Overview

1. Types of syntactic annotation
2. The Penn Treebank
3. Two German Treebanks
4. Creation of Treebanks
5. Exploitation of Treebanks

Types of syntactic information

A linguistically annotated corpus that includes some grammatical analysis beyond the part-of-speech.

- ‘treebank’ vs. ‘parsed corpus’
  - strict: manual annotation or post-editing
  - we will use ‘treebank’ in the broader sense

Detailed descriptions:

  - Amsterdam: John Benjamins.
- Josip Novi (to appear) “Treebanks”. In: Anke Lüdtke and Mirja Kyli, editors, 

Treebanks

Constituent structure

- American structuralism, e.g. Zellig Harris (1951)
- ‘Bracketing’: sentences consist of hierarchically embedded subparts – constituents
  - strings of words that belong together
  - constituency tests
    - substitution, movement, stand-alone test, ...
- Part-whole relations
  - e.g. an NP consists of determiner, adjective and noun

Dependency structure

- First comprehensive theory: Lucien Tesnière (1959)
- Sentences consist of hierarchically structured asymmetric, binary relations between word forms – dependency relations (‘connexions’)
  - governor, dependent(s)
  - closely related to functional analysis
- Relations
  - e.g. determiner and adjective are subordinated to the noun

\[ \text{DET} \rightarrow \text{ADJ} \rightarrow \text{DET} \rightarrow \text{ADJ} \]
Hybrid models

- Combine constituent and functional (dependency) information.
  - function added as additional sub-label to daughter category, e.g. [S [NP-SB ... ]] in Penn Treebank II
  - constituent label as node label, function as edge label, e.g. in TIGER, TüBa

Treebanks and linguistic theory

Three main types of information
- Constituent structure
  - Lancaster Parsed Corpus (BE, part of LOB)
  - Penn Treebank I (skeletal parsing)
- Dependency structure
  - Prague Dependency Treebank (analytical level: Czech)
  - METU-Sabanci Treebank (Turkish)
- Theory-specific annotation
  - Prague Dependency Treebank (textogrammatical level: Functional Generative Description)
  - BiTTreebank (Head Driven Phrase Structure Grammar, Bulgarian)
  - CCG-Bank (Combinatorial Categorial Grammar)

Treebanks and linguistic theory

Hybrid approaches
- combine constituents with functional/dependency information
  - SUSANNE (AE, part of BROWN corpus)
  - Penn Treebank II (AE)
  - Penn Chinese Treebank (Chinese)
  - NEGRA / TIGER treebank (German)
  - TüBa treebanks (German)
  - ARBORETUM (Danish)

Phrases and Chunks

- A phrase is a constituent of a particular category
  - exocentric phrase vs. endocentric phrase
  - a typical chunk consists of a single content word and surrounding constellation of function words
  - the non-recursive core of a constituent which spans the beginning of the constituent up to its lexical head.
  - [the bold man] was sitting [on his suitcase]

Phrases and Chunks

- recursive phrase structure vs. non-recursive chunking

Types of syntactic annot.: References

  - Dordrecht.
  - SASG/1.8 http://www.lc.cmu.edu/EAGLES/InPapers/InPapers1.html
  - Tübingen: Narr, chap. 4.
The Penn Treebank

**Penn Treebank**

- English treebank built at the University of Pennsylvania
- distributed by the Linguistic data consortium (LDC)
  
  - http://www.ldc.upenn.edu

- Phase I (1989 – 1992)
  - skeletal parse
  - 2.6 million words tagged (PoS) material from Dow Jones News Service (Wall Street Journal)
  - therefor over 1.7 million word hand-parsed material
  - first fully parsed version of Brown Corpus (1mil words)
  - tagged and parsed data from Department of Energy abstracts, IBM computer manuals, MUC-3 and ATIS.

**Penn Treebank: POS annotation**

- Modified BROWN tagset
  - avoids lexical redundancies: no tags that are unique to particular lexical items (exception: ‘TO’).
  - encodes word’s syntactic function when possible, e.g. one CD apple vs. the one,s NN
  - allows for multiple tagging: word’s POS cannot be decided or annotator is unsure → avoid arbitrary decisions
  - 98 POS tags, 12 other tags (punctuation, currency symbols)

- *Phase II (1993 – 1995)*
  - enriching part of the original material with grammatical functions and semantic relations
  - null elements, coreference symbols
  - information about non-continuous constituents / dependencies (traces, coreference symbols).

- *Phase III (1996 - 2000)*
  - additional material
  - Switchboard Corpus (telephone conversations): parsed and disfluency-annotated.

**Penn Treebank: References**

- Arctangent: Danish treebank http://corp.hum.dtu.dk/tn/treebank/danish.html
- BU/Treebank: http://www.bu.edu/ltc/treebank
- CoC: Bank
- Lingo Redwood: http://www.stanford.edu/~redwood
- Negra http://www.cis.uni-saarland.de/projects/treebank/negra
- METU: Babtans Turkish Treebank
  
  - http://metal.metu.edu.tr/corpora/treebank.html
- Penn Chinese Treebank
- Penn Treebank: http://www.cis.upenn.edu/treebank
- Prague Dependency Treebank: http://slab1.nlu.nl/dep/depdb.2.0
- Susanne http://www.gnm.psu.edu/psb.html
- TIGER:
  
  - http://www.ims.uni-stuttgart.de/projects/TIGER/
  - TuBa-DZ: http://www.tls.uni-tuebingen.de/en_tubadz.shtml
- TuBa-DS: http://www.tls.uni-tuebingen.de/en_tubads.shtml
Penn Treebank: WH-Question

<s>&lt;w what&gt;&lt;w tim&gt;</s>

Predicate argument structure:
<tim, what>

Penn Treebank: Passive

<s>&lt;w ball&gt;&lt;w chris&gt;

Predicate argument structure:
<thrown(chris, ball)>

Penn Treebank: Control

<s>&lt;w chris&gt;&lt;w chris&gt;

Predicate argument structure:
<wants(chris, throw(chris, ball))>

Penn Treebank: Discontinuous constituent

<s>&lt;w chris&gt;&lt;w chris&gt;

Predicate argument structure:
<knew(chris, catch(Terry, ball))>

Penn Treebank: Pseudo-attachment

<s>&lt;w I&gt;&lt;w I&gt;

Predicate argument structure:
<see(I, man, with(telescope))>

Penn Treebank: References

-M. Marcus, Graeme King, Mary Ann MacClellan, Robert MacHugh, Ann Saxe, Mark Ferguson, Karen Katz, Brita Schaeferger (1994), The Penn Treebank Annotating predicate argument structure.
TIGER and TüBa – two German treebanks

TIGER Treebank

„Linguistic Interpretation of a GERman Corpus“
- 50,000 sentences (1005/12)
- follow up of NEGRA corpus (20,000 sentences)
- German newspaper texts (Frankfurter Rundschau)
- Free license
- Hybrid annotation
- crossing branches for discontinuous constituents
- Stylebook (in German)
- POS-Tags: STTS-Tagset (Schiller et al. 1999).

TüBinger treebanks

TüBA-D/Z
- „TüBinger Baumbank des Deutschen/Zeitungsersprache“
- Written German: newspaper texts („tageszeitung“ (taz))
- 27,000 sentences, 470,000 words (by 2006/07)
- requires licence for taz-CD ~ EUR 50,-

TüBA-D/S
- „TüBinger Baumbank des Deutschen/Spontansprache“
- Spoken German: Verbmbol dialogues
- 38,000 sentences, 360,000 words
- licence free of charge
- POS-Tags: STTS-Tagset (see TIGER)
- Related: TüBinger Partial Parsed Corpus (TüPP), 200 mill. token

TüBA-D/Z: discontinuous constituents

TüBA-D/Z: Topological fields

Field Description
VF initial field (Vorfeld)
LK left sentence bracket (Linke Satzklammer)
NF middle field (Mittefeld)
VC verb complex (Verb complex)
NF final field (Nachfeld)
LV field for resumptive constructions (Linkversetzungsfield)
C complementizer field (C-Feld)
TIGER & TüBA-D/Z: similarities

- based on newspaper texts
- hybrid annotation
  - constituency / phrase structure — nodes
  - non-terminal nodes: category labels
  - terminal nodes: part-of-speech labels
  - empty categories
  - dependency / predicate argument structure
  - edge labels
- employ secondary edges
- PoS tags: STTS tags (Schiller et al. 1999)
- provide morphological information
- created with tool Annotate (Brants & Plaehn 2000)

TIGER & TüBA-D/Z: differences

- primary ordering principle
  - TIGER: predicate argument structure
  - TüBA-D/Z: topological fields (Höhle 1986)
- encoding free word order
  - TIGER: unordered trees (Skut et al. 1997) / syntactic graphs (Lezius and König 2001)
  - TüBA-D/Z: context-free backbone with specialized edge labels and secondary edges.
- structure sharing, e.g. coordination
  - TIGER: secondary edges
  - TüBA-D/Z: no marking

TüBA-D/Z: Representation

TüBA-D/Z: Bracketing format

%% Sent 1330
(SIMPX)
(VF
(RX-ON
(PPER-TO Wir)))
(LK
(VXFIN-BG
(VAFHN-BG sind)))
(MF
(ADJX-PRED
(ADJD-HD begeistert))))
(S ! !)
TüBa-D/Z: Column format

\*NEGRO export format\* of the Annotate tool (Brants 2000).

<table>
<thead>
<tr>
<th>#EOS</th>
<th>1630</th>
<th>26</th>
<th>1047480241</th>
<th>567</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mir</td>
<td>PPER</td>
<td>1p1</td>
<td>HD</td>
<td>500</td>
</tr>
<tr>
<td>sind</td>
<td>VAFIN</td>
<td>1p1</td>
<td>HD</td>
<td>501</td>
</tr>
<tr>
<td>beginusers</td>
<td>ADJD</td>
<td>--</td>
<td>HD</td>
<td>502</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td>--</td>
<td>0</td>
</tr>
<tr>
<td>450</td>
<td>0X</td>
<td>--</td>
<td>OS</td>
<td>503</td>
</tr>
<tr>
<td>4501</td>
<td>VOFL</td>
<td>--</td>
<td>HD</td>
<td>504</td>
</tr>
<tr>
<td>4502</td>
<td>ADXJ</td>
<td>--</td>
<td>PRED</td>
<td>505</td>
</tr>
<tr>
<td>4503</td>
<td>VP</td>
<td>--</td>
<td>--</td>
<td>506</td>
</tr>
<tr>
<td>4504</td>
<td>LK</td>
<td>--</td>
<td>--</td>
<td>506</td>
</tr>
<tr>
<td>4505</td>
<td>ME</td>
<td>--</td>
<td>--</td>
<td>506</td>
</tr>
<tr>
<td>4506</td>
<td>SIMFX</td>
<td>--</td>
<td>--</td>
<td>0</td>
</tr>
<tr>
<td>#EOS</td>
<td>1630</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TüBa-D/Z: Export XML

```xml
<node cmt="10" date="2003031215:44:01" origin="T\*5907.132" comment="">
  <node cmt="SIMFX" func="--" parent="5" comment="">
  
  </node>
</node>
```

German treebanks: References


Creating Treebanks

\*Manual annotation\*
- Word freak
- CLARK

\*Automatic annotation with human post-editing\*
- Collins' Parser
- LoPar / BiPar
- Stanford Parser

\*Interactive annotation\*
- Annotate Tool
  - runs under Solaris and Linux
  - needs the GNU C-Compiler, Tcl/Tk 8.0, Embedded Tk, and an installation of MySQL
  - includes statistical part-of-speech tagger (TnT) and an parser (cascaded Markov models).

Creating Treebanks: Annotate tool

![Creating Treebanks: Annotate tool](image1)

Creating Treebanks: Annotate tool

![Creating Treebanks: Annotate tool](image2)
Exploiting Treebanks: Charniak 1996

- \( r \) rule
- \( |r| \): number of times rule \( r \) occurred in the training corpus
- \( RN(r) \): root node of \( r \) node on the left-hand side of \( r \); the non-terminal that \( r \) expands
- Rule probability

\[
p(r) = \frac{|r|}{\sum_{r' \in R} |RN(r')|}^{|r|}
\]

- Tree/grammar transformations
  - undocumented PoS: ORT, PRT not in guidelines (but in treebank)
  - new PoS: AUX, AUXG to distinguish auxiliaries
  - new cat: S1 as new start symbol (root symbol)
- Resulting PCFG
  - 10,605 rules
  - 3,943 occurred more than once

---

Exploiting Treebanks: Charniak 1996

- Why does a parser not identify the correct parse?
  - The necessary rules are not in the grammar
  - The rules are there but their probability is not correct
  - The probability is there but the tag sequence itself does not provide sufficient information to select the correct parse.
  - The information is sufficient but because the parser could not consider all the possible parses, it did not find the correct parse.
  - It found the correct parse but the treebank 'gold standard' was wrong.

---

Exploiting Treebanks: Charniak 1996

<table>
<thead>
<tr>
<th>Sentence Length</th>
<th>Average Length</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-12</td>
<td>8.7</td>
<td>91.7</td>
<td>97.9</td>
<td></td>
</tr>
<tr>
<td>2-16</td>
<td>11.4</td>
<td>85.0</td>
<td>87.7</td>
<td>94.5</td>
</tr>
<tr>
<td>2-20</td>
<td>13.8</td>
<td>83.5</td>
<td>86.2</td>
<td>92.8</td>
</tr>
<tr>
<td>2-25</td>
<td>16.3</td>
<td>82.0</td>
<td>84.0</td>
<td>90.5</td>
</tr>
<tr>
<td>2-30</td>
<td>18.7</td>
<td>80.8</td>
<td>82.5</td>
<td>89.5</td>
</tr>
<tr>
<td>2-40</td>
<td>21.5</td>
<td>75.8</td>
<td>80.4</td>
<td>87.7</td>
</tr>
</tbody>
</table>

* parsing with subset of rules that occurred more than once did not change much in the result.
* other non-lexicalised parsers were outperformed (especially wrt sentences > 17 words).

---

Problems of Independence Assumption

- **Motivation**
  Structural ambiguities can be better resolved if contextual information is taken into account.

- **Goal**
  Keep context-free architecture but enrich local trees with contextual information.

- **Advantages**
  - Allows for more specific analyses
  - More phenomena can be differentiated in grammar training

- **Drawbacks**
  - Number of rules increases
  - Problem of sparse data

---

How to integrate contextual information?

- Tree transformations
  - Parent Encoding (Johnson 1989)
  - Base NP Marking (Collins 1999)
  - Mark complements (Collins 1997)
  - Add subcategorization information (Carroll and Rooth 1989)
  - Mark long distance dependencies (Collins 1997)
  - Marks bought books
    S(bought) -> NP(Marks) VP(bought)