

Corpus Linguistics (L615)

Syntactic Annotation and Treebanks

Markus Dickinson

Department of Linguistics, Indiana University
Spring 2009

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Treebanks

Theory-dependency

Constituency &
Dependency

Examples

References

Navigation icons

1/38

Treebanks

A **treebank** is a syntactically annotated corpus

As with other corpora, they have several general issues:

- ▶ spoken vs. written language?
 - ▶ Spoken language faces unique structural challenges
- ▶ manual vs. automatic annotation?
 - ▶ Parsers are not more than 90% precise, generally speaking
- ▶ theory-neutral vs. theory-dependent?
 - ▶ Every decision is a theoretical decision

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Treebanks

Theory-dependency

Constituency &
Dependency

Examples

References

Navigation icons

2/38

Treebank issues

And then there are issues specific to treebanks

- ▶ complete analysis vs. partial analysis
 - ▶ Syntactic *chunks* are easier to annotate more reliably and can be used for a variety of purposes
 - ▶ Chunks are generally non-recursive NPs and PPs
- ▶ constituency vs. dependency annotation
 - ▶ Within constituency annotation: should we annotate grammatical functions?

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Treebanks

Theory-dependency

Constituency &
Dependency

Examples

References

Navigation icons

3/38

Some remarks about treebanking

- ▶ 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
 - ▶ try Annotate!
- ▶ a detailed stylebook is essential
- ▶ every time you hire a well-trained linguist, your treebank will get better

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Treebanks

Theory-dependency

Constituency &
Dependency

Examples

References

Navigation icons

4/38

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)))
  .))
```

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Treebanks

Theory-dependency

Constituency &
Dependency

Examples

References

Navigation icons

6/38

Guidelines examples

From Bies et al. (1995, p. 12-13)

1.1.3 Level of attachment

- ▶ The following are attached at S-level: subject NP, highest VP, fronted constituents, initial and final punctuation, and most modifiers that precede the verb phrase. When there is no VP (as in “small clauses”), the predicate is labeled -PRD, and it and any following adjuncts are attached at S-level.
- ▶ VP-level:
 1. Almost all modifiers that follow the verb are attached under the lowest appropriate VP. When there is conjunction and the modifier applies to both VPs, the modifier is attached at conjunction level.
 2. An exception is made for modifiers that are interpreted as appositives to the event or the predicate itself. Such modifiers are adjoined to VP. Some of them may also have a -ADV tag.

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Treebanks

Theory-dependency

Constituency &
Dependency

Examples

References

Navigation icons

7/38

Trebanks for spoken language (cont.)

Things to note:

- ▶ Need to have explicit notation for speakers
- ▶ Have explicit disfluency tags (e.g., -DFL-)
- ▶ Interjections (INTJ) are much more prevalent, so bracketing guidelines need to know where to put them.
 - ▶ Also, questions are much more common than in newspaper text, so annotation scheme needs to have good coverage of questions.

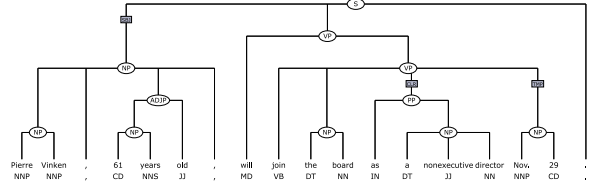
Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Constituent-Based Annotation

- ▶ Constituency annotation describes phrase *structure* and clause *structure*
 - ▶ e.g., noun phrases, adjectival phrases, adverbial phrases, clauses
 - ▶ prominent example: Penn treebank
- ▶ Structures are often recursive
- ▶ For languages like German, this might also include notions such as topological fields

Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Penn WSJ Treebank – Example



Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Properties of constituency annotation

Some properties of (continuous) constituency annotation:

- ▶ **recursive** = a rule can be reapplied (within its hierarchical structure).
 - ▶ NP → NP PP
 - ▶ PP → P NP
- ▶ potentially (**structurally**) **ambiguous** = have more than one analysis
 - (1) I [_{VP} saw [_{NP} [_{NP} the man] [_{PP} with the telescope]]]
 - (2) I [_{VP} saw [_{NP} the man] [_{PP} with the telescope]]
 - ▶ This kind of ambiguity is generally not encoded

There is little notion of being a **generative** grammar that distinguishes grammaticality

- ▶ unlike traditional context-free grammars in linguistics

Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Discontinuous constituents

- ▶ Discontinuous constituents (or equivalents) have been proposed in a wide range of syntactic frameworks (e.g., HPSG, Reape 1993; TAG, Rambow and Joshi 1994; DG, Plátek et al. 2001; other, McCawley 1982)
- ▶ They are also used in the two German treebanks: Verbmobil (Hinrichs et al. 2000) and the TIGER corpus (Brants et al. 2002)
- ▶ German extraposition example (Brants et al. 2002):

Ein Mann kommt , **der lacht**
 a man comes , who laughs

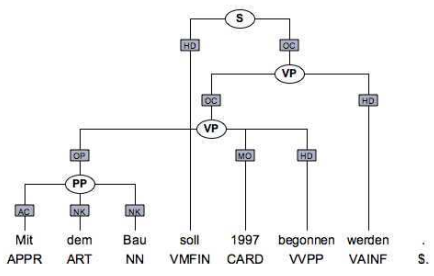
'A man who laughs comes.'

Discontinuous string **Ein Mann der lacht** labeled as NP.

Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Discontinuous constituent example

A sentence with crossing branches in TIGER:



Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Dependency Grammar

Dependency grammar is interested in grammatical relations between words of a sentence, the governing and the dependent words.

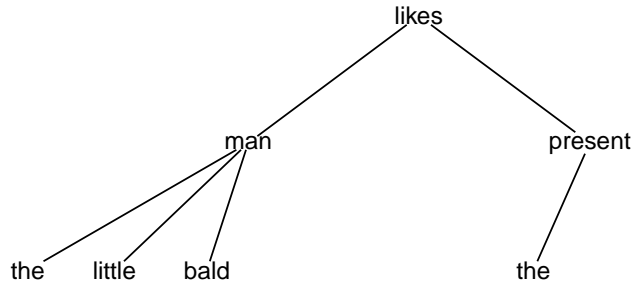
- ▶ PSG describes the *structure* of a sentence

Dependency grammar does not propose a recursive structure but rather a network of relations

- ▶ the verb is the part of the sentence on which ultimately everything depends
- ▶ The direction of a link represents the dependency, the angle represents the word order

Dependency Grammar - An Example

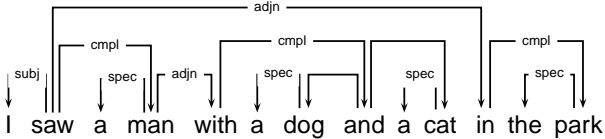
the little bald man likes the present



Extending Dependency Grammar

- ▶ dependency grammars are often extended by labels that denote the grammatical function that the dependent word has with regard to its governor

Example:



from (Lin 1995)

Properties of dependency annotation

- ▶ Generally, you can have non-adjacent arcs
 - ▶ You also get non-**projective** structures, where the dependency arcs cross
 - ▶ These structures generally have correlates with discontinuous constituents
- ▶ Difficult aspects to cover when no constituency is used:
 - ▶ Coordination: what is the head of a coordinate phrase?
 - ▶ Verbal modification: no distinction between sentential & VP adjuncts
- ▶ Some dependency labels are more syntactic, some more semantic

The Prague Dependency Treebank (PDT)

The PDT has different layers to handle syntactic and semantic relations

1. Morphemic layer: tag assigned to each word form (c. 3000 tag values)
2. Analytic tree structures: dependency relations for every word form & punctuation (forms a rooted tree)
3. Tectogrammatical tree structures: underlying sentence representations

Both the analytic and tectogrammatical layers are (formally) dependency structures, but capturing different information

The tectogrammatical level

Tectogrammatical structures are more semantic and have the following aspects:

- ▶ only lexical words serve as tree nodes
 - ▶ Auxiliaries & prepositions are attached as indices to lexical items
- ▶ Nodes are added for surface deletions
- ▶ Non-projectivity is not allowed
- ▶ Analytic functions (e.g., subject) are replaced with tectogrammatical ones
 - ▶ e.g., Actor/Bearer, Patient, Addressee, ...
- ▶ Topic-Focus information is added

Dependency & constituency annotation

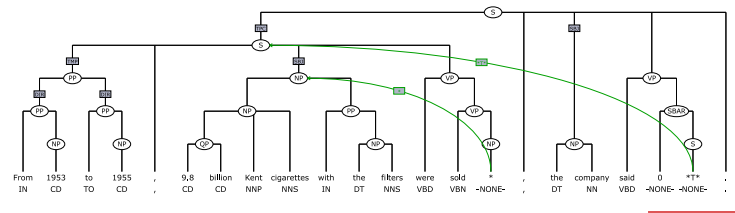
In principle, if a constituency treebank contains information about what the head is, then it can also have dependencies

- ▶ The Talbanken05 corpus of Swedish is a good example (<http://w3.msi.vxu.se/~nivre/research/Talbanken05.html>)
- ▶ They converted a 1976 corpus into one with both kinds of annotation:
 1. Original flat annotation converted to bare phrase structure
 2. Bare phrase structure extended to full phrase structure
 3. Full phrase structure converted to dependency annotation, using grammatical functions as edge labels

On jones at /Volumes/Data/Corpora/sv/Talbanken05

Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

English Treebanks -Penn Treebank



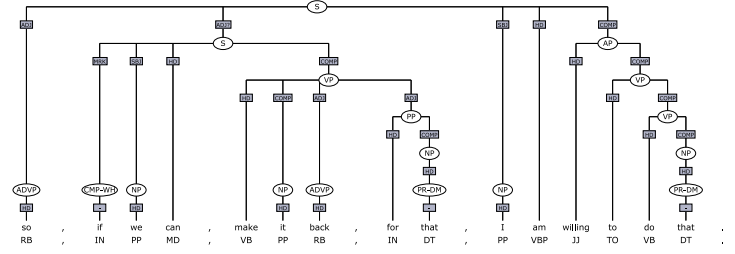
Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

English Treebanks -ICE Treebank

```
[<#6:1> <sent>]
PU,CL(main,montr,pass,pres)
SU,NP
NPHD,PRON(pers,sing) {It}
VB,VP(montr,pres,pass)
OP,AUX(pass,pres) {is}
MVB,V(montr,edp) {chosen}
A,PP
P,PREP(ge) {for}
PC,NP(coordn)
CJ,NP
NPHD,N(com,sing) {comfort}
COOR,CONJUNC(coord) {and}
CJ,NP
NPHD,N(com,sing) {ease}
NPPO,PP
P,PREP(ge) {of}
PC,NP
NPHD,N(com,sing) {washing}
COOR,CONJUNC(coord) {rather than}
CJ,NP
NPHD,N(com,sing) {stylishness}
PUNC,PUNC(per) {.
```

Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

English Treebanks -Verbmobil Treebank



Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Some other 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

Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Treebanks Projects

- ▶ Arabic
 - ▶ Penn Arabic Treebank
- ▶ Bulgarian
 - ▶ HPSG-based Syntactic Treebank of Bulgarian (BulTreeBank)
- ▶ Catalan
 - ▶ CAT3LB project
- ▶ Chinese
 - ▶ Penn Chinese Treebank
 - ▶ Sinica Treebank
- ▶ Czech
 - ▶ Prague Dependency Treebank

Corpus Linguistics
 Syntactic Annotation and Treebanks
 Treebanks
 Theory-dependency
 Constituency & Dependency
 Examples
 References

Trebanks 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)

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ 🔍 ↻

37/38

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Trebanks
Theory-dependency

Constituency &
Dependency

Examples

References

Trebanks Projects (3)

- ▶ Italian
 - ▶ Turin University Treebank TUT
 - ▶ Italian Syntactic-Semantic Treebank
- ▶ Japanese
 - ▶ Verbmobil Treebank of Spoken Japanese (TüBa-J/S)
- ▶ Portuguese
 - ▶ The Floresta Sinta(c)tica project
- ▶ Swedish
 - ▶ Talbanken05, Swedish Treebank
- ▶ Turkish
 - ▶ METU treebank

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ 🔍 ↻

38/38

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Trebanks
Theory-dependency

Constituency &
Dependency

Examples

References

References

- Bies, Ann, Mark Ferguson, Karen Katz and Robert MacIntyre (1995). *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. University of Pennsylvania.
- Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith (2002). The TIGER Treebank. In *Proceedings of TLT-02*. Sozopol, Bulgaria.
- Hinrichs, Erhard, Julia Bartels, Yasuhiro Kawata, Valia Kordoni and Heike Telljohann (2000). The Tübingen Treebanks for Spoken German, English, and Japanese. In Wolfgang Wahlster (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin: Springer, pp. 552–576.
- McCawley, James D. (1982). Parentheticals and discontinuous constituent structure. *Linguistic Inquiry* 13(1), 91–106.
- Plátek, Martin, Tomáš Holan, Vladimír Kuboň and Karel Oliva (2001). Word-Order Relaxations and Restrictions within a Dependency Grammar. In G. Satta (ed.), *Proceedings of the Seventh International Workshop on Parsing Technologies (IWPT)*. Beijing: Tsinghua University Press, pp. 237–240.
- Rambow, Owen and Aravind Joshi (1994). A Formal Look at Dependency Grammars and Phrase-Structure Grammars, with Special Consideration of Word-Order Phenomena. In L. Wanner (ed.), *Current Issues in Meaning-Text-Theory*, London: Pinter. <http://arxiv.org/abs/cmp-lg/9410007>.
- Reape, Mike (1993). A Formal Theory of Word Order: A Case Study in West Germanic. Ph.D. thesis, University of Edinburgh, Edinburgh.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ 🔍 ↻

38/38

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Trebanks
Theory-dependency

Constituency &
Dependency

Examples

References

- Vadas, David and James Curran (2007). Adding Noun Phrase Structure to the Penn Treebank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 240–247.

◀ ▶ ⏪ ⏩ ⏴ ⏵ ⏶ ⏷ 🔍 ↻

38/38

Corpus Linguistics

Syntactic
Annotation and
Treebanks

Trebanks
Theory-dependency

Constituency &
Dependency

Examples

References