**Introduction: the why, what, who**

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

### 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
  
  (ii) **Syntactic constituent analysis**: trees in text format
  
  (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 Sintáctica, 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.

#### 3. Tools

- **Água**: 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-pan**: 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

- **Past**
  
  (a) sister project for 1 million words of Brazilian Portuguese;
  
  (b) improvement and evaluation of parser tool and associated software.

- **Future**
  
  (a) The text will benefit from a qualitative manual revision:
  
  (i) automatic parser (PALAVRAS) and its lexicon were tuned and improved throughout this first phase;
  
  (ii) better revision based on and improved consistency man-machine;

  (b) sister project for 1 million words of Brazilian Portuguese;

  (c) parser testing and comparison;

  (d) improved tools for editing, searching and visualizing trees.