NPFL070
Language Data Resources

Treebank examples
A treebank is a …

- database of syntactic trees
- corpus annotated with morphological and syntactic information
- segmented, part-of-speech tagged, and fully bracketed corpus
- (typically hand-built) collection of natural language utterances and associated linguistic analyses
- collection of morphologically, syntactically and semantically annotated sentences
- database essential for the study of the language due to it provides analyzed/annotated examples of real language
- large collection of sentences which have been parsed by hand
Main Entry: \textbf{parse}

Pronunciation: 'pärz, chiefly British 'pärz
Function: verb
Inflected Form(s): parsed, parsing
Etymology: Latin pars orationis part of speech
transitive senses
1 a : to resolve (as a sentence) into component parts of speech and describe them grammatically b : to describe grammatically by stating the part of speech and explaining the inflection and syntactical relationships
2 : to examine in a minute way : analyze critically <parses appellate court opinions> intransitive senses
1 : to give a grammatical description of a word or a group of words
2 : to admit of being parsed
Penn Treebank

- [http://www.cis.upenn.edu/~treebank/home.html](http://www.cis.upenn.edu/~treebank/home.html)
- 1992 - Release 0.5
- 1.5 MW
- English newspaper texts; the largest subpart taken from the Wall Street Journal
- Bracketing style (constituent syntax)
- Structural reconstructions (traces)
Analysts have been expecting a GM-Jaguar pact that would give the U.S. car maker an eventual 30% stake in the British company.
PropBank

- [http://www.cis.upenn.edu/~ace/](http://www.cis.upenn.edu/~ace/)
- „adding a layer of semantic annotation to the Penn Treebank“
- Rozlišení arguments-adjuncts
- Ukázka anotace: *Mr. Bush met him privately, in the White House, on Thursday.*

Rel: met
Arg0: Mr. Bush
Arg1: him
ArgM-MNR: privately
ArgM-LOC: in the White House
ArgM-TMP: on Thursday
The same sentence, PropBanked

**have been expecting**

**Arg0**

Analysts

**a GM-Jaguar pact**

**Arg1**

(S Arg0 (NP-SBJ Analysts)

(VP have

(VP been

(VP expecting

Arg1 (NP (NP a GM-Jaguar pact)

(SBAR (WHNP-1 that)

(S Arg0 (NP-SBJ *T*-1)

(VP would

(VP give

Arg2 (NP the U.S. car maker)

Arg1 (NP (NP an eventual (ADJP 30 %) stakes)

(PP-LOC in (NP the British company)))

that would give**

**Arg0**

*T*-1

**Arg2**

the US car maker

**Arg1**

an eventual 30% stake in the British company

expect(Analysts, GM-J pact)
give(GM-J pact, US car maker, 30% stake)
NEGRA

- [http://www.coli.uni-sb.de/sfb378/negra-corpus/](http://www.coli.uni-sb.de/sfb378/negra-corpus/)

- NEGRA corpus version 2 consists of 355,096 tokens (20,602 sentences) of German newspaper text, taken from the Frankfurter Rundschau
NEGRA – příklady stromů
Japan wird offenbar die Fusion der Geldkonzerne Daiwa und Sumitomo zur größten Bank der Welt vorbereitet.

$. 0
TIGER Treebank

Abbildung 1: Annotation von Extraposition durch kreuzende Kanten
Abbildung 2: Annotation von Koordination durch sekundäre Kanten
Imageschäden und geschäftlichen Missbrauch ausländischer Unternehmen, die lange und illegalen Transaktionen vorhaben, wird nach vermutlichen Vorfällen die Behörde VAFI N erstmalig international in einem Konsens einer internationalen Behördenvereinigung verhandelt. Vor möglichen Untersuchungen droht dem Daiwa Corporation eine hohe Strafe.
BulTreeBank (I)

• HPSG-based treebank of Bulgarian at Bulgarian Academy of Sciences
• detailed described in a sequence of technical reports in 2004 (www.bultreebank.org)
• source of material - BulTreeBank corpus
  – archive of texts converted from HTML and RTF documents: 90 MW
  – morphologically analyzed corpus: 10 MW
  – disambiguation by hand: 1 MW
• the treebank itself: 200 kW
  – 1.500 kS extracted from Bulgarian grammars to show the variety of syntactic structures
  – 10 kS from the corpus (newspapers, government documents, prose) to show their distribution
BulTreeBank (II)
Morphosyntactic Tagset

• **BTB-TS**
  – derived from MULTEXT
  – positional ("positional") - the first letter (POS) specifies how many positions follow and what categories they express
  – examples:
    • Ncmsd (noun common masc. sing. def.)
    • Amsi (adjective masc. sing. indef.)
She is angry whatever you tell her.
BulTreeBank (IV) Discontinuities
Szeged Treebank (I)

- syntactic structures for Hungarian developed at University of Szeged
- 82 kS, 1.2 MW(+200000 punctuation marks)
- 5 types of text material: fiction, compositions of 14-16-year-old students, newspaper articles, IT texts, law
- manual morphological disambiguation and syntactic annotation
Szeged Treebank (II)

- following generative syntax
- freer word-order in Hungarian -> many traces
張三勸李四參加會見 (ZhangSan persuaded LiSi to attend the meeting)
Conversion of phrase structures into dependencies

• simple recursive procedure:
  
  1. Mark the head child of each non-terminal node
  2. In the dependency structure, make the head of each non-head child depend on the head of the head child