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

Susana Afonso*, Eckhard Bick*, Renato Haber ♣, Diana Santos♣
*VISL project, University of Southern Denmark
Institute for Language and Communication, Campusvej,55, DK-5230 Odense M, Denmark
saf@language.sdu.dk, lineb@hum.au.dk
♣SINTEF Telecom & Informatics, Pb 124, Blindern, NO-0214 Oslo, Norway
renato.haber@pobox.com, Diana.Santos@sintef.no

• Introduction: the why, what, who

Why building a treebank for Portuguese?	**What** is the Floresta Sintá(c)tica?	**Who** are we?
• desire to create a new reasearch tool for the Portuguese language community; • wish to establish a compromise for encoding the syntactic information across different schools of grammar; • need for a Portuguese treebank, similarly to what already exists for other languages (English, German, Czech …); • support for parser improvement and testing; • larger data base for grammar teaching.	• running text, chunked in sentences and syntactically analysed in tree structure; • accessible on the Web (http://cgi.portugues.mct.pt/PaginaFloresta.html or/ and http://visl.sdu.dk/visl/pt/treebank.html), for searching, downloading or tree vizualising; • rich annotation scheme for form & function, handles ellipsis and discontinuous constituents; • manually revised first part; • extensive formal and linguistic documentation, **• the first project of its kind and scope for Portuguese**	• **VISL**, an ongoing teaching and research project with a current development focus on: – internet based grammar teaching tools, games etc. for 16 languages; – taggers, parsers and lexica for 5 languages; – syntactically annotated corpora, semantics, MT. • **Processamento computacional do português** (computational processing of Portuguese) – creating publicly available resources; – fostering evaluation and collaboration in the area of language technology of the Portuguese language.

2. Planting the forest

2.1. Project results

• ***Bosque* –** 1,437 syntactically analysed and revised trees (1,404 distinct sentences, 36,693 tokens, ca. 34,554 words) in the following formats:

(i) Constraint Grammar (CG)- word based dependency grammar

China [China] **PROP** F S @SUBJ>
condiciona [condicionar] <fmc> **V** PR 3S IND VFIN @FMV
.

(ii) Syntactic constituent analysis: trees in text format

A1
STA:fcl
SUBJ:prop('China' F S) China
P:v-fin('condicionar' PR 3S IND) condiciona

(iii) Syntactic graphical tree: java representation or GIF.file

• ***Floresta Virgem* –** the first raw one million words of the CETEMPúblico corpus, 41,406 sentences analysed and automatically annotated without revision (1,072,857 tokens).
• **Associated documentation –** notational and terminological guidelines, linguistic choices taken during the annotation and revision process, formal definition of the *Floresta Sintá(c)tica,* inter-annotator test, grammatical categories used in the project.

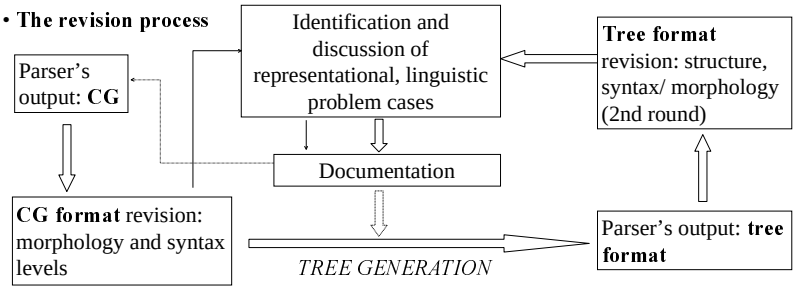
2.2. Phases

•Pre-processing

(i) revising raw corpus layout, specially sentence separation, to have well-defined and meaningful units for tree constituent analysis;

(ii) introduction of new lexemes (8-9,000) to the PALAVRAS lexicon.

• The revision process



3. Tools

• **Águia**: An Internet-based search tool for syntactically annotated corpora in tree format, that allows global searches not only for morphological features and syntactic tags but also for structural features, mother node conditions, etc.
• **Pica-pau**: An Emacs-based tree editor, that facilitates movement, addition or removal of entire nodes or words in text based vertical tree annotation format
• **VISL's tree visualizer**: A Java program allowing step-wise (un)folding, inspection, construction and labelling of syntactic trees from a corpus

4. Summing up and future perspectives

(i) Results of the first phase (approx. one year's work):

- improve, evaluate and document the processes of annotation and revision and the formal linguistic decisions;
- *Bosque* exhaustively revised at all annotation levels; enough syntactic and morphological variation encountered;
- experience in managing and interpreting inter-annotator tests;
- establishment of principles and resolution of descriptional issues.

(ii) Future

- future work will benefit from the exhaustive manual revision:
  a) the automatic parser (PALAVRAS) and its lexicon were tuned and improved throughout this first phase;
  b) better revision base → better consistency man-machine;
- sister project for 1 million words of Brazilian Portuguese;
- parser testing and comparison;
- improved tools for editing, searching and visualizing trees.