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

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
- a detailed stylebook & thorough training are essential

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.

Some syntactic annotation tools

- Brat: http://brat.nlplab.org
- Annotate: http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate
- XCDG: http://nats-www.informatik.uni-hamburg.de/CDG/DownloadPage
- CLaRK: http://www.bultreebank.org/clark/

Can find more at, e.g., http://www.clarin.eu/view_tools
Theory-dependency vs. theory-neutrality

Some treebanks encode a great deal of theory, e.g.,:

- 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

Some treebanks do not adhere to any particular linguistic theory:

- Widely usable
- Less dependent on particular grammatical theory at
Prague Dependeny Treebank
Combinatory Categorial Grammar derivations
 Chunck is generally non-recursive NPs and PPs

Treebanks for spoken language

Here's an example from the Switchboard (swbd) corpus from the PTB:

```plaintext
(CODE (SYM SpeakerA1) (. .))
(INTJ (UH Okay) (. .) (-DFL- E_S) )
(CODE (SYM SpeakerB2) (. .))
(SBARQ
  (INTJ (UH So) )
  (WHNP-1 (WP who) )
  (SQ (BES 's)
   (NP-SBJ (PRP$ your)
    (, ,)
    (INTJ (UH uh) )
    (, ,) (JJ favorite) (NN team) )
   (NP-PRD (NONE- T*1) )
  )
  (-DFL- E_S) )
)```

Partial analysis

In a sense, the PTB annotation is only partial, at least in some spots, and some researchers have added structure e.g., Vadas and Curran (2007) add intenal noun phrase structure

- Instead of this:
  (NP (NN lung) (NN cancer) (NNS deaths))
- we have this:
  (NP (NML (NN lung) (NN cancer) )
   (NNS deaths )

You can get this standoff annotation at:

Various issues for treebanks:

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

Examples:

- Penn Treebank, Negra treebank, Tübingen treebanks

Treebanks for spoken language

- do not adhere to any particular linguistic theory
- encode those grammatical properties that are distinguished by most grammatical frameworks

Advantage:

- Widely usable
- Less dependent on particular grammatical theory at the time when the treebank annotation scheme was determined

Examples:

- Penn Treebank, Negra treebank, Tübingen treebanks

References

Examples

Guidelines

Remarks

Guidelines example (cont.)

(S (NP-SBJ Investors)
  (VP might
    (VP force
      (S (NP-SBJ the government)
        (VP to
          (VP raise
            (NP the coupon)
            (PP-CLR to
              (NP (QP more than 7)
                %)))))))))))

"You can get this standoff annotation at: http://www.cs.usyd.edu.au/~dvadas1/?NounPhrases"
Things to note:

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

Discontinuous constituents

- Discontinuous constituents (or equivalents) have been proposed in a wide range of syntactic frameworks
- 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):

  \[ \text{Ein Mann der lacht} \]

  labeled as discontinuous NP

Dependency grammar

- Dependency grammar is interested in grammatical relations between words of a sentence
- Dependency grammars do not propose a recursive structure but rather a network of relations
  - The verb is the part of the sentence on which everything ultimately depends
  - The direction of a link represents the dependency, the angle represents the word order
**Dependency grammar example**

`the little bald man likes the present`

```
likes
   
man
   
the little bald
   
the
```

**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 without constituency:
  - Coordination: what is the head of a coordinate phrase?
  - Verbal modification: no distinction between sentential & VP adjuncts

**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:**

```
I saw a man with a dog and a cat in the park
```

from (Lin 1995)

**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 analytic & tectogrammatical layers are dependency-based, but capture 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)
- 1976 corpus converted into one with both annotations:
  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, with grammatical functions as edge labels
English treebanks

- Penn Treebank
- Verbmobil Treebank

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

Treebank projects

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

Treebank 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)
Treebank projects (3)

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

References


