"FLORESTA SINTÁ(C)TICA": A PORTUGUESE TREEBANK

Susana Afonso
Eckhard Bick
VISL project, lineb@hum.au.dk, saf@language.sdu.dk
Southern Denmark University

Renato Haber
Diana Santos
Processamento computacional do português
SINTEF Telecom & Informatics

Abstract

The Floresta Sintá(c)tica is a Constraint Grammar based syntactic treebank project for Portuguese, initiated jointly by two research groups, the Odense based VISL project and the Oslo based project "Processamento computacional do português".1 million words from the CETEMPúblico corpus have been automaticallyannotated and made accessible through the internet, both as is" and through specially designed search interfaces. Methodologically, our approach strives to combine the robustness of the CG formalism with the descriptional elegance of syntactic tree structures. About 10% of the corpus was manually revised at both these levels,and inter-annotator differences quantified in a test phase. Special focus was placed on the issues of standardization, documentation and evaluation. This article reviews the project's first year and describes the methodology used, principlesadopted and results achieved

1. INTRODUCTION: MOTIVATION AND OBJECTIVES

There are various good motives for creating a Portuguese treebank, one of them simply being the desire to make a new research tool available to the Portuguese language community, another the wish to establish some kind of workable compromise for the encoding of syntactic information across different schools of grammar. Syntactic treebanks have been emerging for a number of languages, and Portuguese seemed next in line.

So far, only two research groups, in Odense and Oslo, have actively contributed to shaping the *Floresta Sintá(c)tica* treebank, but the authors hope to stimulate a broader discussion in the future. Secondary objectives were the testing and improvement of a pre-existing syntactic parser, PALAVRAS, a feasability-study regarding the effectiveness and speed of human post-processing, and the creation of data for internet based grammar teaching.

The treebank described here, *Floresta Sintá(c)tica*, consists of running text chunked in sentences and syntactically analyzed in tree structures, making use of both automatic parsing and human revision. In order to make our data accessible to a wider public, the treebank has been published on the internet[1], providing both download and search options, as well as graphical tree representation. Thus, we intend to meet the general demand that new linguistic resources should be shared and open to external evaluation (cf. Gaizauskas, 1998 e Hirschman, 1998).

2. ORGANISATION

The present project was initiated as a collaboration between two groups, both with prior experience in the processing and annotation of corpora, and on the background of another succesful joint venture, the AC/DC project (Santos & Bick, 2000).

The VISL project is an ongoing research and teaching project at Southern Denmark University, now in its 6th year. Using a Constraint Grammar framework (Karlsson et.al. 1995) for the development of automatic taggers and parsers, VISL has built an internet based user interface for its linguistic and pedagogical tools and data bases, supporting 16 different languages. VISL's Portuguese system is based on the PALAVRAS parser (Bick 2000), and has been functioning as a role model for other languages. More recently, VISL has moved to incorporate semantic research, machine translation, and corpus annotation proper.

In VISL's teaching ingerface, users can choose between different notational filters incorporating different descriptive paradigms of grammar, allowing, for instance, the interactive manipulation of syntactic trees, or java games where words are coloured, "stamped" or "shot" for form and function.

The project *Processamento computacional do português* (Santos 2000), which recently evolved into a Center for distributed resources in the processing of Portuguese, was initiated by the Portuguese Ministry of Science and Technology in order to further development in this area. One of its primary lines of action is the creation of public resources for the investigation and development in the field of computational processing of Portuguese. Various projects (some in the shape of joint ventures) have been launched to make such resources accessible, such as the AC/DC, the COMPARA, the CETEMPúblico and the Floresta Sintá(c)tica. Another project priority is evaluation.

[1] At http://cgi.portugues.mct.pt/treebank/PaginaFloresta.html and http://visl.sdu.dk/visl/pt/treebank.html.

Given its affiliation to the universities language department, VISL's ultimate interest in the Floresta project is linguistic rather than computational, involving the creation and propagation of linguistic knowledge (more trees, better parsing), rather than the evaluation of computational efficiency as such. The main participation motive for the project Processamento computacional do português, on the other hand, has been the production of an evaluation resource for syntactic analysers and other computational tools, based on public and linguistically validated objects (trees).

During the Floresta project, we have grown to regard these differences in motivation as stimulating and beneficiary, given the fact that both one and the other can be achieved in synergistic ways. Also, an initial experimental phase accompanying the formulation of stringent definitions and specifications was judged useful before launching a wider cooperative venture involving the major part of all syntactic research groups in the Portuguese field.

As the language material used in a treebank has to be copyright cleared, we decided to base the first million words of the Floresta on CETEMPúblico corpus (Rocha & Santos, 2000), while working towards the clearance of a simliar set of data for Brazilian Portuguese.

3. OTHER PROJECTS

Various ongoing and concluded similar projects[2] can be cited, that all aim at the creation of language engineering tools for a number of different tasks. However, considerable differences exist in terms of methodology, annotation principles and even concerning the definition of what a treebank is.

Since the pioneering Penn Treebank (Marcus et al., 1993) and the SUSANNE corpus (Sampson, s.d.), for English, and the Prague Dependency Treebank (Hajic, 1998), for Chech, which all implemented one given linguistic formalism (constituent trees in the former, dependency trees in the latter), many other approaches have been realized, like the recently published sophisticated TIGER corpus (Dipper et al., 2001) for German, which takes into consideration both of the above formalisms at the same time. Regrettably, a detailed discussion of the field would be beyond the scope of this article.

4. PLANTING THE FOREST: DESCRIPTION OF THE PROCESS

4.1. REVISING RAW CORPUS LAYOUT

The production of the Floresta Sintá(c)tica was done in two distinct phases: First a preprocessing phase, and second the annotation phase proper, which again consisted of alternately reiterated rounds of automatic analysis, manual revision and linguistic evaluation, covering both a Constraint Grammar phase and a tree structure phase as successive steps.

One of the objectives of the preprocessing phase was simply sentence separation, i.e. the creation of well defined and syntactically meaningful units for tree constituent analysis. Since punctuation based automatic separation did not produce the desired result, more linguistic criteria were crafted, and manual separation performed (Afonso & Marchi, 2001a), considerably prolonging the preprocessing phase. As an additional advantage, the manual inspection allowed us to mark certain complex text sections (poems), og syntactically "unsentential" lines (soccer results, address lists) as <sic>, to be exempted from syntactic tree analysis.

Also during this preliminary phase, the whole corpus was tokenized and analysed with the PALAVRAS parser, extracting all tokens, where derivation or heuristic analysis had been used by the parser to establish a lemma relation. As a result, 8-9.000 new lexemes were added to the parser's lexicon, improving morphological coverage and establishing a lexical balance between Brazilian and European Portuguese.

4.2. THE ANNOTATION PROCESS

The second phase was dedicated to automatic annotation and manual revision of first CG, then tree structures. Given the large number of categories already used in the PALAVRAS parser, and the interest of the participating groups in creating a corpus adaptable to many different needs and applications, it was decided to perform an exhaustive morphosyntactic revision, both in terms of form, function and structure.

PALAVRAS itself is a lexicon and rule based morphosyntactic dependency parser relying on Constraint Grammar methodology as a means of disambiguating morphological ambiguity and mapping syntactic function in a context dependent way. In previous test runs, PALAVRAS has achieved recall rates of 99% for PoS and 96-97% for syntactic tags, when geared to - almost - full disambiguation (i.e. approaching 100% precision). In order to add constituent tree structure, a PSG-like program was used to bracket words into groups and clauses, moving syntactic function tags from dependency heads onto newly created mother nodes (Bick, 2000). Unlike classical PSG rules, where "terminals" are words, the underlying CG-analysis makes it possible to use existing higher level function tags, like subject and adverbial, as terminals, cutting down on the number of rules necessary, and increasing their descriptive power. It must be stressed that the resulting structural tree *adds* information (and thus, the risk of new errors) to the original CG annotation, having to resolve previously underspecified structural ambiguity (in particular, coordination and the "distance" of postnominal attachment).

4.3. THE REVISION PROCESS

It seemed logical to match these two steps of annotation in the manual revision work, too: First, the revision of CG form and function tags, then - after trees were generated from the revised CG files - the structural revision of

[2] Another ten or so different tree bank projects are mentioned at the TIGER site, http://www.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml.

syntactic trees. Thus, certain errors could be prevented from propagating into the tree generation phase. For instance, a CG function tag error from stage 1, like an additional subject reading in the wrong place, might well prevent the generation of a well formed tree at stage 2, because the PSG rules will not be able to accommodate for the extra subject in any legitimate string of function nodes. Here, a correction at stage 1, may prevent a somewhat larger number of manual interventions at stage 2, augmenting the robustness of the process and making human revision more time and "cost" effective. Also, a bipartition of the process of analysis and revision facilitates the maintenance of the separate CG and PSG grammars, allowing the PALAVRAS author to locate and remedy parsing problems in a transparent way.

In addition, since both automatical analysis and subsequent manual revision were done in chunks of a few hundred sentences at a time, it was possible to correct parsing or lexical errors identified in one round before running the next, minimizing the need for multiple correction of the same error. Moreover, the stage distinction allowed us to discuss and implement new annotational distinctions, as made desirable by the revision process, not only in the Floresta corpus, but also, to a certain degree, in the subsequent automatic analysis.

Though at first sight a simple linear process, the annotation phase repeatedly raised complex linguistic issues. An important reason for this is the text/genre type of the corpus chosen. Thus, journalistic texts, interviews and the like are rich in impromptu formulations, colloquialisms, indirect speech, hesitations, syntactically incomplete sentences, and even outright linguistic errors. Following a policy of minimal corpus intervention, the latter were not corrected, but rather treated as interesting research data illuminating, for instance, issues of language change, performance stability and linguistic variation. Another particularity of journalistic texts is the high incidence of titles, representing a special type of (averbal) syntax. As a consequence of the frequency of these phenomena, certain structural sentence type markers were introduced (e.g., #D for *discourse structure*, #E for *ellipsis*).

4.3.1 REVISING THE CG FORMAT (CONSTRAINT GRAMMAR)

Since Constraint Grammar uses a word and tag based annotation scheme, revision at this stage implied correcting/substituting morphosyntactic tags at word level: Word class (PoS), base form (lemma), inflexion features, syntactic function and dependency markers (the latter two joined in the same tag). PALAVRAS encodes subclause function as an additional tag at the head verb or complementizer of a given subclause, so for some words, two syntactic tags (intra-clause and clausal-external) had to be revised.

Below, an example of CG-annotation is given, first after automatic analysis, then after human linguistic revision (changes in bold face). Note the special dependency markers (< , >), which indicate the direction of the syntactic head of a given dependent, and allow the tree-generator to chunk constituents into groups and clauses.

Os [o] <art> DET M P @>N
que [que] <rel> SPEC M S @SUBJ> @#FS-N<
escolheram [escolher] <fmc> V PS/MQP 3P IND VFIN @FMV
sábado [sábado] N M S @<ADVL
podem [poder] <fmc> V PR 3P IND VFIN @FAUX
começar [começar] V INF @IMV @#ICL-AUX<
cedo [cedo] ADV @<ADVL

Os [o] **<dem>** DET M P **@SUBJ>**
que [que] <rel> SPEC M S @SUBJ> @#FS-N<
escolheram [escolher] <fmc> V PS/MQP 3P IND VFIN @FMV
sábado [sábado] N M S **@<ACC**
podem [poder] <fmc> V PR 3P IND VFIN @FAUX
começar [começar] V INF @IMV @#ICL-AUX<
cedo [cedo] ADV @<ADVL

As mentioned above, a thorough revision at the CG-level will save more than its own worth of work at the tree revision level, where the revision effort concentrates on structural matters.[3]

4.3.2. REVISING THE SYNTACTIC TREE FORMAT

In the Floresta sintá(c)tica, constituent trees are built from the vertical CG-notation by introducing non-terminal nodes and indenting the corresponding daughter nodes or terminals (words) with equal signs, the number of equal signs representing tree depth at a given level. Every node, both terminal and non-terminal, is marked for form and function, and terminals also inherit all morphological and secondary tags from the CG-level. The main focus of revision at the tree level is on constituent boundaries, attachment and indentation depth. Morphosyntactic information is now secondary, but does of course profit from a second revisional pass.

While it is clearly most effective to do morphosyntactic tag revision as early as possible, the same strategy fails for certain structural distinctions, because Constraint Grammar uses a suface oriented "flat" dependency notation. For instance, the postnominal attachment of prepositional phrases is underspecified at the CG-level: The preposition in question would carry a left dependency arrow, but it would not be clear whether attachment is to be made to the first, second or even third nominal head candidate to the left. Consider the corpus quote *O empregado de balcão do café Magestic*, where postnominal PP-attachment is shown with the tag @N<:

O	[o] DET M S @>N
empregado	[empregado] N M S @NPHR
de	[de] PRP **@N<**
balcão	[balcão] N M S @P<
de	[de] PRP **@N<**
o	[o] DET M S @>N

[3] The second pass will, of course, not only address structural issues, but also provide a chance to remedy human revision failures from the first phase of tag revision.

```
café          [café] N M S @P<
Magestic      [Magestic] PROP M S @N<
```

Semantically, there is no real ambiguity, speakers of Portuguese will agree that we are talking about a waiter in the café Majestic, not about the balcony of the café Majestic. Syntactically, however, *do café Magestic* may well attach to either *balcão* or *empregado,* which is exactly what the CG annotation suggests. In our tree notation, this ambiguity has to be resolved, and since the tree generator does not make heavy use of semantic og collocational knowledge, but mostly follows a close attachment strategy, such disambiguation will often take the shape of manual correction:

```
UTT:np
>N:art('o' <artd> M S)  O
H:n('empregado' M S)    empregado
N<:pp
=H:prp('de')    de
=P<:np
==H:n('balcão' M S)     balcão
==N<:pp
===H:prp('de' <sam->)   de
===P<:np
====>N:art('o' <-sam>  M S)     o
====H:n('café' M S)     café
====N<:prop('Magestic' M S)     Magestic
```

Project policy has, in fact, been to rely on human disambiguation, and not to mark formal syntactic ambiguity, whereever an individual sentence (or even its context) provides enough clues for a human to arrive at an unambiguous reading.[4] However, in cases of true ambiguity, our formalism allows the specification of alternative readings, either by adding tag and indentation alternatives for a given node, or even by supplying two ore more complete readings for the entire sentence (A1, A2, etc.) (cf. Afonso et al., 2001).

Many individual acts of revision imply changes in node depth, or the addition/deletion of a node. In order to save repetitive labour in these cases (in particular, changing the indentation or attachment of all involved daughter nodes, too), an emacs based tree manipulation tool, the *Pica-pau*, was developed as part of the project (Haber, 2001). Another very helpful, pre-existing tool, especially for sentences of some complexity, was the graphical tree interface used by the VISL project for tree visualisation and grammar teaching. This tree visualizer is a platform independent java program, and greatly facilitates the detection of attachment errors and constituent boundary irregularities.

In its revision work, the Floresta team was confronted, on an almost daily basis, with many quite complex linguistic and descriptional problems arising simply from the fact that running unfiltered journalistic text was used as input. Elliptic constructions can be mentioned as one of the most recalcitrant problems. Since project policy was that descriptional solutions used in the Floresta would respect the terminological principles already agreed upon in the VISL project, and since the CG-based VISL project discourages zero constituents, elliptic constituents were analysed *as if* complete, that is, daughters were assigned such function as they would have had in the non-elliptic constituent. Consider the following sentence:

Os quatro primeiros temas destinam-se a mostrar o papel de Portugal no mundo e o quinto é justificado por a experiência de Barcelona (Port Aventura), não se pôde considerar tema como constituinte vazio, o que simplificaria a análise:

Here, *tema* would be the candidate for a zero head of *quinto.* Therefore, both *o* and *quinto* were tagged as prenominal dependents (@>N):

```
STA:cu
CJT:fcl
=SUBJ:np
==>N:art('o' M P)   Os
==>N:num('quatro' <card> M P)    quatro
==>N:adj('primeiro' <NUM-ord> M P) primeiros
==H:n('tema' M P)      temas
=P:v-fin('destinar' PR 3P IND)     destinam-
=ACC:pron-pers('se' M 3P ACC)        se
=PIV:pp          a mostrar o papel de Portugal em o mundo
CO:conj-c('e' <co-subj>)    e
CJT:fcl
=SUBJ:np
==>N:art('o' M S)    o
==>N:adj('quinto' <NUM-ord> M S)  quinto
=P:vp            é justificado
=PASS:pp         por a experiência de Port-Aventura (Barcelona)
```

However, since the CG-notation always need a word to attach a function to, this is not what the parser produces. Rather, *quinto* would become the only carrier candidate for the subject function tag (@SUBJ), demanding human revision to arrive at the ellipsis annotation scheme advocated above.

Another VISL principle, seeking to keep syntactic trees as simple as possible, discourages the use of one-daughter nodes. Therefore, in cases like *Comem dois pães ao pequeno-almoço e três Ø ao lanche*, the numeral *três* has to assume all of its constituent's function, and will become the lone carrier of a direct object tag (rather than a prenominal @>N) thus maintaining its CG-tag even at the tree notation level.

In order to facilitate corpus searches aimed at these cases, an ellipsis marker, #E, was added to the sentences in question (with the subdivisions of group ellipsis <Eg>, syntactic ellipsis <Es> and morphological ellipsis <Em>).

5. TOOLS

During its first year, the Floresta project inspired the creation of two tree related tools, one for manipulating (o *Pica-pau*), one for searching syntactic trees (o *Águia*).

[4] Another factor is extra-linguistic knowledge. In spite of the syntactic ambiguity, in the sentence *Em relação ao Iraque, Valeri Progrebenkov (…) desmentiu a existência de uma encomenda de 4000 carros de combate russos, como afirmara o genro de Saddam Hussein* ***que desertou para a Jordânia****, (...)* it is quite clear to the enlightened reader who defected to Jordan.

Somewhat unfortunately, the specification, development and testing of these tools was done in parallel with the annotation work proper. Therefore, no extensive use was made of these tools in the present project phase, and both fruits will be tasted mainly by future users (o *Águia*) or future tree revisors (o *Pica-pau*).

The objective of the *Pica-pau* is to facilitate tree-editing, i.e. the movement, addition and removal of entire nodes, words and punctuation in the vertical tree notation. Its working environment is the editor *emacs*. For more information on individual commands, as well as a detailed description/manual, see Haber (2001)

Targeting both developers and users of the published Floresta corpus, the *Àguia* tool allows internet based searches in the tree corpus, involving not only lexical, but also syntactic and structural search criteria encompassing one or more whole nodes. This tool is accessible to all, but has also the "internal" value of being able to pinpoint and quantify problems in the automatic analysis for later systematic correction, without the use of repetitive manual intervention. The Águia represents not only a natural extension of the AC/DC search interface, which focuses on word based information only, but also a supplement to the VISL interface, which allows the inspection of individual trees rather than sets of trees.

6. PROJECT RESULTS

During its first phase (approximately 1 year's work), the *Floresta Sintá(c)tica* project produced

(a) The *Bosque*, 1.427 syntactically analysed and revised trees (1.405 distinct sentences, 36.408 tokens, ca. 34.256 words)

(b) The *Floresta Virgem*, the raw first million words of the CETEMPúblico corpus, 41.406 trees, analysed and automatically annotated, without revision (41.406 sentences, 1.072.857 tokens).

Each tree in our forests corresponds to three different objects: (i) a word based dependency grammar analysis (CG format), (ii) a syntactic constituent analysis (trees in text format), (iii) a syntactic graphical tree (java-presentation).

Another important project result, essential to the interpretation of these above objects, is the body of associated documentation. In a project like ours, documentation is fundamental for various reasons. First, because of the great amount of information involved, it is necessary to produce different types of documentation for different uses of the data, from web site based general project information to exhaustive formal tag and meta tag definitions for all descriptive categories used in both the automatic analysis[5] and the human-revised sentences[6], and finally discussions of the linguistic descisions taken during the chunking, annotation and revision processes. Only in this way the *Floresta Sintá(c)tica* can be made fully accessible and evaluable to a broader user community.

From an annotator point of view, to document linguistic options and choices also involves a prior phase of reflexion, discussion and data mining. This process would depart from concrete descriptional and annotational problems, irregularities in language use and the like, and aim at ensuring cross sentence and cross annotator consistency for similar cases throughout the corpus.

The linguistic documentation is divided in two distinct parts, one concerning more generic and formal options transcending the *Floresta Sintá(c)tica* in the sense that they would be based on more general guide lines already established in the VISL project. These options mainly involve basic annotation principles. Another part consists of linguistic decisions taken and descriptional problems resolved during the iterative revision process, meant to regularize the formal representation of linguistic phenomena encountered in the CETEMPúblico corpus.

7. INTER-ANNOTATOR TEST

An inter-annotator test is an important means for evaluating revision accuracy and of measuring consistency across different annotators. In the project at hand we focused not only on the overall number of differences, but also on the different types of differences and their causes, such as performance errors, ambiguity and linguistic theory.

The following methodology was adapted: Three annotators had a week to revise 107 syntactic trees in parallel and "in isolation". The revision was done "by hand" and directly in the text file format, without the use of graphical or other special editing tools, consistency checking programs and the like. The three resulting files were then compared two-by-two (**R**(evision)**1** and **R**(evision)**2**; **R1** and **R3**; **R2** and **R3**), using the Unix *diff* command, and differences were listed and categorized according to a prearranged typology scheme. Differences were discussed by the whole annotator team, and either resolved (producing error counts) or maintained (producing ambiguity counts).Due to the two-by-two comparison technique, any given grammatical feature (both category and structure), would produce either three "counts" (if all three annotators were in disagreement), two (where only one annotator disagreed with the others), or zero (in the case of unanimity).

For a complete description of the inter-annotator test techniques and results, evaluation and conclusions, see Afonso (2001).

8. PERSPECTIVES

The *Floresta Sintá(c)tica* is the first corpus project of its kind and scope for Portuguese, so a special effort was made during this first phase to improve, evaluate and document both the processes of annotation and revision and any formal or linguistic decisions motivated either by the corpus/language data involved or by purely methodological needs. The resulting body of information should make further work more effective and allow continued consistency, and thus it is our hope that experiences form this first year will help to guide and smoothen future work on this or other Portuguese tree banks.

[5] Cf. the glossary at: http://cgi.portugues.mct.pt/treebank/glossario.html

[6] Cf. http://cgi.portugues.mct.pt/treebank/BNFfloresta.html.

In terms of direct quantitative results, 1427 sentences were annotated and revised, representing about 10% of the first million word chunk of the CETEMPúblico corpus. Though not yet constituting a real tree bank, these 10% make up for a valuable corpus kernel of "safe" data that were exhaustively revised at all annotational levels involved. The finished part is also big enough to ensure enough syntactic and morphological variation for a satisfying coverage of phenomena likely to be encounted elsewhere in the CETEMPúblico corpus. Principles established here, and descriptional issues resolved, will likely hold for the rest of the corpus, too.

Also, future work will hopefully benefit from the fact that, as manual revision progressed, the automatic parser was tuned and improved along the same lines, thus enabling a better revision base and better consistency between "man and machine". As a matter of fact, the *Floresta Virgem*, i.e. the automatically annotated whole corpus, though not revised (yet), can at least be regarded as a result of *revised principles*, and a kind of extrapolation of the human effort made on the core corpus.

9. THANKS

We would like to mention Ana Raquel Marchi, another *Floresta* team member, who, though unfortunately unable to participate in this article, has done important revision work on the corpus. Throughout the project, we have received important technical and moral support from members of the VISL group at SDU's Institute of Language and Communication.

REFERENCES

Afonso, Susana. ”Na trilha de um teste inter-anotadores”, 2001, http://cgi.portugues.mct.pt/treebank/TrilhaTIA.rtf.

Afonso, Susana & Ana Raquel Marchi. "Critérios de separação de sentenças/frases", 2001a, http://cgi.portugues.mct.pt/treebank/CriteriosSeparacao.html

Afonso, Susana & Ana Raquel Marchi. "A etiqueta <sic> </sic>", 2001b. http://cgi.portugues.mct.pt/treebank/CriteriosSic.html

Afonso, Susana, Eckhard Bick & Ana Raquel Marchi. "Notational and terminological guide-lines", 2001, http://www.visl.hum.sdu.dk/visl/pt/guidelines.html

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

Gaizauskas, Robert. "Evaluation in language and speech technology". *Computer Speech and Language*, **12** (4) (1998), pp.249-62.

Dipper, Stefanie, Thorsten Brants, Wolfgang Lezius, Oliver Plaehn & George Smith. "The TIGER treebank", presented at *Third Workshop on Linguistically Interpreted Corpora*, www.ims.uni-stuttgart.de/projekte/TIGER/paper/linc2001-abstract-tiger.pdf.

Hirschman, Lynette. "The evolution of Evaluation: Lessons from the Message Understanding Conferences", *Computer Speech and Language* **12** (4) (1998), 281-305.

Haber, Renato Ribeiro. "Pica-pau: Um protótipo de ferramenta para visualização e edição de árvores sintáticas", http://cgi.portugues.mct.pt/treebank/Picapau.html.

Hajič, Jan. "Building a Syntactically Annotated Corpus: The Prague Dependency Treebank", in *Issues of Valency and Meaning*, Karolinum, Praha 1998, pp. 106-132.

Karlsson, F., A. Voutilainen, J. Heikkilä & A. Anttila. *Constraint Grammar: ALanguage-Independent Framework for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin / New York, 1995.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz. "Building a large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics* **19** (2), June 1993, 313-30.

Rocha, Paulo & Diana Santos. "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", in Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR’2000* (Atibaia, SP, Brasil, 19-22 de Novembro de 2000), 131-140.

Sampson, Geoffrey. "SUSANNE Corpus and Analytic Scheme", http://www.cogs.susx.ac.uk/users/geoffs/RSue.html.

Santos, Diana. "O projecto Processamento Computacional do Português: Balanço e perspectivas", in Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada, PROPOR’2000* (Atibaia, São Paulo, Brasil, 19-22 de Novembro de 2000), 105-113.

Santos, Diana & Eckhard Bick. "Providing Internet access to Portuguese corpora: the AC/DC project", in Gavriladou et al. (eds.), *Proc. Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), 205-10.