Chapter 1: Introduction

Maybe the most abstract idea contained in this thesis is that two languages do not say the same thing (add quantification according to your liking).

This belief of mine arose as a reaction to the overwhelmingly dominant idea, in machine translation and natural language processing, that languages are only <u>formally</u> different: in other words, they have different forms to convey the same meaning. (Incidentally, the belief that this is wrong was considerably increased when my children began to grow bilingual.) In practical terms, natural language processing is also identified by most people as English language processing, together with localization or translation procedures.

Brought up with Fernando Pessoa's quotation "A minha pátria é a língua portuguesa" ('My homeland is Portuguese'[1]), I never doubted that the best contribution I could make to natural language processing would be to the processing of Portuguese, my native language. Culture and behavioural conventions are inextricably mingled with language, in such a way that to master one tongue is much more than being able to produce grammatical and meaningful sentences in it.

Given that tense is one of the most developed grammatical systems in Portuguese (and known to be one of the most difficult to master by foreign speakers, as well), it is hardly surprising that it appealed strongly to me.

In addition, its (apparent) reference, to time, seemed to be a domain where English and Portuguese could hardly disagree - and, in fact, several transfer-based MT systems suggested an interlingual approach for precisely that domain, cf. van Eynde (1988) and Hauenschild (1988). (When I began this work, I was not aware that precisely the area of tenses had been identified by Whorf, opposing Hopi to English, as a domain characterized by linguistic relativity.)

If I were able to find significant differences in the organization of the two languages, this would be a stronger argument than if I studied parts of language directly connected with social behaviour, for example. On the other hand, if I could use the same underlying analysis for both languages, I might be able to write a dissertation for a broader audience than Portuguese linguists.

Actually, I started with the second goal in mind. In fact, for some time I strove to get a unified model that could handle English and Portuguese. It was only when I started to look carefully at the translations that I realized that my old intuition was right and that a unified approach was the wrong way to go.

This dissertation can then be seen as an investigation into how much (or how little) the languages English and Portuguese differ in the restricted domain of reference to time (in linguistic terms, in the fields of tense and aspect).

In the following sections of this introductory chapter, I present succinctly the problems

---

[1] Bernardo Soares, *O livro do desassossego*.

addressed, before attempting a critical assessment of natural language processing and defining the place of this dissertation in relation to other research in the field. Finally, I describe the empirical data and how the entire thesis is organized.

1.1 Contrast and comparison of languages through translation

The investigation of the differences and similarities between two language (sub)systems through the use of real translations is a complex task, for several reasons: First, because of the quantity and diversity of the material presented in real texts, as opposed to constructed examples. (The use of corpora will be discussed in Section 1.4 below.) Second, because of the interaction of two complex factors that enter into the making of a translated text, namely, translation accuracy and style of the target language.

The interaction referred to above can be cast in the following terms: A translated text is necessarily closer to the original text (in a different language) than to original texts in the target language. Hence, it is difficult to assess to what extent its particular linguistic properties are dependent on the target language, and to what extent they originate from the source language.[2]

This could be termed the "translation paradox": the more literal the translation, the more linguistically faithful, and, therefore, the less problematic a comparison between individual instances of translation. However, the better the translation[3], the more it approaches the stylistic conventions of the target language system, the more problematic the study of individual instances becomes.

This crucial property of translations, namely, that by belonging to the target language they are bound to follow target language norms, ranging from obligatoriness of use of one particular grammatical category to simple obedience to general stylistic trends, implies that translation often does not preserve the meaning of the original (even if it could).

Language difference is thus one key to the understanding of meaning differences in translation (even though meaning differences can arise in the very same language by having two speakers choosing to convey different aspects of the very same situation).

I will, however, be using translations to detect language differences, i.e., I will be looking at the consequences in order to uncover the causes. After discussing several features of both language contrast and translation in Chapters 2 and 3, Sections 3.7 and 3.8 will deal with this issue in detail.

1.2 Tense and aspect

The interest of tense and aspect is obviously not just the fact that Portuguese grammar is

---

[2] In order to circumvent this problem, current research in corpus-based translation studies (e.g. Johansson and Hofland (1994)) advocates the constitution of bilingual corpora in which translated texts are supplemented with original texts in the two languages.

[3] Of course this can be problematized. It is not easy to define what a good translation is, and there are conflicting criteria on this subject as well. Gutt (1991) presents some cases where literal translation has been claimed to be the best translation. Still, some form of this paradox could be claimed to hold in the vast majority of the cases.

particularly rich in that area. Chapter 4 presents a selected overview of what a number of people have said on the subject in particular; here, I only note that the domain of representation of time in natural language has a number of interesting features in general:

First of all, it is a genuine semantic matter (possibly the easiest to identify in human language, besides number). Even though one could in principle select a syntactic phenomenon and delimit a (descriptive linguistic) study in that way, I assume that a <u>system</u> is structured semantically. In other words, that a coherent collection of formal indications, identifiable in syntax, morphology, and/or the lexicon, can only be explained by reference to a common particular conceptual domain. Studying the tense and aspect system then means that I will be studying any features (irrespective of descriptive level) that relate to time. This, incidentally, allows one to contrast / compare systems of two different languages, while it seems ill-founded to compare just forms.

Tense and aspect also constitute a core system, in that it is grammaticalized in the two languages. In consequence, it is not something that those languages convey optionally, or rarely. Tense and aspect markers are pervasive: almost any sentence contains such markers. In fact, it is necessary to account for them while processing almost any natural language utterance, and, therefore, extremely relevant for natural language processing in general.

Last but not least, tense and aspect are related to a concept which plays a crucial role in human life, time, which has been extensively discussed in philosophy, mathematics, and the natural sciences (most notably physics). Time has also been one privileged subject matter in artificial intelligence and in theoretical computer science. A semantic study of tense and aspect can thus be seen as an investigation of the role of time in "natural language metaphysics" and thus carry philosophical as well as linguistic import.

The subject matter is thus old and respectable. What is relatively new (though not unique; see e.g. Matveyeva (1985)) is the approach: the reliance on corpora, and the contrastive perspective. The former amounts basically to considering non-selected actually occurring examples, while the latter provides new facts for which an explanation must be sought. It is interesting that the contrastive approach has recently been defended, although with the opposite motivation, by Kamp (1991:64):

There have been few systematic attempts to explain the differences between the distinct ways in which different European languages use their tenses and temporal adverbs to express what appears to be the same information - at least that have made use of a precise formal framework. [...] explanations of such interlingual distinctions - explanations, in other words, of what might be considered idiosyncrasies of the tense-and-aspect systems of individual languages - require a more refined machinery than one may have seemed to need in the (non-comparative) study of tense-and-aspect systems of single languages. The impression that analyses of individual systems would be able to make do with substantially simpler machinery, however, must, I think, be illusory: Ultimately the additional machinery [...] will be just as indispensable for explaining the interaction of tense and aspect within a single language as it is to account for the features that set one system apart from another. If this is so, then comparative studies of tense and aspect may well be indispensable even when one is interested only in the workings of tense and aspect in the one language of one's choice.

1.3 Natural language processing

What is the relevance of a contrastive study of tense and aspect to natural language processing? To answer this question, I will first describe my view of this discipline, and then argue for what the right approach to it should be.

Natural language processing was, right from the beginning of artificial intelligence (AI), one way of doing AI. In those times, it was called "natural language understanding", and, like her mother AI , had two main ways of being understood by practitioners: One was to do applications that understood (reacted intelligently) to natural language texts (or utterances), a definition by goal; the other was to study the organization of natural language as reasoning and knowledge representation devices, in order to be able to automate and process natural intelligence, a definition by process. While, in theory, these two ingredients had to be there for a computational account of natural language, they drastically diverged in practice. At the same time, the community grew significantly, and separation from AI became more and more conspicuous. Given the easily seen difficulties of understanding natural language, the name of the discipline became natural language processing (NLP) or computational linguistics (CL). (In fact, I think the names NLP or CL have more to do with the previous background of the people involved than with actual work in progress, and I will use them interchangeably.[4])

Computational linguistics has come to encompass a wide range of activities, which can be summarized as four different viewpoints:

1. Linguistic philosophy: the concern with reasoning processes, ontology and epistemology, and logic-oriented computational models. Logicians are the kind of practitioners of this new NLP, and, in fact, they rephrased many of the questions posed by young AI.

2. Linguistic tools: the concern with formalisms, environments, tools for describing language, computability and the like. These matters, at the heart of computer science proper, now embody a large part of the effort in CL.

3. Language description: the traditional concern of grammarians *per se*, now in a computational environment. It should be noted, at this point, that a computational description of language has different requirements than pure linguistic theory. For one thing, it has to be objectively implementable, and, in addition, it must cope with ill-formed input.

4. System development: the creation of programs that do something with language (i.e., that handle, generate, or analyse significant portions of real text/speech). In fact, this is the only part visible outside the community itself. However, and because of the common separation of academy and market (at least in Europe), this part is the least practised by NLP researchers, and so much so that recent efforts have been done to promote it, not least by launching the new "language engineering" discipline.

Language engineering is thus a new branch of NLP, whose aim and primary scope is that

---

[4] Jan Engh has noted that the growing interest in computers in linguistics, having given rise to what he calls "computer-assisted linguistics", is another (unrelated) source for the CL discipline. A similar observation can be made about what is called "literary computing" or "computing in the humanities". I should then say that NLP is a proper subset (the interesting one) of CL. Still, I will be using the two terms without further precision in the text.

of taking the bull by the horns and get at real language. The emphasis of language engineering is thus on the raw material, and its processing, in computer applications (and therefore the term "engineering"). Engineering is, in my opinion, the art of solving problems in context, hopefully resorting to scientific "truths" originating in the lab (decontextualized). The creation of an engineering of language was thus an antidote against a too theoretical NLP/CL discipline. NLP/CL had become too theoretical, too concerned with discipline-internal questions, to be able to produce things of value for the end user, while, on the other hand, there was a wide market potential for such products.

However, an engineering approach alone is not a solution in itself, even though I consider it to be the right way to face natural language processing. There is good engineering -- and bad engineering. Good engineering, in my opinion, requires the use of a good description, and, additionally, good tools with which that description is expressed as well as good insight into the theoretical underpinning. But I hold that the importance of the viewpoints for language engineering is in inverse order to the one presented above. Therefore, in this dissertation I will attach more importance to (linguistic) description than to the way I couch it in formal terms or to the philosophical claims I may make.[5]

Another key concept in language engineering is broad coverage. It is a well-known fact of life that there is always a trade-off (for theories, programs and systems alike) between being encompassing or non-trivial, between breadth or depth of analysis. The NLP / CL community has chosen depth in the overwhelming majority of cases, with obvious damage to the engineering side. Prototypes, fragments, restricted sublanguages are terms often used, which I will not employ in this dissertation. For my NLP creed consists of one single assumption: it is a necessary requirement to handle real natural language. Any attempt to restrict the language accepted immediately turns the study (or the tool) into something else: at best, a model of a quasi-natural language. Even if this newly devised formal language has some theoretical interest, it is not natural language, and it will have no (automatic) practical applicability to real natural language processing (even if it embodies nice prototypes). I contend, in fact, that too much work in the discipline has been wasted in the design and implementation of systems which mimic natural language to some extent, but which can never be used to actually handle it.

Let me now return to the question raised at the beginning of this section. In fact, a contrastive study of tense and aspect is not *per se* relevant or irrelevant to natural language processing. Relevance will derive from the way it is conducted. I claim that some of the methodological options taken, as well as the questions raised, should make this dissertation relevant to NLP.

First of all, it is a study conducted on real text (real translations), which, as I claimed

---

[5] This dissertation includes a fair deal of linguistic description proper, due to the fact that there was no published account of Portuguese I could rely on. If there had been a computational description of Portuguese in the field of tense and aspect (or any other field), I could have proceeded by taking it as basis. However, I had to do most Portuguese monolingual analysis, as well as the contrast with English, from the start.

above, is an essential requirement.

Second, it is based on several measurable (i.e., objective) clues, such as grammatical markers, lexical items, actual translation pairs, even though it resorts to a fair dose of interpretation as well.

Third, a systematization of the results of the study is always attempted. This allows for (at least) partial implementation as a computer program. In particular, significantly more structure is given to what was *a priori* a mere collection of translation pairs.

Fourth, several NLP applications are hinted at based on the study described here.

Last but not least, this dissertation makes the theoretical point (with important practical consequences) that each language has to be described in its own terms. (It does not prove such a statement, but provides, in my opinion, strong evidence for it. To reject this claim, one would need to present a model handling my data without resorting to the sort of language differences I consider.) This implies that practical NLP systems that deal with more than one language cannot succeed without taking these language differences explicitly into account.

A natural conclusion of this section is a presentation of my view of the current situation in NLP, and of its future, positioning this dissertation in the more concrete field of computer-aided translation.

The era of toy systems based on a handful of example sentences is over. This change of perspective has brought with it the acknowledgement, by researchers and developers alike, that much is still unknown about language, i.e., one cannot start coding the language knowledge required, because it turns out that this knowledge is far too complex and strongly resists clean formalization. Therefore, a growing trend among empiricist-minded scholars has been to process corpora, dictionaries and thesauri, which is a very good thing in itself. What is not so good is that, overwhelmed by the complex problems raised by real text, and given that traditional linguistic theory is ill-equipped to handle them, many of those researchers have proposed too simplistic models or systems to handle natural language,[6] casting doubt on their basic adequacy to process language. This has led to a reaction from the more theoretically minded researchers, forming what is called the rationalist camp.[7]

This dissertation is an attempt to arrive at a satisfactory middle ground. On the one hand, it is geared to handle real text examples, being, therefore, strongly empirically based. On the other hand, it tries to preserve much of what is well known about language behaviour from theoretical studies.

The thesis is inspired by the belief that an important part of research and development in machine translation should be devoted to the study of translations already performed, and its purpose is to allow one to look at real examples through a more informed perspective than purely

---

[6] A prototypical example is the pure statistical approach to machine translation followed by the IBM group at T.J.Watson, cf. Brown et al. (1990). Their results are among the best claimed by existing systems.

[7] Such reaction, in the 70's, materialized in the famous ALPAC report (ALPAC, 1966), which was responsible for a large reduction of machine translation research all over the world.

quantitative or statistical processing.

As a general trend, NLP goals have moved from full automation to the development of tools to help in a human task. In translation, fully automated machine translation has given way to computer-aided translation as well as to tools for the translator. I will claim in this thesis that, in order to have a useful tool for handling large bodies of parallel texts, one must resort to semantic and contrastive knowledge. But the tool in turn becomes a very good testbed for exactly semantic and contrastive analysis: the creation of a tool to perform a human task is one of the best ways to learn how to perform that particular task.

This kind of "research bootstrapping" represents, I believe, one of the methodologically sounder conclusions of the research community in natural language processing.

Combining theoretically informed models with useful but not too ambitious applications seems to be a good middle ground between the sceptical attitude of those who hold that natural language is too complicated to handle by a computer, and that the task is therefore not worth trying ("first the theoretical issues must be settled"), and the naive attitude of those who attempt to do everything at once, and whose systems, displaying obvious inadequacies, feed the attitude of the sceptic.

1.4 Presenting the data

I have already stated that this dissertation uses "real text"[8] as its object of study, and have employed the term "corpus-based" to describe it. Here, I describe briefly the corpus and some general questions related to it. Part III provides a very detailed description of both the texts and the studies conducted on them.

As the empirical basis for my study, I used the texts of two books and their translation into the other language. These will be referred to as the EP corpus (i) and the PE corpus (ii) respectively:

(i) one (American) English novel, *The Pearl,* by John Steinbeck, Bantam Books, 1975 (1st edition, 1945). (Each chapter will be described by EPn, *n* ranging from 1 to 6.)

(i') and its translation into (European) Portuguese: John Steinbeck, *A pérola*, Publicações Europa-América, 1977, translated by Mário Dionísio.

(ii) a selection of Portuguese short stories by Jorge de Sena: Jorge de Sena, *Antigas e Novas Andanças do Demónio*, Edições 70, 5ª edição, 1984 (1st edition, 1978). The short stories chosen, together with their date of creation, were:

PE10: *A noite que fora de Natal* (1961);
PE11: *O grande segredo* (1961);
PE12: *Super flumina Babylonis* (1964);
PE3: *Mar de pedras* (1960);

---

[8] Incidentally, the expression "real text" is puzzling for the layman: Are there "unreal" texts, or abstract texts? Here, it contrasts with invented/constructed examples by linguists, i.e., natural language utterances produced outside a normal communicative situation, created with the mere purpose of investigating language. Often, they are syntactically simpler, and much harder to interpret, than "real" sentences. Many linguists are not even shy of producing doubtful or even starred sentences, generally prefixed by '?' or '*'. "Real text" may contain equally rare sentences, but they were produced in a particular communicative situation by a non-linguist.

PE6: *A comemoração* (1946);
PE8: *A campanha da Rússia* (1946-1960);
PE9: *Kama e o génio* (1964).
(ii') and their translation into (American) English: Jorge de Sena, *By the rivers of Babylon and other stories*, edited and with a Preface by Daphne Patai, Rutgers University Press, 1989.
The translators involved were:
PE10: A Night of Nativity (Frederick G. Williams);
PE11: The Great Secret (Daphne Patai);
PE12: By the rivers of Babylon (Daphne Patai);
PE3: Sea of Stone (Christopher C. Lund);
PE6: The Commemoration (Edward V. Caughlin);
PE8: The Russian Campaign (Daphne Patai);
PE9: Kama and the Genie (K.D.Jackson).

It may seem surprising that the study was conducted on literary texts, which seem to be farthest apart from the domain of applicability of natural language processing (notwithstanding the growing awareness in the humanities of the advantages of computer-aided techniques for literary analysis). But this choice was actually motivated by a set of independent considerations:

First, translations of novels and short stories are easily available, which is not the case with other kinds of literature, such as newspaper texts.

Second, literary texts are created with considerable care. After all, both writers and literary translators are the "best" users of natural language, in the sense that their language is considered paradigmatic, and their use of language a form of art.

Third, in the translation of literary text, translators are less tempted to depart radically from the source language text and create something more target language oriented than in other kinds of texts, out of respect for the original author. Even though this goes against the general belief that literary translation is the least literal of all kinds of translations, Monika Doherty (p.c.; but see Doherty (1992, 1995ab)) claims that one is liable to find freer translation pairs in other kinds of texts (such as technical or scientific writing) than in literary works.[9]

Finally, I was specifically interested in studying tense and aspect in narrative (as opposed to, for example, technical text) because one of the claims generally made in the literature on tense and aspect is its relevance to narrative. Undoubtedly, novels and short stories which report a sequence of events are the most developed instances of narrative text that can be found. Conversely, it is well known that technical texts display considerably less variety as far as tense and aspect devices are concerned.

In fact, one could argue that the object of this dissertation is tense and aspect in narrative (and not in general). To some extent, however, other kinds of discourse were also considered, as they were represented in direct speech and in non-narrative sections contained in the texts

---

[9] I should note that by "translation of technical/scientific writing" I mean, for example, a text in chemistry translated by (or with the help of) a specialist in that domain. I do <u>not</u> include catalogue or procedure types of "business translation", e.g. the essential parts of computer or car manuals translated for a company that produces computers or cars. As anyone can appreciate, that kind of translation has indeed to be very literal.

analysed. Even though a study like Caenepeel's (1995) on the influence of text type would be an interesting follow-up of the work reported in this dissertation, I believe that it is the meaning of tenses that produces their particular uses in narrative, and not the other way around, i.e., text type is a consequence and not a cause of tense distribution. People who do not share this opinion can, nevertheless, obviously read this dissertation as dealing exclusively with narrative discourse.

Summing up, I believe that there are good reasons to use literary narrative texts as the object of the study, provided they are not too marked as literary artifacts. I.e., it would have been disastrous to choose James Joyce or José Saramago, for the study of English or Portuguese. By contrast, the texts chosen are relatively short and are characterized by favouring the description of actions. In addition, I believe that none of the original texts can be considered difficult to understand by the ordinary (native) reader.

Other questions that should be taken up are corpus size and variety, and the associated issues of representativity and balance. Let me begin by a rudimentary external description of the corpus, presenting a few basic numbers in Tables 1.1 and 1.2. ("Translation pairs" are the result of the sentence alignment process. If there were only one-to-one sentence correspondences, it would equal the number of sentences (which would then necessarily be the same in both languages). "Tensed translations" are the tensed clauses in the source language which were translated.) Note that the detailed characterization of the texts is given in Chapter 9.

Table 1.1: General description of the EP corpus

Number of words in English	26060
Number of words in Portuguese	23262
Number of sentences in English	1628
Number of sentences in Portuguese	1861
Number of translation pairs	1602
Number of tensed translations	3744

Table 1.2: General description of the PE corpus

Number of words in Portuguese	25174
Number of words in English	27972
Number of sentences in Portuguese	1449
Number of sentences in English	1459
Number of translation pairs	1446
Number of tensed translations	3318

There are two main definitions of corpus in linguistics: one refers to any sufficiently large collection of items, and in most cases could be easily rephrased by the less scientific "the set of instances I observed". Then, there is the recent definition by John Sinclair, one of the most distinguished scholars in corpus research and the "father" of the COBUILD dictionary of

English, a corpus-driven enterprise. According to Sinclair, "A corpus consists of texts or parts of texts [...] selected according to external criteria -- their place in the sociocultural order, so that their linguistic characteristics are, initially at least, independent of the selection process" (Sinclair, 1995).

Unfortunately, neither definition is appropriate for my purposes, the first being too broad, and the second being too restrictive. Radical as it may be, this position is not unusual in the community. In fact, Sinclair himself has to resort to the concept "specialized corpus" to be able to take into account things like corpora of errors. Furthermore, as noted in Gavrilidou & Piperidis (1995:10), "the compilation of a *generic parallel* corpus, which will include many languages and be application independent, seems to be a contradiction in terms". In fact, it is obvious that parallel corpora must be highly specific regarding text type. Finally, I should note that most corpus researchers (I take for example Church and Gale as distinguished representatives) do not handle corpora in Sinclair's sense.

I propose, therefore, another definition of corpus, parameterized by the phenomenon of interest, and possibly also dependent on research goal: A corpus of X is defined as "a collection of appropriate linguistic entities containing X whose selection was not made dependent on X, such that its size is large enough to confidently represent most different instances of X in context". To define "large enough" formally one might need to perform statistical analyses in the domain in question. As a first approximation, I suggest that corpus size should be at least one order of magnitude above the number of possible values of X in context.

This definition, applied to my object of study, would be instantiated as: A corpus of <u>tense and aspect forms and their translation</u> is a collection of <u>pairs of sentences</u> whose selection was not made dependent either on tense and aspect forms or on their translation, such that its size is large enough to confidently represent most different instances of tense and aspect and their translation.

Likewise, Sinclair's definition could be accommodated to mine.[10] At first sight, one might use "texts" substituted for X and "complete[11] texts with classification of type" for the appropriate linguistic entities. The definition would thus become: A <u>text</u> corpus is a collection of <u>complete texts</u> whose selection was not made dependent on the text (actually it was made dependent on text type only) such that its size is large enough to confidently represent most different instances of text. However, some reflection on the use of a corpus by Sinclair and his followers makes me suggest the following simpler version: A corpus of <u>words in context</u> is a collection of <u>texts</u> (implying necessarily words in context), whose selection was not made dependent on the words in question, such that its size is large enough to confidently represent most different instances of words in context.

---

[10] Note, however, that Sinclair does not take up the question of size and/or representativity, so in this respect I am definitely adding something.

[11] By "complete" I mean some sort of coherent part, not necessarily a whole text: one chapter, one scene in a narrative, one abstract of a piece of news, etc.

This general definition has the advantage of making corpus size requirements <u>relative</u> to corpus of what, which is an intuitively desirable property, and which agrees, in fact, with actual practice. In fact, for corpus-based lexical studies the corpus size (in number of words) is drastically larger than for grammatical studies, because one is generally interested in the frequency distribution of a word, which can be modelled as contrasting with all others of the same category. On the other hand, it is foreseeable that grammatical studies, where the contrast is in terms of a finite and usually rather small number of alternatives, require much smaller corpora (in number of words).

For example, in Bacelar do Nascimento et al. (1993), a corpus of original Portuguese text containing 220,000 words was used to study 266 occurrences of the verb *dar* ('give'), while Church and Gale (1991) used a corpus of 890,000 aligned sentences to obtain results for the translation of lexical items with frequencies around several hundred.[12]

By contrast, Hakulinen et al., describing quantitative studies of Finnish, claim that the syntactical freezing point is fairly low: "the sample size above which you cannot really find any significant changes in the parameters and their frequencies is a corpus of a few hundred sentences" (Hakulinen et al., 1980:104, my translation), and, in general, researchers dealing with quantitative information about grammatical devices employ relatively small corpora (cf. e.g. Biber (1988) and Givón (1995)). For example, Biber (1993) claims that for features like the number of present and past tense verbs in English a 1,000-word sample size is enough. As far as the number of different texts per corpus is concerned, Biber considers a collection of 481 texts taken from twenty-three spoken and written registers a "relatively large and wide ranging corpus of English texts" (Biber, 1993:253), which seems to agree nicely with my requirement of an order of magnitude higher (relative to register).

In this dissertation, I have a corpus of more than 7,000 translated tensed clauses, even though -- measured in words -- the material comprises only around 50,000 words (in each language). As a rough estimate, this would allow me to have a convenient collection of instances to study ten different markers in each language (this corresponds to 10x10 = 100 possible different pairs of tenses and translations; one order of magnitude above thus asks for a corpus size of some thousands of instances).

Let me just invoke, on this subject, what Biber (1993) has to say on his paper devoted to corpus representativeness. General as this issue might be, his recommendations are still based on the assumption that the goal of a corpus is to study "the full range of linguistic variation existing in a language" (Biber, 1993:247). Now, this is undeniably one use of corpora, but it is certainly not the only one. He himself acknowledges this when discussing demographic sampling, claimed to be only appropriate for corpus design based on text production or text reception. According to Biber, "these kinds of generalizations, however, are typically not of interest for linguistic

---

[12] Some examples discussed in Church and Gale's paper have the most frequent translation pair for the English word as follows: *houses-maisons* - 159 occurrences; *risk-risque* - 363 instances; *ignore-compte* - 108 cases. Their example of *house-chambre*, appearing 30,584 times, is clearly exceptional.

research" (Biber, 1993:247). Further evidence for my claim that corpus design is goal-dependent, is his conclusion that to study the distribution of present and past tense 1,000-word samples are enough. If one wants to study their <u>grammatical</u> behaviour instead, the number of such verbs is almost an irrelevant question: that is a feature of a text, and tense is above all a feature of a sentence. To study tense in sentences (and not in texts) the sample "text" size should be the sentence, and not 1,000 words.

Even if the reader agrees whole-heartedly with my redefinition of corpus and thus is willing to admit that its size is sufficient, I suppose s/he would still hold that a balance of different authors and translators would be desirable (I note that there is one author for each language, one translator into Portuguese and five into English).

To this objection I have a somewhat weaker response: I indeed assumed that the variation of author was not crucially relevant, on the grounds that, lying at the heart of our native competence, tense choice does not seem to require a conscious decision on the part of the writer. Rather, it seems that it should vary much more significantly with the content of what is described, the linguistic context and the like, than with the subjective language user. However, I recognize that this assumption was left untested in this work.[13]

As far as the variability of the translation due to different translators is concerned, again no study has concentrated specifically on this matter, even though in the case of the translation into English the separation of data per short story of Sena provides some empirical data that could be used to settle the issue. The study of the individual contribution of the translator is, however, too far from my goal in this thesis. But it should be noted that on this issue my position is distinct from the one regarding the author's influence: I do not assume that variation is unimportant. On the contrary, I have carefully used my intuition in every translation analysed, which means that in practice two different translators were always consulted.[14]

To finish, I should like to draw attention to Sinclair's distinction between between corpus-driven and corpus-based studies: While for the former kind of studies it is the corpus that is primary, for the latter kind the corpus is used to test hypotheses, but one's own intuitions are always called into play. Contrary to his position, I believe that one's own intuitions (or judgement) should always be called into play, i.e., the corpus is a tool, and not the engine, of a linguistic investigation. And it is in this spirit, and only in this spirit, that I call my dissertation corpus-based.

1.5 Organization of the thesis

---

[13] The assumption is <u>not</u> that different authors have the same distribution of tenses (or other tense and aspect devices). This is trivially false: it is enough to recall that some authors write action stories while others write descriptions with hardly any action at all. The assumption is that in writing about the same subject, and <u>if</u> they used similar linguistic devices, different authors would come up with the same tense, i.e., the properties of the linguistic context and of what they wanted to convey are more important than their own style for this matter.

[14] Furthermore, and even though, for lack of time, the results of such an investigation could not be made to bear on the present text, I have checked many of my claims as regards the translation from English into Portuguese against an independent translation of the first three chapters of *The Pearl*, for which I am extremely grateful to my mother.

The thesis is organized in three parts, containing theoretical background, practical proposals, and empirical studies, respectively.

Part I, comprising Chapters 2 to 4, provides background for and discusses the general themes of the dissertation: language contrast (Chapter 2), translation (Chapter 3), and tense and aspect (Chapter 4). However, these chapters are not simply introductory, as they are used to present several non-consensual claims, suggesting some analyses as well. For example, Chapter 3 contains my own argument for translation-based corpus studies, as well as a semantic typology of translation pairs, while Chapter 4 presents my own critical assessment of some current tense and aspect theories.

Part II, comprising Chapters 5 to 8, contains the main contributions of the thesis, dealing with English and Portuguese tense and aspect: Chapter 5 introduces the descriptive framework I suggest, namely, a re-working of Moens's (1987) aspectual network for monolingual description, and a descriptive device termed the Translation Network for contrastive analysis. While the English tense and aspect system is also included in Chapter 5, since it contains at most local improvements regarding the literature of tense and aspect in English, the whole of Chapter 6 is devoted to tense and aspect in Portuguese, corresponding to a substantial amount of original work on the subject. Chapter 7, in turn, can be considered the main chapter of the thesis, in that it uses the contrastive model and the monolingual analyses of the two languages to provide an overview of the contrasts between English and Portuguese, based on real translations. Finally, Chapter 8 tries to give a formal description of what was suggested so far.

Part III, comprising Chapters 9 to 14, presents several empirical studies which constitute the background for most of the claims of the dissertation, and especially for Chapter 7, which explains many if not most of the phenomena pointed out. Their presence in this thesis is required to prove that there is solid empirical ground behind the claims and analyses provided. In particular, it shows not only that the examples of Chapter 7 were not made up, but that they were not chosen because they conformed to particular claims I wished to make, either; rather, they were illustrative of tendencies only perceivable after labour-intensive analyses of hundreds of translation pairs. Chapter 9 describes the corpus and the preliminary quantitative studies; Chapter 10 presents a case study of Imperfeito both in the Portuguese tense and aspect system and in what concerns translation into and from English; Chapter 11 displays a contrastive study of perception sentences; while Chapter 12 studies the English present perfect and how it is rendered in Portuguese, and Chapter 13 contrasts the pluperfect in the two languages based on the translation relations found. Finally, Chapter 14 studies the opposition Perfeito/Imperfeito in Portuguese and how it is rendered in English.

Chapter 15 concludes the thesis with a discussion of further work and a short summary of the main points presented in the dissertation.

The reader who wonders whether all chapters are equally worth while reading, in view of his or her particular interests, is advised to proceed at once to Chapter 15, where a more informative summing up of the contributions included in the whole text is attempted.

1.6 A note on the translation of the translation examples

One last observation is required here, regarding the presentation of the corpus examples.

As indicated in the notes on typographical conventions, for almost every translation pair I produced a gloss in English in order to make this dissertation fully accessible to those without knowledge of Portuguese.

This turned out, however, to be far from a minor task, because different translation pairs are presented for different reasons, and, therefore, the kind of translation better suited to the point I make in each case can be radically different as well.

Moreover, for the gloss to be useful at all, it also has to differ in some significant respect from the English text of the translation pair (which itself might range from a rather literal to a very free translation). In fact, if I was not able to produce such a gloss, backtranslation -- or retranslation -- was simply omitted.

My glosses thus range over at least the following cases, according to the nature of the translation and the issue at hand:

- the gloss provides a near to literal translation
- the gloss provides a semantically more appropriate translation
- the gloss provides another possible interpretation of the Portuguese text

In some cases, the criteria for a particular (set of) gloss(es) are made explicit in the text, but not in the overwhelming majority of the times, for obvious reasons.

As noted above, I do not claim in general that the glosses are cast in proper English. I have nevertheless avoided literal translations in the most obvious cases of obligatory differences between the two languages: For example, I always convert a standard Portuguese noun phrase structure into the corresponding English one, and Portuguese punctuation marks for direct speech into English ones. In general, I try to render (more) literally only what is related to tense and aspect. To do it properly is again far from trivial, however, as the reader is invited to try for him/herself.

On purpose, I have not marked in any way blatant errors appearing in the translation (in either direction). A careful reader with knowledge of the two languages can easily detect them. But the present dissertation was not concerned with the subject of translation errors in the first place (except when such errors crucially involved tense and aspect devices, in which case they are commented on in the main text). In addition, to distinguish between a radically different but possible interpretation and a blatant misinterpretation is again more difficult than it seems. The readers must therefore be cautioned against interpreting every case in which a different meaning is expressed in the gloss as a mistranslation (or believing that I do).

As a general rule, I suggest that Portuguese native speakers simply should not read the glosses, though I believe that they can be quite useful for non-native speakers (excepting the most literal glosses). Most importantly, I hope that a non-speaker of Portuguese can get a good grasp of most of what I claim by contenting her/himself with reading the two English versions of

a translation pair. After all, it was for this kind of reader that I produced the glosses in the first place: for my own part, I did miss an English rendering of the Russian translations when reading the studies in Maslov (1985).