The Floresta Sintá(c)tica Project

- Publicly available treebank for Portuguese (PE and PB)
- Collaboration between Linguateca and VISL project
- Began in 2001 (until 2005)
- Re-started in 2007
- TEAM (since 2007):
  - Eckhard Bick (project leader since 2005)
  - Claudia Freitas (linguistic revision)
  - Paulo Rocha (tool development)
- Before 2007: Diana Santos (project leader), Susana Afonso, Raquel Marchi, Anabela Barreiro, Renato Haber, Luís Sarmento, Rui Vilela.

Textual Material

- Based on human revision of the output of PALAVRAS
- FLORESTA
  - Newspaper texts (first million words from CETEM Publico and CETEM Folha)
- BOSQUE
  - Subset of Floresta, fully revised by linguists (~ 9,000 syntactic trees / 180,000 words)
- SELVA NEW!!!
  - (~ 300,000 words)
  - Other kinds of texts (oral interviews, scientific, literary)
  - Partially revised (mainly NP-related functions)
    - Not yet available 😞
    - Few sentences fully revised 😞

Main Aspects (...to a linguist...)

- Linguistic information
- Search interface -> the information is useful if it can be (easily) found
Since 2007, Bosque has been under a “re-revision” process...
- More consistency
- New tags

- Distinction between Adjuntos Adnominais and Complementos Nominais (Adjuncts and arguments of the head noun)
  - Encontrei-o na entrada do prédio... -> AA (N; N)<
  - Os vôos permitem a entrada de insetos. -> CN (N<ARG)
  »..a entrada de insetos (= os insetos entram) N<ARGS
  »..perdi com a venda da casa (=vender a casa) N<ARGO
  »As leituras de J.L.Borges (=Borges leu -> N<ARGS)
  (= ler Borges -> N<ARGO)

- “Procuráveis” (“searchables”):
  - partitive constructions, main verbs of: relative clauses, passive clauses, substantive clauses, concessive clauses, no-subject clauses...

One may agree or disagree with our linguistic choices, but it’s important to have them documented:
- Bíblia Florestal (the main linguistic guide, now in html)
- Guia rapidíssimo da Floresta (a quick guide)
- Complete documentation webpage
http://acdc.linguateca.pt/treebank/DocumentacaoFloresta.html

The search interface

- The tree format does not need to be fully understood...
  - The search interface – ask for what you need, without having to understand the complex annotation.
  - ÁGUIA...
  - MILHAFRE NEW!!!

- Under development
- It still doesn’t do lots of things
The first thing you want is the first thing you ask

- VP + ACC $\neq$ ACC + VP

- VP + ACC:
  - CP85-4: António Guterres *possui* uma boa relação com o teatro.
  - CP87-4: A França e a Grécia *apoiam* a Grã-Bretanha.

- ACC + VP:
  - CP85-4: António Guterres possui *uma boa relação com o teatro*.
  - CP87-4: A França e a Grécia apoiam *a Grã-Bretanha*.
The results... ("frases completas")

<table>
<thead>
<tr>
<th>Search Terms</th>
<th>Results</th>
<th>Milhafre: A good way to understand Floresta’s symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Bela Vista&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>&quot;Floresta&quot;</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>&quot;symbols&quot;</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>&quot;Milhafre&quot;</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Results... ("apenas extrato")

<table>
<thead>
<tr>
<th>Search Terms</th>
<th>Results</th>
<th>Milhafre: A good way to understand Floresta’s symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Bela Vista&quot;</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>&quot;Floresta&quot;</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>&quot;symbols&quot;</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>&quot;Milhafre&quot;</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
Exploring Portuguese Syntax

• Getting started
  – CU that are subjects (compound subjects)

Results (CU that are subjects)

NP that has an adverbial function
Rel. pronouns that are dir. objects

dir. obj.-clauses without subord. conj.

Pcp without verbal function
Verbs used in ACC constructions

Verbs in ACC constructions

Complements of a given verb

Pcp without verbal function
Complements of “concordar”

What it doesn’t do (...yet!)

• Quantitative / “Qualitative” searches:
  – The kind of complements of a given verb (3 DAT, 7 ACC etc etc)
  – “Structural” searches: a NP containing (exactly) ART PROP ADJ.

• ...but ÁGUIA does it!!!

Visit us! Try us! Use us!

• http://www.linguateca.pt
• http://www.linguateca.pt/Floresta/

Acknowledgment

• This work was done in the scope of the Linguateca, contract nº339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union.