Def. Treebank

A treebank is a syntactically annotated corpus.

Issues:
- complete analysis vs. partial analysis
- theory-neutral vs. theory-dependent
- spoken vs. written language
- constituents vs. dependency
- annotate grammatical functions?
- manual vs. automatic annotation

Some Remarks

- treebanking is extremely labor-intensive (i.e. costly)
- good planning is therefore necessary
- good tools are crucial
  - they speed up the process
  - they help with consistency
- a detailed stylebook is essential

Penn WSJ Treebank – Example

( (S (NP-SBJ (NP Pierre Vinken)
  ,
  (ADJP (NP 61 years)
   old)
  ,)
  (VP will
    (VP join
      (NP the board)
      (PP-CLR as
        (NP a nonexecutive director))
    (NP-TMP Nov. 29)))
  ).)

Treebanks for English

- Penn Treebank
- BLLIP Treebank
- The Penn-Helsinki Parsed Corpus of Middle English
- Susanne Corpus and Christine Project
- International Corpus of English ICE
- Lancaster Treebank
- The Redwoods HPSG Treebank

Treebanks Projects

- Basque
  - Eus3LB project
- Bulgarian
  - HPSG-based Syntactic Treebank of Bulgarian (BulTreeBank)
- Catalan
  - CAT3LB project
- Chinese
  - The Chinese Treebank Project
- Czech
  - Prague Dependency Treebank
Treebanks Projects (2)

- Danish
  - Danish Dependency Treebank
- Dutch
  - The Alpino Treebank
- French
  - Project TALANA
- German
  - NeGra Project - NeGra Corpus
  - Project TiGER
  - Verbmobil Treebank of Spoken German (TüBa-D/S)
  - The Tübingen Treebank of Written German (TüBa-D/Z)

The Annotation Scheme

- Should the annotation scheme be dependent on a particular theory?
  - Theory-neutrality doesn’t really exist. Every annotation scheme is at least implicitly theory-dependent.
  - Grounding an annotation scheme in a linguistic theory tends to improve consistency of annotations.

Theory-neutral Treebanks

Theory-neutral treebanks do not adhere to any particular linguistic theory

- encode those grammatical properties that are distinguished by many, if not all grammatical frameworks

Advantage:

- more widely usable
- less dependent on whatever version of a particular grammatical theory may have existed at the time when the treebank annotation scheme was determined

  - examples: Penn Treebank, Negra treebank, Tübingen treebanks

Theory-dependent Treebanks

- Prague Dependency Treebank
  - based on Dependency Grammar
- The Redwoods HPSG Treebank
  - based on Head-Driven Phrase Structure Grammar
- CCGbank
  - translation of the Penn Treebank into a corpus of Combinatory Categorial Grammar derivations

Extracting the CFG Grammar

How does Thursday look for you?
Ambiguity – A Serious Problem

example sentence:

Volker Tegeler, stellvertretender Geschäftsführer des Landesverbandes, sagt:

(Volker Tegeler, vice CEO of the national association, says ...)

How many different analyses?

- training corpus: 15 000 sentences, grammar: ca. 6 000 rules

Ambiguity – Sample Analyses

the most probable parse:

```
ENX-ON  NX-ON  R-SIMPX
NE-n    NX-ON  VC
Volker  NX-g   VXF
Tegeler  VVF
```

```
R-SIMPX-ON  R-SIMPX
NX-ON       NX-ON
R-SIMPX     VC
```

```
ADJX-n    NN     ART-g  NN-g  VVF
Stellv.  Geschäftsf. des Landesv. sagt
```

Ambiguity – Sample Analyses

the almost correct parse:

```
ENX-ON  NX-ON  VC-FIN
NE-n    NX-ON  VC-NH
Volker  NX-g   VVF
Tegeler  VVF
```

```
SIMPX-ON

NX-ON
```

```
ADJX-n    NN     ART-g  NN-g sagt
Stellv.  Geschäftsf. des Landesv.
```
Ambiguity – Sample Analyses

- both are based on newspaper texts: Frankfurter Rundschau, taz
- both use same POS tagset: STTS
- both annotate constituent structure and function-argument structure
- Negra: 20,000 sentences; TüBa-D/Z: 15,000 sentences (version 1), 27,000 sentences (version 3)

- different annotation decisions

Negra and TüBa-D/Z

 Parsing and Treebanks – How do they Relate?

- Progress in parsing has been guided by improvements on the Penn Treebank
  - How much of this is dependent on the annotation scheme?
  - German: has two treebanks → ideal for comparing influence of annotation scheme on parsing

A Tree from Negra

That could be the mask that his character pressed like a stamp on his face.

A Tree from TüBa-D/Z

The convoy of the rehearsal visitors’ cars goes along a street that is still called Lagerstraße.

Main Differences in Annotation

- phrase structure
  - Negra: extremely flat
  - TüBa-D/Z: premodification flat, postmodification high
- clause structure
  - Negra: VP ⇒ crossing branches
  - no unary nodes
  - TüBa-D/Z: topological fields
- long-distance relationships
  - Negra: crossing branches
  - TüBa-D/Z: pure tree structure + special labels
The amateur painter can by all means apply this metaphor also to her life.