Treebanks and Parsing

L645
Dept. of Linguistics, Indiana University
Fall 2009
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)
Treebanks Projects (3)

- Italian
  - Turin University Treebank TUT
  - Italian Syntactic-Semantic Treebank
- Portuguese
  - The Floresta Sinta(c)tica project
- Slovene
  - Slovene Dependency Treebank
- Swedish
  - Swedish Treebank
- Turkish
  - METU treebank
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-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
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
Extracting the CFG Grammar

How does Thursday look for you?
Extracting the CFG Grammar

rules:

- **SIMPX** → VF LK MF VC
- VF → ADJX
- LK → VXFIN
- MF → NX PX PX
- VC → PTKVZ
Extracting the CFG Grammar

SIMPX  →  VF-NH LK-NH MF-NH VC-NH
VF-NH  →  ADJX-PRED
or:    LK-NH  →  VXFIN-HD
MF-NH  →  NX-ON PX-FOPP PX-VMOD
VC-NH  →  PTKVZ-VPT
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 – 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

Answer: don’t know

- but 58(!) root categories (if complex information is used in the nodes)
Ambiguity – Sample Analyses

the most probable parse:

R-SIMPX-ON

NX-ON

ENX-ON

NE-n

NE-n

Volker Tegeler

ADJX-n

NN

ART-g

NN-g

VVFIN

stellv.

Geschäftsf. des Landesv. sagt
Ambiguity – Sample Analyses

the almost correct parse:
an “interesting” parse
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 $\rightarrow$ ideal for comparing influence of annotation scheme on parsing
Negra and TüBa-D/Z

- 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
That could be the mask that his character pressed like a stamp on his face.
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 $\Rightarrow$ 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.
Resolved Structure

Diese Metapher kann die Freizeitmalerin durchaus auch auf ihr Leben anwenden.
Resolved Structure

S → NP VMFIN NP ADV VP
## Results

<table>
<thead>
<tr>
<th></th>
<th>lab. prec.</th>
<th>lab. rec.</th>
<th>lab. F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Negra</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no GF</td>
<td>69.96</td>
<td>69.95</td>
<td>69.95</td>
</tr>
<tr>
<td>all GF</td>
<td>47.20</td>
<td>56.43</td>
<td>51.41</td>
</tr>
<tr>
<td>subject</td>
<td>52.50</td>
<td>58.02</td>
<td>55.12</td>
</tr>
<tr>
<td>acc. object</td>
<td>35.14</td>
<td>36.30</td>
<td>35.71</td>
</tr>
<tr>
<td>dat. object</td>
<td>8.38</td>
<td>3.58</td>
<td>5.00</td>
</tr>
<tr>
<td><strong>TüBa</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no GF</td>
<td>89.86</td>
<td>88.51</td>
<td>89.18</td>
</tr>
<tr>
<td>all GF</td>
<td>75.73</td>
<td>74.93</td>
<td>75.33</td>
</tr>
<tr>
<td>subject</td>
<td>66.82</td>
<td>75.93</td>
<td>71.08</td>
</tr>
<tr>
<td>acc. object</td>
<td>43.84</td>
<td>47.31</td>
<td>45.50</td>
</tr>
<tr>
<td>dat. object</td>
<td>24.46</td>
<td>9.96</td>
<td>14.07</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>lab. prec.</th>
<th>lab. rec.</th>
<th>lab. F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Negra</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no GF</td>
<td>69.96</td>
<td>69.95</td>
<td>69.95</td>
</tr>
<tr>
<td>all GF</td>
<td>47.20</td>
<td>56.43</td>
<td>51.41</td>
</tr>
<tr>
<td>subject</td>
<td>52.50</td>
<td>58.02</td>
<td>55.12</td>
</tr>
<tr>
<td>acc. object</td>
<td>35.14</td>
<td>36.30</td>
<td>35.71</td>
</tr>
<tr>
<td>dat. object</td>
<td>8.38</td>
<td>3.58</td>
<td>5.00</td>
</tr>
<tr>
<td><strong>TüBa</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no GF</td>
<td>89.86</td>
<td>88.51</td>
<td>89.18</td>
</tr>
<tr>
<td>all GF</td>
<td>75.73</td>
<td>74.93</td>
<td>75.33</td>
</tr>
<tr>
<td>subject</td>
<td>66.82</td>
<td>75.93</td>
<td>71.08</td>
</tr>
<tr>
<td>acc. object</td>
<td>43.84</td>
<td>47.31</td>
<td>45.50</td>
</tr>
<tr>
<td>dat. object</td>
<td>24.46</td>
<td>9.96</td>
<td>14.07</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>lab. prec.</th>
<th>lab. rec.</th>
<th>lab. F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Negra</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no GF</td>
<td>69.96</td>
<td>69.95</td>
<td>69.95</td>
</tr>
<tr>
<td>all GF</td>
<td>47.20</td>
<td>56.43</td>
<td>51.41</td>
</tr>
<tr>
<td>subject</td>
<td>52.50</td>
<td>58.02</td>
<td>55.12</td>
</tr>
<tr>
<td>acc. object</td>
<td>35.14</td>
<td>36.30</td>
<td>35.71</td>
</tr>
<tr>
<td>dat. object</td>
<td>8.38</td>
<td>3.58</td>
<td>5.00</td>
</tr>
<tr>
<td><strong>TüBa</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no GF</td>
<td>89.86</td>
<td>88.51</td>
<td>89.18</td>
</tr>
<tr>
<td>all GF</td>
<td>75.73</td>
<td>74.93</td>
<td>75.33</td>
</tr>
<tr>
<td>subject</td>
<td>66.82</td>
<td>75.93</td>
<td>71.08</td>
</tr>
<tr>
<td>acc. object</td>
<td>43.84</td>
<td>47.31</td>
<td>45.50</td>
</tr>
<tr>
<td>dat. object</td>
<td>24.46</td>
<td>9.96</td>
<td>14.07</td>
</tr>
</tbody>
</table>