Butt, Frank, Kuhn:

Development of large scale LFG grammars

ESSLLI 1999, Utrecht

Processing Aspects

Jonas Kuhn

jonas@ims.uni-stuttgart.de

Overview

- **Motivation**
  - Certain decisions in grammar design will have significant consequences on the efficiency of parsing

- **Some background on LFG parsing**
- **The problematic situation**
  - demonstration of a potential complexity problem
- **Modifying the grammar for efficiency – an experiment**

Some background on LFG parsing

- Natural division of the parsing task (following the basic architecture of LFG)
  - **c-structure:** context-free parsing
    complexity (with well-known algorithms like active chart parsing): polynomial in the length of the string (processing time for a string of length n will not grow faster than kn^3 for some constant k)
  - **f-structure:** resolution and propagation of f-equations (constructing a model)
    worst-case complexity: exponential (processing time may grow with k^n) – but see (Maxwell and Kaplan 1989)
- the combination of the two processing steps may cause additional complexity (cf. Maxwell and Kaplan 1993)

Context-free parsing: active chart parsing

A sample grammar

```
S    --> NP VP.

VP   --> V ({ VP | A }).

NP   --> { (A) N'
         | VN (NP) }.

N'   --> N PP*.

PP   --> P NP.
```

Four readings for:

- flying planes with wings of wood can be dangerous

The context-free grammar (slightly simplified)

S $\rightarrow$NP VP
NP $\rightarrow$A N$'$
NP $\rightarrow$N$'$
NP $\rightarrow$N$'$
NP $\rightarrow$VN NP
N$'$ $\rightarrow$N
N$'$ $\rightarrow$N PP
N$'$ $\rightarrow$N PP PP
PP $\rightarrow$P NP
VP $\rightarrow$V VP
VP $\rightarrow$V A

chart initialization (lexical lookup)

prediction:

S → · NP VP

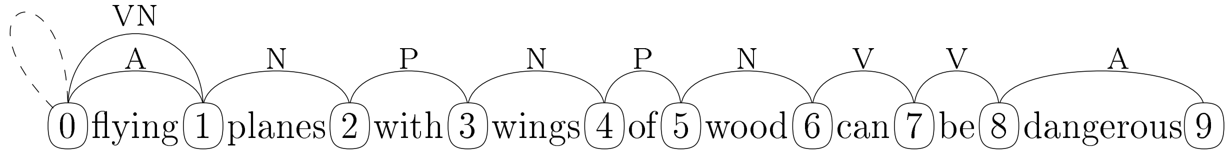
NP → · A N′

NP → · N′

NP → · VN NP

N′ → · N

N′ → · N PP



completion:

NP → A · N′

NP → VN · NP

N′ → N ·

(*inactive*)

N′ → N · PP

NP → N′ ·

(*inactive*)

S → NP · VP

prediction:

NP → · A N′

NP → · N′

NP → · VN NP

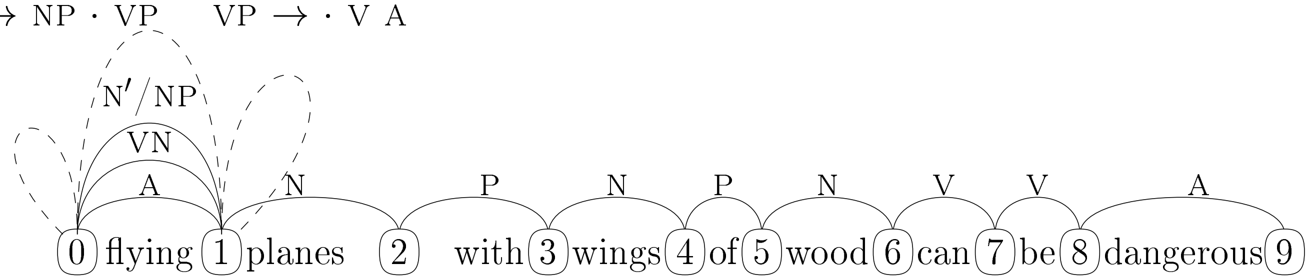
N′ → · N

N′ → · N PP

PP → · P NP

VP → · V VP

VP → · V A

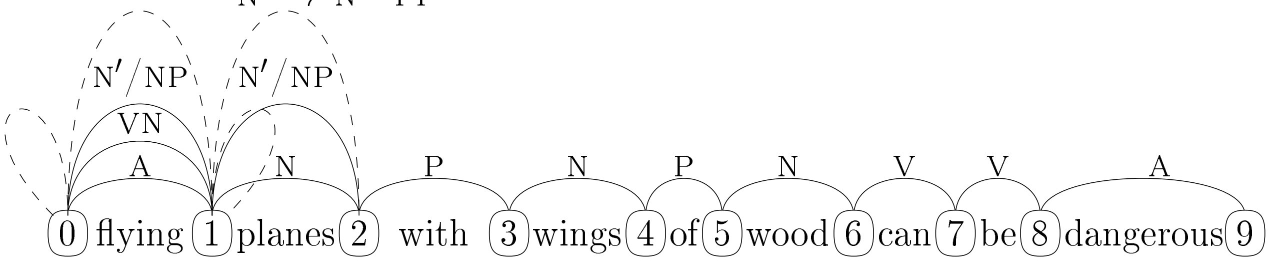


completion:

N′ → N ·

NP → N′ ·

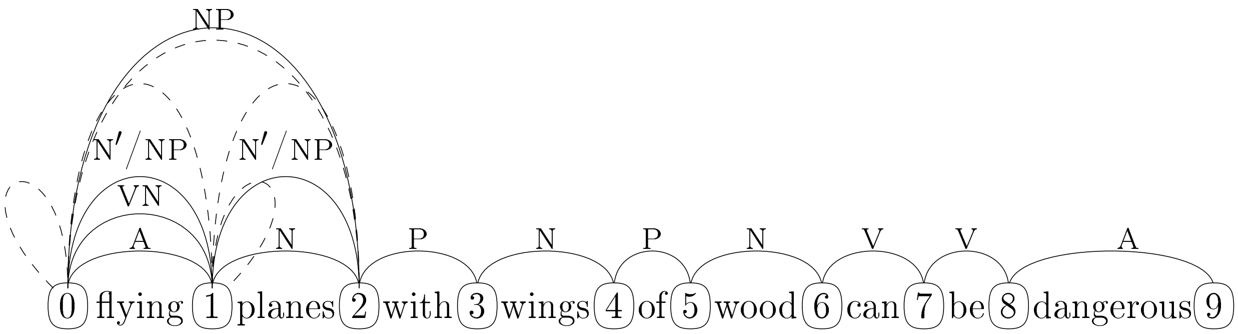
N′ → N · PP



completion:

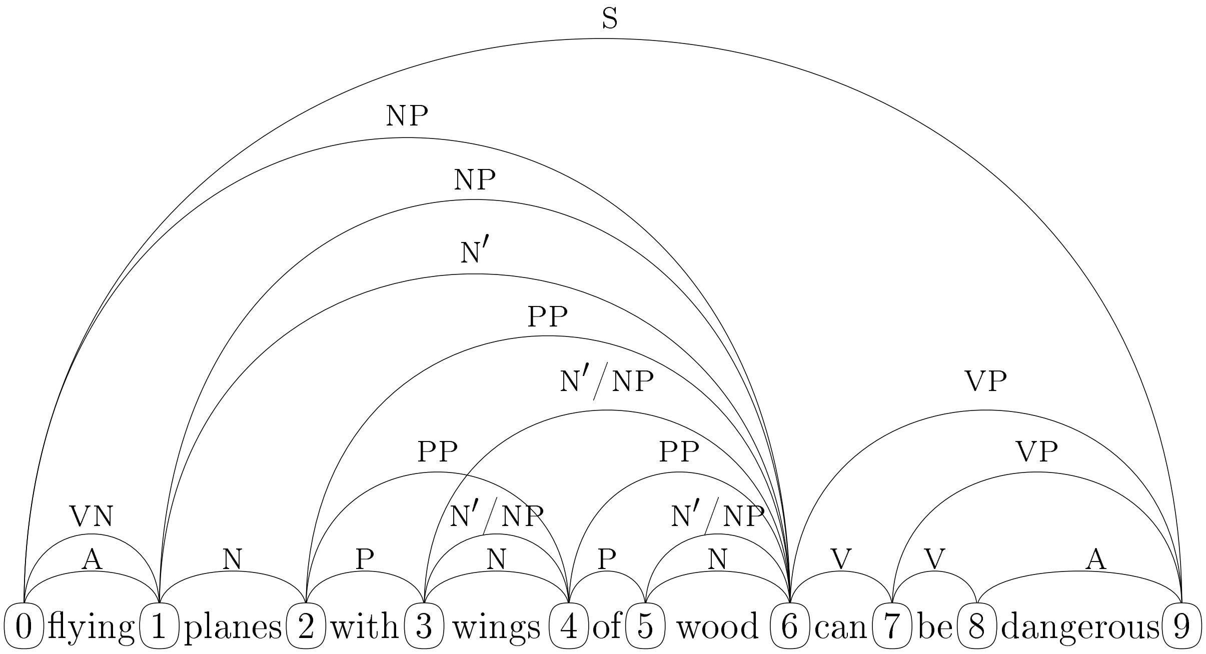
NP → A N′ ·

S → NP · VP



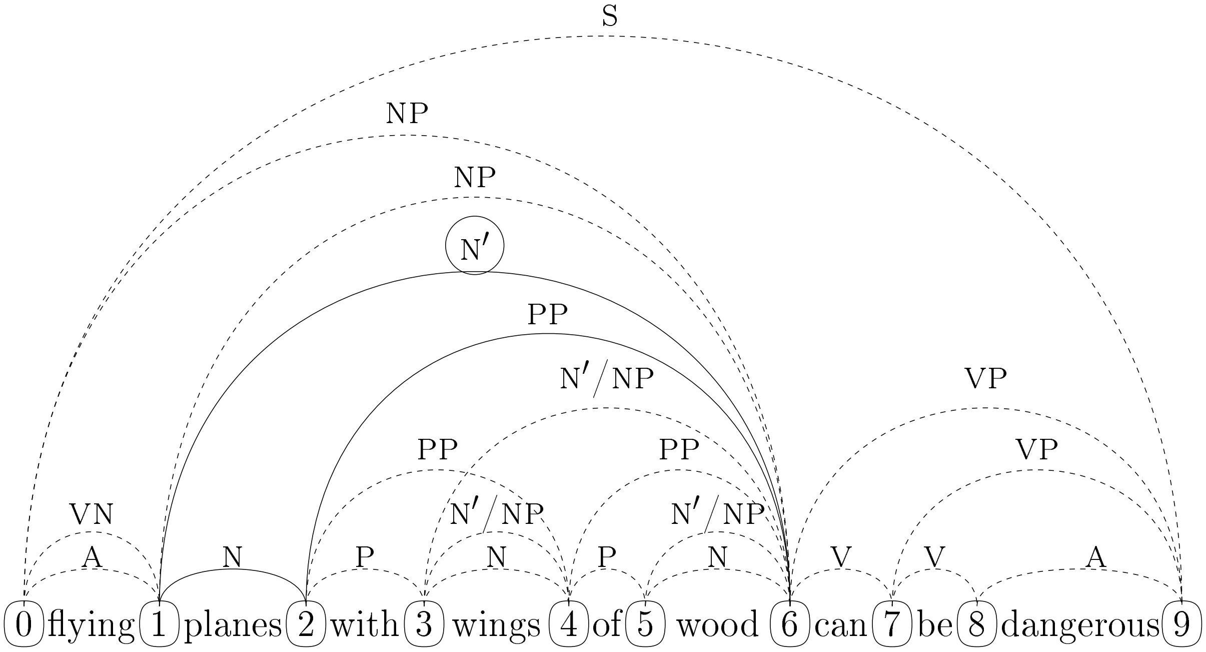
- finished when an edge of the root category is found that spans the complete sentence

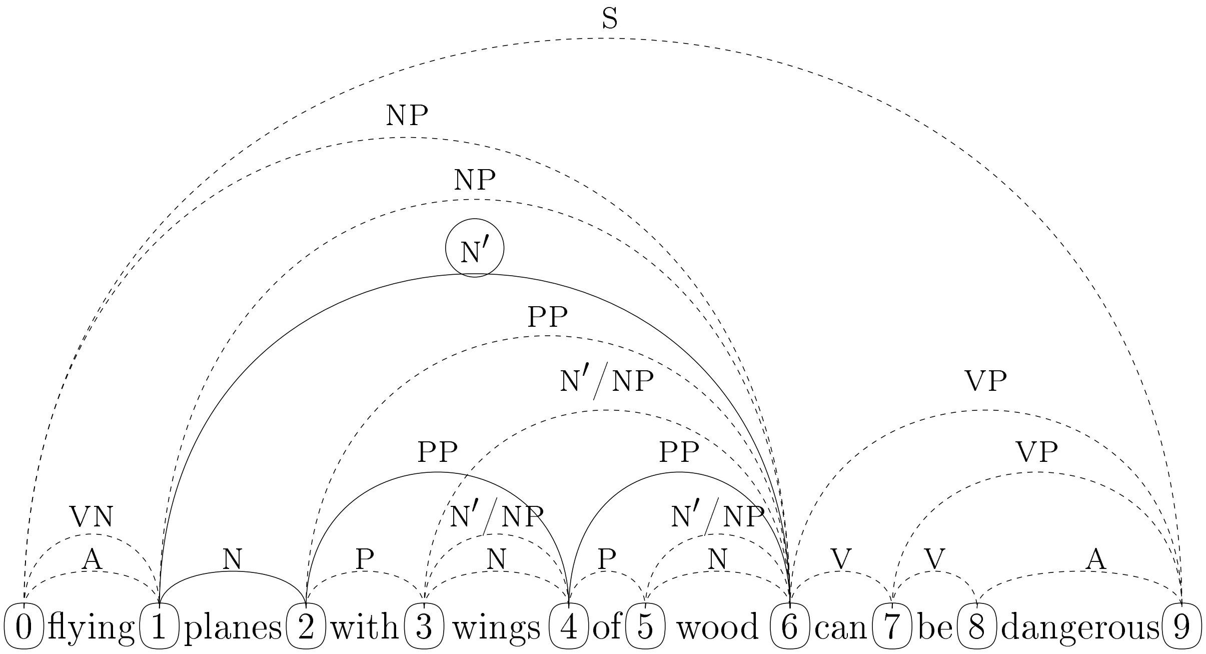
The complete chart (relevant inactive edges)



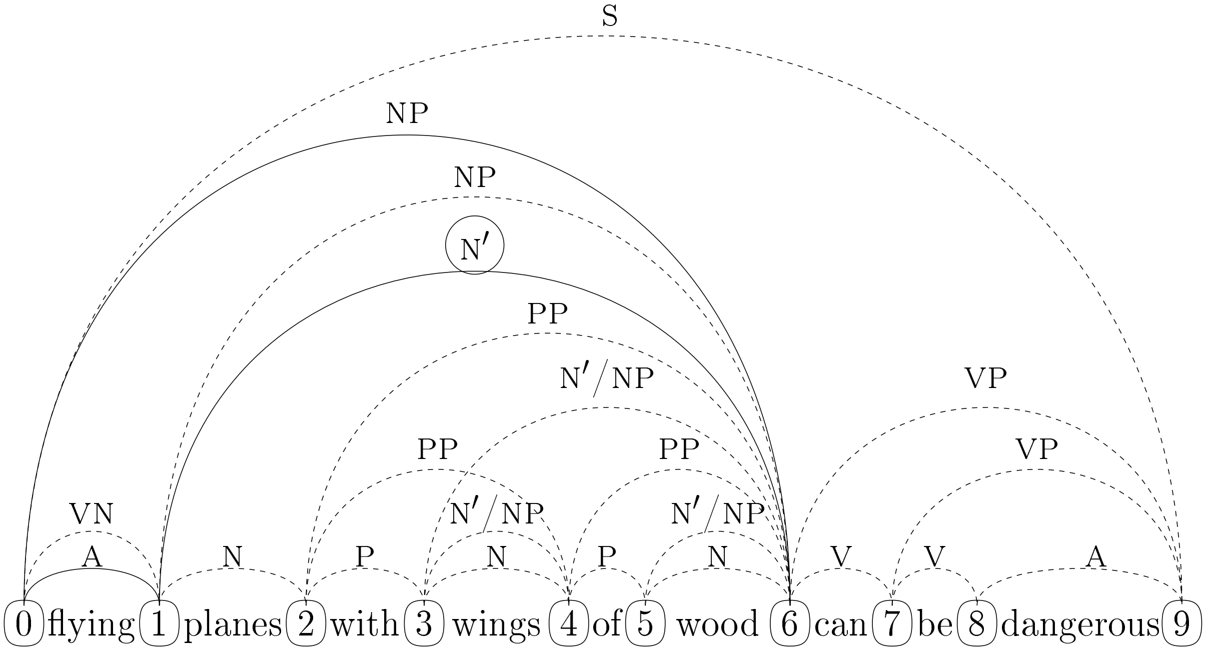
The encapsulation of ambiguities

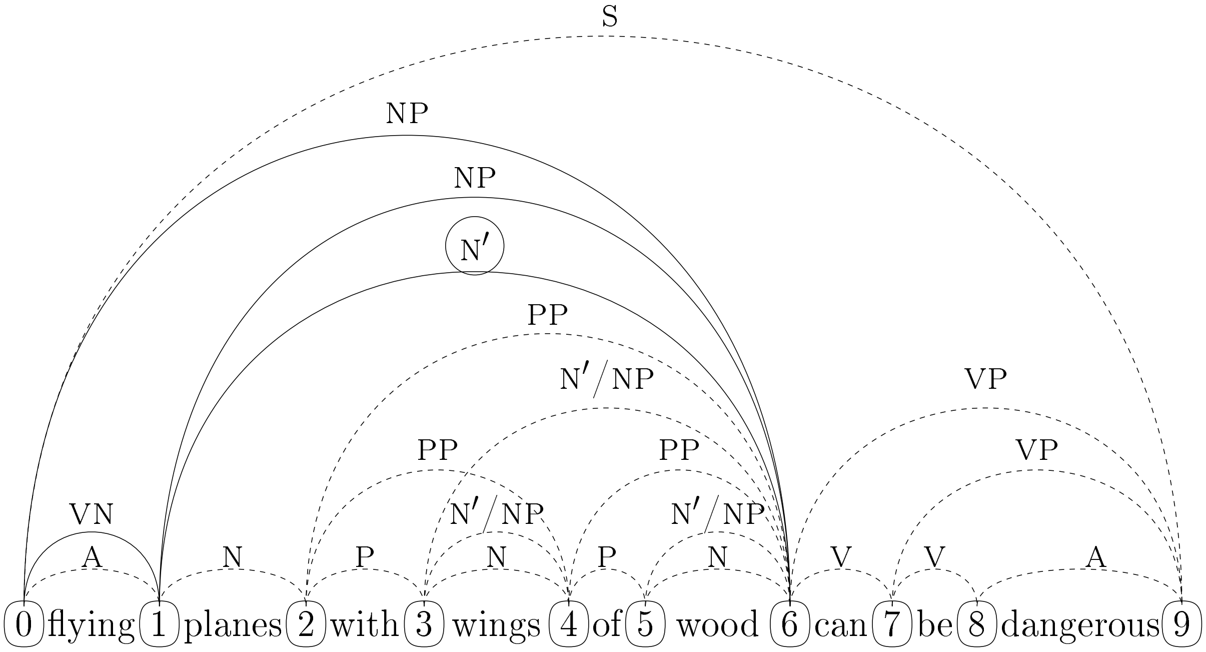
```
N'   --> N PP*.
```





```
NP   --> { (A) N'
         | VN (NP) }.
```


S
NP
NP
N′
PP
N′/NP
VP
PP
PP
VP
VN
N′/NP
N′/NP
A
N
P
N
P
N
V
V
A
0 flying 1 planes 2 with 3 wings 4 of 5 wood 6 can 7 be 8 dangerous 9


S
NP
NP
N′
PP
N′/NP
VP
PP
PP
VP
VN
N′/NP
N′/NP
A
N
P
N
P
N
V
V
A
0 flying 1 planes 2 with 3 wings 4 of 5 wood 6 can 7 be 8 dangerous 9

The encapsulation of ambiguities

- A chart can represent an exponential number of readings in polynominal space
- Equivalent partial trees are collapsed, so externally they look as just one thing
  - the node numbers in the XLE display of the c-structure refer to edges in the chart (you can check it for the sample grammar) – dotted lines in trees point to alternatives
  - there's also a chart display (via the *Commands* pull-down menu in the c-structure window), which is however hard to read for longer sentences

- For the second step of LFG parsing, f-equations have to be resolved and propagated along the tree structure
  - problem: this may require to distinguish what has previously been collapsed
  - algorithms have been optimized to avoid this situation whenever possible, but in the worst case (when all encapsulations have to be "unpacked") the parsing complexity may get exponential

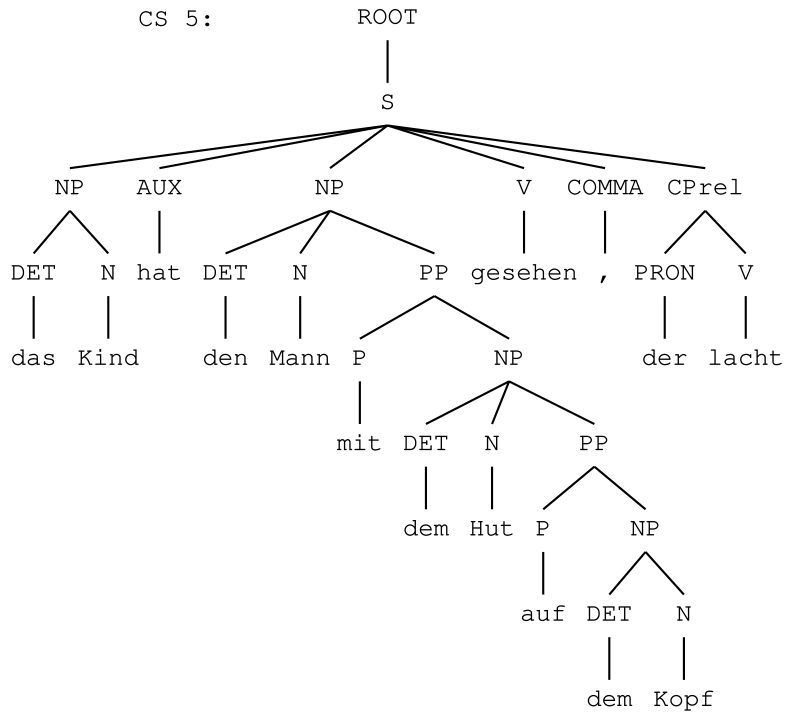
Interaction

An experimental grammar

```
S --> NP: (^SUBJ)=!                 "das Kind    - the child"
          (!CASE)=nom;
      AUX                           "hat         - has"
      NP: (^OBJ)=!                  "den Mann    - the man"
          (!CASE)=acc;
      [PP: !$(^ADJUNCT);]*          "mit dem Hut - with the hat"
      V                             "gesehen     - seen"
      (COMMA
       CPrel:                       "der lacht   - who laughs"
           (^ ({SUBJ|OBJ})
              (ADJUNCT $ [OBJ ADJUNCT $]*) OBJ) = %REF
           ! $ (%REF ADJUNCT)
           (! TOPIC GEND) =c (%REF GEND)).

CPrel -->
       PRON: (! REL) =c +
             (^ TOPIC) = !
             (^ SUBJ)=!;
       V.

NP --> DET N  [PP: !$(^ADJUNCT);]*.

PP --> P NP: (^OBJ)=!.
```

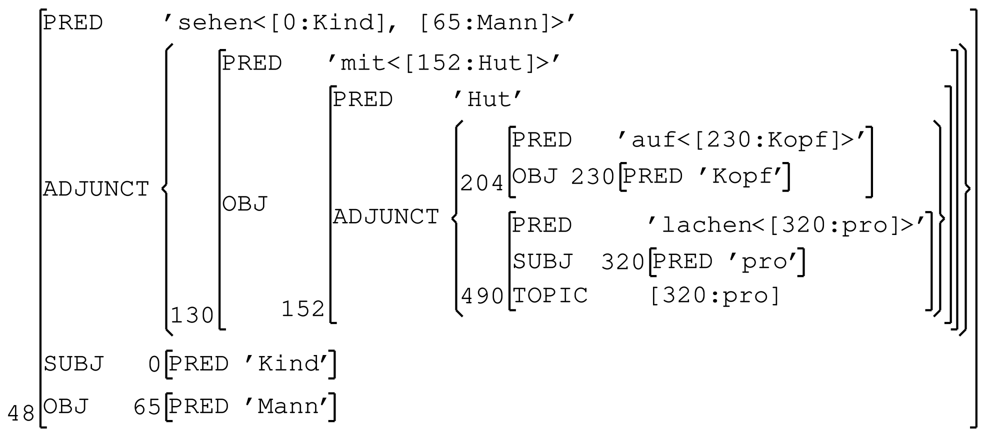
An experimental grammar

CS 5:



Three f-structure readings for this c-structure tree:

"das Kind hat den Mann mit dem Hut auf dem Kopf gesehen, der lacht"



"das Kind hat den Mann mit dem Hut auf dem Kopf gesehen, der lacht"

```
48[ PRED      'sehen<[0:Kind], [65:Mann]>'
    ADJUNCT   { 130[ PRED  'mit<[152:Hut]>'
                     OBJ   152[ PRED     'Hut'
                                ADJUNCT  { 204[ PRED  'auf<[230:Kopf]>'
                                                OBJ   230[ PRED     'Kopf'
                                                           ADJUNCT  { 490[ PRED   'lachen<[320:pro]>'
                                                                           SUBJ   320[PRED 'pro']
                                                                           TOPIC  [320:pro] ] } ] ] } ] ] }
    SUBJ      0[PRED 'Kind']
    OBJ       65[PRED 'Mann'] ]
```

"das Kind hat den Mann mit dem Hut auf dem Kopf gesehen, der lacht"

```
48[ PRED      'sehen<[0:Kind], [65:Mann]>'
    ADJUNCT   { 130[ PRED  'mit<[152:Hut]>'
                     OBJ   152[ PRED     'Hut'
                                ADJUNCT  { 204[ PRED  'auf<[230:Kopf]>'
                                                OBJ   230[PRED 'Kopf'] ] } ] ] }
    SUBJ      0[PRED 'Kind']
    OBJ       65[ PRED     'Mann'
                  ADJUNCT  { 490[ PRED   'lachen<[320:pro]>'
                                  SUBJ   320[PRED 'pro']
                                  TOPIC  [320:pro] ] } ] ]
```

Some examples from a German newpaper corpus

- Ein Stuttgarter Großhändler hatte [fünf Paletten [mit fertigen italienischen **Konserven**]] gestiftet, **die** zwar noch einwandfrei, aber hierzulande nicht mehr zu verkaufen gewesen wären.
- Über diese Verbindung wurde auch [Kontakt [zum **Jermolova-Theater**]] hergestellt, **das** zu den wichtigsten Bühnen Moskaus gehört.
- Escobar hatte am Wochenende [die Freilassung [von zwei **Journalisten**]] angeordnet, **die** vor acht Monaten von der Kampforganisation der Drogenhändler Medellins entführt worden waren.
- Das Großkraftwerk Mannheim (GKM) hat vor [Belastungen [durch die von der Bundesregierung geplante Kohlendioxydabgabe [für **Großfeuerungsanlagen**]]] gewarnt, **die** der Abgabe gar nicht ausweichen könnten.
- Diese Gruppe hat [das Schicksal [von **Zwangsarbeiterinnen**]] erforscht, die in Calw während der letzten Kriegsmonate in Nachtschichten für Hitlers Rüstung gearbeitet haben.

Referent is in an intermediate position:

- Heftig wurde beispielsweise [der Streit [über das **Nachzugsalter** [von ausländischen Familienangehörigen]]] ausgetragen, **das** Zimmermann auf sechs Jahre herabsetzen wollte.
- (der Student ... hat ...) [nach einer Lösung [für das gesamte **Quartier** [bis vor zum Albplatz]]] gesucht, **das** heute kaum mehr als planerisches Flickwerk ist.
- Das UN-Team hatte [weitere Informationen [über ein **Programm** [zur Anreichung [von Uran]]]] verlangt, **das** Irak zunächst verheimlicht hatte.
- Dem britischen Komitee für Arzneimittelsicherheit seien [17 Fallberichte [über schwere **Zwischenfälle** [mit Herzrhythmusstörungen]]] zugegangen, **die** dem Wirkstoff Terodilin zugeordnet würden.

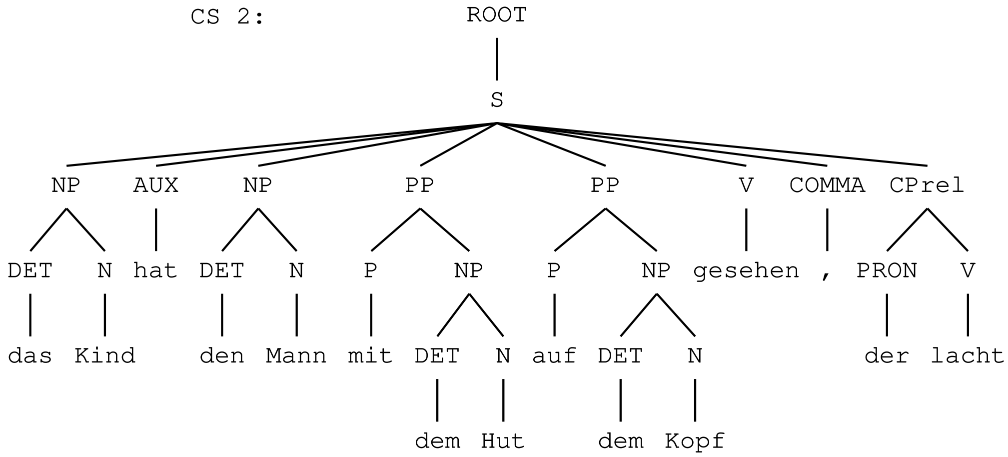
Referent in the lowest one of several PPs:

- Schließlich ist auch [eine Änderung [in der Struktur [der **Landesanstalt** für Kommunikation (LfK)]]] vorgesehen, **die** ja grundsätzlich verändert werden soll.
- Außerdem wurden [ehemalige Müllplätze [in der Nähe [von **Industriebetrieben**]]] untersucht, **die** möglicherweise diese Deponien benutzt haben, um ihre giftigen Produktionsrückstände kostengünstig loszuwerden.
- Wie Staatsminister Helmut Schäfer erklärte, sind darin [die Kosten [von 165 Millionen Mark] [für ein **Anti-Raketensystem Patriot**]] enthalten, **das** die USA an Israel liefern.
- Mehrheitlich hat der Bauausschuß des Ludwigsburger Gemeinderats [grünes Licht [für die Suche [nach einer möglichst optimalen Trasse [für die sogenannte **Westumgehung**]]]] gegeben, **die** zur Entlastung des westlichen Stadtgebiets eine direkte Verbindung zwischen der Schwieberdinger Straße und der Bundesstraße 27 in Eglosheim bringen soll.

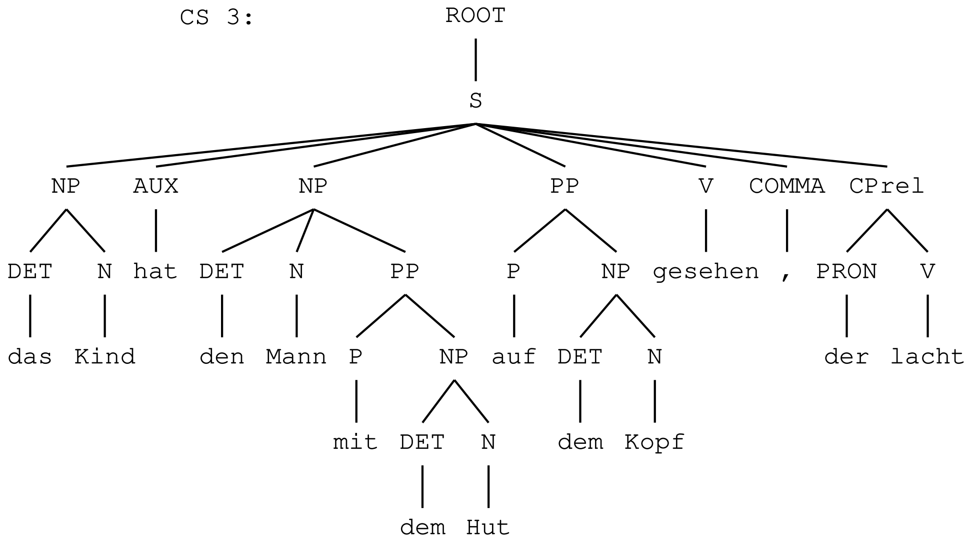
Exponentially growing number of c-structure readings:

Five readings for a sentence with three PPs:

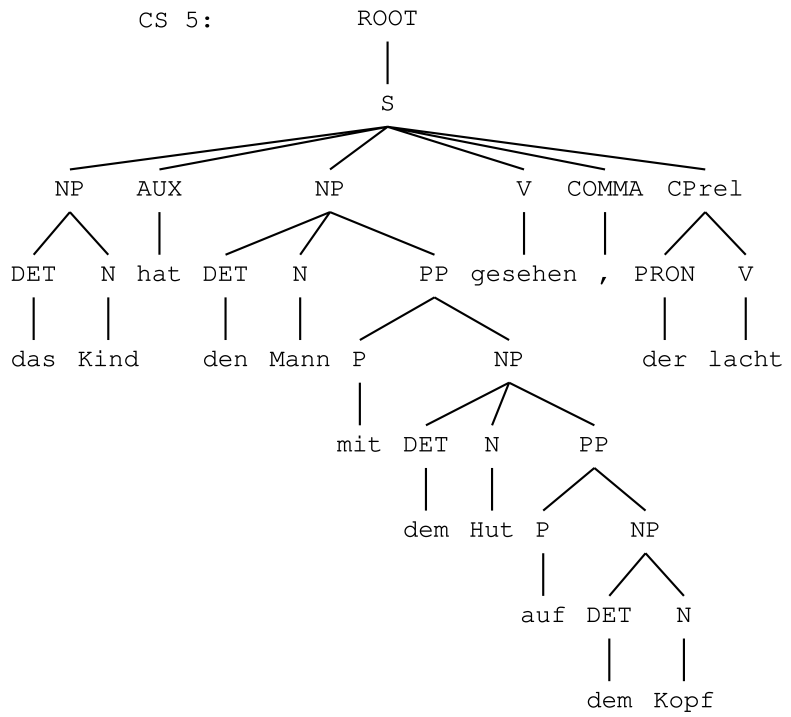
CS 2:



CS 3:



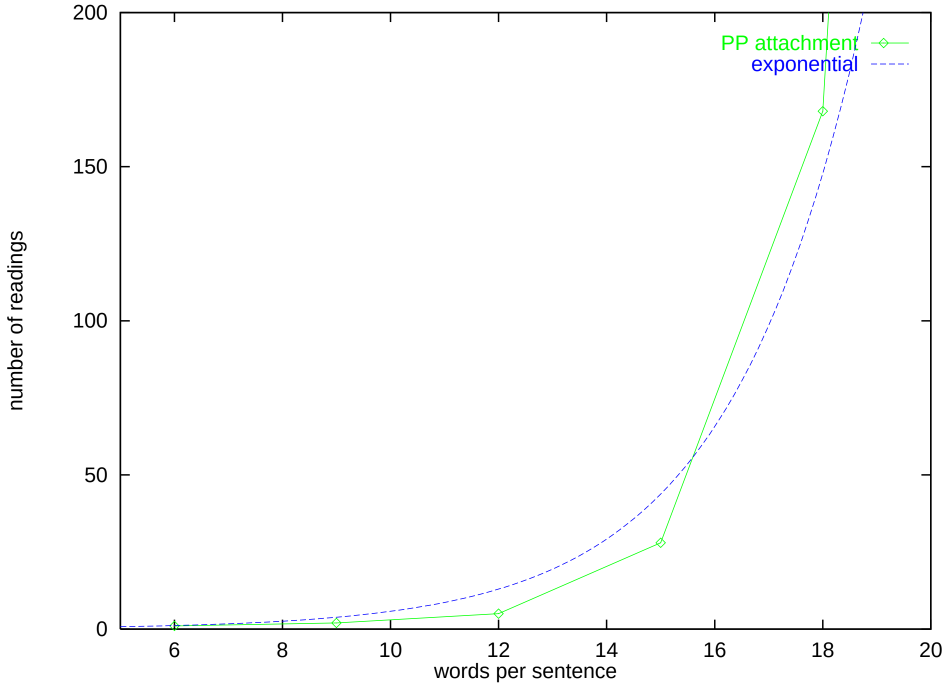
⋮



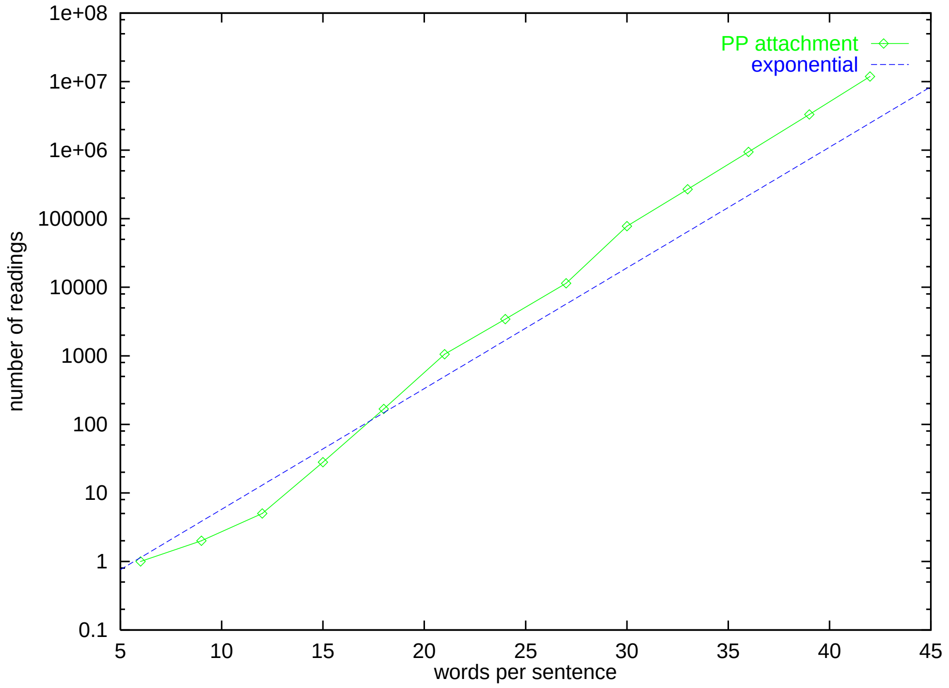
Exponentially growing number of c-structure readings:

number of PPs	c-structure readings
1	1
2	2
3	5
4	28
5	168
6	1,056
7	3,432
⋮	
13	11,886,400

Number of readings



Logarithmic scale



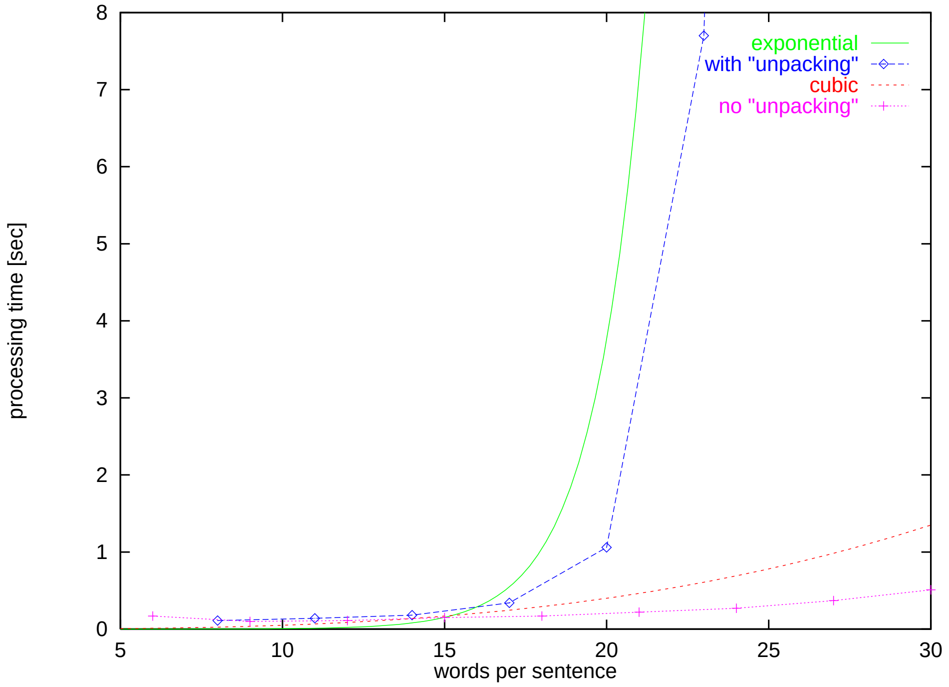
Interaction with f-constratins

- The functional uncertainty equation attached to the extraposed relative clause will look into any one of the different f-structures built up

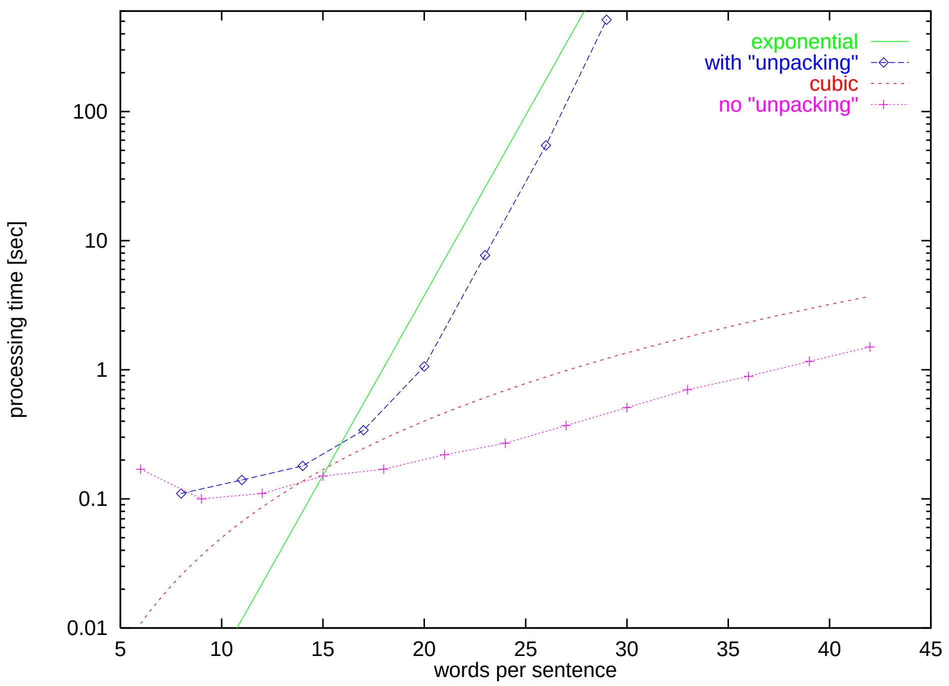
⇒ the compact representation is effectively "unpacked"

```
S --> NP: (^SUBJ)=!                "das Kind    - the child"
          (!CASE)=nom;
      AUX                          "hat         - has"
      NP: (^OBJ)=!                 "den Mann    - the man"
          (!CASE)=acc;
      [PP: !$(^ADJUNCT);]*         "mit dem Hut - with the hat"
      V                            "gesehen     - seen"
      (COMMA
       CPrel:                      "der lacht   - who laughs"
           (^ ({SUBJ|OBJ})
              (ADJUNCT $ [OBJ ADJUNCT $]*) OBJ) = %REF
           ! $ (%REF ADJUNCT)
           (! TOPIC GEND) =c (%REF GEND)).
```

Processing time



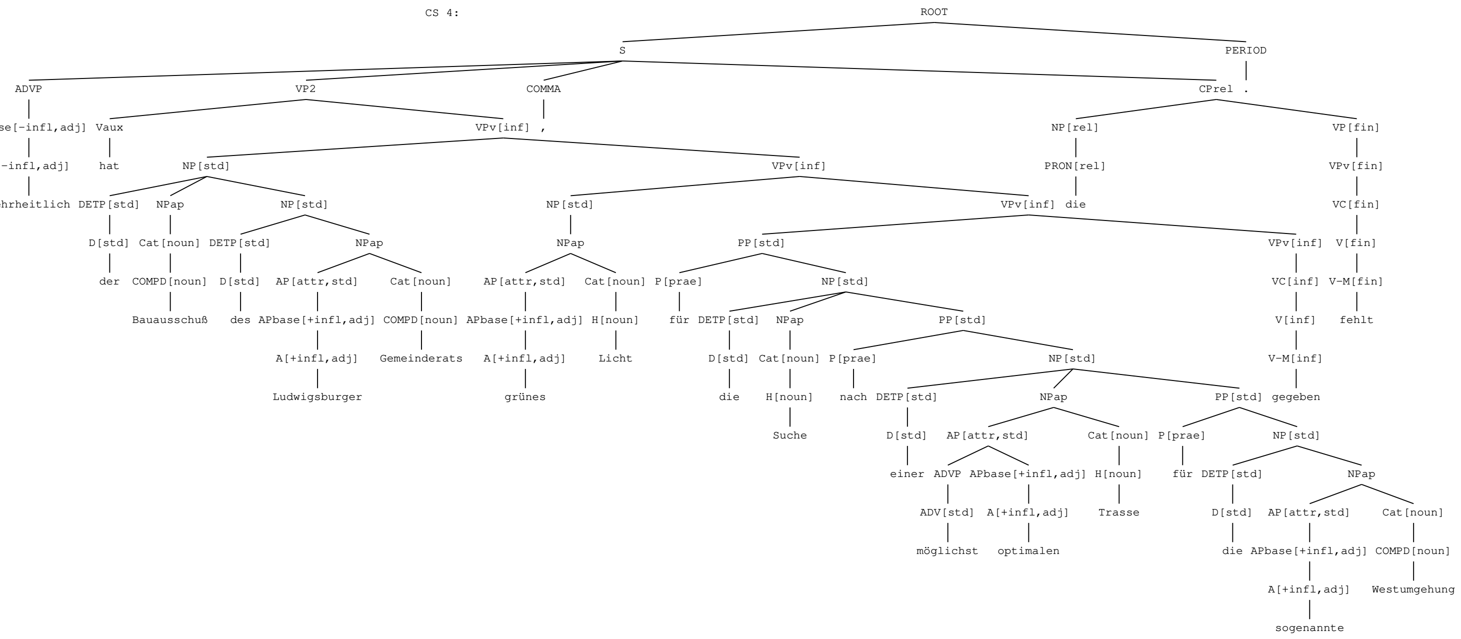
Logarithmic scale



Interaction

- The worst-case scenario may not be of high practical relevance

```
parse {Mehrheitlich hat der Bauausschuß des
Ludwigsburger Gemeinderats grünes Licht für die Suche
nach einer möglichst optimalen Trasse für die
sogenannte Westumgehung gegeben, die fehlt.}
14 solutions, 49.28 CPU seconds, 2005 subtrees
unified
```



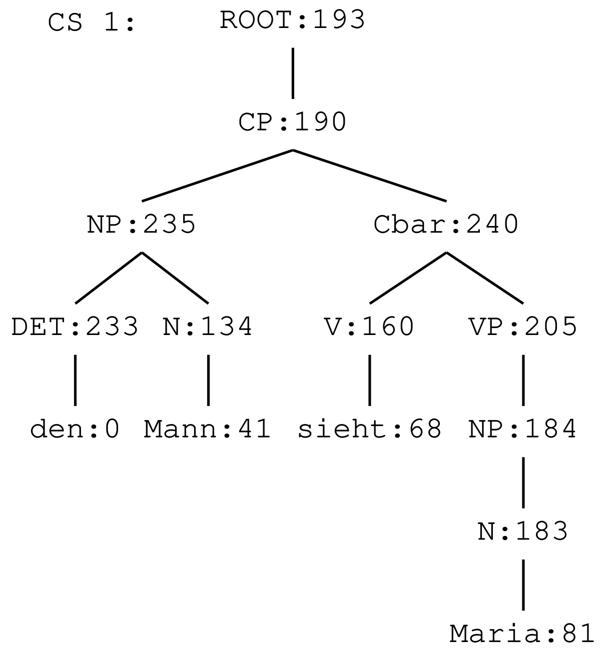
However: lesson learned

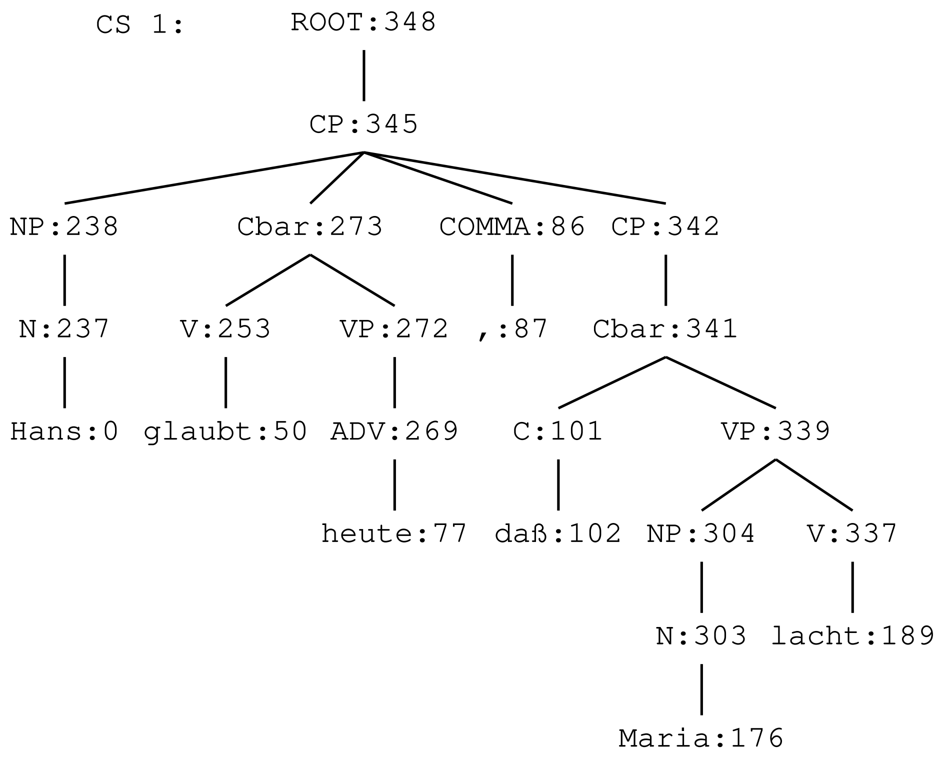
- Where possible, exponential c-structure ambiguity should be avoided

Avoiding exponential c-structure ambiguity

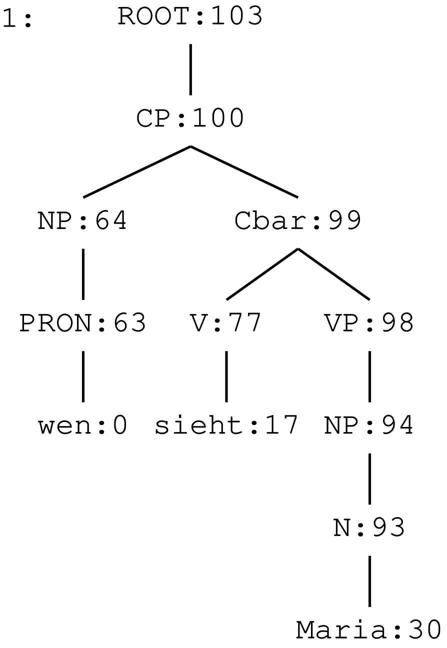
- Specifying a grammar in a very general way may go along with a highly unconstrained c-structure
  - The straightfowrard way of making generalizations explicit is by localizing distinctions at the level of f-structure (this makes it possible to leave the respective information unspecified)
- In the light of the processing considerations, this is a problem
  - Distinctions that are made already at the level of c-structure help to prune the search space very early, avoiding

Example: generalized CP-analysis of German clauses

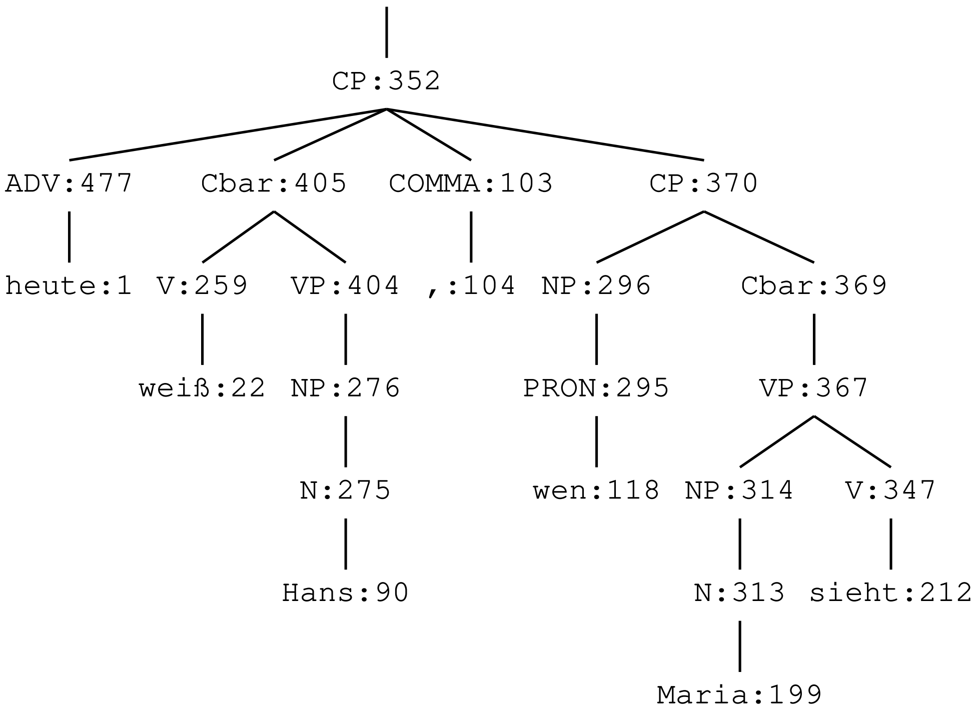




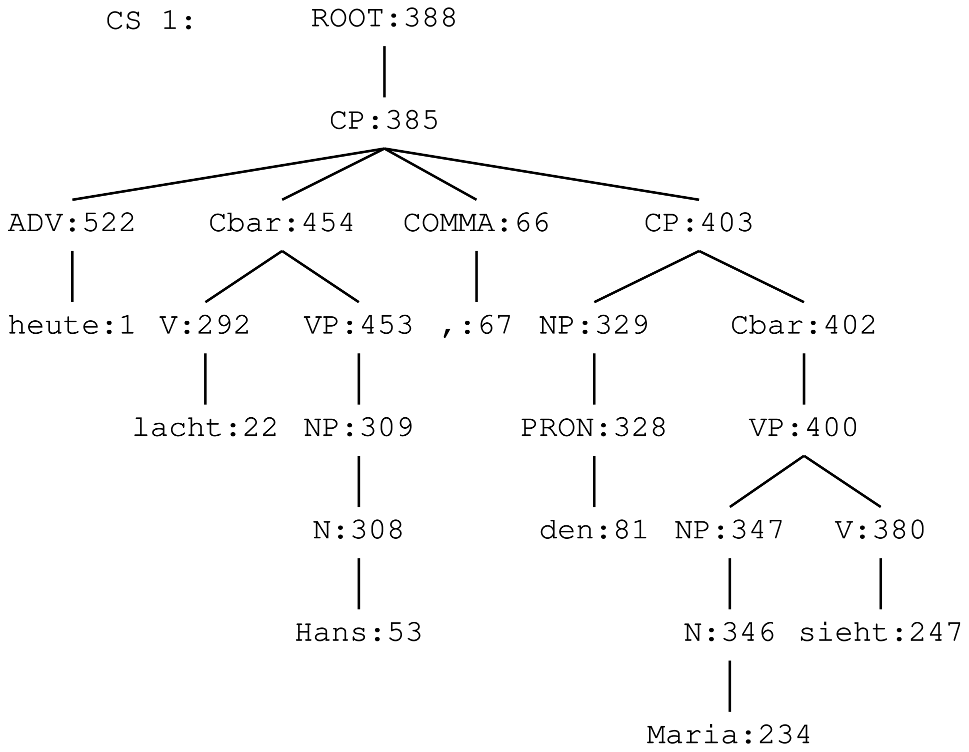
CS 1:


ROOT:103
CP:100
NP:64
Cbar:99
PRON:63
V:77
VP:98
wen:0
sieht:17
NP:94
N:93
Maria:30

CS 1:


ROOT:355
CP:352
ADV:477
Cbar:405
COMMA:103
CP:370
heute:1
V:259
VP:404
,:104
NP:296
Cbar:369
weiß:22
NP:276
PRON:295
VP:367
N:275
wen:118
NP:314
V:347
Hans:90
N:313
sieht:212
Maria:199

CS 1:


ROOT:388
CP:385
ADV:522
Cbar:454
COMMA:66
CP:403
heute:1
V:292
VP:453
,:67
NP:329
Cbar:402
lacht:22
NP:309
PRON:328
VP:400
N:308
den:81
NP:347
V:380
Hans:53
N:346
sieht:247
Maria:234

```
CP --> {
         "Vorfeld:"
         { NP: (^TYPE)=(!TYPE)
               (^ TOPIC) = !
               (^ { SUBJ
               | PATH { OBJ | OBJ2 }} ) = !;
         | .....
         | e: (^TYPE) ~= rel;               }

         "Verbklammern and Mittelfeld:"
         Cbar

         "Nachfeld:"
         [ COMMA
          CP: { (!TYPE) =c rel
                (^ { SUBJ
                | PATH {OBL OBJ|OBJ|OBJ2}} ) = %REF
                ! $ (%REF ADJUNCT)
                (! TOPIC GEND) =c (%REF GEND)
              | (!TYPE) = std
                (^ XCOMP* COMP)=!
              | (!TYPE) =c int
                (^ XCOMP* WH-COMP)=!
                (!VPOS)~=v2 "excl. emb. v2 quest." }]*
       | @(Coord CP)}.
```

```
Cbar -->
         { C: (^TYPE) =std
         | V: ^=!
              (^VFORM)=fin
              (^TYPE) ~=rel
              (^VPOS)=v2;
         | e: (^TYPE) ~= std;
              "linke Verbklammer empty in rel. clause and
                                  embedded int. clause"
         }
         (VP).


 VP -->  { NP:  { (^ NGF)=!
                  (^ VFORM) = fin
                | (^ {OBJ|OBJ2})=! "no SUBJ"
                  (^ VFORM) = inf}
          | .....
          | VP: (^XCOMP)=!
          | ADV: ! $ (^ADJUNCT)
                }*
          { V
          | e: (^ VFORM) = fin }.
```

Avoiding exponential c-structure ambiguity

Rule parametrization with complex category symbols

- Parametrized rules:

```
PP[_var] --> { P        NP[_var]
             | NP[_var] Ppost    }.
```

  expands to (for instance)

```
PP[-wh] --> { P       NP[-wh]
            | NP[-wh] Ppost   }.
PP[+wh] --> { P       NP[+wh]
            | NP[+wh] Ppost   }.
```

- Expansion depends on usage of `PP[xyz]` categories elsewhere in the grammar.

Usage for generalizing over rules:

```
XP[_C]    -->  YP: (^DF)=!;   Xbar[_C]: ^=!.
Xbar[_C]  -->  X[_C]: ^=!;    YP: (^GF)=!;*.
```

Using complex category symbols and rule parametrization in CP grammar

```
CP[_type] -->
    {
       "Vorfeld:"
       { NP[_type]: (^ TOPIC) = !
                    (^ { SUBJ
                       | PATH { OBJ | OBJ2 }} ) = !;
       | .....
       | e: _type ~= rel;                         }

      "Verbklammern and Mittelfeld:"
      Cbar[_type]

      [ COMMA
        { CP[rel]:
             (^ { SUBJ
                | PATH {OBL OBJ|OBJ|OBJ2}} ) = %REF
             ! $ (%REF ADJUNCT)
             (! TOPIC GEND) =c (%REF GEND)
        | CP[std]: (^ XCOMP* COMP)=!
        | CP[int]: (^ XCOMP* WH-COMP)=!
               (!VPOS)~=v2 "excl. emb. v2 quest."}]*

     | @(Coord CP[_type]) }.
```

```
Cbar[_type] -->
         { C: _type =std
         | V: ^=!
              (^VFORM)=fin
              _type ~= rel
              (^VPOS)=v2;
         | e: _type ~= std;
              "linke Verbklammer empty in rel. and
                          embedded int. clause"
         }
         (VP).
```

```
VP --> { NP[std]: { (^ NGF)=!
                    (^ VFORM) = fin
                  | (^ {OBJ|OBJ2})=! "no SUBJ"
                    (^ VFORM) = inf}
       | .....
       { V
       | e: (^ VFORM) = fin }.
```

Comparing performance

Experiment (cf. Kuhn 1999; Butt et al. 1999)

- systematic comparison of the grammar variants
  - (A) f-structure level distinction of clause type
  - (B) parametrization of c-structure categories according to clause type
  - (C) additional parametrization of NP-internal categories with morphosyntactic information (number, declension class)
- Test suite: 25 sentence (8-18 words)
  examples:
  - Maria soll den neuen Film sehen und Otto sieht den, der seit einem Jahr läuft.
  - der alte Mann empfiehlt heute schnellen Autos neue.
  - der Student sieht das kleine Kind mit dem kleinen Teleskop auf dem Berg.
  - Maria weiß, welcher Mann an sie denkt und wann Otto glaubt, daß er den Fisch sieht, der schwimmt.

Parametrization within the NP rules (excerpt)

```
NP[_type] -->
    { "plural nouns with strongly declined adjective"
      { DET[_type]  "determiner optional"
      | e: _type=std }
      ADJ[pl,st]: ...;*
      { N[pl]
      | ADJ[pl,st]: ...;
        e: (^PRED)='pro' "no overt noun"}

    | "plural nouns with weakly declined adjective"
      DET[_type]    "determiner obligatory"
      ADJ[pl,wk]: ...;*
      { N[pl]
      | ADJ[pl,wk]: ...;
        e: (^PRED)='pro' "no overt noun"}

    | "singular (count) nouns"
      DET[_type]       "determiner obligatory"
      ADJ[sg,_]: ...;*
      { N[sg]
      | ADJ[sg,_]: ...;
        e: (^PRED)='pro' }
    }
    [...].
```

Experiment

Grammar version	(A)	(B)	(C)
Results based on entire testsuite (25 sentences)			
parsing time			
average [sec]	> 300	> 75	1.19
std deviation [sec]	> 370	> 220	1.13
median [sec]	> 130	> 3	0.94
quartile dist. [sec]	> 730	> 14	1.01
maximal time [sec]	> 900	> 900	5.6
# sent. beyond timeout	6	1	-
Results based on the 19 "easiest" sentences			
parsing time			
average [sec]	121.19	4.83	0.92
standard deviation [sec]	163.38	6.82	0.65
median [sec]	69.89	2.15	0.88
quartile dist. [sec]	161.62	4.70	0.69
maximal time [sec]	720.76	29.2	5.6
# sent. beyond timeout	-	-	-

- timeout after 900 seconds

Conclusion

- From the point of view of performance, it is often better to express certain constraints at the level of c-structure
  - the search space will be pruned early in parsing
  - the negative effect of having to “unpack” exponential c-structure ambiguity during the resolution of f-equation can be reduced
- Normally, generalizations over c-structure rules cannot be made as explicit as generalizations residing at the level of f-structure
- Complex category symbols are one way of overcoming this problem
  - they allow explicit generalization over different c-structure rules
  - modifications and extensions made to such a parametrized rules will thus generalize to all instances automatically

References

Butt, M., T. King, M.-E. Niño, and F. Segond (1999). *A Grammar-Writer's Cookbook*. Number 95 in CSLI Lecture Notes. Stanford, CA: CSLI Publications.

Kuhn, J. (1999). Meta-descriptions of rules for generalization in constraint-based grammar design. Ms. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, available at `http://www.ims.uni-stuttgart.de/~jonas/`.

Maxwell, J. and R. Kaplan (1989). An overview of disjunctive constraint satisfaction. In *Proceedings of the International Workshop on Parsing Technologies*, Pittsburgh, PA. (also appears in Dalrymple et al. 1995.).

Maxwell, J. and R. Kaplan (1993). The interface between phrasal and functional constraints. *Computational Linguistics 19*(4), 571–590. (also appears in Dalrymple et al. 1995.).