Treebank Troubles

Eckhard Bick
Southern Denmark University
lineb@hum.au.dk

1. Theory & Teaching

- Floresta Sintáctica in parallel theory dependent formats:
  - Constraint Grammar, Constituent Grammar, Dependency Grammar
    (the first 2 are implemented, a prolog program is being written at VISL to
    create the latter from the CG-version)
- Create filter programs for different formalisms within one super-family of
  theories, e.g. for generative constituent grammar:
  create word nodes for 1-constituent groups
  create S:np PRED:vp daughters for finite clauses
  create AUX constituents
  replace indentation with tabs and brackets etc.
- Use graphical front-ends for teaching, with user-driven interactive formatting
- Document both "filterable" and "un-filterable" theory-clashes
- Offer simplifying filters and user-driven long-forms for the tags used

2. Descriptive Linguistics

- Adapt search tool to user needs (Águia)
  and/or filter format to other existing tools (XML-tools, Tiger ...)
  e.g. qualitative/quantitative, conditioned multiple searches
- Mark points of special interest explicitly in the treebank
  (now e.g. ellipsis, errors, averbal constructions etc.)
- What people's special interests?
- How to reach the linguists?
- Balance the data in terms of genre (now only news texts):
  speech data, dialectal data, fiction and science data ....
- Balance the data in terms of language variety (Lusitan - Brazilian)
- Add section for historical Portuguese?

An example

STA:cu
CTF:k1
sSUB:np
==NAbicats' <card> M P) Os
==N:spaintos 's -<filter> MP) spainto
==NAdm'prenantos ~<SVM-ent> M P:prenantos
==NAdm'prenantos M P: puroan
===`is=det' destinar' PR 3P IND) destinar-
==ACC:pro-pen' <card> 's -<filter> M SP ACC) =
==P:sp; a mostra o papel de Portugal em o mundo
CD:com'p a exposição de Portugal em o mundo
==P:sp; a exposição de Portugal em o mundo
==P:sp; a exposição de Portugal em o mundo

3. Parser development

- Sparse data problem
  Increase "Mata virgem", using a simplified (thus safer) tag set
  Revise manually only "verbal" tags (cave: parser dependent?)
- For training of probabilistic / automated learning systems:
  "simplified" tag set, e.g. as in VISL-lite (only one type of group, only 2
  types of group constituents, H[dependent] and H[ependent])
- Compatibility problem
  Create user-specific tags from implicit information (e.g. N[oun] from
  Hasj in noun phrase, VT [transitive] from co-daughtering of ACC
  Create user-specific layout (p. 2.)
4. Evaluation & measurability

- **Tag set incompatibility**
  Compile set of *core categories* (PoS, syntactic function, attachment link)
  Build set of “unify” programs to handle *tag synonyms*
  Simplify tags to the smallest common denominator
  (e.g. free adverbials, object adverbials, adverbial predicates, prepositional objects all as ADVL, or N<PRED/APP into D)
  Translate implicit information (e.g. on the mother) into explicit tags

- **Researcher incompatibility** *(you know what I mean...)*
  Prior to joint evaluation, cross-revision of to-be-used Floresta-chunks by members of all participating research groups

- **Inter-annotator disagreement**
  Identify “soft categories”, with high inter-annotator disagreement, then
  a) either fuse them into neighboring “hard categories”, or
  b) ignore them in the evaluation

---

A call to arms

- The Floresta Sintáctica may not be the perfect resource, but it IS a resource, possibly the only one of its kind for Portuguese (in terms of information richness)
- Therefore, let’s make the most of it
- If it can’t immediately be used for a given purpose, let’s create filters and other solutions as discussed before, rather than not use it
- In order to make the Floresta more palatable to new users, allow them a say in - which data or genre will be used - which information will be incorporated - which tag set or formalism the treebank will be filtered into
- A sparse data problem is bad, but a sparse linguist problem is worse…
  So, let’s invite everybody to contribute with data, categories & revision