The *Floresta* experience

Diana Santos

Linguateca

SINTEF Telecom and Informatics

Plan

1. Motivation
2. A typology of treebank concerns
3. Treebank workflow + result
4. Treebank evaluation
5. Treebank tools

Motivation

- *Working with Portuguese in the Nordic countries?*
- *Honest attempt to address other languages as well*

- Reuse of the results
- Sharing experience

Note: This is a personal view of the experience gathered, and does not necessarily reflect the opinion of the Floresta team

The Floresta Sintá(c)tica

http://acdc.linguateca.pt/treebank/

- A collaboration project between VISL (Southern Denmark University) and Linguateca (SINTEF); project leaders: Eckhard Bick & Diana Santos
- *Bosque:* 1,427 syntactically analysed and revised trees (1,405 distinct sentences, 36,408 tokens, ca. 34,256 words), automatically created
- Started October 2000, stopped December 2001, some research still being done as of today

See Afonso et al. (2002) at LREC'2002

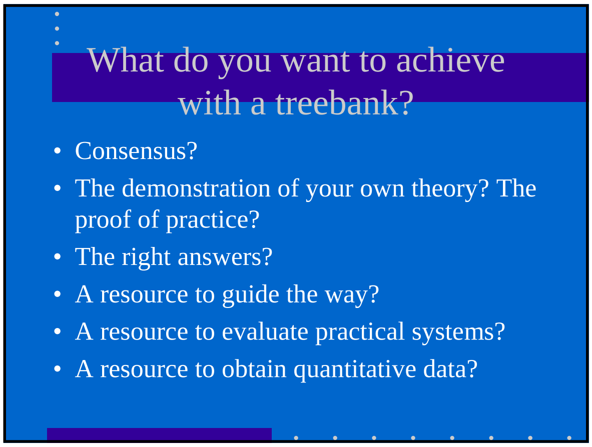
What is a treebank?

- A syntactically analysed corpus (generally in the form of trees)
- Revised and corrected (human intervention)
- Reflecting a consensus
- Publicly available
- Well documented; maintained; with associated tools
- With an evaluation purpose

What is a treebank?

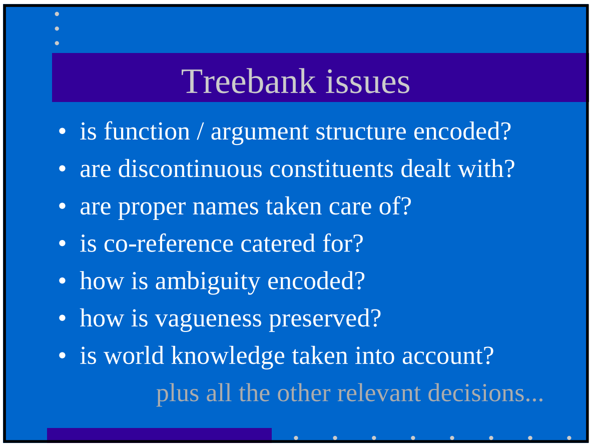
- An evaluation resource
- Point of departure for creating parsers, taggers, etc
- Point of departure for doing (quantitative) linguistics

Conflicting requirements?



What do you want to achieve with a treebank?

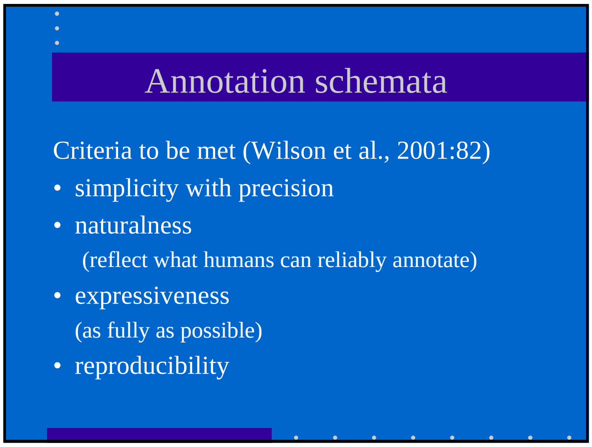
- Consensus?
- The demonstration of your own theory? The proof of practice?
- The right answers?
- A resource to guide the way?
- A resource to evaluate practical systems?
- A resource to obtain quantitative data?



Treebank issues

- is function / argument structure encoded?
- are discontinuous constituents dealt with?
- are proper names taken care of?
- is co-reference catered for?
- how is ambiguity encoded?
- how is vagueness preserved?
- is world knowledge taken into account?

plus all the other relevant decisions...



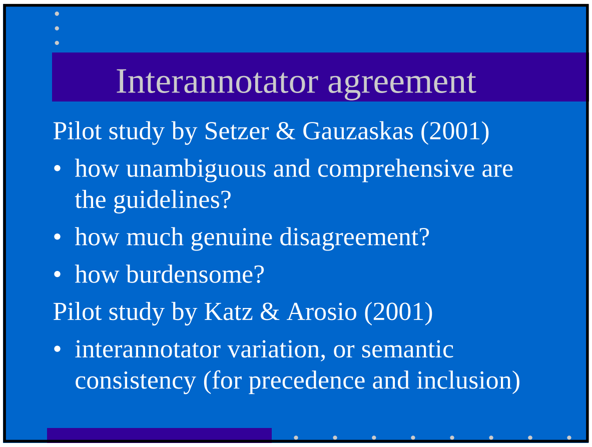
Annotation schemata

Criteria to be met (Wilson et al., 2001:82)

- simplicity with precision
- naturalness

  (reflect what humans can reliably annotate)
- expressiveness

  (as fully as possible)
- reproducibility



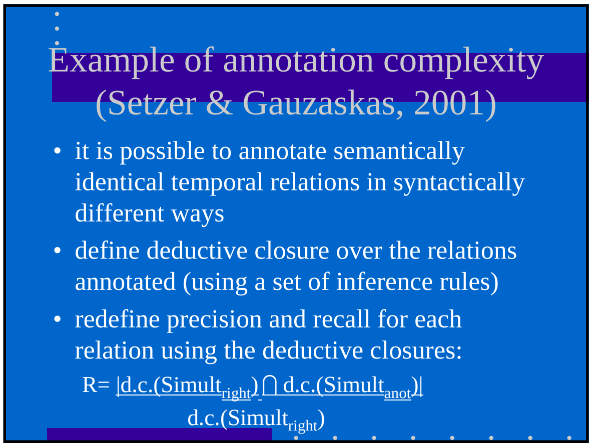
Interannotator agreement

Pilot study by Setzer & Gauzaskas (2001)

- how unambiguous and comprehensive are the guidelines?
- how much genuine disagreement?
- how burdensome?

Pilot study by Katz & Arosio (2001)

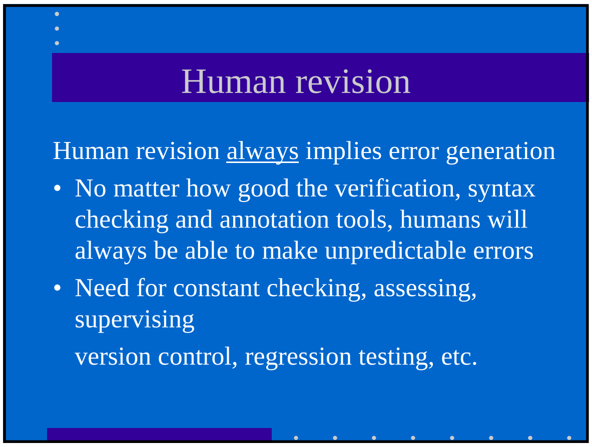
- interannotator variation, or semantic consistency (for precedence and inclusion)



Example of annotation complexity (Setzer & Gauzaskas, 2001)

- it is possible to annotate semantically identical temporal relations in syntactically different ways
- define deductive closure over the relations annotated (using a set of inference rules)
- redefine precision and recall for each relation using the deductive closures:

$$R= \frac{|d.c.(Simult_{right}) \cap d.c.(Simult_{anot})|}{d.c.(Simult_{right})}$$

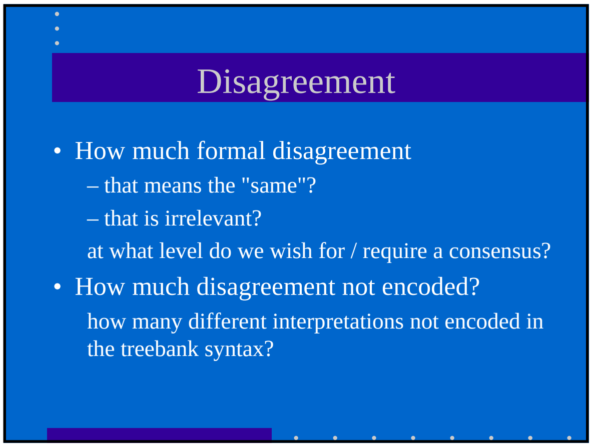


Human revision

Human revision always implies error generation

- No matter how good the verification, syntax checking and annotation tools, humans will always be able to make unpredictable errors
- Need for constant checking, assessing, supervising

  version control, regression testing, etc.

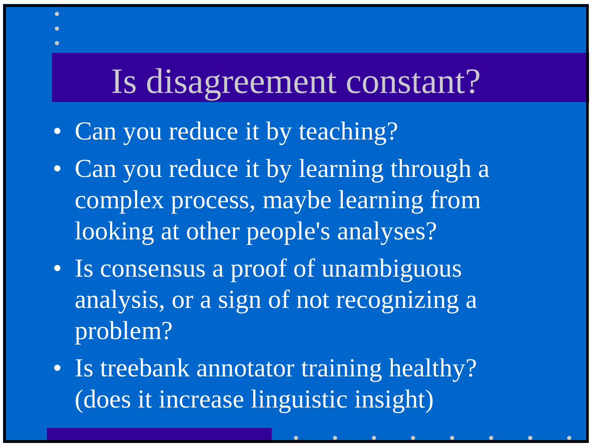


Disagreement

- How much formal disagreement
  - that means the "same"?
  - that is irrelevant?

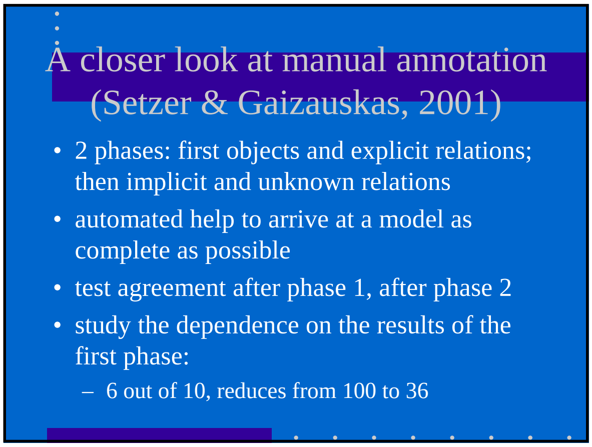
  at what level do we wish for / require a consensus?
- How much disagreement not encoded?

  how many different interpretations not encoded in the treebank syntax?



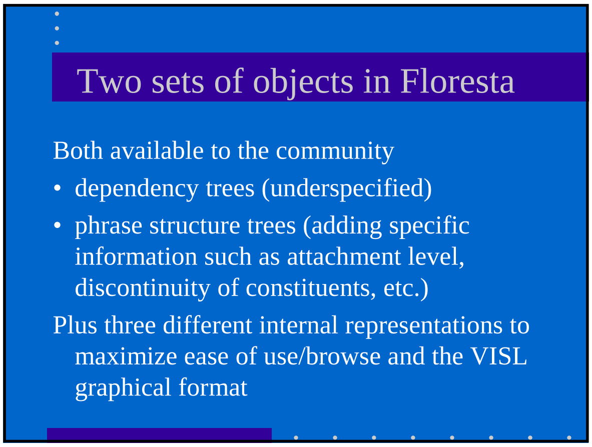
Is disagreement constant?

- Can you reduce it by teaching?
- Can you reduce it by learning through a complex process, maybe learning from looking at other people's analyses?
- Is consensus a proof of unambiguous analysis, or a sign of not recognizing a problem?
- Is treebank annotator training healthy? (does it increase linguistic insight)



A closer look at manual annotation (Setzer & Gaizauskas, 2001)

- 2 phases: first objects and explicit relations; then implicit and unknown relations
- automated help to arrive at a model as complete as possible
- test agreement after phase 1, after phase 2
- study the dependence on the results of the first phase:
  - 6 out of 10, reduces from 100 to 36

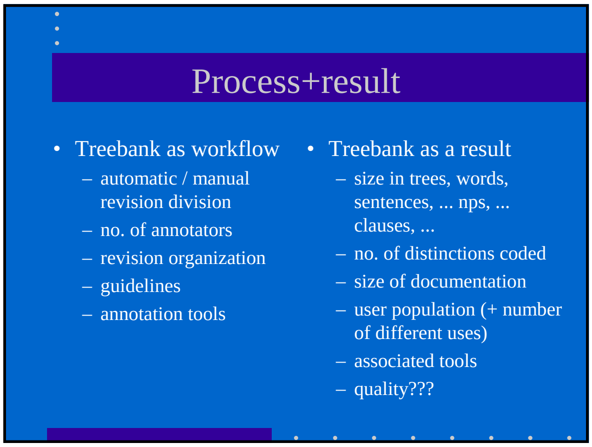


Two sets of objects in Floresta

Both available to the community

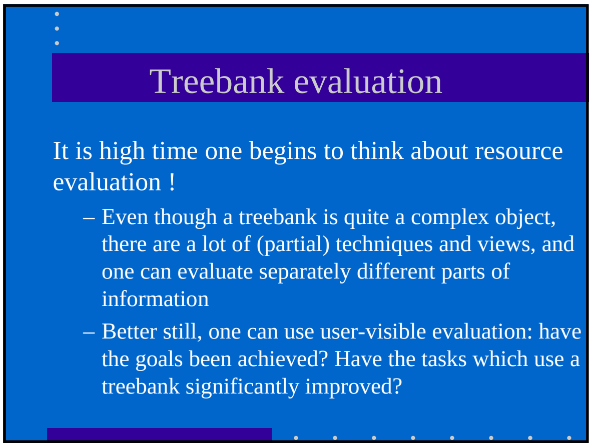
- dependency trees (underspecified)
- phrase structure trees (adding specific information such as attachment level, discontinuity of constituents, etc.)

Plus three different internal representations to maximize ease of use/browse and the VISL graphical format



Process+result

- Treebank as workflow
  - automatic / manual revision division
  - no. of annotators
  - revision organization
  - guidelines
  - annotation tools
- Treebank as a result
  - size in trees, words, sentences, ... nps, ... clauses, ...
  - no. of distinctions coded
  - size of documentation
  - user population (+ number of different uses)
  - associated tools
  - quality???



Treebank evaluation

It is high time one begins to think about resource evaluation !

- Even though a treebank is quite a complex object, there are a lot of (partial) techniques and views, and one can evaluate separately different parts of information
- Better still, one can use user-visible evaluation: have the goals been achieved? Have the tasks which use a treebank significantly improved?

Treebank tools

- to help create
  - speed up the process
  - constrain the choices / reduce the errors
  - allow multiple views and change of opinion
  - to aid documentation
  - to feed inter-annotator tests
- to browse
  - to access the information
  - to find problematic patterns
  - to have a quantitative overview
  - to be able to understand the trees without having created them
  - to focus only on parts of trees

Treebank browsing tools

- To search
  - tree patterns
  - text patterns
  - labelled patterns
- To visualize
  - tree objects
  - text
  - labels

Águia: our treebank browsing tool

http://acdc.linguateca.pt/treebank/TreeSearch.html

- based on the IMS Corpus Workbench
- allows looking for complex patterns both in terms of form and of function
- allows presentation of distribution
- text is the main output format: it provides concordances / text as the search result, not trees (this is a design feature: maximally readable)

Why use the Águia tool

- we want to be partners
- easier to experiment with and redesign than do everything from scratch
- learn from previous experience, have something to improve upon
- (remote...) comparing Swedish with Portuguese, looking for "relative clauses modifying proper nouns": **comparable treebanks**

References

Afonso, S., E. Bick, R. Haber & D. Santos. "Floresta sintá(c)tica": a treebank for Portuguese. Rodríguez & Araujo (eds.), *Proceedings of LREC'2002* (las Palmas, May 2002), ELRA, pp.1698-1703

In *Proceedings of the Worskhop for Temporal and Spatial Information Processing* (Toulouse, July 7th 2001), EACL-ACL 2001, ACL:

- Katz, G. & F. Arosio. The Annotation of Temporal Information in Natural Language Sentences. pp. 104-111
- Setzer, A. & R. Gaizauskas. A Pilot Study on Annotating Temporal Relations in Text. pp. 73-80
- Wilson, G. , I. Mani, B. Sundheim & L. Ferro. A multilingual approach to annotating and extracting temporal information. pp 81-87