"Floresta Sintá(c)tica": A treebank for Portuguese

Susana Afonso*, Eckhard Bick*, Renato Haber†, Diana Santos†

*VISL project, University of Southern Denmark
Institute of Language and Communication, Campusvej, 55, 5230 Odense M, Denmark
lineb@hum.au.dk, saf@language.sdu.dk

†SINTEF Telecom & Informatics,
Pb 124, Blindern, NO-0314 Oslo, Norway
renato.haber@pobox.com,Diana.Santos@sintef.no

Abstract

This paper reviews the first year of the creation of a publicly available treebank for Portuguese, Floresta Sintá(c)tica, a collaboration project between the VISL and the Computational Processing of Portuguese projects. After briefly describing the main goals and the organization of the project, the creation of the annotated objects is presented in detail: preparing the text to be annotated, applying the Constraint Grammar based PALAVRAS parser, revising its output manually in a two-stage process, and carefully documenting the linguistic options. Some examples of the kind of interesting problems dealt with are presented, and the paper ends with a brief description of the tools developed, the project results so far, and a mention to a preliminary inter-annotator test and what was learned from it.

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 our two groups 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 feasibility 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 Web[1], providing both download and search options, as well as graphical tree representation. Thus, we intend to meet the general demand that linguistic resources should be shared and open to external evaluation (Gaizauskas, 1998; 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 successful joint venture, the AC/DC project (Santos and 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 interface, 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.

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). The main reason for participation for the Computational Processing of Portuguese project, 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

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

words of the Floresta on the CETEMPúblico corpus (Rocha and 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 Penn Treebank (Marcus et al., 1993) and the SUSANNE corpus (Sampson, s.d.), pioneers for English, and the Prague Dependency Treebank (Hajič, 1998), for Czech, all implementing one given linguistic formalism (constituent trees in the former, dependency trees in the latter), many other approaches have been followed. For example, the recently announced TIGER corpus (Dipper et al., 2001) for German, a highly sophisticated approach taking into consideration both the above formalisms. 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 was considered not to produce the desired result, more linguistic criteria were crafted, and manual separation performed (Afonso and Marchi, 2001a), considerably prolonging the preprocessing phase. As an additional by-product, the manual inspection allowed us to mark certain complex text sections (poems), and 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 (Bick, 2000) 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 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. 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 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, newspaper text, including interviews and the like, is rich in impromptu formulations, colloquialisms, indirect speech, hesitations, syntactically incomplete sentences, and even outright 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 newspaper text is the high incidence of

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

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 Constraint Grammar format

Since Constraint Grammar (CG) uses a word and tag based annotation scheme, revision at this stage implied correcting or replacing 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[3] of CG-annotation is given, first after automatic analysis, then after human 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.

```
Queremos     [querer] <fmc> V PR 1P IND VFIN @FMV
que          [que] KS @SUB @#FS-<ACC
especialistas [especialista] N M/F P @<ACC
internacionais [internacional] ADJ M/F P @N<
e            [e] <co-postnom> KC @CO
nacionais    [nacional] ADJ M/F P @N<
pensem       [pensar]  V PR 3P SUBJ VFIN @FMV
em           [em] PRP @<PIV
as           [o] <artd> DET F P @>N
possibilidades [possibilidade] N F P @P<
que          [que] KS @SUB @#FS-<ACC
existem      [existir] V PR 3P IND VFIN @FMV
de           [de] PRP @<ADVL
abordagem    [abordagem] N F S @P<
de           [de] PRP @N<
o            [o] <artd> DET M S @>N
tema         [tema] N M S @P<
em           [em] PRP @N<
o            [o] <artd> DET M S @>N
contexto     [contexto] N M S @P<
de           [de] <sam-> PRP @N<
a            [o] <-sam> <artd> DET F S @>N
sociedade    [sociedade] N F S @P<
de           [de] <sam-> PRP @N<
a            [o] <-sam> <artd> DET F S @>N
informação   [informação] N F S @P<
:
```

```
Queremos     [querer]  <fmc> V PR 1P IND VFIN @FMV
que          [que] KS @SUB @#FS-<ACC
especialistas [especialista] N M/F P @SUBJ>
internacionais [internacional] ADJ M/F P @N<
e            [e] <co-postnom> KC @CO
nacionais    [nacional] ADJ M/F P @N<
pensem       [pensar] V PR 3P SUBJ VFIN @FMV
em           [em] <sam-> PRP @<PIV
as           [o] <-sam> <artd> DET F P @>N
possibilidades [possibilidade] N F P @P<
que          [que] <rel> SPEC F P @SUBJ> @#FS-N<
existem      [existir] V PR 3P IND VFIN @FMV
de           [de] PRP @N<
abordagem    [abordagem] N F S @P<
de           [de] <sam-> PRP @N<
o            [o] <-sam> <artd> DET M S @>N
tema         [tema] N M S @P<
em           [em] <sam-> PRP @N<
o            [o] <-sam> <artd> DET M S @>N
contexto     [contexto] N M S @P<
de           [de] <sam-> PRP @N<
a            [o] <-sam> <artd> DET F S @>N
sociedade    [sociedade] N F S @P<
de           [de] <sam-> PRP @N<
a            [o] <-sam> <artd> DET F S @>N
informação   [informação] N F S @P<
$:
```

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.[4]

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 revision pass.

While it is clearly most effective to do morphosyntactic tag revision as early as possible, this strategy fails for certain structural distinctions, because Constraint Grammar uses a surface 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 specified whether attachment is to be made to the first, second or even third nominal head candidate to the left. Consider the corpus quote *No quiosque vendem-se dessas revistas de viagens que agora proliferam e que perpetuam as fantasias sobre ilhas exóticas*, where the attachment of the postnominal relative clause (compound unit) is shown with the tag @#FS-N<:

(…)

```
essas        [esse] <-sam> <dem> DET F P @>N
revistas     [revista] N F P @P<
de           [de] PRP @N<
viagens      [viagem] N F P @P<
que          [que] <rel> SPEC F P @SUBJ> @#FS-N<
agora        [agora] <kc> ADV @ADVL>
proliferam   [proliferar] V PR 3P IND VFIN @FMV
```

[3] This and all the following examples in the article are taken from the CETEMPúblico newspaper corpus.

[4] 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.

```
e           [e] <co-vfin> <co-fmc> KC @CO
que         [que] <rel> SPEC F P @SUBJ> @#FS-N<
perpetuam   [perpetuar] <fmc> V PR 3P IND VFIN @FMV
as          [o] <artd> DET F P @>N
fantasias   [fantasia] N F P @<ACC
sobre       [sobre] PRP @N<
ilhas       [ilha] N F P @P<
exóticas    [exótico] ADJ F P @N<
$.
```

Semantically, there is no real ambiguity, speakers of Portuguese will agree that we are talking about the magazines, not about the trips themselves. Syntactically, however, *que proliferam e que perpetuam as fantasias sobre ilhas exóticas* may well attach to either *revistas* or *viagens*, 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 or collocational knowledge, but mostly follows a close attachment strategy, such disambiguation will often take the shape of manual correction:

```
A1
STA:fcl
(…)
ADVL:pp
=H:prp('de' <sam->)                 de
=P<:np
==>N:pron-det('esse' <-sam> <dem> F P)      essas
==H:n('revista' F P)                revistas
==N<:pp
===H:prp('de')  de
===P<:np
====H:n('viagem' F P)               viagens
====N<:cu
=====CJT:fcl
======SUBJ:pron-indp('que' <rel> F P)       que
======ADVL:adv('agora' <kc>)                agora
======P:v-fin('proliferar' PR 3P IND)       proliferam
=====CO:conj-c('e' <co-vfin> <co-fmc>)      e
=====CJT:fcl
======SUBJ:pron-indp('que' <rel> F P)       que
======P:v-fin('perpetuar' PR 3P IND)        perpetuam
======ACC:np
=======>N:art('o' F P)              as
=======H:n('fantasia' F P)          fantasias
=======N<:pp
========H:prp('sobre')              sobre
========P<:np
=========H:n('ilha' F P)            ilhas
=========N<:adj('exótico' F P)              exóticas
.
```

Project policy has, in fact, been to rely on human disambiguation, and not to mark formal syntactic ambiguity, wherever an individual sentence (or even its wider context) provides enough clues for a human to arrive at an unambiguous reading.[5] However, in cases of true ambiguity, our treebank language allows the specification of alternative readings, either by adding tag and indentation alternatives for a given node, or by supplying two or more complete readings for the entire sentence (A1, A2, etc.) (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, *Pica-pau,* was developed in the framework of the project (Haber, 2001).

Another very helpful tool, already existing prior to project launching, was the graphical tree interface used by the VISL project for tree visualisation and grammar teaching, which was invaluable for sentences of some complexity. 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 newspaper text was used as input.

Elliptic constructions can be mentioned as one of the most difficult problems. Since the project policy was to respect, in the Floresta, the descriptive principles already agreed upon in the VISL project, and the 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 a corresponding 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).* 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 needs a word to attach a function to, this is not what the parser produces. Rather, *quinto* would become the only carrier

[5] 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.

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, prevents 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 the function of the constituent formed by itself, and must become the lone carrier of a direct object tag (rather than a prenominal @>N as ellipsis would have it).

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 (*Pica-pau*), one for searching syntactic trees (*Águia*). 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 (*Águia*) or future annotators (*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 Emacs editor. For more information on individual commands, as well as a detailed description/manual, see Haber (2001)

Targeting both developers and users of the public Floresta corpus, the *Águia* 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 one 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 or GIF file).

Another important project result, essential to the interpretation of the 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 general project information on a Web site to formal definition of the treebank objects[6], as well as an exhaustive and readable description of the import and meaning of all descriptive categories used[7], and finally discussions of the linguistic decisions 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.

Second, 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 start from concrete descriptive problems for the annotation, 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 generic options transcending the *Floresta Sintá(c)tica* in the sense that they are based on the more general guidelines already established in the VISL project. These options mainly involve basic annotation principles. The other part consists of linguistic decisions taken and descriptive problems solved 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

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

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.

In terms of direct quantitative results, 1,427 sentences were annotated and revised, representing about 10% of the first million word chunk of the CETEMPúblico corpus. These 10% make up for a valuable corpus kernel of "safe" data that were exhaustively revised at all annotation 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 one-million 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. Acknowledgements

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.

10. References

Afonso, Susana, Eckhard Bick, Renato Haber and Diana Santos. (2002). Floresta sintá(c)tica: um treebank para o português. In *Actas do XVII Encontro da Associação Portuguesa de Linguística*. Lisboa: APL.

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

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

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

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

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

Gaizauskas, Robert. (1988). Evaluation in language and speech technology. *Computer Speech and Language*, 12(4), 249-62.

Dipper, Stefanie, Thorsten Brants, Wolfgang Lezius, Oliver Plaehn and George Smith. (2001). The TIGER treebank. *Third Workshop on Linguistically Interpreted Corpora*, www.ims.unistuttgart.de/projekte/TIGER/paper/linc2001-abstract-tiger.pdf.

Hirschman, Lynette. (1988). The evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*, 12(4), 281-305.

Haber, Renato Ribeiro. (2001). Pica-pau: Um protótipo de ferramenta para visualização e edição de árvores sintáticas, http://cgi.portugues.mct.pt/treebank/Picapau.html.

Hajič, Jan. (1998). Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning* (pp. 106--32). Praha: Karolinum.

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

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz. (1993). Building a large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, 19(2), 313--330.

Rocha, Paulo and Diana Santos. (2000). 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* (pp. 131-140). Atibaia, SP, Brasil.

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

Santos, Diana. (2000). 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* (pp. 105--113). Atibaia, SP, Brasil.

Santos, Diana and Eckhard Bick. (2000). 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* (pp. 205--210). Athens: ELRA.