Annotating COMPARA, a Grammar-aware Parallel Corpus

Diana Santos, Susana Inácio

Linguateca, www.linguateca.pt, Oslo node at SINTEF ICT
SINTEF ICT, Pb 124 Blindern, N-0314 Oslo, Norway
Diana.Santos@sintef.no, susinacio@sapo.pt

Abstract

In this paper we describe the annotation of COMPARA, currently the largest post-edited parallel corpora which includes Portuguese. We describe the motivation, the results so far, and the way the corpus is being annotated. We also provide the first grounded results about syntactical ambiguity in Portuguese. Finally, we discuss some interesting problems in this connection.

1. Introduction

COMPARA (www.linguateca.pt/COMPARA/) is a large parallel corpus based on a collection of Portuguese-English and English-Portuguese literary source texts and translations, which has been developed and post-edited ever since 1999 (Frankenberg-Garcia & Santos, 2003). COMPARA has been designed with a view to be an aid in language learning, translation training, contrastive and monolingual linguistic research and language engineering.

This paper has two aims: first, to present for the first time the syntactic annotation of COMPARA and its intellectual revision (or post-edition), after its automatic annotation with PALAVRAS (Bick, 2000) and a post-processing similar to the one used in the AC/DC project (Santos & Bick, 2000). We document and describe the new functionalities to a wider public. Second, we provide a quantitative assessment of the PoS disambiguation required, as well as give some measures of categorial ambiguity in Portuguese.

2. Motivation

As of today, COMPARA – through the DISPARA system (Santos, 2002) – 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). Santos & Frankenberg-Garcia (subm.) provide a user study of these capabilities.

However, one of the most sought after options ever since launching COMPARA was the possibility to make queries (also) based on part of speech, lemma, morphological and syntactical information – well known from both the BNC (Aston & Burnard, 1996) for English and the AC/DC (Santos & Sarmento, 2003) for Portuguese.

This requires obviously an annotated corpus, and therefore annotation of COMPARA started November 2003 and has proceeded at a steady pace the way it is documented here and in the current set of precise annotation guidelines (Inácio & Santos, in progress).

As of today, we can announce that the Portuguese side of COMPARA is annotated with part of speech (PoS), lemma, morphological information (mostly revised) and syntactic function (not revised). The annotation was automatically performed by the PALAVRAS parser developed by Eckhard Bick, followed by a thorough intellectual revision which has resulted in parser improvements and new versions being applied to the same material and to new additions to COMPARA. Work on annotating the English side of COMPARA, using the CLAWS PoS tagger (Rayson & Garside, 1998) is just starting.

Let us give the readers a flavour of what COMPARA offers as new search functionalities: it is possible to look for the part of speech distribution of forms known to be ambiguous between grammatical categories, as well as select concordances of only one grammatical interpretation. Conversely, one can get all forms of a given verb occurring in COMPARA by just selecting its lemma, or obtain the distribution of forms or lemmas in a particular tense or in a particular syntactic context. Table 1 provides some examples, expressed in plain IMS Corpus WorkBench (CWB) syntax (Christ et al., 1999).

Request	(Truncated) output
Search expression: *criado* Chosen output: PoS distribution	noun 93 verb18
Search expression: [lema="ter"] Chosen output: form distribution	*tinha* 3131 *ter* 2171 *tem* 1436
Search expression: [lema="ter"] Chosen output: tense distribution	Imperfeito 4605 simple present 3482 infinitive 2359...
Search expression: [lema="comer"] Chosen output: concordance	Every translation pair with verb *comer* (but no homographs)

Table 1: Examples of search in annotated COMPARA.

3. Implementation: the revision process

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. Post-editing is thus required in order to create a COMPARA that contains trustworthy information and can therefore be used to perform reliable contrastive studies or language learning or teaching activities.

The first issue we dealt with was to arrive at a consistent and correct PoS assignment. The section "Ambiguity

measures" below gives a quantitative measure of the work involved and the dimension of the problem.

3.1. Workflow

A list of all forms (or lemmata) was created per part of speech (see Table 2 for the current top words), and one proceeds by revising all contexts in which these words occur (starting from the top of the list, the most frequent first).[1]

Nouns	Adjectives	Verbs	Adverbs
coisa 3039	grande 2020	ser 28,813	não 20,072
casa 2462	bom 1775	ter 13,706	já 2870
vez 2375	novo 1140	estar 10,768	depois 2665
dia 2204	pequeno 978	dizer 7951	só 2230
tempo 2148	próprio 638	ir 6587	ainda 2151
mão 1976	velho 632	fazer 5825	também1808
homem 1947	cheio 568	poder 4988	nunca 1788

Table 2: Most common lemmata in COMPARA 6.7.1.

The revision process in context is done using a special version of COMPARA's Complex Search Web interface – with some added functionalities, but basically with just no limit on the number of concordances returned – but the actual changes are performed in the annotated text files (see the annotation page[2] for details).

Let us take a closer look at the revision process. The human reviewer will, in principle, look only at the occurrences which seem to her PoS ambiguous: for example, the word *janela* doesn't need to be revised, since it can only be a noun. [3] On the contrary, the disambiguation of the words *casa* and *criado* needs to be looked into, since out of context they can be verbal or nominal forms.

3.2. Revision problems

In this process, many other things (which we will not have room to discuss in detail here, but which are being carefully documented elsewhere) have to be tackled. The tokenization issue is relevant in three different cases:

- proper noun delimitation (PROP), and the associated question of capitalization: how to deal with the syntactic interpretation of capitalized common nouns;
- verbs with enclitics, still a major headache for parsers of Portuguese (Santos, 1999), which is unfortunately not yet completely solved by PALAVRAS, and is the source of ill-formed parses which have obviously consequences for the other words in the clause as well);
- and multiword expressions, another thorny subject on which there is no theoretical agreement among linguists: for examples, see (Santos, Costa & Rocha, 2003; Santos & Gasperin, 2002). For this issue we took a very conservative path, providing both PoS to the individual constituents and a conjoined PoS to the MWE itself, trying to satisfy different plausible requirements.

Since roughly half of Portuguese COMPARA consists of translated text, issues of translationese, i.e., use of English words or conventions (Gellerstam, 1986) and, in general, foreign quotations and named entities are inescapable and add to the number of grey zones we have to provide some solution for. As COMPARA is a corpus of fiction, we also have to deal with the many problems of rendering direct speech in literary text, especially when in translation from languages with different traditions and rules. That this is a major problem even for consistent encoding of sentences or translation units in a parallel corpus has already been argued in (Santos, 1998).

Finally, it should be noted that COMPARA also includes texts with non standard typographical conventions which lead to unproper tokenization and sentence separation and therefore bring additional problems both for automatic and human parsing. A concrete example are the two texts by José Saramago, the Portuguese Nobel prize winner, known by his peculiar and unstandard way of writing in what concerns punctuation. Likewise, texts by authors from Portuguese speaking African countries also pose more challenges to annotation and post-editing, given that the lexical richness and other writing conventions of African Portuguese are not yet encompassed by printed dictionaries and PALAVRAS alike.

To make our claims more concrete, let us present an (unstructured) list of interesting details about the parsing of Portuguese, to our knowledge not yet settled by Portuguese or Brazilian grammarians:

- What is the best way to analyse *obrigada* ("thank you", feminine form) or *se calhar* ("if it happens", "maybe")?
- How to deal with the lemmata of forms which have changed spelling (like *vôo* to *voo*) in the course of the history of Portuguese language, or which are different in different varieties of Portuguese?
- Should two prepositions on a row (or a preposition followed by an adverb), as *até a* or *por sobre*, be considered a (prepositional or adverbial) multiword?
- What is the best way to deal with the constituents of expressions such as *em vista* ("prospective", "desired", "wished for") or *de imediato* ("at once")? [4] Should one assign different "syntactic meanings" to the words involved, or rather not assign such labels if the forms cannot occur alone with such a meaning, considering them for all purposes similar to *entanto* in *no entanto* ("however"), which can only occur in this expression? In (Santos & Gasperin, 2002), such

[1] In fact, this is a truth with modifications: this workflow is true about COMPARA version 5.7 (first 54 pairs of texts). From then on, revision proceeded, for the new pairs, only for the current lexical items being revised. Only when this list was ready did revision of the full list of adjectives begin, now in version 6.7 with 70 pairs. A list of all nouns in the remaining files that have since been added was prepared, to be revised after finishing the adjective list.

[2] http://www.linguateca.pt/COMPARA/Annotation.html

[3] But then it has to be checked whether it is not part of a proper name (PROP) or has been assigned a different PoS somewhere in the corpus.

[4] The ordinary translations of *vista* as a noun are *view*, *vision*, *sight*, and as a past participle it is translated by *seen*. One might consider it here as a metaphorical use of the sight meaning. However, *imediato* as a noun is a rank in a boat, and *imediato* as an adjective means "immediate". Not only it is controversial to assign an adjective reading after a preposition, but if one considers it a noun, its reading in this case is obviously unrelated to the noun meaning.

problems are amply discussed in the context of evaluating a parsed corpus.

- Are the words *ao* and *à* in the following (made-up) sentence *Ao chegar, cumprimentou todos à la Robert de Niro* ("As he arrived, he greeted all like Niro") still to be analysed as contractions?

And many, many other questions have to be dealt with, and documented, as we go along, which is why the guidelines have to be considered work in progress.

3.3. Dealing with indeterminacy

While performing this extensive revision, we note that many cases have also had to be taken into account, and accordingly classified and annotated, such as ellipses and vagueness: in fact, we believe (Santos, 1997), that a constitutive property of natural language is that it has vague categories and often does not require one to choose. Although we tried to find sensible rules and consistent decisions, we did preserve cases of pure vagueness (where human annotators could not decide or agree) by using vague categories of the kind A_B (N_ADJ, ADJ_V, ADJ_ADV, N_PROP, etc.). Real examples of these situations are provided in Figure 1, where the Pos classifiation was added for readibility.[5]

«Ele é um **velho**$_{N\_ADJ}$ **sábio**$_{N\_ADJ}$.	`He's a wise, elderly man.
Tardieu parou à frente de uma casa com uma pequena tabuleta **pintada**$_{ADJ\_V}$ e colocada sob o candeeiro por cima da porta: «Pension Bellegarde.»	Tardieu halted in front of a house with a small painted sign under a light over the door, *«Pension Bellegarde»* .
A esposa Aguiar, comovida, apenas pôde responder logo com o gesto; só instantes depois de levar o cálix à boca, acrescentou, em voz **meia**$_{ADJ\_ADV}$ surda, como se lhe custasse sair do coração apertado esta palavra de agradecimento: -- Obrigada.	The wife was so moved she could respond only with answering gesture. It was not until several moments after raising the glass to her lips that she added in a muffled voice, as if her full heart kept back the words, «Thank you.»
Céu$_{N\_PROP}$ e **Inferno**$_{N\_PROP}$ são concepções sociais para uso da plebe -- e eu pertenço à classe média.	Heaven and Hell are social concepts created for the sole use of the lower classes and I belong to the middle classes.

Figure 1: Maintaining vague PoS assignment

3.4. Annotation progress

Since every other month (in average) there are new files added to COMPARA – check the Contents page[6] for update frequency and actual dates, and see Figure 2 for size increase –, the post-edition process is never ready (rather, it has to start afresh for the new pairs). Still, it is convenient to indicate post-edition progress by reporting on how many words in each category have already been revised.

Table 3 gives a quantitative overview of post-edition progress in terms of major PoS categories (counting separately capitalized and non capitalized forms).

PoS	Distinct words (types)	Words (tokens)	Revised distinct words	Revised total words
N	21,017	266,636	19,189	219,927
ADJ	10,020	64,493	116	19,408
V	39,504	281,923	-	-
ADV	1,584	67,380	-	-
PROP	7,972	39,612	-	-

Table 3: Revision progress end February 2006

Table 3 may be misleading if the reader infers that no adjectives or verbs have so far been changed or post-edited, which is obviously not true. A lot of post-edition has occurred for many of the forms that are now marked as ADJ or V. In fact, what is still lacking is a systematic check of V forms. The same is also true of proper nouns (PROP),[7] where considerable work has been put into their correct delimitation, but which have not yet been fully revised as a set. The numbers in Table 3 are only indicative, since a PoS change in any of the yet unrevised forms will also change another row in the table.

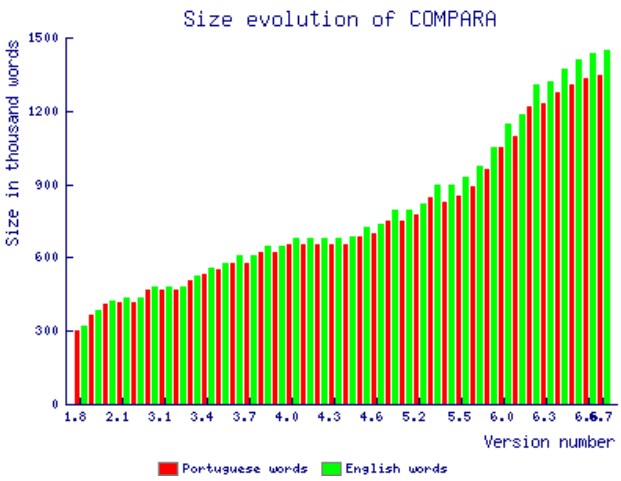

Figure 2: Size in words in COMPARA until version 6.7.

4. Ambiguity measures

A byproduct of our revision work is that it allows us to measure parsing difficulty in Portuguese, and provide language-specific measures which are hard to find, in fact, for any language. In order to give some idea of what is involved, table 4 quantifies the most common cases of categorial ambiguity in Portuguese, as present in COMPARA, version 6.7.1.

In theory, there are three ways of measuring ambiguity in a corpus (as sample of language): the most open is the one that takes into account the general rules of the language and makes use of a minimal number of forms (grammatical or "start words", to make the analogy with IR "stopwords"), and produces a theoretical upper limit of

[5] Although we present the usual bilingual concordance as provided by COMPARA, note that we are only considering the Portuguese text for syntactical annotation.

[6] http://www.linguateca.pt/COMPARA/Contents.html.

[7] Most proper nouns are multiword, so the column "tokens" in Table 3 has been normalized by size in words, that is, *Mrs. Robinson* is counted as one, not two words.

how ambiguous a given text/language can be.

PoS	Types	Tokens	ambiguous types		ambiguous tokens	
			#	%	#	%
nouns	19,845	266,636	1378	6.943	48,193	18.07
adjectives	9,641	64,493	1378	14.29	48,193	74.72
total	28,108	331,129	1378	4.902	48,193	14.55
PoS	Types	Tokens	ambiguous types		ambiguous tokens	
			#	%	#	%
adverbs	1,343	67,380	49	3.648	10,784	16.00
adjectives	9,641	64,493	49	0.508	10,784	16.72
total	10,935	131,873	49	2.42	10,784	8.177
PoS	Types	Tokens	ambiguous types		ambiguous tokens	
			#	%	#	%
verbs	35,253	281,925	671	1.903	17,969	6.37
adjectives	9,641	64,493	671	6.959	17,969	27.86
total	44,223	346,418	671	1.517	17,969	5.19
PoS	Types	Tokens	ambiguous types		ambiguous tokens	
			#	%	#	%
verbs	35,253	281,925	1149	3.259	82,787	34.57
nouns	19,845	266,636	1149	5.789	82,787	31.04
total	53,949	548,561	1149	2.129	82,787	15.09

Table 4: Measure of grounded ambiguity in COMPARA, using case-insensitive comparison.

The most common is the one that uses regular lexical knowledge (the information that *casa* ("house", "marries") can be V or N but *janela* ("window") can only be noun) and assesses what we may call dictionary-based ambiguity. Both these measures, based on a small PoS-annotated corpus, were first presented for Portuguese in (Medeiros, Marques & Santos, 1993), but different ways of getting at this kind of information are also reported e.g. in (Santos, Costa & Rocha, 2003).

However, this "dictionary-based ambiguity", or potential ambiguity given a particular lexicon, does not take into account the fact many of these ambiguous forms reflect arcane or at least extreme rare uses. After all, there is no general purpose dictionary. Any dictionary, as lexicologists know, has a purpose and a user model, and most probably no appropriate dictionary to do this ambiguity measurements has ever been built.

What we are calling here "grounded ambiguity" provides clearly a lower limit of ambiguity in that it only considers forms which actually have been assigned different PoS in the corpus. This means, for example, that only forms with more than one occurrence can be considered ambiguous, and that many consensually ambiguous forms will not be labelled as such because they happen to have been used in COMPARA in only one of the several grammatical interpretations.

In table 4, we counted (for pairs of categories) how many forms were ambiguous between the two categories. We did not count the cases of ambiguity with proper nouns or named entities (such as names of works of art) in order not to artificially increase the ambiguity measures. In order to have an estimate not only of the lexical ambiguity (in types) but also of the textual ambiguity (in tokens) we provide the ambiguity values for both.

The counts were based on the full corpus (some of it revised, some as yet unrevised, as seen above). This helps to bring both human knowledge (supplementing the parser and its lexicon) and parser knowledge (and therefore the information in its lexicon) into the picture. In fact, these numbers can be read from two perspectives: characterizing the language in terms of information-theoretic contents, and measuring parser's work (how many decisions are required – and thus how many revisions by the human expert as well). Even though they are preliminary (or rather, based in a not fully revised corpus yet) they help to assess the work involved.

5. Comparison with the Floresta Sintá(c)tica project

Another project performing human revision of the output of PALAVRAS is the Floresta Sintáctica project (Afonso et al., 2002; Afonso, 2004-2006)[8]. By highlighting the differences we hope to clarify the strengths and weaknesses of either approach.

5.1 Genre and availability

COMPARA includes original and translated published fiction, while Floresta is composed of newspaper text. For this reason, and due to the more stringent requirements of copyright holders, only online access to concordances (and distribution figures) is allowed for COMPARA, while Floresta, in addition to online search (Santos, 2003), is also freely available for download.

Whether genre will bring substantial differences to the annotation process and result, for example in terms of PoS classification, is an empirical question, about which we have no results to present but wish to investigate later.

5.2 Methodology

The most interesting difference from our point of view is the way the revision proceeds: while for COMPARA we used a breadth-first strategy, dealing with the revision of the major PoS assignments in the whole corpus at first, in

[8] http://www.linguateca.pt/Floresta/.

Floresta a depth-first approach was followed, revising completely (every syntactic property of) every sentence.
It is easy to compare numbers, since the sizes of the resources created by these two projects stand roughly in a 1:2 relationship.[9] According to the Floresta site, Floresta in its current (7.3) version features 40,522 revised nouns, an order of magnitude less than those covered in COMPARA.
An obvious consequence of this difference in approach becomes apparent in the documentation associated to each project: while Floresta has striven to document every piece of syntactic information available – while not being able to provide detailed procedures for testing and/or consistency checking of the material (see at this respect Sampson (2000), who claims that *constructions which [a]re individually very rare [a]re collectively quite common*), documentation of the annotation of COMPARA has produced a wealth of detail about very specific subjects.
For example, we have developed detailed guidelines about how to choose a given PoS in context (or whether vagueness should be preserved), that will allow the measurement of the weight of several heuristics and, more importantly, allow users to agree or disagree with the criteria, which are made explicit. See examples in Figure 3 and in (Inácio & Santos, 2006).

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**

Figure 3: An heuristic to decide between N and ADJ.

On the other hand, while the Floresta team has primarily dealt with syntactic vagueness or ambiguity (involving more than one token), in COMPARA we have exclusively dealt with PoS vagueness or ambiguity so far, as reported above.
In any case, in order to reduce the burden of documentation, we have referred to the Floresta documentation whenever options taken in COMPARA are shared, and have precisely documented whenever different choices have been made.

6. Concluding remarks

Consistent and error-free PoS annotation of large bodies of Portuguese text is something that is hard work but also theoretically and computationally interesting, as beautifully argued by (Sampson, 2000). In fact, we agree with him that defining a precise annotation scheme for Portuguese is a worthier goal than creating treebanks or annotated corpora.[10]

With this project, we are produced two kinds of publicly available material relevant to those interested in the parsing of Portuguese: the documentation, which is the distillation of the actual annotation work, and the corpus, offering the ability to query a sizeable corpus with revised grammatical annotation.
In this paper, we give a flavour of the complex issues involved, as well as try to characterize the richness of the resource we are building. Readers are welcome to use it and suggest improvements.
In addition, this work allowed us to estimate (and thus produce relevant empirical data about) PoS ambiguity of Portuguese, which we believe is very relevant in order to evaluate general parser's performance later on, providing ceiling data, i.e, what are the upper limits to judge parser performance above which there is no human consensus.
The corpus and the debates around its annotation also constitute an ideal testbed to discover and document unsolved problems of Portuguese syntactical description, and to start more encompassing contrastive grammatical studies as proposed in (Santos, 2004).

7. Acknowledgments

This work was partially supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI. We are grateful to Ana Frankenberg-Garcia and the other members of the COMPARA team, who all contribute to the development of this resource.

8. References

Afonso, Susana. (2004-2006). Árvores deitadas: Descrição do formato e das opções de análise na Floresta Sintáctica. In progress. http://www.linguateca.pt/Floresta/ArvoresDeitadas.pdf

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. (2002). "Floresta sintá(c)tica": a treebank for Portuguese. In M.G. Rodríguez & C.P.S. Araujo (Eds.), Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation (Las Palmas, 29-31 May 2002) (pp. 1698--1703). ELRA.

Aston, Guy & Lou Burnard. (1996). The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Bick, Eckhard. (2000). The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus: Aarhus University Press.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann & Esther Koenig. (1999). The IMS Corpus Workbench: Corpus Query Processor (CQP): User's Manual. Institute for Natural Language Processing. University of Stuttgart, March 8, (CQP V2.2). http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/.

Frankenberg-Garcia, Ana & Diana Santos. (2003). Introducing COMPARA, the Portuguese-English parallel translation corpus. In F. Zanettin, S. Bernardini and D. Stewart (Eds.), Corpora in Translation Education (pp. 71--87). Manchester: St. Jerome Publishing.

Gellerstam, Martin. (1986). Translationese in Swedish novels translated from English. In Lars Wollin & Hans

[9] Floresta Sintáctica is supposed to annotate 2 million words (one in the Portuguese and the other in the Brazilian variety); while COMPARA at start of the annotation process reported here ca. 1 million words (in four varieties of Portuguese).

[10] Although Sampson notes: *At present, most computational linguists see the point of an annotated corpus, but few see the point of putting effort into refining schemes of annotation.*

Lindquist (Eds.), Translation studies in Scandinavia, (pp. 88--95). Lund: CWK Gleerup.

Inácio, Susana & Diana Santos. (2005-2006). Documentação da anotação da parte portuguesa do COMPARA. In progress. First version: 9 Dec. 2005. http://www.linguateca.pt/COMPARA/DocAnotacaoPortCOMPARA.pdf.

Inácio, Susana & Diana Santos. (2006). Syntactical nnaotation of COMPARA: workflow and first results. In Vieira et al. (Eds.), VII Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2006) (Itatiaia, RJ, 13-17 May 2006). Springer.

Medeiros, José Carlos, Rui Marques & Diana Santos. (1993). Português Quantitativo. In Actas do 1.º Encontro de Processamento de Língua Portuguesa (Escrita e Falada) - EPLP'93 (Lisboa, 25-26 February 1993) (pp. 33--38). Lisbon.

Rayson, Paul & Roger Garside. (1998). The CLAWS Web Tagger. ICAME Journal 22, 121--123.

Sampson, Geoffrey. (2000). The role of taxonomy in language engineering. Philosophical Transactions of the Royal Society (Mathematical, Physical and Engineering Sciences) 358 (4), 1339--1345.

Santos, Diana. (1997). The importance of vagueness in translation: Examples from English to Portuguese. Romansk Forum 5, 43--69.

Santos, Diana. (1998). Punctuation and multilinguality: Reflections from a language engineering perspective. In J.T. Ydstie & A.C. Wollebæk (Eds.), Working Papers in Applied Linguistics 4/98 (pp. 138--160). Oslo: University of Oslo.

Santos, Diana. (1999). Toward Language-specific Applications. Machine Translation 14 (2), 83--112.

Santos, Diana. (2002). 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) (pp. 209--218). LNAI 2389, Springer.

Santos, Diana. (2003). Timber! Issues in treebank building and use. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (Eds.), Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings (pp. 151--158). Springer Verlag.

Santos, Diana. (2004). Translation-based corpus studies: Contrasting English and Portuguese tense and aspect systems. Amsterdam/New York, NY: Rodopi.

Santos, Diana & Eckhard Bick. (2000). 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). ELRA.

Santos, Diana & Frankenberg-Garcia. (submitted) The corpus, its users and their needs: a user-oriented evaluation of COMPARA.

Santos, Diana & Caroline Gasperin. (2002). Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation. In M.G. Rodríguez & C.P.S. Araujo (Eds.), Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation (Las Palmas, 29-31 May 2002) (pp.597--604). ELRA.

Santos, Diana & Luís Sarmento. (2003). O projecto AC/DC: acesso a corpora / disponibilização de corpora. In Amália Mendes & Tiago Freitas (Eds.), Actas do XVIII Encontro da Associação Portuguesa de Linguística (Porto, 2-4 de Outubro de 2002) (pp. 705--717). Lisboa: APL.

Santos, Diana, Luís Costa & Paulo Rocha. (2003). Cooperatively evaluating Portuguese morphology. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso & Maria das Graças Volpe Nunes (Eds.), Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, 26-27 June 2003, Proceedings (pp. 259--266). Springer.