Abstract

This dissertation is on tense and aspect and on translation, and was inspired by the desire to make their analysis computationally possible. It can be read from two complementary perspectives:

On the one hand, it concerns the investigation of the tense and aspect systems of Portuguese and English by means of real translations, taken as objective semantic data. On the other hand, it investigates translation between English and Portuguese through a detailed analysis of their respective tense and aspect systems.

The text is organized in three parts:

The first provides critical background for the three main subjects addressed in the thesis, the comparison/contrast of languages, the analysis of real translations, and the study of tense and aspect semantics. The main claims of this part are that (i) studies based on real translations are the only sound method for contrastive analysis, albeit (ii) languages are so different that they do not say the same things even when they are supposed to (as is the case of translation); (iii) tense and aspect semantics requires finer discriminations than those usually found in the formal literature.

The second part uses the ideas discussed in the first part to propose (i) a descriptive model for tense and aspect systems, (ii) and for their translation -- the Translation Network; (iii) a description of the English and Portuguese systems, respectively; and (iv) a significant number of contrasts between them, based on real translation pairs. Finally, it presents some steps towards a formal model of what was accomplished. As far as I know, this part presents the first systematic description of a significant portion of Portuguese tense and aspect grammar, as well as of its contrast regarding English.

Finally, the third part describes a set of empirical studies performed on aligned texts in English and Portuguese, through which the intuitions and knowledge presented in the first two parts of the thesis were obtained. The presentation of these studies constitutes the empirical justification for the claims put forward previously, while at the same time providing the reader with a basis for alternative analyses of the data.

Keywords: Tense, Aspect, Semantics, Portuguese, English, Translation, Contrastive studies, Natural Language Processing, Parallel corpora, Computer assisted translation.

Resumo

Os temas focados nesta tese são simultaneamente o tempo e aspecto e a tradução. Subjacente à sua criação esteve o desejo de tornar possível a análise computacional de ambos. Este texto pode pois ser lido sob duas perspectivas complementares:

Por um lado, o trabalho pode ser considerado como a investigação dos sistemas de tempo e aspecto do português e do inglês, através da análise de traduções, encaradas como dados semânticos objectivos. Por outro lado, pode ser visto como focando a problemática da tradução entre o inglês e o português através de uma análise detalhada dos respectivos sistemas de tempo e aspecto.

O texto está organizado em três partes:

A primeira contém uma panorâmica crítica das três questões gerais abordadas pela tese, ou seja, a comparação ou contraste entre duas línguas, a análise de traduções, e o estudo da semântica do tempo e do aspecto. As ideias principais aventadas podem ser resumidas do seguinte modo: (i) a análise contrastiva deve ser baseada no estudo de traduções independentes, apesar do facto de que (ii) duas línguas em geral são tão diferentes que não exprimem a mesma informação mesmo quando em princípio o deveriam fazer, como é o caso da tradução; (iii) a semântica do tempo e do aspecto requer distinções mais finas do que as geralmente feitas na literatura que emprega métodos formais.

A segunda parte, desenvolvendo as ideias expostas na primeira, (i) apresenta um modelo descritivo de um sistema de tempo e aspecto -- a rede aspectual--, e (ii) propõe um modelo para exprimir a tradução entre dois sistemas -- a rede de tradução. Além disso, (iii) fornece uma descrição dos sistemas do inglês e do português e (iv) sugere uma explicação para um número significativo de contrastes entre as duas línguas, sistematizados a partir de numerosos pares original-tradução reais. Finalmente, (v) alguns passos são dados na direcção de um modelo formal das propostas feitas. Esta parte constitui a primeira descrição sistemática (de grande parte) da gramática portuguesa do tempo e do aspecto, assim como do seu contraste em relação ao inglês, de que tenho conhecimento.

Finalmente, a terceira parte descreve um conjunto de estudos empíricos, realizados sobre textos paralelos alinhados em inglês e português. Foi através destes estudos que o conhecimento apresentado nas duas primeiras partes da tese foi alcançado. Por isso, a sua apresentação visa não só fornecer uma justificação empírica das propostas apresentadas mas também proporcionar ao leitor a possibilidade de formular análises alternativas dos mesmos dados.

Palavras chave: Tempo, Aspecto, Semântica, Português, Inglês, Tradução, Estudos contrastivos, Processamento de Linguagem Natural, Corpora paralelos, Tradução assistida por computador.

Foreword

When I started work on my dissertation in 1990, my intention was to write a study in semantics, focussing on the tense and aspect system of my native language, Portuguese.

Very soon, though, it became clear that the best way to study tense and aspect in Portuguese was to compare it with English, given the overwhelming size of the literature on tense and aspect in the latter language, which I could not overlook in any case.

However, rather than attempting to apply models devised for English to Portuguese, or depart radically from the existing models on the grounds that they were based on English and thus inappropriate for Portuguese, I chose to study systematically the differences and similarities between the two languages.

When Lauri Carlson, my supervisor, introduced the idea of using real translations to check the empirical adequacy of the models I was devising, early in 1992, I had no idea that my study was going to become so heavily involved with corpus studies and translation.

The path I followed in my studies, in fact, led from a formal theoretical approach to the study of natural language to a primarily empirical preoccupation with descriptive adequacy. This makes my dissertation more akin to language engineering, whose goal is to process real language, than to formal semantics or studies in language, logic and information, which was what I actually had in mind when starting my work.

Even though this situation may reflect to a large extent my own personal tastes and capabilities, I believe that this path is justified methodologically as well. For, as Sandström put it, "finding out what to formalize is a task that is logically prior to, and a necessary ingredient in, the development of a formal theory" (Sandström, 1993:2). It is therefore my hope that the work described here can be useful for formally-minded and descriptively-minded researchers alike.

Throughout the years, I became indebted to a great many people in the course of my studies. Undoubtedly, the one I owe most to is Jan, as a tireless reader, a careful linguist and an outstanding husband. It is no doubt due to his supportive attitude and continuous work at home that I was able to finish the dissertation despite two wonderful daughters and a time-consuming position as INESC's Natural Language Group leader.

I am also grateful to Lauri Carlson for considerably more than his supervision: he has had to put up with more trips to Portugal than he ever wished to, as well as babies and small girls around his office and home when I visited him. In addition, he has had to concern himself with a language hitherto unknown to him, Portuguese, as well as dealing with my unclear intuitions for years. His patience and constructive criticism are deeply appreciated.

Amilcar Sernadas deserves my gratitude as well, for accepting to support my PhD application as co-supervisor, on the part of the University, together with the corresponding work it involves, and for his readiness to help in formal matters whenever I asked him to.

Stig Johansson helped me obtain a visiting researcher status at the Department of British and American Studies at the University of Oslo, and allowed me to participate in the several seminars and workshops on contrastive and English linguistics there. In addition, he provided me twice with a learned audience for previous versions of some of the material presented here, and he also commented on and read carefully several drafted chapters, supporting me with encouraging words on a number of occasions. I truly consider him, unofficially, a "third supervisor".

In addition, I would also like to mention several other people I am indebted to:

Karen Jensen, George Heidorn and Stephen Richardson deserve special mention. They were my first mentors in natural language processing and from them I learned some fundamental attitudes, namely, the broad-coverage approach, and the belief that, ultimately, language is right. This inspires the desire to look at and handle language even when its behaviour is not sanctioned by any fashionable current theory. I am only sorry that circumstances have not allowed me to benefit from their advice or comments for the present thesis. I hope, however, that they will sympathize with most of my claims.

As far as studies of Portuguese are concerned, I thank Henriqueta Costa Campos for giving me all her writings on the subject of tense and aspect, as well as granting me access to her considerable stock of references. I am also grateful to Maria Fernanda Bacelar do Nascimento, for sharing with me her knowledge of Portuguese corpus processing and for general encouragement regarding the ideas I defend in this thesis. Finally, I am indebted to Kåre Nilsson for his generally positive appreciation of Chapter 6, as well as for insightful remarks and counterexamples, and to Signe Oksefjell, who provided me with an extremely thorough revision of Chapters 6 and 7 and suggested several improvements to them.

During the last stages of the writing process, several people helped with relevant comments and with the general revision: My special gratitude goes to Cristina Sernadas, who provided detailed comments by e-mail on a draft the very same day I had sent it, to Isabel Trancoso who sent several notes while proceeding with a careful reading, to Marc Moens who sent encouraging signals after getting a draft of the first seven chapters, and to Mona Flognfeldt who read a preliminary version of Chapter 4 and discussed it in detail with me.

At different times, I have had discussions with several people, which were beneficial to a deeper understanding of what I was doing. In addition to the afore-mentioned ones, I thank Bergljot Brynildsen, Östen Dahl, Monika Doherty, Eva Ejerhed, Sylviane Granger, Arne Larsson, José Carlos Medeiros, Amália Mendes, Fátima Oliveira, Josef Schmied, Dan Slobin, Kjell-Johan Sæbø and Kay Wikberg. One person that deserves separate mention is Wilfried Meyer Viol, by whose hand I became aware of the world of generics, and whose stay for a month in Lisbon considerably improved my knowledge of such and other matters.

I am also grateful to Antonietta Alonge, Dorit Bar-On, Monika Doherty, Helge Dyvik, Eva Ejerhed, Sylviane Granger, Edward Keenan, Judith Klavans, Arne Larsson, Isabel Leiria, Birger Lohse, Marc Moens, Kåre Nilsson and Dan Slobin for having provided me with copies of their

work and/or the work of their students, very difficult for me to get otherwise.

Historically, I am grateful to João Pavão Martins and Ernesto Morgado for having introduced artificial intelligence disciplines, and particularly natural language processing, in the curriculum at IST at the time I did my Master's degree; and to Jorge Ferreira Pinto and Luís Vidigal for launching the IBM-INESC Scientific Group in 1987 that made it practically possible for me to work in NLP ever since.

Since the dissertation is written in English and to some extent on English, I must mention my only teacher of English at school, Madalena Donas Botto, by far the most exacting but probably the most dedicated as well, who taught us a lot more about English grammar and British culture than the curriculum required, as well as George Lind-Guimarães, my teacher at the British Institute, who always managed to provide interesting classes. Since then, my English has considerably improved, especially due to the help of all readers of previous drafts of the present text. Still, I feel that it is to them that I owe my greatest debt in the matter. This is why I decided to write in British English, even though my corpus is composed of American English only.

Last but never least, I acknowledge my mother's support and influence: she transmitted to me her love of reading, and her preoccupations as a literary translator made me acquainted with the complexity of the task of translation from an early age. For this work, she volunteered to translate the English source text for me to obtain more data, as well as giving me as much practical assistance as she could.

I finally acknowledge the financial support I received for the work on my dissertation:

Junta Nacional de Investigação Científica e Tecnológica, JNICT, granted me a four-year scholarship (through the programs CIENCIA, three years, and PRAXIS XXI, one year part-time) and supported four trips and corresponding stays for the annual meetings with my supervisor.

IBM Portugal fully financed approximately the first six months of my studies.

The Finnish Ministry of Foreign Affairs granted me a three-month scholarship in Helsinki in the beginning of 1991.

The University of Oslo has granted me computer access, and excellent library facilities, in the last one and a half years.

During my second maternity leave (three months in 1993), I was paid solely by INESC, which also paid the corresponding part-time of the fourth year supported by JNICT. INESC also gave me a working place with computer facilities for most of the period I worked on this dissertation, and paid the kilos of paper that a dissertation costs.

In return, I supervised the work of the NLP Group in several projects. This work, although not directly related to my thesis, contributed to a broader view of the aims and methods of computational linguistics, as well as for a deepening of my understanding of Portuguese. The same can be said of the opportunity I had, in the autumn of 1993, to devise the curriculum and teach the discipline of Natural Language Processing at IST, thanks to the invitation of João

Pavão Martins. Although it certainly delayed the conclusion of the dissertation, it indirectly added to its content.

Typographical conventions

Quotations are presented inside double quotes, followed by their reference. Deleted material from quotations will be marked by [...]. Square brackets will also be used to specify any comments of my own added to quoted material. Alternatively, when the quotation is fairly long, it begins a new paragraph in a smaller font size, as is displayed below.

Names of Portuguese tense and aspect forms are capitalized, while those of English are not typographically marked. For instance, I will be discussing Imperfeito and simple past. When a name is composed of many words, only the first is capitalized, as in Mais que perfeito. Names of tenses in other languages are not graphically signalled, either, but are always preceded by the name of the language they belong to, e.g. French passé simple.

A particular sequence of text in either language is represented in *italics*. When the text is in Portuguese, I present a gloss or translation in simple quotes between parentheses ('like this').

The few translations suggested by me are presented as *original -> translation*.

On the contrary, translation pairs actually occurring in my corpus are presented as

original sentence
translated sentence
'gloss of whichever sentence is in Portuguese'

This gloss is what Baker calls "back-translation": "Back-translation [...] involves taking a text (original or translated) which is written in a language with wich the reader is assumed to be unfamiliar and translating it as literally as possible into English" (Baker, 1992:7). Furthermore,

> The quality of the English that appears in a given back-translation is not meant to reflect the quality of the translation itself. [...] the English used in the back-translation is not necessarily correct and is not to be confused with natural English.

And I add: whenever appropriate, I depict in the back-translation another interpretation of the Portuguese sentence than the one conveyed by the English sentence of the translation pair, without implying that it is the best or only correct interpretation. In cases of translation pairs involving a fairly literal translation, back-translation may be dropped.

Italicized boldface is used to emphasize particular parts of the translation pairs being discussed, while <u>underlining</u> is employed for emphasis in the text.

Finally, SMALL CAPITALS are used whenever reference is made to nodes of aspectual or translation networks.