# 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

# 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