistic perspective, in an attempt to uncover, on the one hand, the writers' willingness to reveal or obscure the presence of the reported authors and their awareness of clarity, while on the other hand, the diversity of the reporting practice of the CLE and NSE groups and how well reporting serves CLE's purposes.

The corpus contains 13 MA theses altogether, with seven written by CLE and the remaining six by NSE, all of which were randomly collected from the Internet. The disciplines covered range from language and communication to biology etc. The CLE sub-corpus contains 100,519 words, while the NSE component amounts to 100,218 words. After constructing the corpus, we read it line by line and inserted tags after each reporting phenomenon (excluding self-reporting). The annotated corpus was analyzed with the Wordsmith Tools, applying the Chi-square test for statistical significance.

2. Reporting forms

The reporting form is a functional category, considered as revealing or obscuring the authors' presence in texts and the writers' willingness for clarity. Based on this function, four categories are identified in the present study: integral vs. non-integral reporting, non-citation, and stative reporting.

2.1. Integral and non-integral reporting

The choice between integral and non-integral reporting indicates different focuses of writers, i.e. on the researchers who have put forward those findings, or on the findings that have been made (Thompson 2000).

As Thompson put it, if the reporting is in the form of an author name followed by the publication year, typically the name of the cited author will be incorporated as an integral part of the syntax of the sentence, and will not be separated by brackets. This is called integral reporting. Obviously, integral reporting appears within sentences. For example:

> Kent and Taylor (1998) ***suggest*** that dialogic communication involves a relational interaction.

On the contrary, non-integral reporting is separated from the sentence by brackets and it plays no explicit grammatical role in the sentence. Reporting can also take the form of a number, rather than author name and publication year (Thompson 2000), telling the reader where the information (verbal or numerical) or idea comes from. The function of the reporting is that of attribution. For instance:

> Additionally, research has found that 74% of consumers said that they were prepared to switch brands if a similar brand was associated with a worthy cause (Adkins & Kowalska 1997; New Zealand Marketing Magazine 2000, as cited in McAlister & Ferrell 2002).

Table 24-1. Frequency of reporting forms in the CLE and NES sub-corpora

Type	Freq. in CLE's sub-corpus	Freq. in NSE's sub-corpus	Difference
Integral	268	161	107**
Non-integral	99	143	-44*
Non-citation	80	98	-18
Stative	25	34	-9
Total	472	436	36

(Note: * significant difference of normalized frequency at .01 level, ** significant difference at .0001 level)

Table 24-1 shows that there is much more integral reporting in CLE's sub-corpus, with 268 in CLE's sub-corpus against 161 in NSE's sub-corpus. On the other hand, the normalized frequency of non-integral reporting is 99 in CLE's sub-corpus in contrast with 143 in NSE's sub-corpus (see Figure 24-1). These differences in frequencies are significant at .0001 or .01 level.

The fact that CLE used more integral reporting indicates their willingness to reveal, rather than obscure, the reported authors' presence. CLE theses pay more attention to the individuals who have developed approaches, formulated equations, or articulated complex models, so that these individuals play explicit grammatical roles within sentences.

On the contrary, NSE theses use more non-integral reporting, which suggests NSE's tendency to obscure the reported authors' presence. They tend to focus on previous findings or suggestions rather than on the researchers that have made the findings or suggestions, which suggests a more "impersonal" style in the NSE writing.

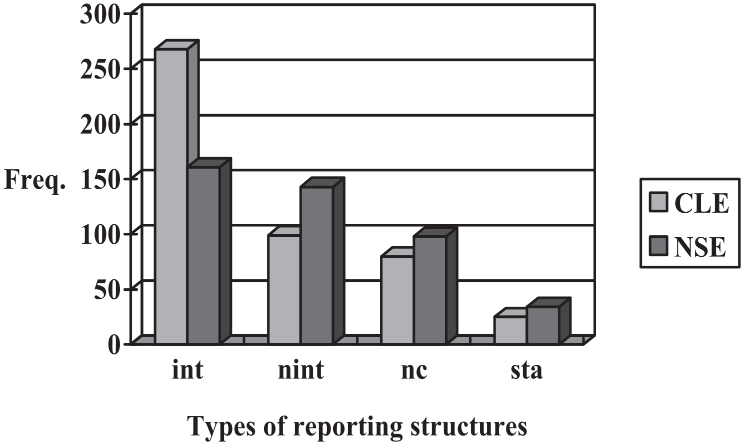


Figure 24-1. Distribution of reporting structures in the CLE and NSE sub-corpora (Note: In this figure, "int" signifies integral reporting; "nint"—non-integral reporting; "nc"—non-citation; "sta"—stative reporting.)

2.2. Non-citation and stative reporting

Non-citation and stative reporting are less frequently used than integral and non-integral reporting; however, these two forms are also important in that they serve as indicators of writers' awareness of clarity.

In the reporting form of non-citation, an occurrence of the author name in the text does not appear as a citation, i.e. no year, or page, reference attached to the name (Thompson 2000). These "non-citations" occur, of course, after the researcher has already been reported. For example:

> Daugherty (2000) suggests that contributions of employee time and talents provide more public relations and marketing benefits. She ***notes*** that forms of voluntarism are seen as more sincere and provide additional benefits for employees who share a unified goal outside of the workplace.

In this example, the writer used *notes* to elaborate the reported author Daugherty's suggestion after an integral reporting and there is no year or page, or reference attached to the reporting. This is how a typical non-citation is employed, providing more specific information after a prior reporting, thus increasing the clarity of the reporting (at least it is an indicator of the writer's intention to make it clearer).

Exceptions to this are instances where names are used to identify a convention, an established fact or a theory or other such commonly recognized construct. Since such convention or theory etc. is generally known, or at least known to the target readers according to the writer's presumption, the reporting form with no year or reference does not prevent readers from understanding the content. In this case, this is also counted as non-citation.

It can be seen from Table 24-1 that there are 80 cases of non-citation in CLE's sub-corpus and 98 cases in the NSE data, and no significant difference is reported. It suggests that CLE also took care of the clarity of the writing, and tried to reduce ambiguity by providing more information after a previous reporting by elaboration.

The fourth reporting form is stative reporting, in which there is also no year, or page, or reference attached to the author name, but it refers to previous research in a general and independent way. By "independent", we mean that there is no directly related reporting previously mentioned. In other words, there is no specific source for the stative reporting; consequently, excessive use of this form of reporting increases ambiguity. For example:

> Finally, some researchers simply described a group of writers' products to aid in pedagogy.

The frequency of the stative reporting is slightly lower in CLE's sub-corpus than in NSE's sub-corpus (25 versus 34), with no significant difference found (Table 24-1). The frequencies in both datasets are low, which might suggest that CLE and NSE did not take it as a usual form of reporting; the similar frequencies signify that CLE postgraduates might be aware of or subconscious of the function of stative reporting, since stative reporting is general and independent and it does not provide the information source as most reporting cases do.

3. Reporting structures

Reporting structures are an indicator of writing diversity, including finite reporting verbs (RVs), RVs in present participle and in past participle, reporting adjuncts, reporting nouns and reporting adjectives (Thompson 1994).

Table 24-2. Distribution of reporting structures in CLE's and NSE's sub-corpora

Type	Freq. in CLE's sub-corpus	Freq. in NSE's sub-corpus	Difference
Finite RVs	496	451	45
Reporting adjuncts	59	158	-99**
Reporting nouns	15	53	-38**
RVs in present participle	11	59	-48**
RVs in past participle	10	39	-29**
Reporting adjectives	7	7	0
Total	598	767	-169**

(Note: **significant difference of normalized frequency at .0001 level)

A total of 598 reporting cases in CLE's sub-corpus and 767 cases in the NSE data are extracted, which indicates significantly higher density of reporting in general in NSE's sub-corpus. Table 24-2 displays the contrastive distribution of reporting structures in the two sub-corpora. It is clear from the table that finite RVs prevail over reporting cases in both sub-corpora. Numbers of other reporting structures are relatively close. Figure 24-2 shows the distribution more clearly. CLE's bar starts high but drops sharply, with NSE's bar remaining relatively stable, which suggests some reporting structures like reporting nouns, RVs in present participle and in past participle occur in CLE's sub-corpus in a very limited number. It is interesting to note that the numbers of reporting adjectives adopted in both sub-corpora are the same.

The analysis of reporting structures is revealing. Interesting variations in the reporting structures were found between CLE's and NSE's sub-corpora. NSE appeared to employ a much larger number of reporting cases in their theses, with more diversity in each reporting structure. The relatively high use of various reporting structures may result from their better mastery of the language skills and awareness to use diverse reporting structures to avoid monotony. A closer investigation in the following will provide detailed information.

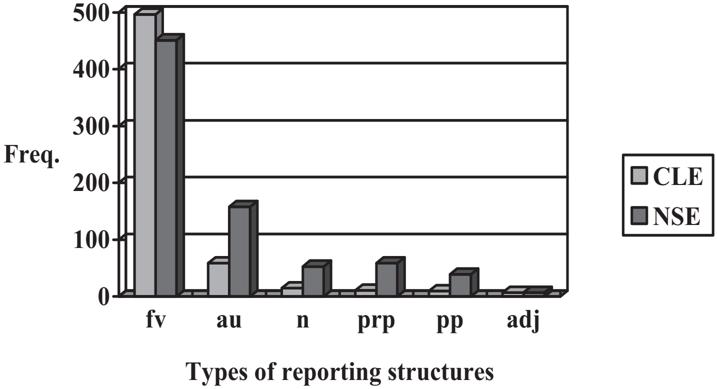


Figure 24-2. Distribution of reporting structures in CLE’s and NSE’s sub-corpora (Note: In this figure, “fv” signifies finite verbs; “au”--reporting adjuncts; “n”--reporting nouns; “prp”--RVs in present participle; “pp”--RVs in past participle; “adj”--reporting adjectives.)

3.1. Finite reporting verbs

Finite reporting verbs, including both active and passive forms, are the most common type in both sub-corpora. Hence, they are the focus of this study. Their frequency is as high as 496 in CLE’s sub-corpus and 451 in NSE’s one, accounting respectively for 82.9% and 58.8% of all reporting instances in the two sub-corpora, which suggests that this is the most common reporting structure adopted by both CLE and NSE. This result is consistent with He’s (2000) findings. In his research, finite reporting verbs account for 85.55% of total reporting instances. However, the difference in percentages between the two sub-corpora (82.9% vs. 58.8%) suggests that finite reporting verbs were CLE’s most frequently used reporting structure, while NSE did not attach such great importance to it as CLE did. Examples are abundant, e.g.

> Esrock and Leichty (1998) ***discovered*** that of the selected Fortune 500 corporations, only 19% placed prominence on general community service or responsibility to the local community.

It should be noted that all frequencies given in results are based on the counts of token occurrences. For example, for the occurrences of the finite verb *suggest*, its variations like the past form or the past participle *suggested* and the present participle *suggesting* are also included, whether active or passive, only if these verbs function as predicates. Furthermore, the reporting verbs under investigation are only finite verbs identified, excluding verbs in past participle and present participle and those in reporting adjuncts.

3.2. Reporting adjuncts

Reporting adjuncts, together with reporting nouns and reporting adjectives, are completely new concepts proposed by Thompson (2000: 91-101). The reporting adjuncts are usually overlooked in traditional grammar, though it is easy to perceive their functions. In this chapter, reporting adjuncts include reporting adverbs such as *apparently*, *reputedly*, *allegedly*, *reportedly* and *supposedly*, prepositional phrases such as *due to*, *according to* and *in accordance with*, infinitive phrases like *to quote,* subordinate finite clauses like *as far as…is concerned*, and *as…points out/said/admitted*.

Reporting adjuncts are the second favourite reporting structure by both CLE and NSE, with a frequency of 59 and 158 in the two sub-corpora. Examples found in the corpus data are varied. The following is an example:

> However, ***as Martha Kolln points out*** regarding Elbow's notion of the death-grip of grammar […]

3.3. Reporting nouns

Reporting nouns are another type of reporting structure which has not received much attention in traditional grammar. Yet the majority of reporting nouns, with only a few exceptions, are closely related to their relevant counterparts - reporting verbs (RVs); the *that/whether/wh*-clause following the nouns are the message carriers, giving complementary information about the reporting content. Commonly used reporting nouns include examples such as *apology*, *admission*, and *description*.

Altogether 15 occurrences of reporting nouns are found in CLE's sub-corpus and 53 in the NSE data, which enables us to conclude confidently that there are significant differences between the two sub-corpora. CLE used much fewer reporting nouns than NSE did. One example of reporting noun is given below.

> There is general ***acceptance*** that good ICT programs can benefit participants in developing good relationships with hosts and completion of assignments rather than premature termination (Brislin and Yoshida 1994).

In this sentence, the reporting content is embedded in the word *acceptance*. Here it is similar to say ***It is generally accepted*** *that…*, but the difference lies in that nominalization is reportedly related to more formal and objective writing style and more abstract thinking pattern (Zhang 2008).

3.4. Reporting verbs (RVs) in present participle

Accordingly, RVs in present participles are counted as a type of reporting when RVs do not act as predicates in sentences. Except for finite reporting verbs, RVs in present participle are far from being neglected. RVs in present participle are found to be about five times as frequent in NSE's sub-corpus as in CLE's sub-corpus, with a ratio of 59 to 11, thus reaching a significant level of .0001. That means NSE applied much more RVs than CLE did. Here is an example:

> Halliday and Hasan provide a comprehensive taxonomy of cohesion, ***classifying*** it into five types: reference, substitution, ellipsis, conjunction, and lexical cohesion.

3.5. Reporting verbs (RVs) in past participle

Separate past participles are counted as a type of reporting when they do not act as predicates in sentences, most of which are complements to modify or to give further specific information. The frequency of RVs in past participle in NSE's sub-corpus is four times as frequent as in CLE's sub-corpus (i.e. 39 versus 10). The numbers are small, since we have only about 200,000 words in the corpus, but it does not mean the results are not significant. For example:

> One of the definitions most often referred to is the one ***provided*** by Tarone 1980, who considers communication strategies.

3.6. Reporting adjectives

This is also a new category suggested by Thompson. Those words are mostly used to report one's feeling. A common reporting adjective is *so-called*, in which reporting meaning is embedded. There are a small number of reporting adjectives extracted from the corpus, such as *so-called,* and *aware*, with the same frequency of 7. For example:

> Therefore, quite a number of teachers (Graham 1997) are ***aware*** of the need to boost students' confidence, both in terms of oral participation and confidence in their general linguistic abilities, and sociolinguistic competence.

There are 7 reporting adjectives in both sub-corpora. The small number of reporting adjectives may be due to the function of reporting adjective itself. Its functions are, to quote Thompson (2000: 86), "to demonstrate the purpose of words, or reflect the content of a real conversation." During research article writing, writers pay more attention to the information conveyed than to the way of conveying the information. That may explain the small number of reporting adjectives. The low frequency of reporting adjectives in the two sub-corpora suggests that both CLE and NSE do not regard them as a common reporting method.

4. Reporting functions: A case study

Discussions in previous sections have demonstrated some differences and similarities in reporting by CLE and NSE. However, what has been done is largely restricted to quantitative analysis of the frequencies of reporting cases used, but the actual usage, i.e. how well they are used, is not discussed. For an adequate analysis and description of reporting, especially on how RVs function in MA theses, a case study may be more revealing.

During the line-by-line reading of the corpus, we discovered that there were some cases that demonstrate the uncertainty in the use of reporting by some CLE postgraduates. For limit of space, we only choose some paragraphs from CLE's sub-corpus and look closely at those reporting cases employed.

Here are some continuous paragraphs excerpted from CLE's sub-corpus in which so many instances of *argue* caught our eyes. The original paragraphs go like this:

> In the past thirty years a substantial amount of research has accumulated regarding the nature and prevalence of communication apprehension (CA). Defined by McCroskey (1977a) as "the fear or anxiety associated with either real or anticipated interaction with others", several researchers argue [1] that no other variable in communication research has received as much attentions (e.g., see Levine & McCroskey 1990; Lustig & Andersen 1991; Payne & Richmond 1984). Other constructs related to communication apprehension have been studied extensively as well…
>
> Normative data indicates that approximately 15 to 20 percent of the United States population experiences high levels of trait CA: that is, anxiety with either real or anticipated interaction with others. McCroskey & Richmond (1996) **argue** [2] that virtually 100 percent of the population experiences one of the four contextual types of CA at some point. Buss (1980) **argues** [3] that some of the salient situational features leading to increased anxiety include novelty, unfamiliarity, and dissimilarity … Based on Buss's (1980) criteria, initial interaction with someone, or interacting with strangers, may produce heightened anxiety in persons. Berger and Calabrese (1975) **argue** [4] that "whenever two people come together and interact for the first time, they have a very limited amount of information about each other. In such circumstances, considerable uncertainty exists. High levels of uncertainty lead to increased anxiety". Berger and Calabrese (1975) **argue** [5] that "in such situations the primary goal of the interactants is to reduce uncertainty and to reduce uncertainty and to increase the predictability about the other…
>
> One type of communication situation that is potentially replete with novelty, unfamiliarity, dissimilarity, and uncertainty is intercultural communication. Gudykunst and Kim (1997) **argue** [6] that when individuals are confronted with cultural differences they tend to view people from other cultures as strangers. Strangers are unknown people who are members of different groups. Anyone entering a relatively unknown or unfamiliar environment falls under the rubric of stranger. In their conceptualization, Gugykunst and Kim (1997) content [7] that interaction with people from cultures other than our own tend to involve the highest

> degree of strangeness and the lowest degree of familiarity. Thus, there is greater uncertainty in initial interaction with strangers than with people who are familiar. In such circumstances not only is uncertainty high but so is anxiety. According to Gudykunst and Kim (1997) [8], actual or anticipated interaction with members of different groups (e.g., cultures different from our own to anxiety).

A prominent feature of these paragraphs is the excessive use of the word *argue*. There are seven tokens of RVs altogether, among which 6 are *argue*. Some people would think that this might be an extreme example caused by individual writing style, which might be true. And this exerted much influence on the results obtained when we were counting the tokens and frequencies of reporting structures and reporting forms.

However, it can not be denied that the overuse of certain RVs is not occasional. The findings of Hu and Jiang (2007) showed that CLE tend to overuse the RVs they prefer, resulting in the unusually high frequency of a certain word, such as *find* and *analyze*; on the contrary, the RVs used by NSE were found to be more diversified. For example, the top 15 RVs employed by CLE reached 247 tokens, while the top 15 RVs by NSE only amounted to 162 tokens. The excessive use of some reporting verbs indicates that CLE lack flexibility and awareness to avoid monotony in writing.

Besides the overuse of a few RVs, this piece of writing represents another typical weakness of CLE in using reporting: inappropriate function of some RVs.

The word under discussion is, again, *argue*. By using the word *argue*, from dictionary explanation, the speaker or writer is trying to persuade others by reasoning and giving evidence when there are other people holding opposite points of view. For example, *I argued with her for a long time, but she refused to listen to reason.*

Hence, we can see inappropriate usage of *argue* here besides obvious grammatical mistakes. Example [1] is a little confusing when "several researchers argue that no other variable in communication research has received as much attentions (e.g., see…" The literal meaning of this sentence indicates that some people said there were some variables receiving more attention than CA, while these researchers did not agree with them; but judging from the context, we can say that the writer actually meant that "the nature and prevalence of CA" had received much attention, as was displayed in those researchers' work.

In Example [3], those "salient situational features" brought forward by Buss never occurred previously, neither did its opposite findings. So it is with Example [4], [5] and [6]. The writer did not make reporting verbs serve her purpose well, employing one verb while actually meaning another as can be inferred from the context, thus leading to confusing or misleading indication of the original meaning.

The only exception is Example [2], in which the two researchers expressed different opinions from previous ones and the word *argue* is just the right word to describe this situation. The lacking accuracy of the word *argue* reflects the weakness of some CLE in their use of reporting verbs.

In this case, the context suggests that reporting verbs such as *suggest*, *present* and *state* could be used instead of *argue* in Example [1], [3], [4], [5] and [6].

This inadequate variety is also represented in reporting structures. In this piece of continuous writing, there is only one reporting adjunct (Example [8]), with all other 7 occurrences being RVs. The insufficient employment of various reporting structures mirrors the lack of variety in the total number of reporting structures in CLE's sub-corpus (see Table 24-2).

The overuse of some reporting forms and the inappropriate function of reporting could result from CLE's different proficiency levels of the language itself (if there are), awareness of writing skills, L1 transference of Chinese in second language learning, different thinking patterns as well as cultural factors (Kaplan 1966), or different epistemologies in which these students have been trained (Cadman 1997). It is confirmed that EFL learners' native language and culture have great impact on their thesis writing, as Yu (1998) and Shaw (1992) pointed out. Moreover, the survey results suggest that this language transfer may even exist at the advanced level as well, after the writers have already acquired a high level of English language proficiency. Therefore, long-term and in-depth investigations are needed to further determine how much and in what ways language and cultural differences impact on CLE's academic writing.

5. Conclusion

The present study finds that there are similarities as well as remarkable differences in the use of reporting between CLE's and NSE's sub-corpora.

Specifically, in terms of reporting forms, CLE prefer integral reporting, i.e. to incorporate the reported authors' names into an integral part of the syntax of the sentence, which indicates the writer's willingness to reveal the reported authors' presence. In contrast, NSE use more non-integral reporting, which suggests the writer's inclination to obscure the reported authors' presence, thus implying a more "impersonal" style in NSE's writing. Besides, an analysis of non-citation and stative reporting shows that CLE also take care of clarity in writing, and try to reduce ambiguity by providing more information after a previous reporting by elaboration.

In terms of reporting structures, NSE's sub-corpus displays higher density of total reporting cases and richer variety as well. Finite reporting verbs prevail over reporting cases in both sub-corpora, but some reporting structures like reporting adjuncts, reporting nouns, RVs in present participle and in past participle occur in CLE's sub-corpus with a very low frequency.

The case study shows some overuse and inappropriate function of reporting in CLE's sub-corpus. Like some CLE, the writer in the case study did not make reporting verbs serve her purpose well, employing one verb while actually meaning another as can be inferred from the context. As a result, it led to confusing or misleading indication of the original meaning. This reflects CLE's lack of flexibility and awareness to avoid monotony in academic writing.

Generally speaking, the difference between CLE's and NSE's theses is significant and it is suggested that CLE should improve their writing in the following aspects: more impersonal and objective reporting forms, more diversity in reporting structures, more awareness to avoid monotony, and more flexibility in making reporting serve their purposes.

The pedagogical implications are that this study can help teachers to become aware of where to put their focuses in the teaching of academic writing. It also helps writers to improve their awareness of reporting in reading and writing academic articles.

As a direction for future research, the reporting pattern in different sections of theses can be investigated, for example, by comparing the frequencies of reporting forms and reporting structures used in Introduction and Literature Review etc., and investigating whether there are any significant differences in CLE's and NSE's sub-corpora. It is also suggested to examine each article instead of the whole corpus so that we can have a clearer picture of the distribution of each element under discussion, and reduce those factors caused by writers' individual preference.

References

Cadman, K. (1997), "Thesis writing for international students: A question of identity?". *English for Specific Purposes* 16 (1): 3-14.

Charles, M. (2003), “Evaluation in report sources: Uncovering disciplinary differences in theses”. Paper presented at Evaluation in Academic Discourse. Certosa di Pontignano, 14-16 June 2003.

Chen, M. F. (2006), *Reporting in Literature Reviews of English Dissertations.* PhD thesis, Xiamen University

He, C. W. (2000) *A Corpus-based Study of Reporting in Academic Research Articles*. Wuhan: Department of Foreign Languages, Huazhong University of Science and Technology.

Hu, Z. Q. and Jiang Y. C. (2007), “A contrastive study of reporting verbs in English MA theses”. *Studies in Language and Linguistics* 27 (3).

Huang, Y. J. (2001), “Subject-predicate collocation in reporting sentences”. *Foreign Language Education*.

Hyland, K. (1999), “Academic attribution: Citation and the construction of disciplinary knowledge”. *Applied Linguistics* 20(3): 341-367.

—. (2000), *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Longman.

—. (2001), “Humble servants of the discipline? Self-mention in research articles”. *English for Specific Purposes* 20: 207-226.

—. (2002), “Authority and invisibility: Authorial identity in academic writing”. *Journal of Pragmatics* (34): 1091-1112.

—. (2005), “Stance and engagement: A model of interaction in academic discourse”. *Discourse Studies* 7(2): 173-192.

Jia, Z. H. (2000), “Reported speech and its pragmatic function”. *Journal of Foreign Languages* 2: 35-41.

Kaplan, R. (1966), “Cultural thought patterns in inter-cultural education”. *Language Learning* 16: 1–20.

Leech, G. N. and Short, M. (1981), *Style in Fiction*. London: Longman.

Okamura, A. (2003), “How do British and Japanese scientists use ‘we’ and verbs in biology, chemistry and physics papers”. *The Economic Journal of Takasaki City University of Economics.*

Peng J. W. (2001), “A cognitive-pragmatic analysis of language report in English and Chinese”. *Foreign Language Teaching and Research* 5: 359-366.

—. (2003), *A Cognitive Analysis of Language Reports.* PhD thesis, Fudan University

Shen, D. (1991), “Different speech expressions in novels”. *Foreign Language Teaching and Research* 1.

Shaw, P. (1992), “Reasons for the correlation of voice, tense, and sentence function in reporting verbs”. *Applied Linguistics* 13(3): 302-319.

Swales, J. (1986), “Citation analysis and discourse analysis”. *Applied Linguistics* 7(1): 39-56.

—. (1990), *Genre Analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tang Q. Y. (2004), "Reporting in academic texts". *Foreign Languages and Their Teaching* 2.

Thomas, S. and Hawes, T. P. (1994), "Reporting verbs in medical journal articles". *English for Specific Purposes* 13(2): 129-148.

Thompson, G. (1994), *Collins Cobuild English Guides 5: Reporting*. London: Harper Collins Publishers.

—. (1996), "Voices in the text: Discourse perspectives on language reports". *Applied Linguistics* 17: 501-530.

Thompson, P. (2000), "Citation practices in PhD theses", in L. Burnard and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective*, 86-101. Frankfurt: Peter Lang.

Thompson, G. (2001), "Interaction in academic writing: Learning to argue with the reader". *Applied Linguistics* 22: 58-78.

Xin, B. (1998), "A critical analysis of reported speech in news reports". *Foreign Language Teaching and Research* 2: 9-14.

Yu, R. D. (1998), "Non-native graduate students' thesis/dissertation writing in science: Self-reports by students and their advisors from two U.S. Institutions". *English for Specific Purposes* 17(4): 369-390.

Zhang, G. Y. (2008), *A Contrastive Study on Nominalization in English and Chinese.* Beijing: China Social Sciences Press.

NOTES ON CONTRIBUTORS

Marco Baroni is a tenured researcher in computational linguistics at the University of Trento, Italy. His research interests concern the extraction of distributional information from verbal input, the practical applications of research on concept induction and the creation of electronic resources (such as corpora and lexica) and related computational tools. He has taught courses in computational lexicography, lexical semantics, text processing and statistics for linguists.

Silvia Bernardini teaches translation from English into Italian at the School for Translators and Interpreters of the University of Bologna at Forlì, Italy. Her research interests are in the areas of corpus-based translation studies (especially collocational regularities in translated language), corpus use in the classroom and corpora from the Web. She is co-editor of the contrastive linguistics journal *Languages in Contrast* (John Benjamins).

Yun-hui Chen is currently studying in an MA program on literature at National Sun Yat-Sen University in Kaohsiung, Taiwan. She has collaborated with Professor Hui-Chuan Lu on a project investigating restrictive relative clauses on the basis of corpus data.

Hercules Dalianis is an associate professor (docent) and tenured lecturer at the Department of Computer and Systems Sciences (DSV) at KTH/Stockholm University, Sweden. With more than 20 years of experience of research in human language technology, Dalianis currently works in the area of text mining and electronic health informatics, focusing on electronic patient records.

Umar Dawut is an associate professor of linguistics at Xinjiang University in China. His main research interests include Uyghur language and culture, applied linguistics, bilingual education, and corpus linguistics.

Carmen Dayrell is currently a post-doctorate research fellow in the Department of Modern Languages at the University of São Paulo, Brazil. She received her PhD from the Centre for Translation and Intercultural

Studies of the University of Manchester, UK. Her main research field is applied linguistics, focusing on the use of corpus resources in foreign language teaching and translation.

Bart Defrancq is an assistant professor at University College Ghent, Belgium. His main research areas are contrastive linguistics and translation studies. His research focuses on clause linkage in French, Dutch and English.

Yan Ding is a PhD candidate in the School of English, the University of Hong Kong. Her main research interest is cognitive linguistics. She is currently working on diachronic and stylistic variations of emotion metaphors.

Adriano Ferraresi is a PhD candidate in English for Special Purposes at the University of Naples 'Federico II', Italy. His main research interests concern the exploitation of Web data to build reference and specialized corpus resources and the development of automatic methods for extracting phraseology from specialized corpora.

Ernest Wei Gao is an associate professor of translation and interpreting studies in Lanzhou University of Technology, China as well as a professional translator and interpreter. He is currently a PhD candidate at Heriot-Watt University, UK. His research interests cover cognitive linguistics, translation and interpreting studies, business English and English for science and technology.

Federico Gaspari has a background in translation studies and holds a PhD from the University of Manchester, UK. A former lecturer at the universities of Manchester and Salford, he is a research fellow at the University of Bologna at Forlì and teaches English language, technical translation and translation theory at the University of Macerata in Italy.

Kim Gerdes is an associate professor at the Sorbonne Nouvelle in Paris. He has studied Mathematics in Berlin, Paris, and Utrecht and was awarded his PhD in Linguistics in Paris. Dr Gerdes works in various fields of formal and computational linguistics including questions of word order in natural language generation, statistical and rule based analyses of corpora, and formalization of oral and written syntactic phenomena.

Martin Hassel is a senior researcher at the Department of Computer and Systems Sciences (DSV) at KTH/Stockholm University, Sweden. His main expertise lies in efficient and flexible models of language use, and his current interest is mainly in medical informatics, in particular modelling and mining of electronic health records from an information extraction perspective.

Lianzhen He is Professor of English and Dean of the School of International Studies, Zhejiang University in China. Dr He is also on the editorial board of the international journal *Language Assessment Quarterly*. Her main research interests cover TEFL, language testing, corpus linguistics, as well as discourse analysis, in which areas she has published extensively.

Huaqing Hong is a research associate at the National Institute of Education / Nanyang Technological University, Singapore. Dr Hong's main research areas cover corpus linguistics and language education.

Zhiqing Hu is Professor at the School of Foreign Languages, Huazhong University of Science and Technology, China. He is mainly interested in stylistics and English for Specific Purposes.

Yuechun Jiang is a doctoral student at Beijing Foreign Studies University and a lecturer at Beijing City University, China. Her research interests include second language acquisition, corpus linguistics, and language teaching.

Defeng Li is Reader in Translation Studies and Director of the Centre for Translation Studies at the School of Oriental and African Studies (SOAS), University of London. He is also holding a concurrent appointment as Professor of Translation Studies at Shandong University. Dr Li conducts research in both Translation Studies and Second Language Education and has published in journals such as *TESOL Quarterly*, *Target*, *Meta*, *Perspectives*, *Babel* and *International Journal of Applied Linguistics*. His most recent publications were two books on translation of financial and journalistic texts respectively, both by Hong Kong University Press.

Hui-Chuan Lu is Professor at the Department of Foreign Languages and Literature in National Cheng Kung University, Taiwan. She obtained her PhD in Spanish linguistics from UCLA in 1994. Her major research interests include Spanish syntax and corpus linguistics. Dr Lu has researched in learner Spanish on the basis of corpus data for many years.

Samat Mamitimin is a lecturer at the Humanities School of Xinjiang University, China. He was awarded his PhD in computational linguistics in 2009. Dr Mamitimin's main research interests cover computational linguistics, corpus linguistics, translation studies, and language teaching.

Jun Miao is a PhD candidate of Translation Studies in ESIT (École Supérieure d'Interprètes et de Traducteurs) of Paris 3 - Sorbonne Nouvelle University. Her current research interests include Translation Studies (translator, translator's style), parallel text processing, and intertextual textometric exploration of bilingual text corpora of Chinese and French.

Guiling Niu is an associate professor at Zhengzhou University, China and a research associate at Nanyang Technological University, Singapore as well as a member of Singapore Association for Applied Linguistics. She has pursued a wide variety of academic interests in corpus linguistics, discourse analysis, world Englishes and second language acquisition and has published over ten journal papers in these fields.

Dirk Noël teaches English linguistics in the University of Hong Kong, after a long association with the University of Ghent's contrastive grammar research group (Contragram). He has mainly published on the complementation patterns of *believe*-type verbs in a (contrastive / diachronic) construction grammar and grammaticalization perspective and has co-authored Contragram's corpus-based Dutch-French-English contrastive verb valency dictionary.

Masahiko Nose is an assistant professor in the College of Foreign Studies at Reitaku University, Japan. He received his PhD on "Passive constructions in Hungarian: in terms of transitivity" from Tohoku University in Japan. His research interests are primarily in Finno-Ugric and Oceanic languages as well as linguistic typology.

Giovanni Picci works as an expert lexicographer (bilingual dictionaries) at the Larousse publishing company in France.

Hongwu Qin, PhD, is Professor of Linguistics at the School of Foreign Languages, Qufu Normal University, China. His research interests include linguistic theory and corpus translation studies.

Marco Rocha is a lecturer at Universidade Federal de Santa Catarina, Brazil. His research focuses on anaphoric relations, using a corpus-based approach. The analysis of anaphora in parallel and comparable corpora is mostly concentrated in the English-Portuguese language pair, with a concern for possible applications in human language technology.

André Salem is Professor at ILPGA (Institut de Linguistique et Phonétique Générales et Appliquées) of Paris 3 - Sorbonne Nouvelle University. His work is within several research areas covering lexicometrics, automatic text and language processing, and knowledge representation and knowledge acquisition from corpora. He is the chief editor of online journal *Lexicometrica* and the principal organizer of the International Conference on the Statistical Analysis of Textual Data (JADT).

Gert De Sutter is a lecturer of Dutch linguistics at University College Ghent Translation Studies, Belgium. He studied Germanic languages (Dutch and German), with a specialization in descriptive linguistics. In 2005, he was awarded a PhD in linguistics from the University of Leuven for his dissertation on word order variation in Dutch verb clusters. His current research focuses on the underlying principles that govern syntactic choices in translated vs. non-translated languages.

Marc Van de Velde is Professor of German at University College Ghent Translation Studies, Belgium. He obtained a PhD in Germanic languages at the University of Ghent (1979). His research concentrates on German syntax, partly in contrast to Dutch syntax, and on translation studies (German-Dutch).

Sumithra Velupillai is a PhD candidate at the Department of Computer and Systems Sciences (DSV) at KTH/Stockholm University, Sweden. She is also affiliated with the National Graduate School of Language Technology (GSLT) in Sweden. With a background in computational linguistics, her main interests cover computational linguistics in general as well as text clustering, text mining, information retrieval, information extraction, text categorization, biomedical and clinical natural language processing.

Jianxin Wang is a PhD candidate at the University of Auckland, New Zealand and Professor Emeritus of English at Beijing University of Posts and Telecommunications, China. His research interests include corpus linguistics, comparative studies of English and Chinese, research methodology, and college English teaching methodology.

Kefei Wang, PhD, is Professor of Linguistics at the National Research Centre for Foreign Language Education, Beijing Foreign Studies University, China. His main research interests include corpus linguistics and translation studies. Dr Wang is currently editor of *Foreign Language Teaching and Research*, a top journal of language and linguistics in China.

Hans-Georg Wolf is Chair Professor for the Development and Variation of the English Language at the University of Potsdam, Germany. His research interests include sociolinguistics, cognitive linguistics, corpus linguistics, and pragmatics. Currently, his main research focus lies on the application of cognitive sociolinguistics to the study of varieties of English.

Yun Xia is a PhD candidate at Shandong University and an associate professor at Qufu Normal University, China. Her research field is translation studies.

Richard Xiao is Professor in Linguistics at Zhejiang University in China as well as Senior Lecturer and Programme Leader of English and Chinese Studies at Edge Hill University in the UK. With a PhD in linguistics from Lancaster University (2002), his major research interests cover corpus linguistics, English linguistics, Chinese linguistics as well as contrastive and translation studies of the two languages. Richard's recent books include *Aspect in Mandarin Chinese: A Corpus-Based Study* (2004), *Corpus-Based Language Studies: An Advanced Resource Book* (2006), *A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners* (2009), and *Corpus-Based Contrastive Studies of English and Chinese* (2010).

Hui Yin is a linguist at Zi Corporation, Canada. His research areas range from cognitive linguistics, corpus linguistics, and psycholinguistics to syntax, semantics and applied linguistics.

Chunling Zhang is a PhD candidate at the University of Alberta, Canada, where she specializes in acoustic phonetics and speech recognition. Being a language teacher of many years, she also takes a keen interest in translation studies and language education.

Chong Zhu teaches at the School of Foreign Languages, University of Electronic Science and Technology of China. His main research interests include corpus-based language studies and computer-aided language teaching and learning.

INDEX