Treebanks

Martin Volk
Universität Zürich
Eurospider Information Technology AG

Topics

- type of annotation
- format
- intended usage
- tools for annotation
- tools for verification
- tools for accessing

The Penn Treebank

- a treebank for English built at the University of Pennsylvania
- Phase 1 (1989-1992)
  - 3 million words (Wall Street Journal and others)
  - bracket representation with PoS labels and node labels

Penn Treebank Example from 1991

```
( bd0011sx .)
( (S (NP *)
    (VP Show
      (NP me)
      (NP (NP all)
        the nonstop flights
        (PP (PP from
              (NP Dallas))
            (PP to
              (NP Denver)))
        (ADJP early
            (PP in
              (NP the morning)))))  .)  )
```

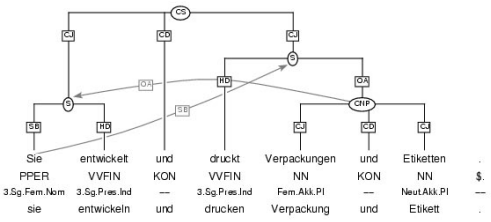

The Penn Treebank

- Phase 2 (1993-1995)
  Enriching part of the original material with
  - syntactic functions
  - traces, null elements, coreference symbols
- Phase 3 (1996-2000)
  - additional material annotated
    - Brown Corpus
    - Switchboard corpus (telephone conversations)

The NEGRA Treebank

- 20'000 sentences
- from the Frankfurter Allgemeine Zeitung
- annotated with the help of the ANNOTATE Treebanking Tool
  - with built-in PoS-Tagger and Parser
- allows crossing branches
- allows secondary edges

The NEGRA Treebank



The NEGRA Treebank

Annotations

- PoS-Tags (STTS)
- Morphological information
- Syntactic nodes (NP, PP, VP, ...)
- Syntactic functions (Subject, Object, Adverbial, ...)

The CZ Treebank

- 3000 sentences with PPs in ambiguous position
- from the 1996 ComputerZeitung
- annotated at the University of Zurich
- following the NEGRA guidelines

Annotation formats

- NEGRA export format
  - line based
  - nesting via pointers
- TIGER XML format
  - XML tags make information explicit

Annotation formats

- Advantages of the NEGRA format over XML
  - well suited for relational databases (= tables)
  - more concise in textual representation
  - crossing branches need pointers

Extraction of PP 5-Tuples

- conversion of nested structure into Prolog
- problem: needed nesting information and immediate precedence information
- extraction of:
  - verb
  - real reference noun
  - possible reference noun
  - preposition
  - core of the PP
  - function of the PP

Extraction of PP 5-Tuples

Some Issues

- Separated verb prefixes and reflexive pronouns
- Multiword proper nouns
- Coordinated NPs and PPs
- Coordinated full verbs
- Postnominal apposition