

Introduction to Corpus Resources, Annotation and Access: Syntactic Annotation

Sabine Schulte im Walde
Universität des Saarlandes

Heike Zinsmeister
Universität Tübingen

Foundational Course
18th European Summer School in Logic, Language and Information
Málaga, Spain
July 31 - August 4, 2006

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

Treebanks

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:

- Anne Abeillé ed. (2003) "Treebanks. Building and Using Parsed Corpora". Dordrecht, Boston, London: Kluwer Academic Publishers.

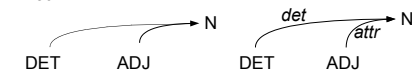
- Joakim Nivre (to appear) "Treebanks". In: Anke Lüdeling and Merja Kytö, editors, "Corpus linguistics: An International Handbook", Berlin: Mouton de Gruyter.

Constituent structure

- American structuralism, e.g. Zelig 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
- [_{NP} [_{DET}] [_{ADJ}] [_N]]

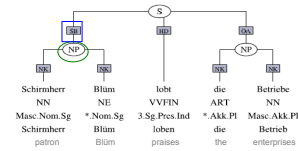
Dependency structure

- 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



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-Sabancı Treebank (Turkish)
- Theory-specific annotation**
 - » Prague Dependency Treebank (tectogrammatical level: Functional Generative Description)
 - » BulTreebank (Head Driven Phrase Structure Grammar, Bulgarian)
 - » CCG-Bank (Combinatory Categorical Grammar)

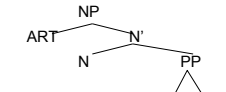
Treebanks and linguistic theory

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

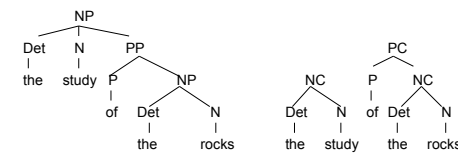
- A **phrase** is a constituent of a particular category
 - » exocentric phrase vs. endocentric phrase



- 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



- Fully-fledged analysis vs. chunking (also 'partial parsing').

Types of syntactic annot.: References

- Anne Abeillé ed. (2003) Treebanks. Building and Using Parsed Corpora. Dordrecht, Boston, London: Kluwer Academic Publishers.
- Steven Abney (1991). Parsing By Chunks. In: Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.
- Geoffrey Leech, B. Barnett, Peter Kahrel. (1996). EAGLES. Recommendations for the Syntactic Annotation of Corpora. EAGLES Document EAG-TCWG-SASG/1.8 <http://www.lic.cnr.it/EAGLES96/segsasg1/segsasg1.html>
- Lothar Lemnitzer & Heike Zinsmeister (2006). *Korpuslinguistik. Eine Einführung*. Tübingen: Narr, chap. 4.
- Joakim Nivre (to appear) Treebanks. In: Anke Lüdeling and Merja Kytö (eds.), *Corpus linguistics: An International Handbook*, Berlin: Mouton de Gruyter.

Types of syntactic annot.: References

Online resources

- ArboRetum: Danish treebank http://corp.hum.sdu.dk/tgrepeye_da.html
- BuTTreebank: <http://www.bultreebank.org>
- CCG-Bank:
- LinGo Redwoods <http://lingo.stanford.edu/redwoods>
- Negra: <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus>
- METU-Sabancı Turkish Treebank <http://il.metu.edu.tr/~corpus/treebank.html>
- Penn Chinese Treebank
- Penn Treebank: <http://www.cis.upenn.edu/~treebank>
- Prague Dependence Treebank: <http://ufal.mff.cuni.cz/pdt.2.0>
- Susanne: <http://www.grsampson.net/RSue.html>
- Tiger: <http://www.ims.uni-stuttgart.de/projekte/TIGER/>
- TüBA-D/Z: http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml
- TüBA-D/S: http://www.sfs.uni-tuebingen.de/en_tuebads.shtml

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 mill words tagged (PoS) material from Dow Jones News Service (Wall Street Journal)
 - thereof over 1,7 mill word hand-parsed material
 - first fully parsed version of Brown Corpus (1mill words)
 - tagged and parsed data from Department of Energy abstracts, IBM computer manuals, MUC-3 and ATIS.

Penn Treebank

- 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: 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 ones_NN
 - allows for multiple tagging: word's POS cannot be decided or annotator is unsure → avoid arbitrary decisions
 - 36 POS tags, 12 other tags (punctuation, currency symbols)

Penn Treebank: POS annotation

1. CC Coordinating conj.	25. TO to
2. CD Cardinal number	26. UH Interjection
3. DT Determiner	27. VB Verb, base form
4. EX Existential there	28. VBD Verb, past tense
5. FW Foreign word	29. VBG V, gerund/pres. participle
6. IN Preposition/subord. conjunction	30. VBN Verb, past participle
7. JJ Adjective	31. VBF V, non-3rd ps. sing. present
8. JJR Adjective, comp.	32. VBZ V, 3rd ps. sing. present
9. JJS Adjective, superl.	33. WDT wh-determiner
10. LS List item marker	34. WP wh-pronoun
11. MD Modal	35. WP Possessive wh-pronoun
12. NN Noun, sg. or mass	36. WRB wh-adverb
13. NNS Noun, plural	37. # Pound sign
14. NNP Proper noun, singular	38. \$ Dollar sign
15. NNPS Proper noun, plural	39. . Sentence-final punctuation
16. PDT Predeterminer	40. , Comma

Penn Treebank: POS annotation

17.POS Possessive ending	41. : Colon, semi-colon
18.PRP Personal pronoun	42. (Left bracket character
19.PP Possessive pronoun	43.) Right bracket character
20.RB Adverb	44. " Straight double quote
21.RBR Adverb, comparative	45. ' Left open single quote
22.RBS Adverb, superlative	46. " Left open double quote
23.RP Participle	47. ' Right close single quote
24.SYM Symbol (mathematical or scientific)	48. " Right close double quote

Penn Treebank: syntactic annotation

1. ADJP	Adjective phrase
2. ADVP	Adverb phrase
3. NP	Noun phrase
4. PP	Prepositional phrase
5. S	Simple declarative clause
6. SBAR	Clause introduced by subordinating conjunction or 0 (zero 'that')
7. SBARQ	Direct question introduced by wh-word or wh-phrase
8. SINV	Declarative sentence with subject-aux inversion
9. SQ	Subconstituent of SBARQ excluding wh-word or wh-phrase
10. VP	Verb phrase
11. WHADVP	Wh-adverb phrase
12. WHNP	Wh-noun phrase
13. WHPP	Wh-prepositional phrase
14. X	Constituent of unknown or uncertain category

Penn Treebank: Skeletal parsing

```
( (S
  (NP Martin Marietta Corp.)
  was
  (VP given
    (NP a
      $ 29.9
      million Air Force contract
    (PP for
      (NP low-altitude navigation
        and
        targeting equipment))))))
.)
```

Penn Treebank: Syntactic tagset II

Null elements

1. * ``Understood'' subject of infinitive or imperative
2. 0 Zero variant of 'that' in subordinate clauses
3. T Trace---marks position where moved wh-constituent is interpreted
4. NIL Marks position where preposition is interpreted in pied-piping contexts

Penn Treebank: Functional tagset

Text categories

-HLN headlines and datelines
 -LST list markers
 -TTL titles

Grammatical functions

-CLF true clefts
 -NOM non NPs that function as NPs
 -ADV clausal and NP adverbials
 -LGS logical subjects in passives
 -PRD non VP predicates
 -SBJ surface subject
 -TPC topicalized and fronted constituents
 -CLR closely related

Penn Treebank: Functional tagset

Semantic roles

-VOC vocatives
 -DIR direction and trajectory
 -LOC location
 -MNR manner
 -PRP purpose
 -TMP temporal phrases

Pseudo-attachment

ICI Interpret Constituent Here
 PPA Permanent Predictable Ambiguity
 RNR Right Node Raising
 EXP Expletive

Penn Treebank: WH-Question

```
(S BARQ (WHNP-1 What)
  (SQ is
    (NP-SBJ Tim)
    (VP eating
      (NP *T*-1)))
  ?)
```

Predicate argument structure:

```
eat(Tim, what)
```

Penn Treebank: Passive

```
(S (NP-SBJ-1 The ball)
  (VP was
    (VP thrown
      (NP *-1)
      (PP by
        (NP-LGS Chris))))
  .)
```

Predicate argument structure:

```
throw(Chris, ball)
```

Penn Treebank: Controll

```
(S (NP-SBJ-1 Chris)
  (VP wants
    (S (NP-SBJ *-1)
      (VP to
        (VP throw
          (NP the ball))))))
  .)
```

Predicate argument structure:

```
wants(Chris, throw(Chris, ball))
```

Penn Treebank: Discontinuous constituent

```
(S (NP-SBJ Chris)
  (VP knew
    (SBAR *ICH*-1)
    (NP-TMP yesterday)
    (SBAR-1 that
      (S (NP-SBJ Terry)
        (VP would
          (VP catch
            (NP the ball))))))
  .)
```

Predicate argument structure:

```
know(Chris, catch(Terry, ball))
```

Penn Treebank: Pseudo-attachment

```
(S (NP-SBJ I)
  (VP saw
    (NP (NP the man)
      (PP *PPA*-1))
    (PP-CLR-1 with
      (NP the telescope)))
  .)
```

Predicate argument structure:

```
see(I, man, with(telescope))
```

```
see(I, man)
```

Penn Treebank: References

•Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz (1993) Building a large annotated corpus of English: the Penn Treebank. In: Computational Linguistics, Vol.19. (reprinted in: Susan Armstrong (ed.) Using Large Corpora, Cambridge/London: MIT Press, 273-290)
<http://www ldc.upenn.edu/Catalog/docs/treebank2/c193.html>

•Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, Britta Schasberger (1994). The Penn TREEBANK: Annotating predicate argument structure.
<http://www ldc.upenn.edu/Catalog/docs/treebank2/arpa94.html>

•Overview: Ann Taylor, Mitchell Marcus, Beatrice Santorini (2003). The Penn Treebank: an overview. In: Anne Abeillé (ed.) *Treebanks: building and using parsed corpora*. Dordrecht: Kluwer, 5-22

•Survey of Penn Treebank II annotation:
<http://www ilc.cnr.it/EAGLES96/synlex/node26.htm#SECTION00435000000000000000>

•POS-Tagset with examples <http://www.comp.leeds.ac.uk/amalgam/tagsets/upenn.html>

•Penn Treebank Online Search
<http://www ldc.upenn.edu/ldc/online/treebank/>

TüBa-D/Z: Column format

„NEGRA export format“ of the Annotate tool (Brants 2000).

```
#BOS 1630 26 1047480241 567
Wir PPER np*1 HD 500
sind VAFIN 1pis HD 501
begeistert ADJD -- HD 502
! $. -- -- 0
#500 NX -- ON 503
#501 VXFIN -- HD 504
#502 ADJX -- PRED 505
#503 VF -- -- 506
#504 LK -- -- 506
#505 MF -- -- 506
#506 SIMPX -- -- 0
#EOS 1630
```

TüBa-D/Z: Export XML

```
<sentence editor="26" date="2003031215:44:01" origin="T990507.132" comment="">
<node cat="SIMPX" func="-" parent="0" comment="">
<node cat="VF" func="-" comment="">
<node cat="NX" func="ON" comment="">
<word form="Wir" pos="PPER" morph="np*1" func="HD" comment="">
</node>
<node cat="LK" func="-" comment="">
<node cat="VXFIN" func="HD" comment="">
<word form="sind" pos="VAFIN" morph="1pis" func="HD" comment="">
</node>
<node cat="MF" func="-" comment="">
<node cat="ADJX" func="PRED" comment="">
<word form="begeistert" pos="ADJD" morph="-" func="HD" comment="">
</node>
</node>
<word form="!" pos="$. " morph="-" func="-" parent="0" comment="">
</sentence>
```

German treebanks: References

Thorsten Brants (1997). "The NeGra Export Format". CLAUS Report #98. Saarland University, Computational Linguistics, Saarbrücken.

Thorsten Brants and Oliver Plaehn (2000). "Interactive Corpus Annotation" in "Second International Conference on Language Resources and Evaluation" (LREC-2000), Athens, Greece.

Online resources:

STTS overview:

- <http://www.ilc.cnr.it/EAGLES96/TT-rep/node54.html>,
- tagging manual (in German)

<http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf>

TüBa-D/Z: http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml

TüBa-D/S: http://www.sfs.uni-tuebingen.de/en_tuebads.shtml

Annotate: <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html>

Creating Treebanks

- Manual annotation
- Word freak
- CLaRK
- Automatic annotation with human post-editing
 - Collins' Parser
 - LoPar / BitPar
 - 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 a parser (cascaded Markov models).

Creating Treebanks: Annotate tool

The screenshot shows the Annotate tool interface. At the top, there are fields for 'Document' (TUEBADZ0001) and 'Sentence' (Herk 2). Below these are buttons for 'Save', 'Print', 'Exit', and 'Options'. The main area contains a text editor with the sentence: 'Die männliche Trinket sei gut erforscht'. Below the text is a dependency graph showing nodes and edges representing the syntactic structure. At the bottom, there are fields for 'Move', 'Dependency', and 'Erasor'.

Creating Treebanks: Annotate tool

This screenshot is identical to the previous one, showing the Annotate tool interface with the same text and dependency graph.

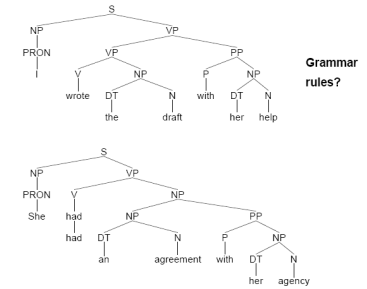
Creating Treebanks: References

Thorsten Brants and Oliver Plaehn (2000), "Interactive Corpus Annotation. In Proceedings of LREC-2000. Athens, Greece.

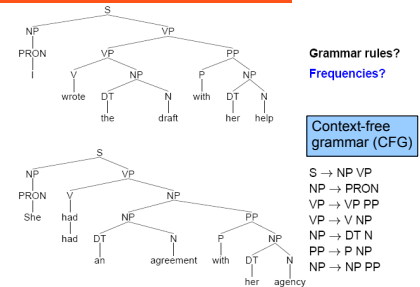
Online resources

- Annotate: <http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html>
- ClARK: <http://www.bultreebank.org/clark/doc/contents.html>
- Collins' Parser: <http://people.csail.mit.edu/mcollins/code.html>
- BitPar: <http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/BitPar.html>
- LoPar: <http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/LoPar.html>
- Stanford Parser: <http://www.nlp.stanford.edu/software/lex-parser.shtml>
- Wordfreak: <http://wordfreak.sourceforge.net>

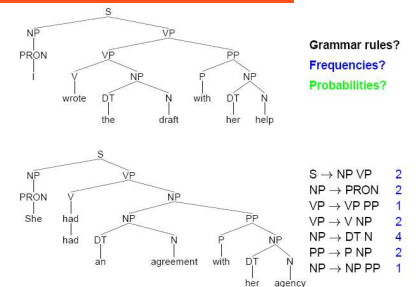
Exploiting Treebanks: Parser Training



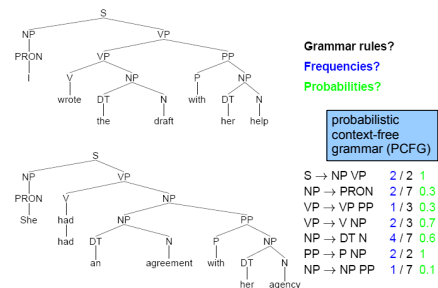
Exploiting Treebanks: Parser Training



Exploiting Treebanks: Parser Training



Exploiting Treebanks: Parser Training



Exploiting Treebanks: Charniak 1996

- **Motivation**
'common wisdom': treebank grammars do not work well
- **Result**
simple treebank-based PCFGs perform as good as other non-lexicalized grammars.
- **Setting**
 - standard chart parser
 - preliminary version of Penn Treebank (Wall Street Journal)
 - test corpus: ~ 30,000 words (1-40 words/sentence)
 - training corpus: ~ 300,000 words
- **Grammar extraction**
 - read rules off all training sentences
 - trace elements were ignored

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

$$p(r) = \frac{|r|}{\sum_{r' \in \{r' | RN(r') = RN(r)\}} |r'|}$$

Exploiting Treebanks: Charniak 1996

- 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

Sentence Length	Average Length	Precision	Recall	Accuracy
2-12	8.7	88.6	91.7	97.9
2-16	11.4	85.0	87.7	94.5
2-20	13.8	83.5	86.2	92.8
2-25	16.3	82.0	84.0	90.8
2-30	18.7	80.6	82.5	89.5
2-40	21.9	78.8	80.4	87.7

- 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).

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.

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)
- Add lexical information (Carroll and Rooth 1993, Collins 1999, Dubey and Keller 2003)
 - Marks bought books
 - S(bought) -> NP(Marks) VP(bought)