Treebanks

Yvonne Adesam

2013

Outline

What are treebanks?

Treebank examples

What are parallel treebanks?

Creating treebanks

Min bakgrund

- Disputerade 2012
    - Avhandling om att skapa högkvalitativa parallella trädbanker
    - Flerspråkiga parallella trädbanken Smultron
- Forskare på Språkbanken
    - Historiska resurser (MAÞiR 2014-2016)
    - Högkvalitativ korpusannotering (Koala 2014-2016)

What is a treebank?

A *corpus* is “a body of naturally-occurring (authentic) language data which can be used as a basis for linguistic research” (Leech and Eyes, 1997, p. 1).

A *treebank* is “a linguistically annotated corpus that includes some grammatical analysis beyond the part-of-speech level” (Nivre et al., 2005; Nivre, 2008).

What are treebanks?

Treebank examples

What are parallel treebanks?

Creating treebanks

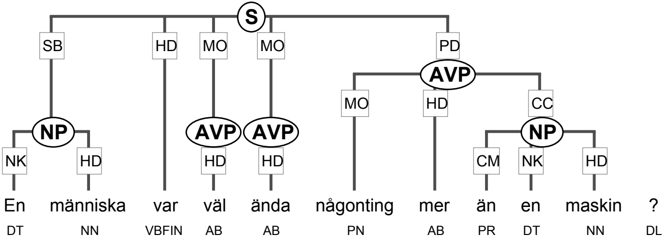
References

What is a Treebank? (2)

- A corpus with linguistic annotation beyond the word level
- The annotation is typically
  - a syntax tree and
  - manually checked and corrected.
- Treebank vs parsed corpus

What is a syntax tree?

Each sentence is mapped to a graph, which represents its syntactic structure.



Why Treebanking?

- Training material for Machine Learning → NLP systems
- Gold Standards for the evaluation of NLP systems
- Linguistic empiricism
- Human grammar exploration and learning

Creating treebanks is still an art, not a science.

The history of treebanks

- Penn Treebank (English; Phase 1: 1989-1992)
- Forerunners:
  - Talbanken (Swedish; Lund 1970s)
  - Ellegård (English; Gothenburg 1978)
  - Tosca (English; Nijmegen 1980s)
  - LOB (Lancaster-Oslo-Bergen) Treebank (Engl.; late 1980s)
  - SynTag (Swedish; Gothenburg 1986-1989)
- Followers
  - NEGRA / TIGER Treebanks (German; 1997-2000s)
  - Prague Dependency Treebank (Czech; 2000s)
  - Svensk trädbank (Swedish; 2007)
  - Bulgarian, Danish, Dutch, French, Chinese, Japanese, Arab, Hebrew, Turkish . . .

The Penn Treebank

- Treebank for English built at the University of Pennsylvania
- Phase 1 (1989-1992)
  - 3 million words (Brown Corpus and others)
  - bracket representation with PoS labels and node labels
- Phase 2 (1993-1995)
  - Enriching part of the original material with
    - syntactic functions
    - traces, null elements, coreference symbols
- Phase 3 (1996-2000)
  - additional material annotated
    - Wall Street Journal
    - Switchboard corpus (telephone conversations)

Penn treebank

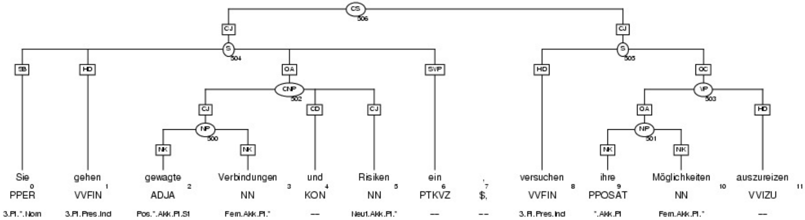
Penn Treebank Example from 1991

```
( bd0011sx .)
( (S (NP *)
     (VP Show
       (NP me)
       (NP (NP all)
           the nonstop flights
           (PP (PP from
                   (NP Dallas))
               (PP to
                   (NP Denver)))
           (ADJP early
                 (PP in
                     (NP the morning))))) .) )
```

The NEGRA Treebank

- 40’000 sentences
- from the Frankfurter Allgemeine Zeitung
- Annotations
  - PoS-Tags (STTS)
  - Morphological information
  - Syntactic nodes (NP, PP, VP, ...)
  - Syntactic functions (Subject, Object, Adverbial, etc)
  - allows crossing branches
  - allows secondary edges

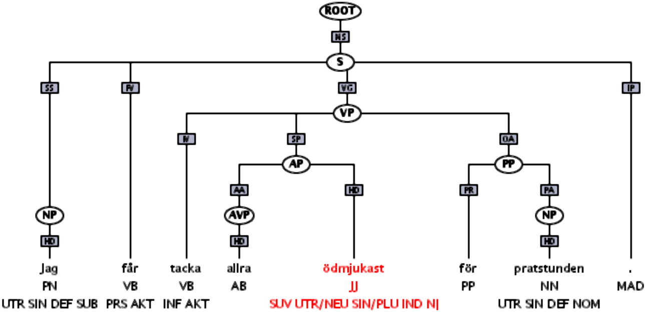
The NEGRA Treebank


CS
506
CJ
CJ
S
504
S
505
SB
HD
OA
SVP
HD
OC
CNP
502
VP
503
CJ
CD
CJ
OA
HD
NP
500
NP
501
NK
NK
NK
NK
Sie 0 PPER 3.Pl.*.Nom
gehen 1 VVFIN 3.Pl.Pres.Ind
gewagte 2 ADJA Pos.*.Akk.Pl.St
Verbindungen 3 NN Fem.Akk.Pl.*
und 4 KON —
Risiken 5 NN Neut.Akk.Pl.*
ein 6 PTKVZ —
, 7 $, --
versuchen 8 VVFIN 3.Pl.Pres.Ind
ihre 9 PPOSAT *.Akk.Pl
Möglichkeiten 10 NN Fem.Akk.Pl.*
auszureizen 11 VVIZU —

The Swedish Treebank

- Developed in Uppsala and Växjö
- Harmonizing two resources:
  - Talbanken: Swedish written and transcribed spoken language from the 1970s, manually annotated with syntactic information according to a traditional Scandinavian analysis tradition (cf. Diderichsen's field analysis)
  - SUC (Stockholm Umeå Corpus), a morphosyntactically annotated (part-of-speech and lemma), balanced corpus of published Swedish written language from the 1990s
- Talbanken annotated with SUC morphosyntactic in a semi-automatic process
- Both Talbanken and SUC automatically syntactically annotated with phrase structure version of Talbanken's original syntax analysis

The Swedish Treebank



Parallel Treebanks

- Corpus of translated (parallel) texts
- Manually or semi-automatically annotated
  - Each language syntactically annotated treebank
  - Alignment
- Useful for
  - word-sense disambiguation
  - bilingual dictionaries
  - machine translation
  - cross-language information retrieval
  - translation studies
  - foreign language pedagogy

Parallel Treebanks

A *parallel corpus* is a collection of naturally-occurring language data consisting of texts and their translations.

Parallel treebanks are treebanks over parallel corpora, i.e., the ‘same’ text in two or more languages.

Parallel Treebanks

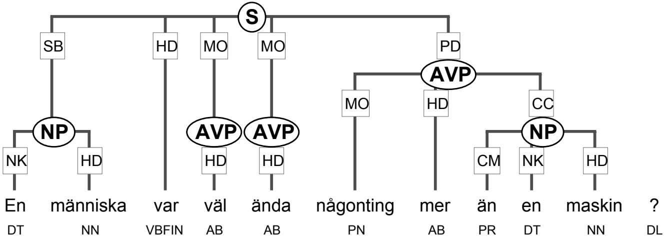
En människa var väl ända någonting mer än en maskin ?

Parallel Treebanks

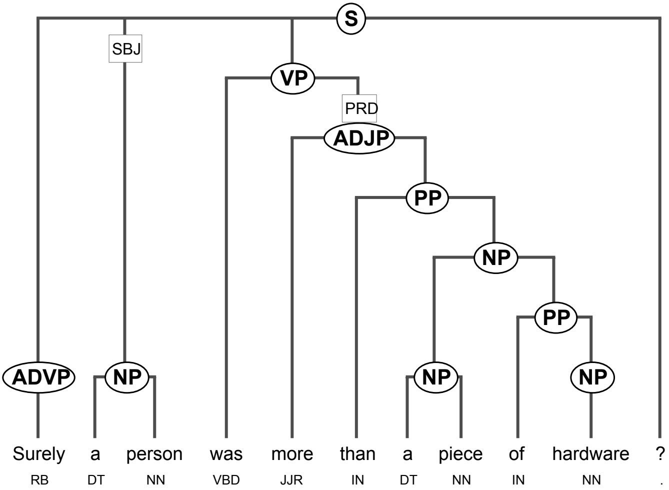
En	människa	var	väl	ända	någonting	mer	än	en	maskin	?
DT	NN	VBFIN	AB	AB	PN	AB	PR	DT	NN	DL

Parallel Treebanks

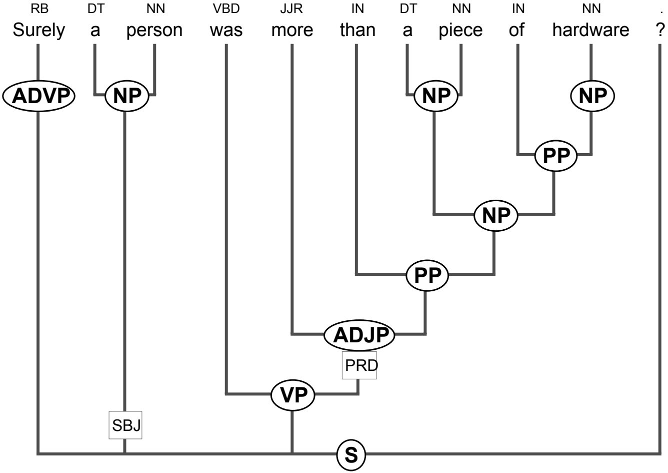
What are treebanks?
Treebank examples
What are parallel treebanks?
Creating treebanks
References



Parallel Treebanks



Parallel Treebanks



GÖTEBORGS
UNIVERSITET

Parallel Treebanks

Yvonne
Adesam

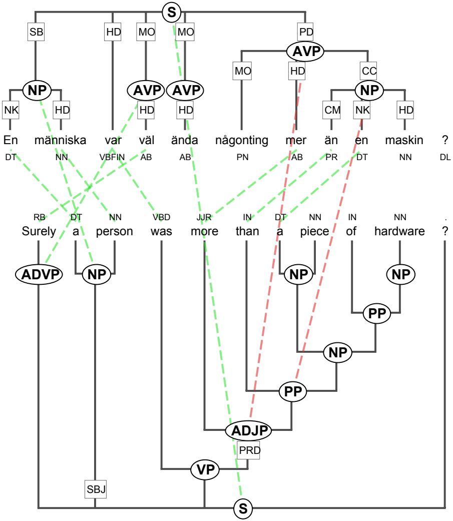
What are treebanks?

Treebank examples

What are parallel treebanks?

Creating treebanks

References



Språk
BANKEN

Smultron

Stockholm MULtilingual TReebank (v1.0)

English, German, Swedish

Over 1 000 sentences, around 18,000 tokens

- The novel Sophie's World (∼530 sentences)
  - The first two chapters
- Economy texts (∼500 sentences)
  - Annual report from a bank (SEB)
  - Quarterly report from a multinational company (ABB)
  - Banana Certification Program (Rainforest Alliance)
- v3.0: more texttypes, more languages, 2'500 sentences in 12 treebank files combined by 9 alignment files

Available from
www.cl.uzh.ch/research/parallelcorpora/paralleltreebanks/smultron_en.html

Treebanking – How To? I

1. Define the purpose
2. Select a corpus
   - written or spoken language?
   - one text genre or many?
   - copyright
3. Choose annotation format
   - what linguistic phenomena to represent
   - theoretical framework (e.g. constituents vs. dependencies)
   - depth of annotation
   - representation type
4. Choose annotation tool (tree editor)
5. Start the annotation (definition phase)
   - start annotation
   - write and revise annotation guidelines
6. Select and adapt support tools
   - PoS tagger

Treebanking – How To? II

- (shallow) parser

7. Annotate the data (production phase)
    - instruct annotators
    - annotation control by cross-checking
    - discussion of critical cases
8. Check the annotation and make corrections
    - completeness check
    - consistency check
9. Distribute the treebank

Treebank Annotation

High-quality corpus development

- Sentence splitting
- Tokenization
- Part-of-speech tagging
- Lemmas
- Morphological annotation
- Parsing/chunking
- Semantic annotation
- Named entity recognition
- Co-reference
- Alignment (sentence/phrase/word)
- (. . .)

Challenges in Treebank Annotation

- Ambiguity

Challenges in Treebank Annotation

- Morphological ambiguity (*lemurhelvete*: le, lem, mur, ur, hel, vete, ve, te)
- Syntactic ambiguity (*He saw a man with a telescope*; *I saw her duck*)
- Lexical ambiguity (*bank*; *saw*; *run*)
- Semantic ambiguity (*every boy loves his mother*; *A and B bought a house*)
- Discourse ambiguity (*A called B, she was sick*)

Challenges in Treebank Annotation

- Ambiguity
- Multiword units (including names)

Challenges in Treebank Annotation

- Ambiguity
- Multiword units (including names)
- Discontinuous units

Challenges in Treebank Annotation

- Ambiguity
- Multiword units (including names)
- Discontinuous units
- Foreign language expressions

Challenges in Treebank Annotation

- Ambiguity
- Multiword units (including names)
- Discontinuous units
- Foreign language expressions
- Symbols, numbers, and abbreviations

Challenges in Treebank Annotation

- Ambiguity
- Multiword units (including names)
- Discontinuous units
- Foreign language expressions
- Symbols, numbers, and abbreviations
- Meta-information (e.g. XML tags)

Challenges in Treebank Annotation

- Ambiguity
- Multiword units (including names)
- Discontinuous units
- Foreign language expressions
- Symbols, numbers, and abbreviations
- Meta-information (e.g. XML tags)
- Trade-off rich annotation vs (manual) labour

Current Treebank Annotation

- Sentence-based
- Mostly written language
- Syntax-centered
- Little work on semantic and discourse treebanks

Constituents vs Dependencies

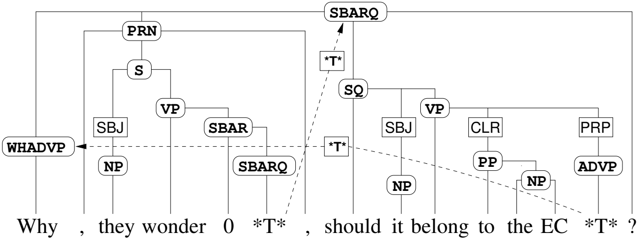


Figure 1: A constituent tree from the Penn Treebank.

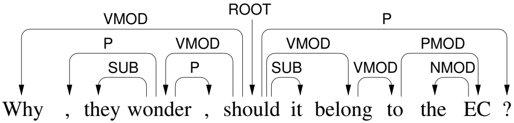


Figure 2: Dependency tree by PENN2MALT.

Constituents vs Dependencies

- Constituents
  - phrase structure
  - words building blocks of larger units
- Dependencies
  - syntactic dependencies
  - grammatical functions of words

Annotation

http://spraakbanken.gu.se/korp/annoteringslabb/

Treebank Quality

- Well-formedness

- Consistency

- Soundness

Treebank Quality

- Well-formedness
  Each token and each non-terminal node is part of a sentence-spanning tree, and has a label.
- Consistency
  The same sequence (of tokens/part-of-speechs/constituents) is annotated the same way given the same context.
- Soundness
  Conform to sound linguistic principles.

Quality Control

- Errors create problems for computational and theoretical linguistic uses of corpora
  - unreliable training and evaluation of NLP
  - detrimental to queries for rare linguistic phenomena
  - error propagation through layers
- Both automatic and human annotation contains errors
- Good guidelines, well-trained annotators, easy-to-use annotation tools, search tools etc
- Inter-annotator agreement should be monitored throughout the project
- Detecting annotation errors using NLP tools
- Feedback from the user

Corpus Development

- Treebanking is
  - time-consuming
  - labour-intensive
- Most applications require large amounts of data

Corpus Development

- Treebanking is
  - time-consuming
  - labour-intensive
- Most applications require large amounts of data
- Use automatic annotation methods to reduce manual work
  - **enlarge** annotated data
  - **guide** quality checking
  - **improve** annotation before quality checking

Why manual work?

Accuracy of most annotation tools depend on

- set of labels
- training data
- language

Part-of-speech tagging: accuracy normally above 95-96%. Example: HunPoS 97% accuracy when trained on SUC (Megyesi, 2009) An error in every second sentence!

Parsing: accuracy varies considerably across languages Example: CoNLL shared task 2007: LAS 84-90: Catalan, Chinese, English, Italian LAS 76-80: Arabic, Basque, Czech, Greek, Hungarian, Turkish

What are treebanks?

Treebank examples

What are parallel treebanks?

Creating treebanks

References

Summary

- A treebank is a corpus with grammatical analysis beyond the word level
- Often large treebanks are needed
- Use automatic tools as much as possible
- Add manual work and manual quality checks
- Important: text type, language, annotation type

- Next time: Historical corpora

References I

Leech, G. and Eyes, E. (1997). Syntactic annotation: treebanks. In Garside, R., Leech, G., and McEnery, T., editors, *Corpus Annotation - Linguistic Information from Computer Text Corpora*, chapter 3, pages 34–52. Addison Wesley Longman Limited.

Megyesi, B. (2009). The open source tagger HunPoS for Swedish. In Jokinen, K. and Bick, E., editors, *Proceedings of the Nordic Conference on Computational Linguistics (Nodalida)*, volume 4 of *NEALT Proceedings Series*, pages 239–241, Odense, Denmark.

Nivre, J. (2008). Treebanks (Article 13). In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter.

Nivre, J., de Smedt, K., and Volk, M. (2005). Treebanking in Northern Europe: A white paper. In Holmboe, H., editor, *Nordisk Sprogteknologi. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*. Museum Tusculanums Forlag, Copenhagen.