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

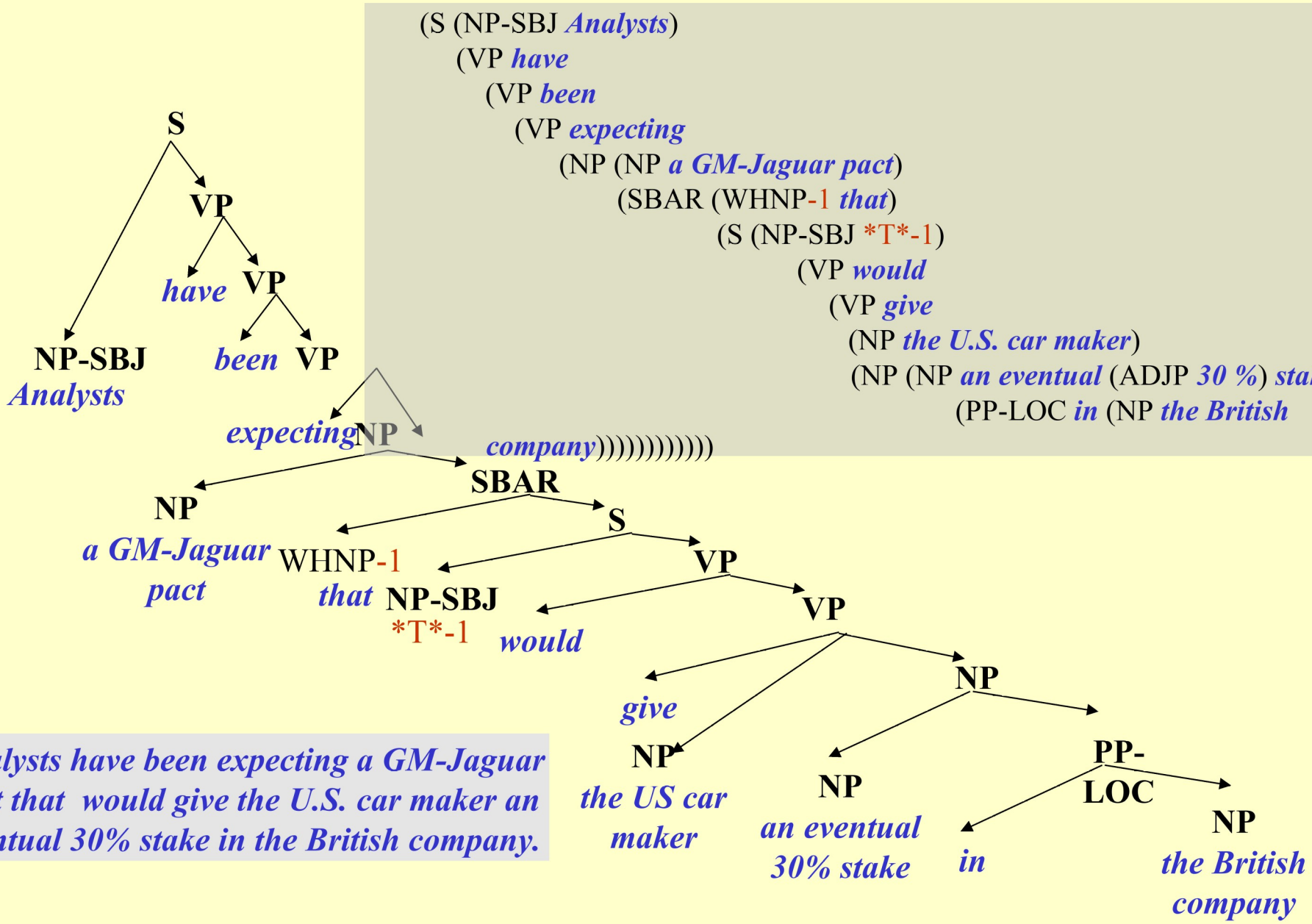
Parsed?

Main Entry: [1]**parse**
Pronunciation: *'pärs, chiefly British 'pärz*
Function: *verb*
Inflected Form(s): **parsed**; **pars·ing**
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
- 1992 - Release 0.5
- 1.5 MW
- English newspaper texts; the larges subpart taken from the Wall Street Journal
- Bracketing style (constituent syntax)
- Structural reconstructions (traces)

A PennTreeBanked Sentence

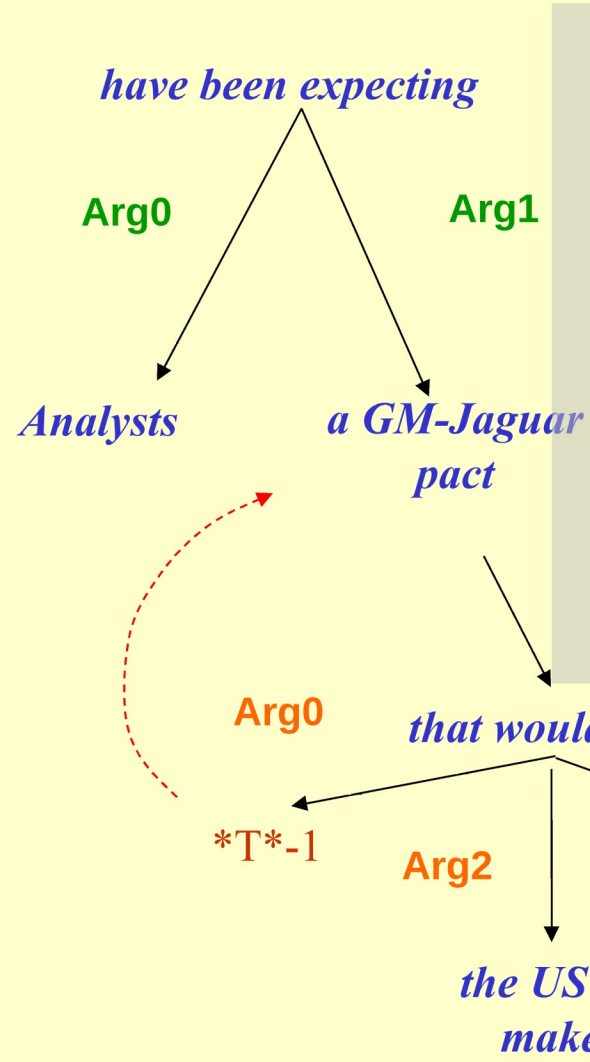


PropBank

- 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



(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 %*) *s*
(PP-LOC *in* (NP *the British company*)))))))))))))

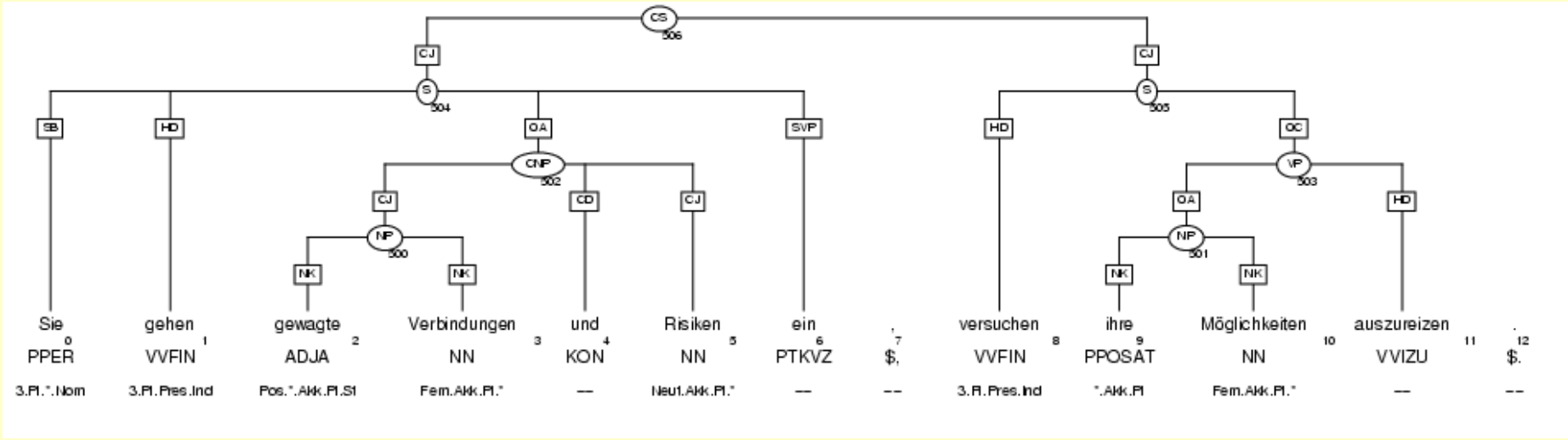
expect(Analysts, GM-J pact)
give(GM-J pact, US car maker, 30% stake)

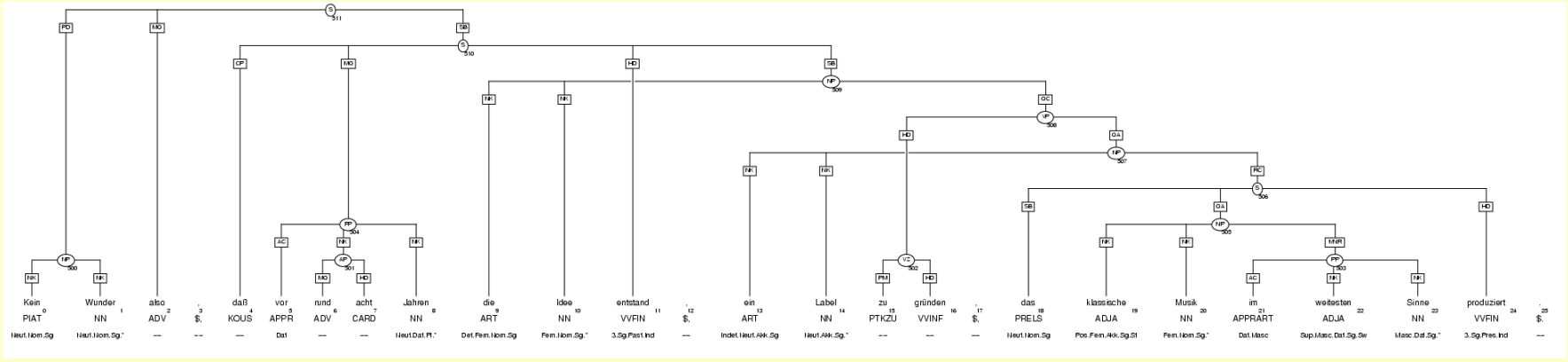
NEGRA

- http://www.coli.uni-sb.de/sfb378/negra-corpus/

- NEGRA corpus version 2 consists of 355,096 tokens (20,602 sentences) of German newspaper text, taken from the Frankfurter Rundschau

NEGRA – příklady stromů





NEGRA Export Format

```
#BOS 4228 101 991063314 18 %% @ST2AV@(source: t_v_janbettina 64)
In                  APPR    --              AC      500
Japan               NE      --              NK      500
wird                VAFIN   --              HD      507
offenbar            ADJD    --              MO      504
die                 ART     --              NK      506
Fusion              NN      --              NK      506
der                 ART     --              NK      503
Geldkonzerne        NN      --              NK      503
Daiwa               NE      --              CJ      501
und                 KON     --              CD      501
Sumitomo            NE      --              CJ      501
zur                 APPRART --              AC      505
größten             ADJA    --              NK      505
Bank                NN      --              NK      505
der                 ART     --              NK      502
Welt                NN      --              NK      502
vorbereitet         VVPP    --              HD      504
.                   $.      --              --      0
#500                PP      --              MO      504
#501                CNP     --              NK      503
#502                NP      --              AG      505
#503                NP      --              AG      506
#504                VP      --              OC      507
#505                PP      --              MNR     506
#506                NP      --              SB      507
#507                S       --              --      0
#EOS 4228
```

TIGER Treebank

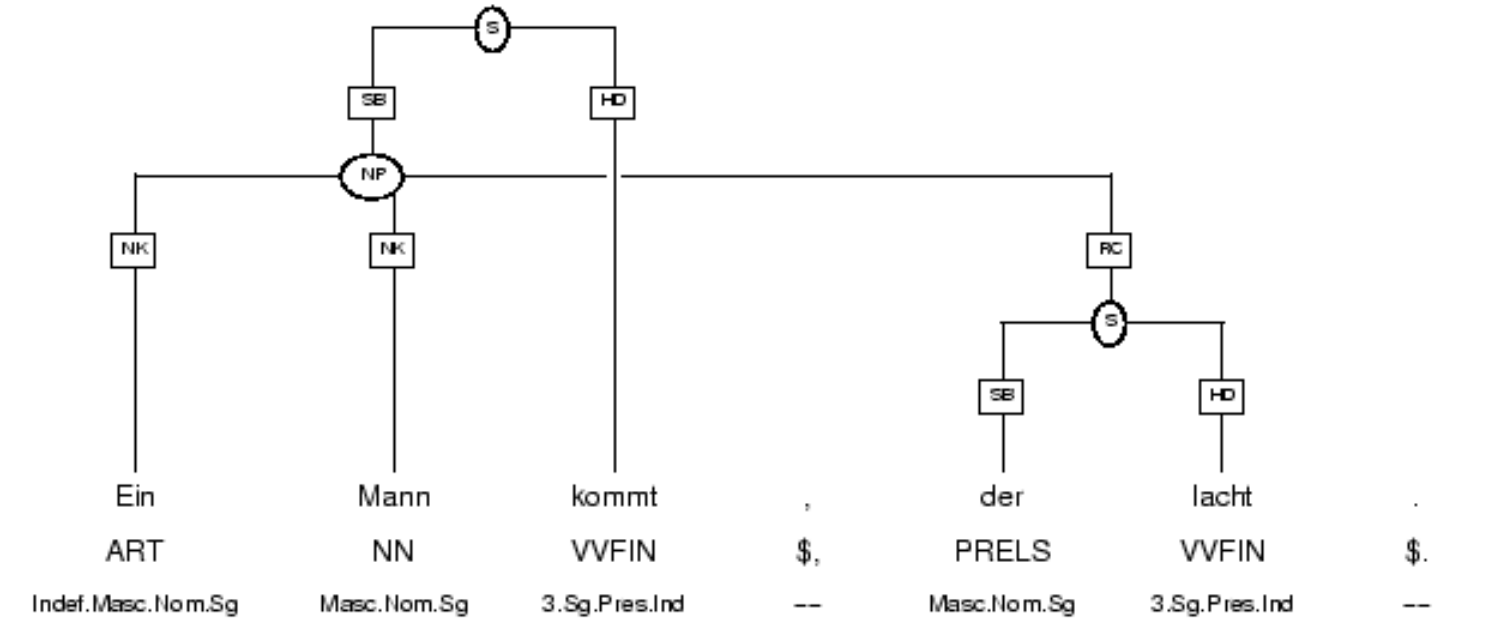


Abbildung 1: Annotation von Extraposition durch kreuzende Kanten

TIGER

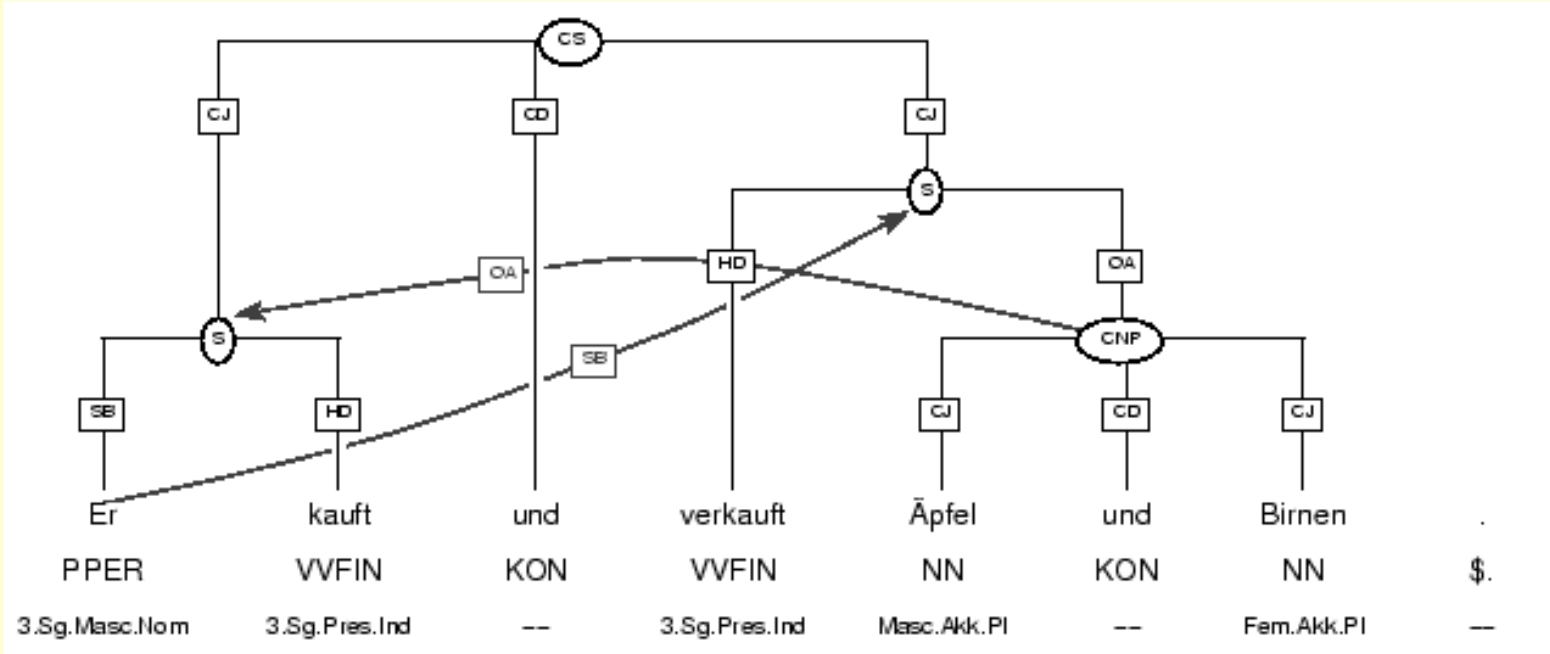


Abbildung 2: Annotation von Koordination durch sekundäre Kanten

TIGER-XML format

```
<s id="s4229">
 <graph root="s4229_509">
  <terminals>
   <t id="s4229_1" word="Hintergrund" pos="NN" />
   <t id="s4229_2" word="sind" pos="VAFIN" />
   <t id="s4229_3" word="die" pos="ART" />
   <t id="s4229_4" word="geschäftlichen" pos="ADJA" >
   <t id="s4229_5" word="Einschränkungen" pos="NN" />
   <t id="s4229_6" word="und" pos="KON" />
   <t id="s4229_7" word="Imageschäden" pos="NN" />
   <t id="s4229_8" word="," pos="$," />
   <t id="s4229_9" word="die" pos="PRELS" />
   <t id="s4229_10" word="Daiwa" pos="NE" />
   <t id="s4229_11" word="nach" pos="APPR" />
   <t id="s4229_12" word="mutmaßlich" pos="ADJD" />
   <t id="s4229_13" word="illegalen" pos="ADJA" />
   <t id="s4229_14" word="und" pos="KON" />
   <t id="s4229_15" word="zudem" pos="PROAV" />
   <t id="s4229_16" word="lange" pos="ADV" />
   <t id="s4229_17" word="vor" pos="APPR" />
   <t id="s4229_18" word="den" pos="ART" />
   <t id="s4229_19" word="Behörden" pos="NN" />
   <t id="s4229_20" word="vertuschten" pos="ADJA" />
   <t id="s4229_21" word="US-Transaktionen" pos="NN" />
   <t id="s4229_22" word="international" pos="ADJD" />
   <t id="s4229_23" word="drohen" pos="VVFIN" />
   <t id="s4229_24" word="." pos="$." />
  </terminals>
  <nonterminals>
   <nt id="s4229_500" cat="NP">
    <edge label="NK" idref="s4229_4" />
    <edge label="NK" idref="s4229_5" />
   </nt>
   <nt id="s4229_501" cat="AP">
    <edge label="MO" idref="s4229_12" />
    <edge label="HD" idref="s4229_13" />
   </nt>
   <nt id="s4229_502" cat="PP">
    <edge label="AC" idref="s4229_17" />
    <edge label="NK" idref="s4229_18" />
    <edge label="NK" idref="s4229_19" />
   </nt>
   <nt id="s4229_503" cat="CNP">
    <edge label="CD" idref="s4229_6" />
    <edge label="CJ" idref="s4229_7" />
    <edge label="CJ" idref="s4229_500" />
   </nt>
   <nt id="s4229_504" cat="AP">
    <edge label="MO" idref="s4229_15" />
    <edge label="MO" idref="s4229_16" />
    <edge label="HD" idref="s4229_20" />
    <edge label="MO" idref="s4229_502" />
   </nt>
   <nt id="s4229_505" cat="CAP">
    <edge label="CD" idref="s4229_14" />
    <edge label="CJ" idref="s4229_501" />
    <edge label="CJ" idref="s4229_504" />
   </nt>
   <nt id="s4229_506" cat="PP">
    <edge label="AC" idref="s4229_11" />
    <edge label="NK" idref="s4229_21" />
    <edge label="NK" idref="s4229_505" />
   </nt>
   <nt id="s4229_507" cat="S">
    <edge label="SB" idref="s4229_9" />
    <edge label="DA" idref="s4229_10" />
    <edge label="MO" idref="s4229_22" />
    <edge label="HD" idref="s4229_23" />
    <edge label="MO" idref="s4229_506" />
   </nt>
   <nt id="s4229_508" cat="NP">
    <edge label="NK" idref="s4229_3" />
    <edge label="NK" idref="s4229_503" />
    <edge label="RC" idref="s4229_507" />
   </nt>
   <nt id="s4229_509" cat="S">
    <edge label="SB" idref="s4229_1" />
    <edge label="HD" idref="s4229_2" />
    <edge label="PD" idref="s4229_508" />
   </nt>
  </nonterminals>
 </graph>
</s>
```

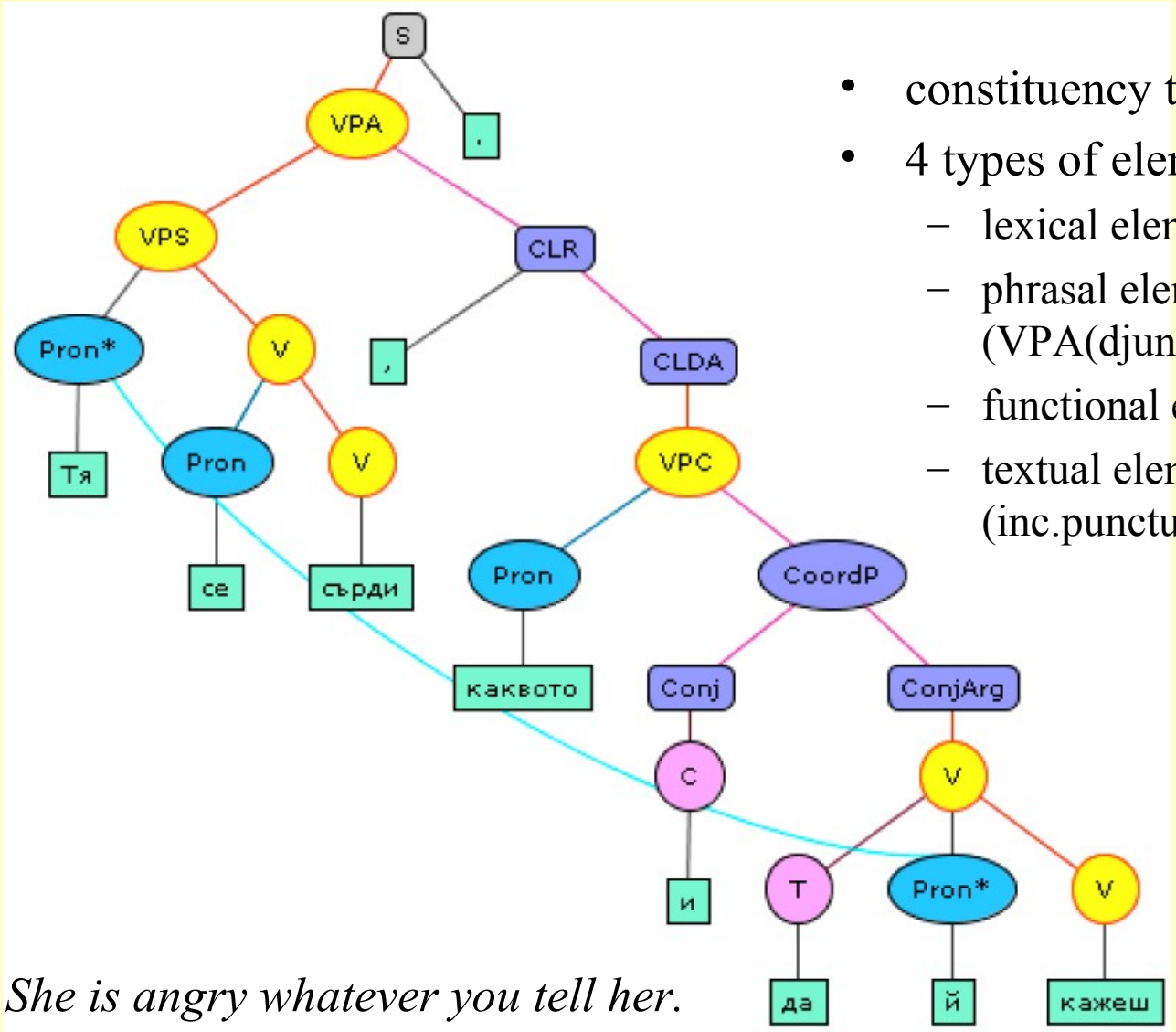
BulTreeBank (I)

- 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.)

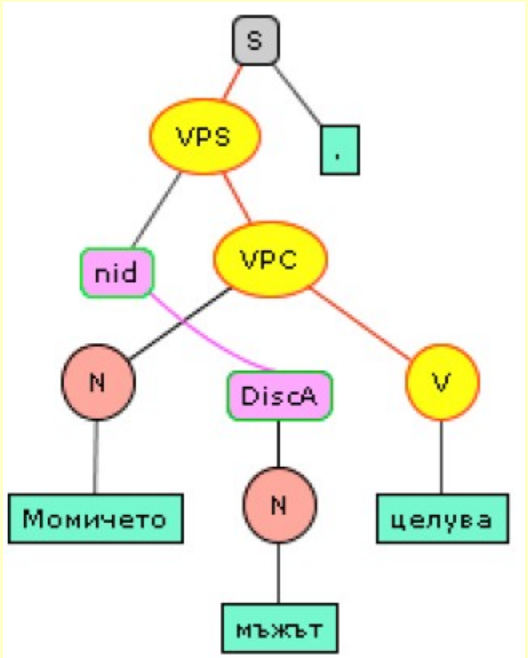
BulTreeBank (III)
Syntactic Structures

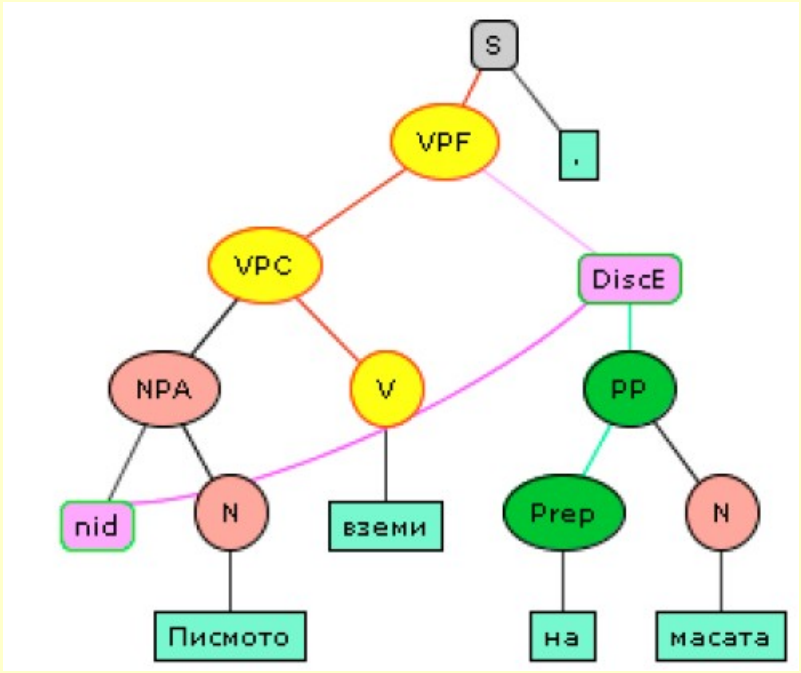


She is angry whatever you tell her.

- constituency trees
- 4 types of elements
  - lexical elements (N,V,Prep,...)
  - phrasal elements (VPA(djunct),NPC(omplement)...)
  - functional elements (Conj,ConjArg,...)
  - textual elements - the actual strings (inc.punctuation)

BulTreeBank (IV) Discontinuities



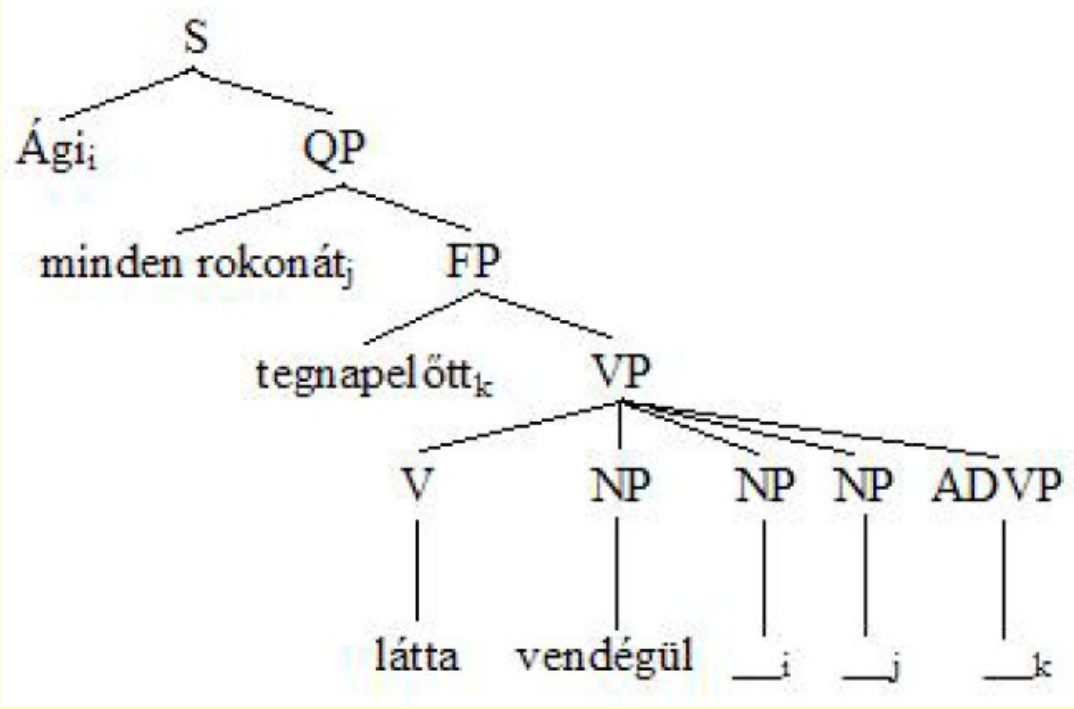


Szeged Treebank (I)

- 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)

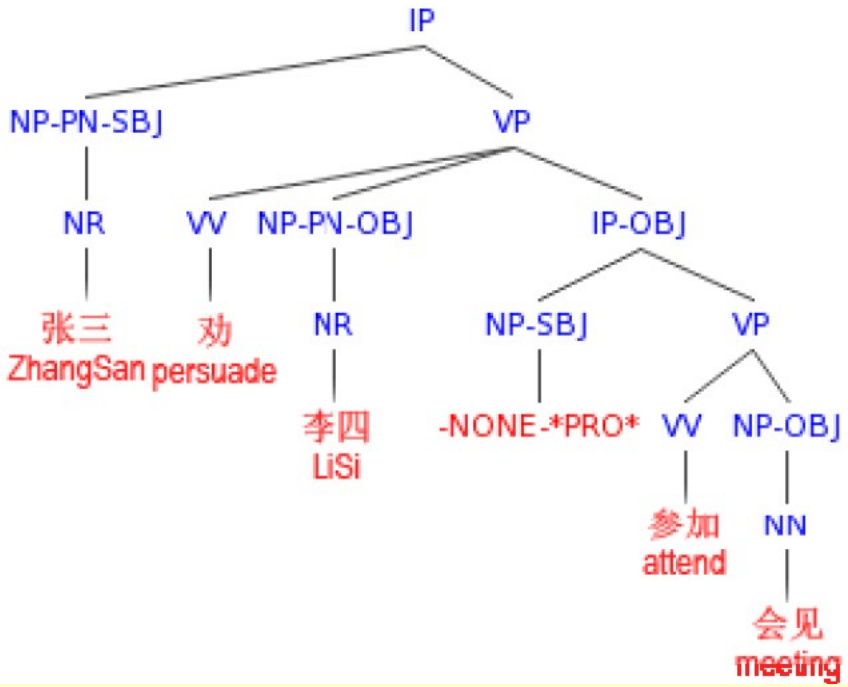
- following generative syntax
- freer word-order in Hungarian -> many traces



Penn Chinese Treebank

张三劝李四参加会见 (ZhangSan persuded LiSi to attend the meeting)

```
(IP (NP-PN-SBJ (NR 张三))
    (VP (VV 劝)
        (NP-PN-OBJ (NR 李四))
        (IP-OBJ (NP-SBJ (-NONE- *PRO*))
                (VP (VV 参加)
                    (NP-OBJ (NN 会见))))))
```



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