Timber! Treebank building and use

Diana Santos, www.linguateca.pt

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

Inception

- Once upon a time...
- We had the idea of creating a treebank, joining people with different creeds and wishes (VISL and Linguateca)
- The Floresta Sintáctica project was launched
- We created a resource
- We asked the community to participate

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

And so...

- We created a resource
- Public, freely downloadable and with special query capabilities (Águia)
- 1,500 sentences, 35,000 words
- The first 300 CETEMPúblico extracts (Portugal)
- Working now with CETENFolha (Brazil)
- Stumbling with all sorts of problems
  - Distributed location
  - All sorts of different backgrounds
- One year intensive development, two years with little work
- We need feedback to know how and whether to proceed
- Discussion session at AVALON’2003

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

Paper

- Represents the author’s experience
- What did I learn
  - About the “treebank” concept
  - About the process
  - About the result
- I have presented a general overview of the problems involved in the general process of treebank building in Växjö
- Here, I want to illustrate specific problems in relationship with Portuguese

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

How to deal with real text

- NP, NP, what’s the function of the second NP?
- Predicative adject (N+PRED) or apposition (APP)?
  - Although there are clear prototypical definitions of either in real text it has proved extremely difficult to decide
- Three ways to go about (reflected in annotator guidelines)
  - Only change when you are sure, let the parser decide otherwise
  - Do not make the distinction
  - Mark the doubtful cases as doubtful, annotate the text with either marker
- What kind of treebank do we want?

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

PoS assignment: what does it mean?

- There is a tacit assumption that PoS assignment is easier than phrase boundary detection, which in turn is easier than role assignment, etc...
- But this is wrong. Very often people have intuitions about higher order phenomena, while PoS are technicalities

```
Eun vitim am Colares
Eun vitim non arredados de Lávios
Eun vitim non Olakas
```
- Gender and number of locations
- Detecfiñacions suprases, eneñes modelos e mulharens polícia

http://www.linguateca.pt/Floresta/
The same happens with all description levels
- It is not even guaranteed that there is a right answer
- Conflicting requirements
- Relative pronouns/adverbs playing different roles
  - ? onde, ele, foi
  - Fui onde, ele se escondeu
- Direct object vs prepositional object
- And how to encode errors, mistakes or deviant language?
  - correct
  - enlarges/relax the grammar
  - take away from the treebank

Presentation of Águia
- Encourage people to use Águia
  - As a way to see people’s real interests
  - To use a public resource which may well offer more detailed information than the other treebank projects and associated tools
- Short tutorial on the Web
  - Freely available, no need to register
  - Why not go and try for yourself ???
- Threshold lowest possible
  - Look in text
  - Give support to anyone who poses questions
  - A set of examples

Some comments related to Morfolimpiadas
- It is hard to agree
  - X different thinking heads, X different opinions
  - Several problems swept under the carpet
    - tokenization
    - sites and contractions handling
  - Can we use it as an evaluation resource?
    - at least as a form of measure disagreement
    - at least for the subset where there is tokenization agreement
- Session in Avalon 2003
  - Why is it much less used than CETEMPúblico?
  - What is required to make it really useful?

What can Floresta be used for
- Allows a linguist to pose more complex queries
- Allows a language engineer to test application improvement
- Allows a grammar developer to bootstrap her/his grammar
- Allows the parsing community to start not from scratch
- clarifying or defining important concepts
- communicating wrt real already analysed examples
- eventually giving origin to one or several evaluation resources