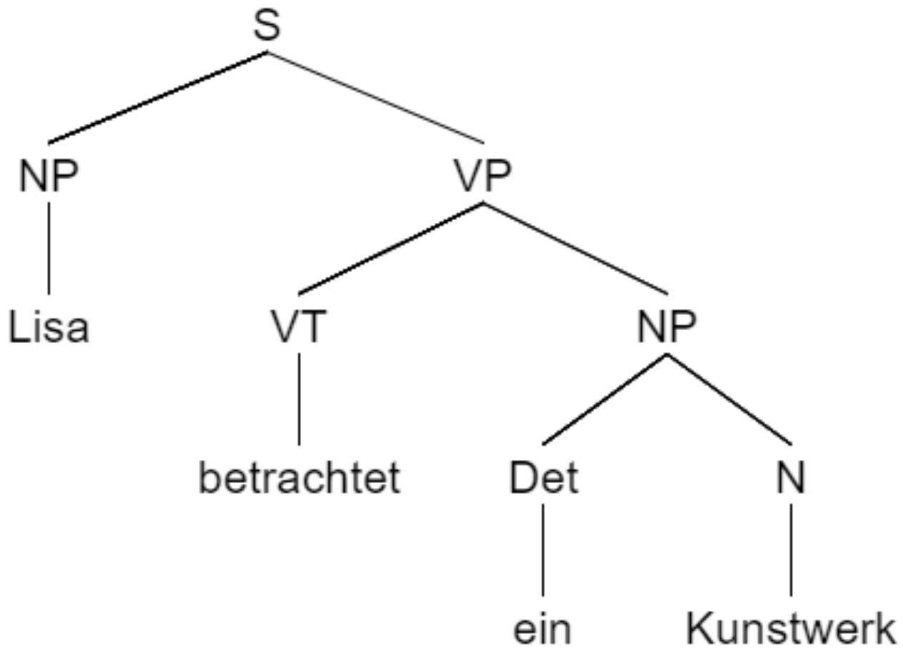
Phrasenstrukturparsing

PROSEMINAR 'TOOLS'

WINTERSEMESTER 2017/18

HANNA KEBER

Was ist das?



Inhalt

- Probleme
- Akkuratheit
- Effizienz
- Übungen

Probleme

- PCFGs stellen keine strukturellen Abhängigkeiten dar
  - ist eine NP Subjekt oder Objekt im aktuellen Kontext?
- Schnelligkeit

Akkuratheit

- Kann durch verschiedene Methoden erhöht werden:
  - Parent Annotation
  - Tag Splitting
  - Split-Merge-Zyklus

Parent Annotation (Stanford)

	Pronoun	Non-Pronoun
Subject	91%	9%
Object	34%	66%

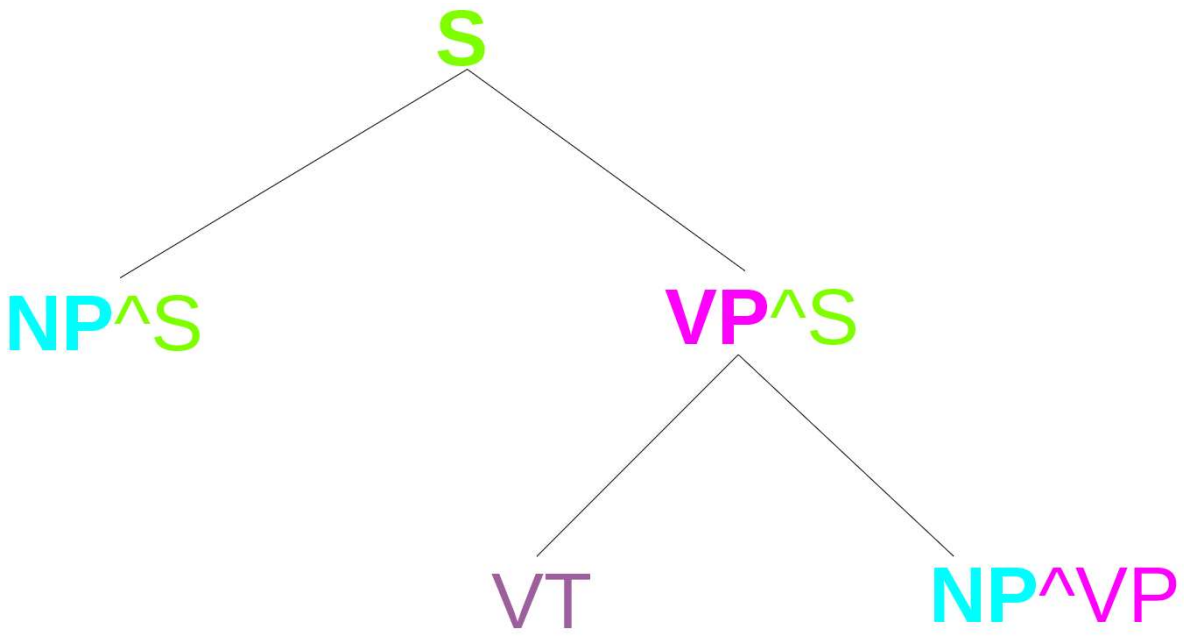
- Wahrscheinlichkeiten der Regeln

NP → Pron [?]

NP → Det N [?]

Abb. aus Jurafsky& Martin

Parent Annotation (Stanford)



NP^S → Pron [0,91]

NP^VP → Det N [0.66]

Tag Splitting (Stanford)

- Parent Annotation von Preterminals

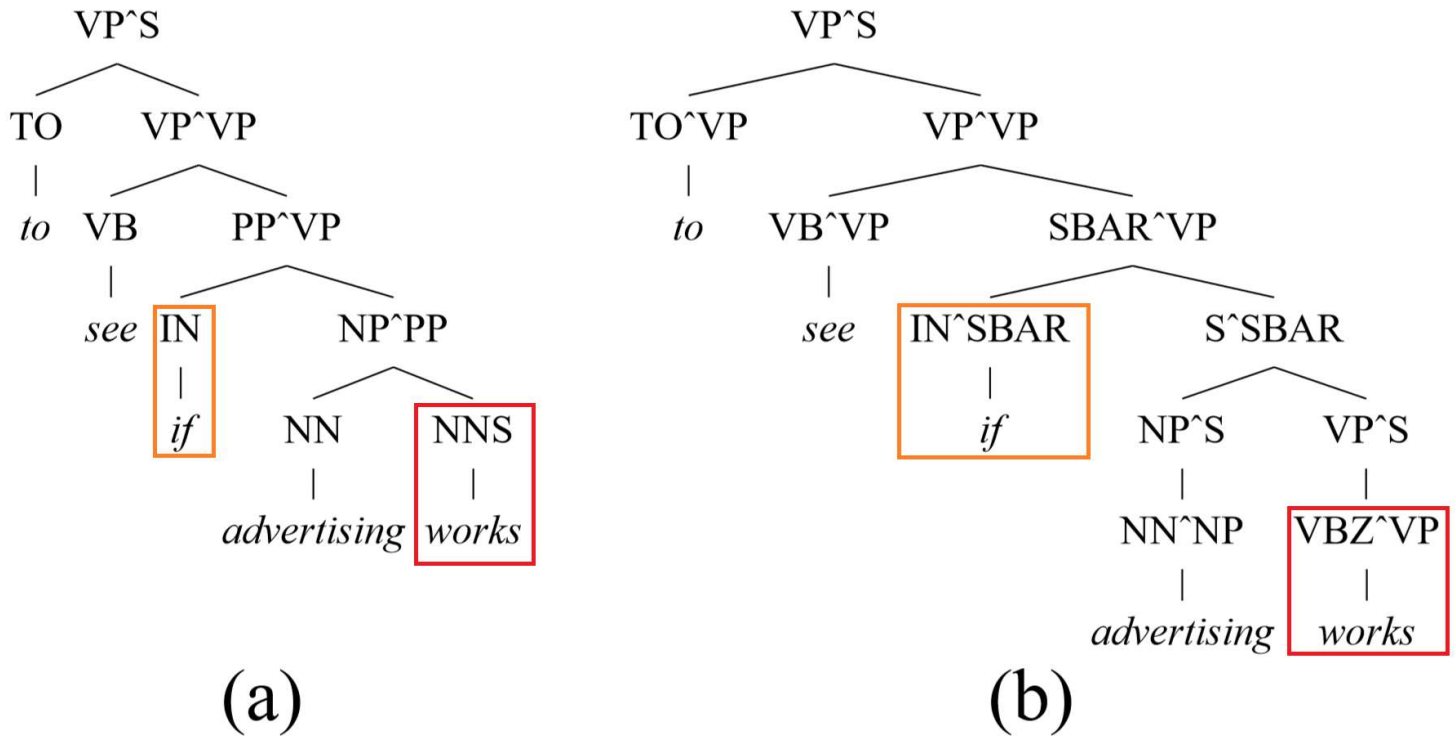
- Beobachtung: bestimmte Wörter treten häufig in gleichen Phrasen auf

  Adverbien in ADVPs: also, now

  Adverbien in NPs: only, just

  ...

Tag Splitting (Stanford)



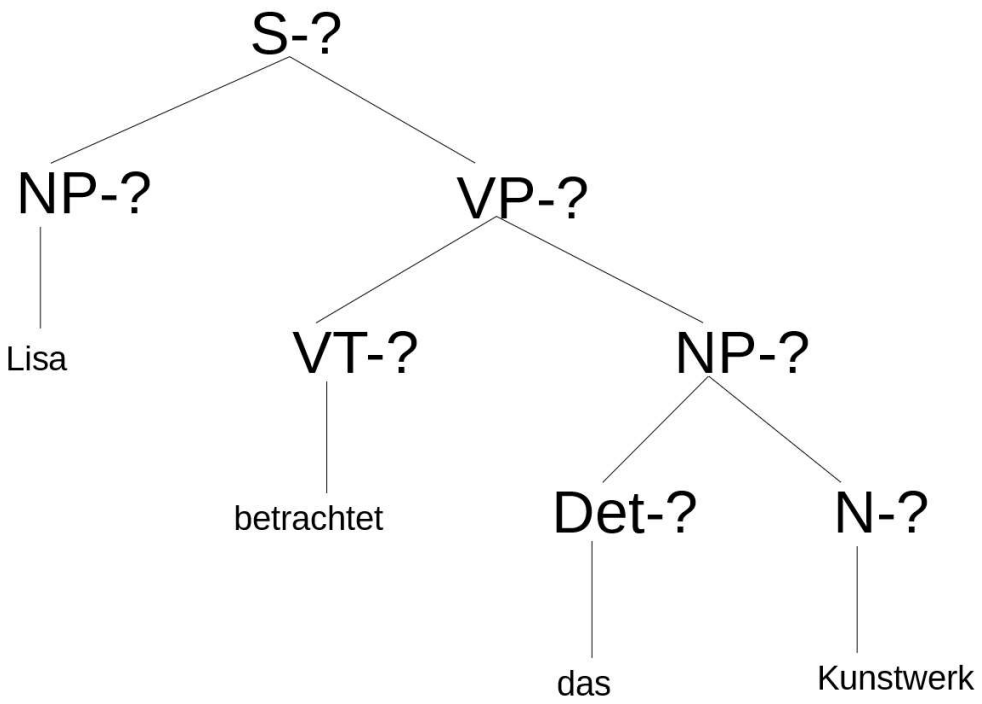
- Bsp: PA und SPLIT-IN Annotation

Abb. aus Klein & Manning

Berkeley und Stanford Parser

- Stanford: Regeln manuell erstellt
- Berkeley: automatisches Erlernen der Regeln

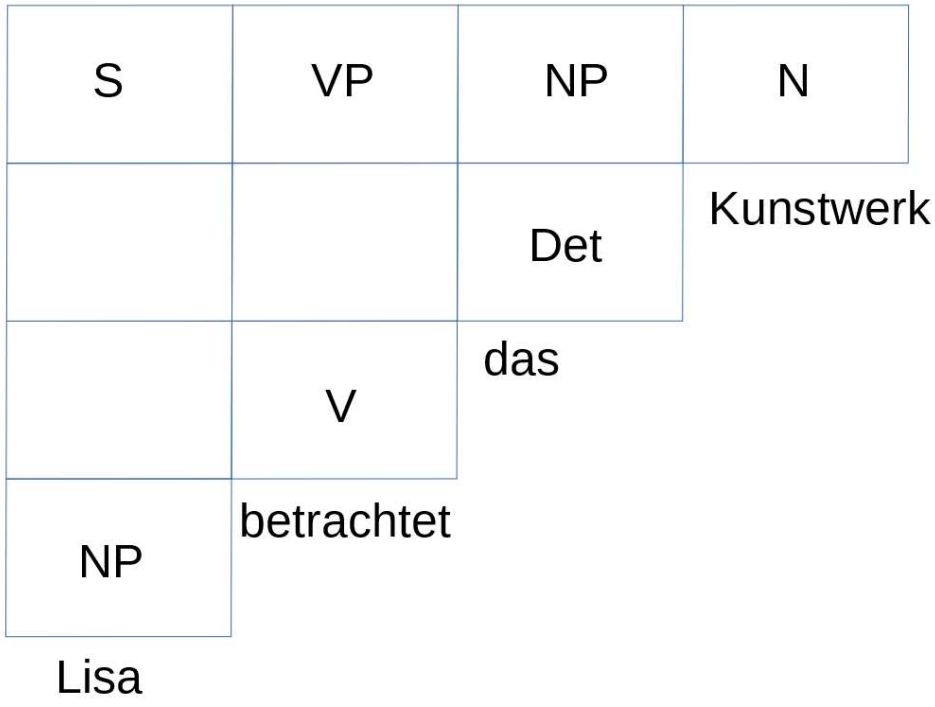
Split – Merge – Zyklus (Berkeley)



- Splits: ? = NP-1 oder NP-2 usw.
- Grammatik wird vergrößert:
  S1 → NP1 VP1
  S1 → NP2 VP1
  S1 → NP2 VP2
  …
- Mit EM Gewichtung lernen
- Mergen reduziert G um unnötige Splits

Effizienz

- Größere Grammatik → höherer Zeitaufwand → größere CYK-Chart



Effizienz

- Pruning verkleinert Chart:
  - Weniger Einträge pro Zelle erlauben: keine Nonterminals mit geringer Wahrscheinlichkeit
  - Abarbeitungsreihenfolge der Agenda beeinflussen

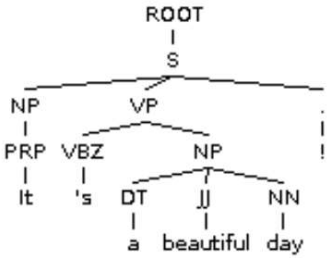
- Optimierung möglich:
  - Stanford: A*-Algorithus
  - Berkeley: coarse-to-fine

Übung 1

- Online Demo des Berkeley Parsers:

http://tomato.banatao.berkeley.edu:8080/parser/

```
(ROOT
  (S
    (NP (PRP It))
    (VP (VBZ 's)
      (NP (DT a) (JJ beautiful) (NN day)))
    (. !)))
```



It's a beautiful day!

Parse!

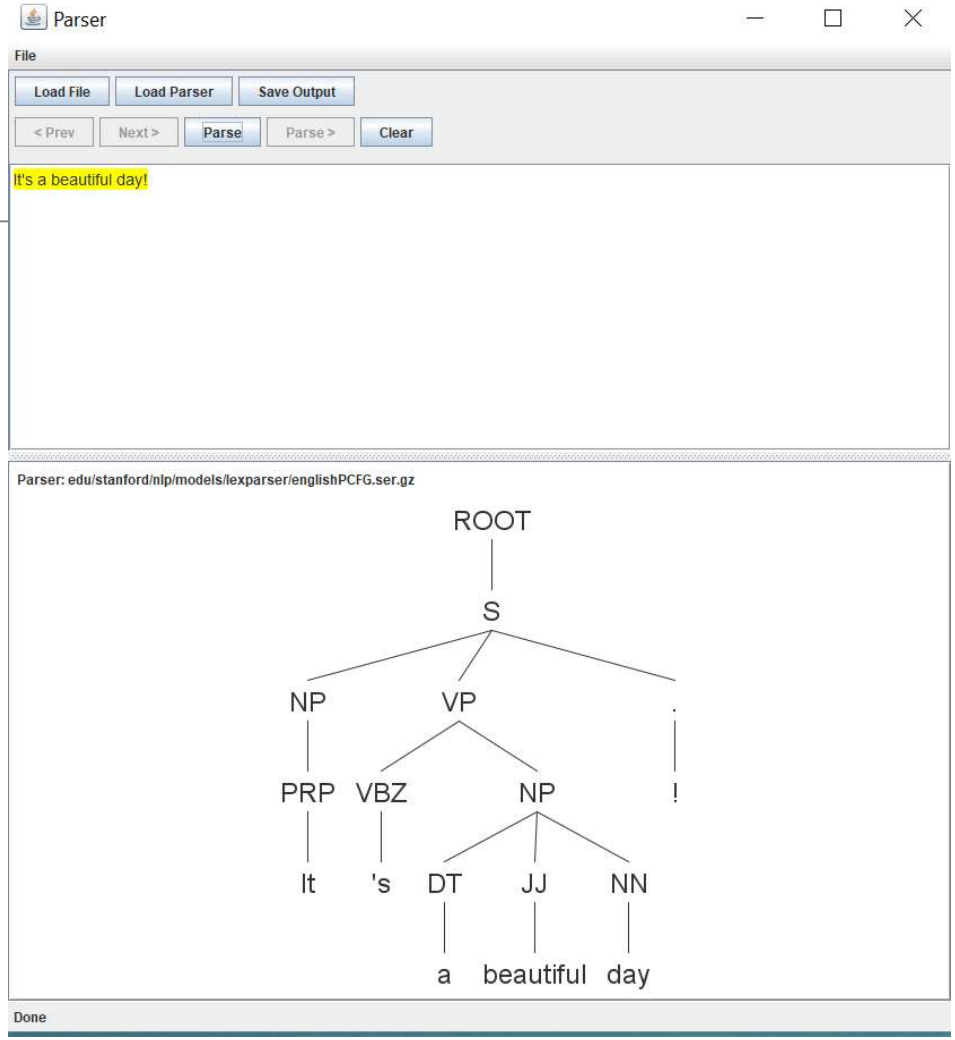
The parser is online.

Übung 1

- Stanford GUI

Für Windows:
stanford-parser.jar

Für Linux/ Unix:
Lexparser-gui.sh



Übung 1

- Parst mit beiden den Satz

  "Look at the dog with one eye"

- Was fällt auf?

Übung 2

- Berkeley Parser auf größeren Texten anwenden

```
java -mx 600M -jar BerkeleyParser-1.7.jar -gr <grammar file> -
inputFile <xy>
```

```
C:\Users\hanns\Documents\Berkeley Parser>java -mx600M -jar BerkeleyParser-1.7.jar -gr eng_sm6.gr -inputFile bsp_Text.txt

( (S (NP (NNP Finland)) (VP (VBD was) (VP (VBN inhabited) (SBAR (WHADVP (WRB when)) (S (NP (DT the) (JJ last) (NN ice) (
NN age)) (VP (VBP ended,) (NP (QP (RB approximately) (CD 9000)) (NNS BCE.[10])) (NP (NP (DT The) (JJ first) (NNS settler
s)) (VP (VBN left) (PP (IN behind) (NP (NP (NNS artifacts)) (SBAR (WHNP (WDT that)) (S (VP (VBP present) (NP (NP (NNS ch
aracteristics)) (VP (VBN shared) (PP (IN with) (NP (NP (DT those)) (VP (VBN found) (PP (IN in) (NP (NP (NNP Estonia,) (N
NP Russia,)) (CC and) (NP (NNP Norway.[11])))))))))))))))) (SBAR (SBAR (S (S (NP (DT The) (JJS earliest) (NNS people)) (V
P (VBD were) (ADJP (JJ hunter-gatherers,)) (S (S (VP (VBG using) (NP (NN stone) (NN tools.[12])) (S (NP (DT The) (JJ fir
st) (NN pottery)) (VP (VBD appeared) (PP (IN in) (NP (NP (CD 5200) (NNP BCE,)) (SBAR (WHADVP (WRB when)) (S (NP (DT the)
 (NNP Comb) (NNP Ceramic) (NN culture)) (VP (VBD was) (ADJP (JJ introduced.[13])))))))) (S (NP (NP (DT The) (NN arrival))
 (PP (IN of) (NP (DT the) (NNP Corded) (NNP Ware) (NN culture))) (PP (IN in) (NP (JJ southern) (JJ coastal) (NNP Finland
))) (NP (QP (IN between) (CD 3000) (CC and) (CD 2500)) (NNP BCE))) (VP (MD may) (VP (VB have) (VP (VBD coincided) (PP (I
N with) (NP (NP (NP (DT the) (NN start)) (PP (IN of) (NP (NN agriculture.[14])))) (NP (DT The) (NNP Bronze) (NNP Age) (C
C and) (NNP Iron) (NNP Age))))))))))) (VP (VBD were) (VP (VBN characterised) (PP (IN by) (NP (NP (JJ extensive) (NNS co
ntacts)) (PP (IN with) (NP (NP (JJ other) (NNS cultures)) (PP (IN in) (NP (DT the) (NNP Fennoscandian) (CC and) (JJ Balt
ic) (NNS regions))))))))))) (CC and) (S (NP (DT the) (JJ sedentary) (NN farming) (NN inhabitation)) (VP (VBN increased)
 (PP (IN towards) (NP (NP (DT the) (NN end)) (PP (IN of) (NP (NNP Iron) (NNP Age.))))) (PP (IN At) (NP (NP (DT the) (NN
time)) (SBAR (S (NP (NNP Finland)) (VP (VBD had) (NP (NP (CD three) (JJ main) (JJ cultural) (NN areas,) (NNP Finland) (F
W proper,) (NNP Tavastia)) (CC and) (NP (NNP Karelia,))))))))))) (SBAR (WHNP (WDT which)) (S (VP (VBZ is) (VP (VBN refl
ected) (PP (IN in) (NP (JJ e.g.) (JJ contemporary) (NN jewellery.[15])))))))))))))))) )
```

Grammar files:
eng_sm6.gr
ger_sm5.gr

Übung 2

- Stanford Parser auf größeren Texten anwenden

für Windows:

```
lexparser.bat <input file>
```

für Linux/Unix:

```
lexparser.sh <input file>
```

Übung 2

```
C:\Users\hanns\Documents\stanford-parser-full-2017-06-09>lexparser.bat bsp_Text.txt
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
Parsing file: bsp_Text.txt
Parsing [sent. 1 len. 14]: Finland was inhabited when the last ice age ended , approximately 9000 BCE .
(ROOT
  (S
    (NP (NNP Finland))
    (VP (VBD was)
      (VP (VBN inhabited)
        (SBAR
          (WHADVP (WRB when))
          (S
            (NP (DT the) (JJ last) (NN ice) (NN age))
            (VP (VBD ended) (, ,)
              (ADVP (RB approximately)
                (NP (CD 9000) (NNS BCE))))))))
    (. .)))

nsubjpass(inhabited-3, Finland-1)
auxpass(inhabited-3, was-2)
root(ROOT-0, inhabited-3)
advmod(ended-9, when-4)
det(age-8, the-5)
amod(age-8, last-6)
compound(age-8, ice-7)
nsubj(ended-9, age-8)
advcl(inhabited-3, ended-9)
advmod(ended-9, approximately-11)
nummod(BCE-13, 9000-12)
nmod:npmod(approximately-11, BCE-13)

Parsing [sent. 2 len. 24]: -LSB- 10 -RSB- The first settlers left behind artifacts that present characteristics shared w
ith those found in Estonia , Russia , and Norway .
(ROOT
  (S
    (NP (NNP -LSB-) (CD 10))
    (VP (VBD -RSB-)
      (SBAR
```

Besser in Datei schreiben lassen ;)

Übung 2

- Probiert verschiedene Texte aus
- Was fällt euch auf?

Zum Nachlesen

- Johnson, Mark (1999): *PCFG Models of Linguistic Tree Representations*

http://comp.mq.edu.au/~mjohnson/papers/johnson-97.pdf

- Jurafsky Daniel, Martin James H. (2008, 2.Auflage): *Speech and Language Processing*
- Klein Dan, Manning Christopher D. (2003a): *Accurate Unlexicalized Parsing.*

https://nlp.stanford.edu/manning/papers/unlexicalized-parsing.pdf

- Klein Dan, Manning Christopher D. (2003b): *A* Parsing: Fast Exact Viterbi Parse Selection*

https://nlp.stanford.edu/manning/papers/pcfg-astar.pdf

- Petrov, Slav et al. (2006): *Learning Accurate, Compact and Interpretable Tree Annotation*

http://www.petrovi.de/data/acl06.pdf