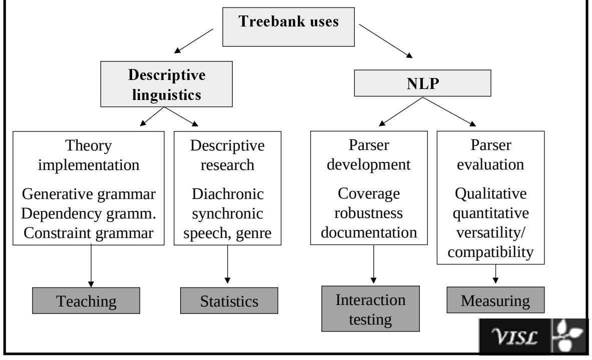
# Treebank Troubles

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

## Science or fiction: A treebank for everybody?



## 1. Theory & Teaching

- Floresta Sintá(c)tica in parallel theory dependent formats:
  Constraint Grammar, Constituent Grammar, Dependency Grammar
  (the first 2 are implemented, a prolog program is being writen 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

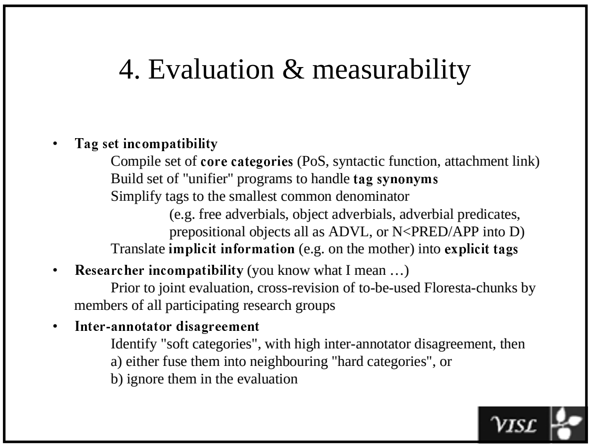
- Adapt search tool to user needs (Águia)
  and/or filter format to other existing tools (XML-tools, Tiger …)
  e.g. qualitative/quantitative, conditioned multiple searches
- Mark points of special interest explicitly in the treebank
  (now e.g. ellipsis, errors, averbal constructions etc.)
  Problems: What *are* 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
CJT:fcl
=SUBJ:np
==>N:art('o' <artd> M P)   Os
==>N:num('quatro' <card> M P)    quatro
==>N:adj('primeiro' <NUM-ord> M P) primeiros
==H:n('tema' M P)      temas
=P:v-fin('destinar' PR 3P IND)     destinam-
=ACC:pron-pers('se' <refl> M 3P ACC)        se
=PIV:pp    a mostrar o papel de Portugal em o mundo
CO:conj-c('e' <co-subj>)    e
CJT:fcl
=SUBJ:np
==>N:art('o' M S)    o
==H:adj('quinto' <NUM-ord> <#E> M S)  quinto
=P:vp        é justificado
=PASS:pp por a experiência de Port-Aventura (Barcelona)
```

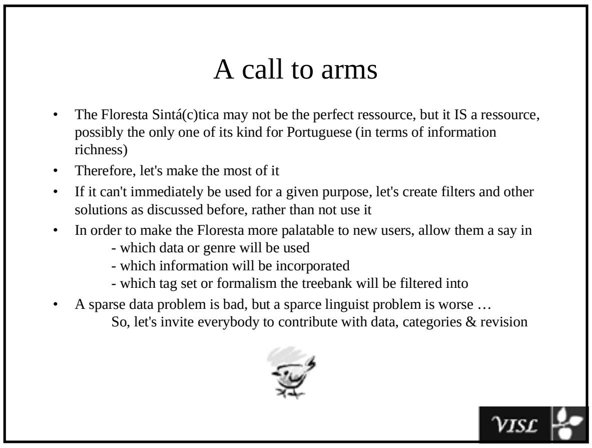
## 3. Parser development

- Sparse data problem
  Increase "Mata virgem", using a **simplified** (thus safer) **tag set**
  Revise manually only **"crucial" tags** (cave: parser dependent?)
- For training of probabilistic / automated learning systems:
  **Fuse tag strings** into units (a la CLAWS)
  **Simplify tag set**, e.g. as in VISL-lite (only one type of group, only 2 types of group constituents, **H**[ead] and **D**[ependent])
- Compatibility problem
  Create user-specific tags from implicit information (e.g. **N**[oun] from **H:adj** in noun phrase, **VT**[ransitive] from co-daughtering **@ACC**)
  Create user-specific layout (jf. 2.)



# 4. Evaluation & measurability

- **Tag set incompatibility**
  Compile set of **core categories** (PoS, syntactic function, attachment link)
  Build set of "unifier" 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 neighbouring "hard categories", or
  b) ignore them in the evaluation



# A call to arms

- The Floresta Sintá(c)tica may not be the perfect ressource, but it IS a ressource, 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 sparce linguist problem is worse …
  So, let's invite everybody to contribute with data, categories & revision