Translating anaphoric *this* into Portuguese: a corpus-based study

Marco Rocha

Universidade Federal de Santa Catarina

Abstract

The paper presents a study on the translation into Portuguese of the demonstrative pronoun *this*, whenever used as an anaphor. The study relies on the analysis of a sample of tokens collected in a parallel English-Portuguese bilingual corpus containing literary texts and international law texts, as well as technical and scientific materials. Only tokens appearing in texts originally produced in English were considered. Each case of anaphoric *this* was classified according to four properties, namely: syntactic function; reference type (anaphoric, cataphoric or deictic), according to the relation each token holds with the antecedent; explicitness of the antecedent; and antecedent phrase structure. Explicitness assigns the values of either explicit or implicit to antecedents of anaphoric *this* tokens, as identified by the analyst, whereas phrase structure classifies those antecedents as either nominal – a noun phrase – or textual – a discourse chunk that cannot be treated as a single noun phrase. Relations between these variables are analysed with a view to revealing semantic relations of a textual nature, and renderings into Portuguese are also discussed so as to establish the way these relations are expressed by means of anaphora in each language. Collocational analysis of the tokens was used to establish patterns of translation for the anaphor in the English-Portuguese language pair. 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 contributes to a better understanding of textual aspects in translation, possibly extensible to other Romance languages, and also provides subsidies for the improvement of machine translation systems.

Key words: corpus linguistics; anaphora; parallel corpora; textual semantics.

1. Introduction

Since the revival in the eighties, corpus-based research grew 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 techiniques 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 paper 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 as retrieved from a parallel English-Portuguese corpus built for the purposes of the present research, including tokens collected from COMPARA (2001), which contains literary texts and their translations; tokens from international law documents from the Organization of American States (OAS); and tokens from material of the Euroepan Medicines Agency (EMEA). The remainder of the paper is organized as follows: the second section briefly presents related work; the third section describes the methodology; the fourth section presents, analyzes and discusses results; the fifth section summarizes implications of findings to textual semantics and machine translation, pointing out future developments of the

analytical approach.

2. Sources of the analytical approach

This short review of related work describes how the analytical approach adopted in the study was built. It should not be construed as a full description of work commented, as this is not intended. 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 paper. Anaphoric phenomena have been approached 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, since the computational processing of anaphoric linkage is a very difficult problem. 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 within textual cohesion, 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 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 describe the function of these items well. 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.

Botley (2000) created 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 a model of the anaphora "world". 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 source text. Differently from the approach used here though, 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 chosen anaphor *this* is used for referring. In particular, there is an ultimate effort better to understand 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 the account 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 and their translation images back into English, 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 textual semantics may uncover useful patterns for the study of translation and subsidies 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 one of the two languages separately. The paper then proceeds to describe a number of translation pairs in detail, with particular concern for 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 each language 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, at least in what regards the use of anaphoric demonstratives. The properties included in the analytical approach are used as a way of characterizing referring systems so as to enable cross-linguistic comparison. Differently from Santos though, the approach used here focuses on 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

The first methodological step is the analysis of *this* tokens in a concordance extracted from the corpus. Cases of anaphoric *this* as a determiner modifying a noun phrase head were removed. COMPARA contained 361,852 English words when the sample was retrieved. The full concordance of *this*, as informed automatically, contains 1033 tokens of *this* (0.28% of the corpus), out of which 1000 were randomly selected by the query-handling interface in COMPARA site. After removing determiners, 171 tokens remained, a proportion of 16.55%. The OAS corpus holds 95,052 tokens in English and is thus relatively small. 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 tokens in documents of the English language 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 4768 tokens of anaphoric *this* in the EMEA corpus. A random sample of 46 tokens of anaphoric *this* from the EMEA corpus would be added, amounting to a total of 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, object and prepositional complement, 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. Grammatical classification is often the starting point for machine translation and anaphora resolution systems, using the output of a parser, a processing module that can be reasonably expected to be accurate with present-day technology. Each category is detailed below.

3.1.1. Lexical verb subject

The classification is assigned to cases to which the notion would apply according to grammar textbooks. One example is given below[1]. The anaphoric demonstrative token is shown in bold.

[1] All examples are taken from the sample collected for the purposes of the study.

(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 presented problems for the classification, a fact of life in corpus analysis. Consider example (2) below.

(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.

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, like in example (4) below. The concept of prepositional object used is the one in Greenbaum (1996, p.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. One aspect involved in this decision was that, whenever the subject predicative is not an adjective phrase, as tokens of pronominal *this* are not, the characterization of a subject predicative relies essentially on its position following the verb. Subject and subject predicative can often swap places without significant changes in meaning. Secondly, there were only two cases in the sample which could have their grammatical function classified as subject predicatives. Both could be classified as copular verb subjects in an inverted construction. The grouping with copular verb subjects simplified the organization of data in tables. In example (6) below, classifying *this* as either subject or subject predicative seems to be perfectly acceptable.

(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 more detail further below. First, the typical cases will be 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 when demonstratives are taken into consideration. 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.

Although the identification of the antecedent in the example above is not without its problems, it is about as straightforward as referring gets with anaphoric demonstratives. There is a specific noun phrase (*name and identity of the petitioner*) referring to two discourse entities of easy identification, although the lack of agreement is a processing difficulty. However, anaphoric demonstratives refer to textual antecedents very often, if compared to other forms of anaphor, and precise delimitation of the antecedent may be a great deal 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 above, *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 also no principled way of deciding which is best, and also, regarding the minor adaptations involved in a substitution-based specification of antecedents, whether this counts as 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 (10) 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. There are several tokens of this kind in the sample. They were all classified as cataphoric, since the antecedent cannot be fully identified before the chunk is fully read. Once again, there is a degree of arbitrariness. Decisions regarding the type of reference quite invariably require guesses on how the reference was processed by a reader. There is no hard evidence regarding anaphora processing by readers. Anaphoric *this* which refers to a discourse chunk may also require information given after the occurrence 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 phrase *the wrong response*, subject predicative in the copular sentence, plays a crucial role in the identification of the antecedent, 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. This study classifies such cases as anaphoric.

3.2.3. Deictic reference

Deictic references are typically associated to 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 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 as well, but characters move in a fictional setting which is revealed in the text by various means. Therefore, decitic references do occur if characters are seen as the processors of anaphoric relations. A definition of standards used to 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, although the contents of the box are not mentioned in the context preceding the anaphor. A degree of inference ─ based on information presented both by direct reproduction of dialogues and narrator intervention ─ 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 to achieve understanding. The matter is made more complex by tokens such as example (13) below.

(13) “And **this** is Zoe.”

This type of occurrence does not require any previous mention of a referent 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. 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 carries the idea of situation description, allowing adequate identification of the vague implicit antecedent, and this 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, based on a combination of linguistic patterns with the described situation and a lexical clue.

(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 to 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 as well, since both preceding and subsequent text provide enough 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; textual 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.

Tokens from (12) to (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 analyst agonizing, since it may be hard to ascertain whether the antecedent is implicit. 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, like in example (20) further below.

(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.

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. One way to solve the difficulty would have been to create a separate category to classify these cases, which was considered unnecessary.

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 distance distinction is not strict, quite in the same way as in English. The distinction between *isso*/*aquilo* and *esse*/*aquele* has no equivalent in English.

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 ahead. Renderings from 6 to 12 are not classified as

demonstratives in standard textbooks. Target texts use different forms of anaphoric reference or rephrase the source text so as to make it unnecessary.

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, 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* as 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 in the analytical approach cross-tabulated by text types. The following subsection presents results and analysis of the cross-linguistic data, followed by a discussion on aspects of textual semantics uncovered by the analytical work.

4.1. Frequency tables

Table 1 below shows frequencies for the grammatical function variable. Percentages are rounded in all tables.

Table 1 – Distribution of grammatical functions

Category	COMPARA	OAS	EMEA	Total	Percent
Prepositional complement	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
Lexical verb subject	11 (7.0%)	22 (51.16%)	26 (56.52%)	59	23.07
Copular verb subject	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 English originals could be said to be balanced, if totals are considered. Percentages in the last column of Table 1 vary from 18.07% to 30.76%, thus close to a 25% even distribution in four parts. However, numbers for each corpus vary widely. Prepositional objects are rare in both OAS and EMEA, but nearly 25% of COMPARA occurrences, which also shows higher numbers for verb objects. Conversely, lexical verb subjects are rare in COMPARA and amount to more than half of tokens in OAS and EMEA. The assertive informational nature of text in law and medical documents seems to favor reference in which antecedents are linked to new information expressed by lexical verbs. The presence of colloquial forms in dialogues and reproduced thought of characters seems to use

copular structures, the most common in COMPARA, to evaluate what has been said up to a given point (*is this a trick?*) and to other purposes in textual semantics typical of spoken language.

Forms of *be* are by far the most common collocation in copular verb subjects. This holds for all text types. All but one (*this seems*) token 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, *do-this* forms alone adding up to a little over 14%. Tokens appear predominantly as subjects of passive constructions – classified as objects ─ in EMEA (6 out of 8) and OAS (6 out of 9). 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* contributes with 31.91% of the tokens classified as prepositional complements and should receive special attention.

Table 2 – Distribution of reference types

Category	COMPARA	OAS	EMEA	Total	Percent
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 reference, with the other two possible classifications adding up to less than 10% of the total. All cases of deictic reference are in COMPARA. This is also true of cataphoric reference, 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 a law document. Corpus data show that both cataphoric and deictic reference are associated to spoken language as presented in dialogues in literary text. Table 3 presents results antecedent explicitness.

Table 3 – Distribution of antecedent explicitness

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 here is also strongly skewed towards the explicit category. There are no implicit antecedents in OAS, perhaps as a result of the need to be transparent in law documents. There are two cases of implicit antecedent in EMEA. These 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 had not been 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 very 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, a syntactic function that by definition often expresses a property or characteristic of the subject. Again, analyst agonizing is unavoidable, but 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 disconsidered. The adjustment was seen as too deep a transformation of previous text. Minor syntactic changes were seen as acceptable to classification as an explicit textual antecedent, but alterations of semantic content were not. Data for antecedent phrase structure are presented in Table 4 below.

Table 4 – Distribution of antecedent phrase structures

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	100.0

The second dichotomic property in the model shows that the anaphoric demonstrative *this* has predominantly textual antecedents, differently from personal pronouns and the "classical" concept of anaphoric reference. Proportionally, the tendency is stronger in OAS data and weaker in EMEA, but percentagens 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 which bear in mind computational processing. Although a degree of underspecification might be acceptable, actual procedures to achieve some form of operationally feasible delimitation of these antecedents should be devised, so that the resolution of anaphoric demonstratives is not left to speculative guesses. This is also true from a psycholinguistic point of view. This frame of mind is a major factor in this study.

4.2. Cross-linguistic analysis

The list of renderings in subsection 3.5 is used to present results for anaphoric *this* translations. In section 4.3, an attempt is made to establish translational patterns associated with collocations in the source language and categories in the variables. Findings are summarized in algorithmic form. Table 5 below shows frequencies for renderings in cross-tabulation by subcorpus.

There are only four classes which appear in all three subcorpora, 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 *ta*l. In OAS, there are also only three tokens that do not belong to these four. In all subcorpora, *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, mostly because NPs are almost as frequent as *isso*-group. 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, but e*sse*-group comes

third and *assim* appears in fourth, one token over NPs. The subsequent discussion tries to reveal motivations for these patterns on the basis of collocations and textual semantics, but the variety shown in COMPARA involves a high degree of translator's choice.

Table 5 – Distribution of anaphoric *this* translations

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

A general pattern for anaphoric *this*, on the basis of the properties analysed, 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 (76.74%) and 31 tokens of EMEA (72.09%). Literary text seems again to allow greater variety of usage, while patterns in law and medical documents conform to a general expectation that characterizes the anaphor.

4.3. Discussion

The discussion is carried out by grammatical category, beginning with prepositional objects. There are only three tokens in EMEA. Two tokens conform to the general pattern in all three variables. The textual antecedent is the full preceding sentence in both (as in example (24)), a common textual antecedent. Both tokens are translated by semantically generic noun phrases (*este facto* and *este assunto*). The standard rendering *isso*-group would be acceptable. The third one, shown in example (25), has a nominal antecedent, *diet*, the object of the verb in the subordinate clause, but the translation omits the object. Omission of subjects and, less often, objects is not rare in Portuguese. Pending further investigation, there seems to be a group of verbs which favor the omission of objects and *continuar* (*continue*) belongs to this group. The standard *isso*-group translation feels idiomatically inadequate in (25b) because object omission is expected and the antecedent is nominal.

(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 facto** e aconselhados a redobrar a atenção ao conduzir ou utilizar máquinas…

(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 para diabéticos, deve continuar enquanto estiver a tomar 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 used as 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 to the most frequent collocation, *like this* (15 tokens). The anaphor in the collocation is omitted in two Portuguese renderings; it is translated by the preposition *como* + *esse*-group in two cases and by *isso*-group in one case.

There is evidence that the attachment of the prepositional phrase plays a role in translation. Even when it was not the translator's choice, *like this* could be translated as *assim* whenever attached to a verb phrase in source or only in 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. The other collocations of anaphoric *this* as a preposition 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).

Beginning with the last and most frequent collocation, all five tokens 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, since there are two translations for one of the tokens in COMPARA. The omission seems to be a translator's choice rather than a pattern. All tokens of *beyond*, *behind*, *for*, and *than* with *this* as 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 by *responder* (*answer*) and the prepositional phrase is omitted, probably a pattern for such usages of *say*. One token of *at this* is translated as *isso*-group, whereas the second one is omitted. The translator chose to use *rir*, the Portuguese equivalent of *laugh*, with the object omitted, common usage in Portuguese which seems acceptable in English too.

Translation patterns for anaphoric *this* as a preposition 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 (deictic reference).
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 *at this*, the anaphor may or may not be omitted if the verb is 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 to 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, which does not take a subject. A degree of syntactic rearrangement is a recurring pattern in omissions. The second case of omission for verb objects in the subcorpus also involves syntactic rearrangement. Nominal antecedents are either the object or the subject of main clauses in preceding sentences. 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, translated by 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 subject of a verb in passive form. A token of *this* as the object of *deem* in source is translated by *assim*. There is probably a translational pattern associated to 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 source text, namely, *describe* and *enjoy*. Their Portuguese translations, *descrever* and *gozar*, seem not to take *isso* as an object in standard usage. Plain omissions occur when *this* is the object of verbs *know*, *notice* and *write* in the source text. Corresponding verbs in Portuguese (*saber*, *notar* e *escrever*) belong to the group of verbs that accept omission of the object well whenever it seems to be readily inferrable from the preceding or subsequent text. Since there also cases in which the omission does not occur and *isso*-group is used, it is hard to uncover a pattern.

One case of omission is the source text phrase *made this an oportunity*, 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 to collocation data. However, the attempt is presented below.

1. If the Portuguese translation verb is *descrever* ou *gozar*, translate as NP such as *cena* ou *situação*.
2. If the source text verb is *know*, *notice* or *write*, omission is possible.
3. If the source text phrase is *make this*-IObj NP-Dobj, omission is possible, along with translation of *make* as a verb semantically related to 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 the 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 structure of the source language is essentially retained in the target text with the anaphor omitted. The use of generic noun phrases, such as *esta situação* e *este facto*, appears along with more technical domain-specific summary terms such as *estes sintomas*. 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*, 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 pronoun. The text type seems to be a crucial aspect for this sort of usage.

Dialogues and literary texts favour deictic references, whereas technical and law 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 (26) below. The *isso*-group rendering is from a second translation of the same book:

(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 **o quadro** nada revelaria a quem, porventura, o visse.

(26c) Estava perfeitamente convencido de que **isto** nada revelaria a ninguém.

The tendency to reduce underspecification may be a characteristic of translated text, rather than a language-specific feature. The variation noun phrase/*isso*-gorup/omission requires further investigation in search of stable patterns. The algorithmic systematization for lexical verb subjects attempts to reflect this element of choice.

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 sentence is translated as a non-finite sentence, consider omission as fits syntax.
3. If the target text changes a textual reference to a nominal antecedent, use repetition.
4. If the antecedent is nominal, consider generic noun phrases such as *essa área* (*this area*).

5. In all other cases, translate as *isso*-group.

Copular verb subjects are the most frequent type of grammatical function in COMPARA. Thus, as a subject, anaphoric *this* is clearly associated to copular verbs in literary texts. Translations show more variety than for any other category. There are 20 *isso*-group renderings, 17 *esse*-group and 16 omissions. This means that, except for two *like this* tokens, all *esse*-group translations come from this category. There are environments which favor *esse*-group renderings. Nominal antecedents linked to an anaphoric subject in a copular construction favor *esse*-group. They are usually subjects or objects in preceding main clauses or independent sentences. This also holds for introductions (example (13)) and collocations such as *this is the case*. The translation choice may be tested in examples (13) to (16) above, by checking whether rephrasing with the noun phrase head as part of the subject (that is, rephrasing *this is my conclusion* as *this conclusion is my conclusion*) makes sense. The rephrasing is odd and semantically distinct in constructions such as *this is a funeral*, which prefer *isso*-group translations. Adjectival subject predicatives favor translations by *isso*-group. Omissions occur interchangeably in *isso*-group but not *esse*-group environments, except for the *this is the case* collocation in a conditional clause.

There are three translations as *assim* in which there is a clausal indirect question as subject predicative in the source text. Translations as demonstrative *o* are associated to the relative pronoun *que* in clauses with a conclusive meaning. There are two 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 the source text by placing *this* in 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 by *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.

In OAS, there are ten tokens of *this* in the grammatical category, and translations include the four basic classes. 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 choice of translation for two tokens. One is translated as a noun phrase (*foi por esta razão que*...). The variation seems unmotivated. The collocation *this is where* is wholly omitted in the target text. The place reference seems irrelevant in the target text. There are three other cases of omission involving syntactic rearrangement. In one of them, the impersonal verb form *trata-se de* is used instead of *this is*. The two others 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*, the Portuguese equivalent of *for the first time*. Translations by *esse*-group involve nominal antecedents which are objects of the verb in a preceding main clause or independent sentence. The copular structure includes a hyperonimic classifying subject predicative.

There are nine cases of copular verb subjects in EMEA, also covering the four basic translation classes. Omissions occur in three tokens. Two of them occur in subordinate conditional clauses; one uses the *trata-se* pattern, 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 to source text collocations *this is the case* and *this is expected to*. Translators apparently choose NPs (*esta situação* and *esta tendência*) out of formality requirements, since copular structures are not retained. The third NP choice seems to aim at

reducing underspecification with the use of *estes efeitos* 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. The rephrasing test described above works as expected. Favorable environments for each choice are confirmed. Summing up:

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 carries a conclusive meaning, 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, use the rephrasing test and translate as appropriate by *esse*-group or *isso*-group.
7. If the subject predicative is an adjective, translate as *isso*-group or omission.
8. If informality or undespecification are to be avoided, consider the *trata-se* pattern and a NP-lexical verb structure as options.

5. 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 to grammatical categories and collocations in the source text 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 to position of the anaphor in the source text and to translation as a specific NP in the target text. The rephrasing test seems useful for choosing between *esse* and *isso*. Some aspects of textual semantics, such as evaluation, underspecification and formality, are difficult to gauge, 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 information available regarding the interaction of ideational, lexicogrammatical and cross-linguistic aspects is not fully mapped and, to the level that it is mapped, it is poorly understood. Anaphoric phenomena, especially 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 samples to expand more quickly to confirm or reject

possibilities of pattern definition. No matter how tentative and laborious, however, the approach may eventually pay off in terms of relevant findings.

References

Botley, S., (2000). *Corpora and discourse anaphora: using corpus evidence to test theoretical claims*. Ph.D. 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*. Amsterdam: Rodopi, pp.51-112.

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

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, 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*. Amsterdam: Rodopi, pp.319-342.

Tymoczko, M., 1998. "Computerized corpora and the future of translation studies". *Meta*, XLIII, 4. Montreal: Montreal.