Syntactical annotation of COMPARA: workflow and first results

Susana Inácio and Diana Santos

Linguateca, SINTEF ICT

Abstract. In this paper we present the annotation of COMPARA, currently the largest parallel corpora which includes Portuguese. We describe the motivation, give a glimpse of the results so far, and the way the corpus is being annotated, as well as mention some studies based on it.

1 Introduction

COMPARA (www.linguateca.pt/COMPARA/) is a large parallel corpus based on a collection of Portuguese-English and English-Portuguese fiction texts and translations, which has been developed and post-edited (or revised) ever since 1999 [1]. COMPARA has been designed with a view to be an aid in language learning, translation training, contrastive and monolingual linguistic research and language engineering.

In this paper, we present for the first time the syntactical annotation of COMPARA and its intellectual revision (or post-edition), after its automatic annotation with PALAVRAS [2] of Eckhard Bick and a post-processing similar to the one used by the AC/DC project [3].

We suggest how this work can be used to measure [4] both PoS annotation entropy and/or perplexity of the Portuguese language, and the amount of work involved in automatic annotation and its intellectual revision. We also mention other kinds of studies or applications that could profit from this annotation.

2 Motivation

As of today, COMPARA offers a lot of functionalities that we believe are original and useful, namely (a) kinds of search (according to alignment type, for translator's notes, reordered units, foreign words and expressions, etc.); (b) kinds of output provided (concordances, several kinds of distribution, parallel snapshot, etc.); and (c) kinds of subcorpus selection (language variety, individual texts, dates). A full description of the DISPARA system is provided in [5].

However, one of the most sought after options – well known from both the BNC [6] for English and the AC/DC for Portuguese – was the possibility to

make queries also based on part of speech, lemma, morphological and syntactical information.

After working since November 2004 in annotating COMPARA, and with a set of precise guidelines [7] in place, albeit still under development, we can now announce that (the majority) of the Portuguese side of COMPARA contains (revised) PoS, lemma, and morphological information, and that annotation of the English side, using the CLAWS tagger [8], is planned to start soon.

Let us present some examples of new search functionalities, to give some flavour of what is now possible: for forms ambiguous between grammatical categories, it is possible to (1) ask for their part of speech distribution, or (2) select (bilingual) concordances only of one grammatical interpretation. One can (3) get all forms of a given verb occurring in COMPARA by just selecting its lemma, as well as (4) obtain the distribution of forms or lemmas in a particular tense or in a particular syntactical or translational context. [9] presents contrastive examples where different syntactic realizations are relevant.

3 Kinds of studies allowed by annotated COMPARA

Already in 1993 the first quantitative studies about PoS ambiguity in Portuguese were published by Medeiros et al. [10] and work in that direction has continued, under different projects, reported in [11], [12] and [13]. Actual data related to annotation of COMPARA can be found in [4].

There are several ways to define (part-of-speech, or morphological) ambiguity: in the lexicon, out of context (as was done in [10-12] using the knowledge embedded in morphological analysers), providing therefore a measure of the work required by a parser; or in running text (in a large enough corpus), where one only considers as ambiguous forms which happen to have more than one interpretation in the corpus [4]. Obviously, these two kinds of measures provide superior and inferior limits to the ambiguity in practice.

Another kind of studies that COMPARA now allows is quantitative studies of translation patterns [14], until now difficult and time consuming, since they required manual selection and annotation.

Finally, we believe COMPARA to be large enough to furnish evaluation material for several NLP tasks such as word or sentence alignment, word sense disambiguation and even machine translation.

4 Workflow and comparison with Floresta Sintá(c)tica

In order to have the corpus return reliable information, it is necessary to check the output of automatic systems that attempt to do the complex job of assigning in context the right syntactical information to texts in natural language.

There are, however, many ways to perform such revision task, so it is interesting to document the way we are working, contrasting it with another project also concerned with human annotation of text in Portuguese, Floresta

Sintáctica [15,16]. Basically, we can say that the annotation of Floresta has proceeded in a depth-first way, with every syntactic detail checked and eventually corrected starting from the first sentence in the corpus, while the annotation of COMPARA took a breadth-first approach, starting with PoS annotation and proceeding from the most frequent to the least frequent items.

These choices were of course motivated by the different intended user models of the two corpora: people interested in Portuguese syntax and/or quantitative studies or training of parsers for Floresta, while a much broader range of users for COMPARA, probably interested in (contrastive) lexical studies as well.

A list of all forms (or lemmata) was created per major part of speech and one proceeds by revising all contexts in which these words occur (starting from the top of the list, the most frequent first). This results also in a very different documentation activity: while for Floresta every piece of information present has to be documented, and note that *constructions which [a]re individually very rare [a]re collectively quite common* [17], in COMPARA we were instead concerned with other kinds of information such as guidelines about how to decide on a particular PoS in context, which, as far as we know, have never been published for Portuguese before. Grammars tend to describe phenomena with clearcut cases, while heuristic rules, such as the following, document how decisions were taken in a particular annotation task.

When one form can be both nominal and adjectival, choose noun:
- when it functions as a vocative: PPEQ2(741): E disse-me ele: «Que quer você, **amigo**?
- when it refers to a profession or activity: PBMA3(555): -- No tempo em que eu era **administrador**....
When one form can be both verbal and adjectival, choose adjectival:
- when senses are different: EBDL3T1(773): Mentiroso extraordinariamente convincente, o Boon: mesmo após anos de convívio **chegado** conseguia levar-nos, ...
- when it is modified by an adverb: EBDL1T1(1350): Ela adormeceu com um ar bastante **satisfeito**.

Also, the Floresta team has primarily dealt with syntactic vagueness or ambiguity (involving more than one token), while in COMPARA we have exclusively dealt with PoS vagueness or ambiguity [18].

Acknowledgement

This work was partially supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

References

1. Ana Frankenberg-Garcia & Diana Santos. "Introducing COMPARA, the Portuguese-English parallel translation corpus", in F. Zanettin, S. Bernardini and D. Stewart (eds.), *Corpora in Translation Education*, Manchester: St. Jerome Publishing, 2003, pp. 71-87.
2. Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.
3. Santos, Diana & Eckhard Bick. "Providing Internet access to Portuguese corpora: the AC/DC project", in Gavriladou et al. (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), pp. 205-210.
4. Diana Santos & Susana Inácio. "Annotating COMPARA, a grammar-aware parallel corpus", *Proceedings of LREC 2006*, Genoa, Italy, May 2006.
5. Diana Santos. "DISPARA, a system for distributing parallel corpora on the Web", in Elisabete Ranchhod & Nuno J. Mamede (eds.), *Advances in Natural Language Processing (Third International Conference, PorTAL 2002, Faro, Portugal, June 2002, Proceedings)*, LNAI 2389, Springer, 2002, pp.209-218.
6. Guy Aston & Lou Burnard. *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, 1996.
7. Susana Inácio & Diana Santos. "Documentação da anotação da parte portuguesa do COMPARA". In progress. First version: 9 December 2005. http://www.linguateca.pt/COMPARA/DocAnotacaoPortCOMPARA.pdf
8. Rayson, Paul & Roger Garside. "The CLAWS Web Tagger". *ICAME Journal* **22**. HIT-centre - Norwegian Computing Centre for the Humanities, Bergen, pp. 121-123.
9. Diana Santos. "Breves explorações num mar de língua". *Ilha do Desterro* (2006).
10. José Carlos Medeiros, Rui Marques & Diana Santos. "Português Quantitativo", *Actas do 1.º Encontro de Processamento de Língua Portuguesa (Escrita e Falada) - EPLP'93* (Lisboa, 25-26 February 1993), pp. 33-8.
11. Diana Santos. "Português Computacional", in Inês Duarte & Isabel Leiria (orgs.), *Actas do Congresso Internacional sobre o português, 1994, Volume III*, Lisboa: Edições Colibri / APL, Junho de 1996, pp. 167-84.
12. Diana Santos, Luís Costa & Paulo Rocha. "Cooperatively evaluating Portuguese morphology", in Mamede et al. (eds.), *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003*, Springer, 2003, pp. 259-66.
13. Luís Costa, Paulo Rocha & Diana Santos. "Organização e resultados morfolímpicos". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. No prelo.
14. Santos, Diana. *Translation-based corpus studies: Contrasting English and Portuguese tense and aspect systems*. Amsterdam/New York, NY: Rodopi, 2004.
15. Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. ""Floresta sintá(c)tica": a treebank for Portuguese", in M.G. Rodríguez & C.P.S. Araujo (eds.), *Proceedings of LREC 2002*, (Las Palmas 29-31 May 2002), ELRA, 2002, pp.1698-1703.
16. Afonso, Susana. Árvores deitadas: Descrição do formato e das opções de análise na Floresta Sintáctica. http://www.linguateca.pt/Floresta/ArvoresDeitadas.pdf. In progress. First version, 2004.
17. Sampson, Geoffrey. "The role of taxonomy in language engineering", *Philosophical Transactions of the Royal Society (Mathematical, Physical and Engineering Sciences)* **358**, 4, 2000, pp. 1339–5.
18. Santos, Diana. "The importance of vagueness in translation: Examples from English to Portuguese", *Romansk Forum* **5** (1997), Junho 1997, pp. 43-69.