Lexicalized Stochastic Modeling of Constraint-Based Grammars using Log-Linear Measures and EM

Stefan Riezler, Detlef Prescher, Jonas Kuhn
Institut f. Maschinelle Sprachverarbeitung
Universität Stuttgart

Mark Johnson
Cogn. & Ling. Sciences
Brown University

ACL 2000 • Hong Kong • October 6, 2000

Motivation: Statistical Disambiguation for SUBGs

Possible applications of linguistically deep UBG-parsing presuppose a precise disambiguation of parse ambiguities, e.g., by stochastic models on UBGs.

Problem: Common wisdom is that stochastic grammars require large manually annotated treebanks for estimation in order to yield accurate results. However, to date large UBG-annotated **treebanks** are **unavailable**!

Our solution: Take a second look at statistical estimation from large corpora of **unannotated data** and make **incomplete-data estimation usable!**

- **Hand-annotation** of large corpora of specific domains of specific languages with specific parses is **dispensable** for SUBGs!
- **Disambiguation precision** of SUBGs is **improvable significantly** by augmenting a small parsebank by a large set of unannotated data in estimation!

SUBGs: Log-Linear Models on UBGs

Log-linear distribution $p_\lambda(x)$ on set of analyses $\mathcal{X}$ of a UBG:

$$p_\lambda(x) = {Z_\lambda}^{-1} e^{\lambda \cdot \nu(x)} p_0(x)$$

where $Z_\lambda = \sum_{x \in \mathcal{X}} e^{\lambda \cdot \nu(x)} p_0(x)$ is a normalizing constant,

$\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$ is a vector of log-parameters,

$\nu = (\nu_1, \ldots, \nu_n)$ is a vector of property-functions $\nu_i : \mathcal{X} \to \mathbb{R}, i = 1, \ldots, n$,

$\lambda \cdot \nu(x)$ is the vector dot product $\sum_{i=1}^{n} \lambda_i \nu_i(x)$,

p_0 is a fixed reference distribution.

Two Optimization Problems: Property Design and Parameter Estimation

- **Property Design:** Incorporate salient statistical qualities of the training data as property-functions ν into the log-linear model!
  - Define **linguistically motivated pool** of property-functions ν, and use cutoff on number of parses for which property-function has positive value for **simple property selection**.
- **Parameter Estimation:** Statistical inference of λ from training sample!
  - **Maximum likelihood estimation** (MLE) from **incomplete data**: Find λ^* that maximizes the incomplete-data log-likelihood L for sample of sentences $y \in \mathcal{Y}$ where $L(\lambda) = \ln \prod_{y \in \mathcal{Y}} p_\lambda(y)^{\tilde{p}(y)} = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \ln \sum_{x \in X(y)} p_\lambda(x)$.

Incomplete-Data MLE for Log-Linear Models

IM Algorithm (Riezler PhD thesis 1999):

Input Reference model p_0, properties ν with constant $\nu_\#(x) = \sum_{i=1}^{n} \nu_i(x)$, parses $X(y)$ for each sentence y in sample $\mathcal{Y}$.

Output MLE model p_{λ^*} on $\mathcal{X} = \sum_{y \in \mathcal{Y} |\ \tilde{p}(y)>0} X(y)$.

Procedure

Until convergence do

Compute p_λ based on $\lambda = (\lambda_1, \dots, \lambda_n)$,

For i from 1 to n do

$$\gamma_i := \frac{1}{\nu_\#} \ln \frac{\sum_{y \in \mathcal{Y}} \tilde{p}(y) \sum_{x \in X(y)} p_\lambda(x|y) \nu_i(x)}{\sum_{x \in \mathcal{X}} p_\lambda(x) \nu_i(x)},$$

$$\lambda_i := \lambda_i + \gamma_i,$$

Return $\lambda^* = (\lambda_1, \dots, \lambda_n)$.

Searching for Order in Chaotic Convergence

Problem: Despite guaranteed convergence of IM iterates to critical points of L (see Riezler PhD thesis 1999), the maximization process can be chaotic in that the final L value is extremely dependent on the starting values of λ.

Duality: maximum likelihood/minimum divergence:
The model p_{λ^*} in the parametric family of log-linear distributions that maximizes the likelihood $L(\lambda) = q[\ln p_\lambda]$ of the training data is the probability distribution p^* that minimizes the divergence $D(p||p_0)$ subject to the constraints $p[\nu_i] = q[\nu_i], i = 1, \ldots, n$.

Solution: Setting $\lambda_i = 0$ for all $i = 1, \ldots, n$ yields an **initial model which minimizes the divergence to the reference model** p_0 over the set of models to which the constraints yet have to be applied. Note: For uniformly distributed p_0 the minimum divergence model is a maximum entropy model!

Experiments: Property Design for a German LFG

190 **basic configurational properties**, referring to characteristics of the c(onstituent)-structure and the f(eature)-structure of the LFG parses , e.g., to

- c-structure nodes, corresponding to standard production properties,
- c-structure subtrees, indicating argument versus adjunct attachment,
- f-structure attributes, corresponding to grammatical functions used in LFG,
- atomic attribute-value pairs in f-structures,
- complexity of the phrase being attached to, indicating high and low attachment,
- non-right-branching behavior of nonterminal nodes,
- non-parallelism of coordinations.

(See Johnson, Geman, Canon, Chi, Riezler ACL'99; Johnson, Riezler NAACL'00).

Class-Based Grammar Lexicalization

45 **lexicalization properties** for the central grammatical relations in each parse. Let $v_r(x)$ be the verbal head and $n_r(x)$ the nominal head of grammatical relation r in parse x. Then a lexicalized property ν_r for grammatical relation r is

$$\nu_r(x) = \begin{cases} 1 & \begin{array}{l} \text{if } f_c(v_r(x), n_r(x)) \geq \\ f_c(v_r(x'), n_r(x')) \ \forall x' \in X(y), \end{array} \\ 0 & \text{otherwise.} \end{cases}$$

The ν_r **pre-disambiguate** the parses $x \in X(y)$ of a sentence y according to

$$f_c(v, n) = \max_{c \in C} p(c|v, n)(f(v, n) + 1),$$

the frequency of a (v, n)-pair weighted by the best estimate of the class-membership probability $p(c|v, n)$ of a dyadic clustering model (Prescher, Riezler, Rooth COLING '00; Rooth, Riezler, Prescher, Carroll, Beil ACL'99).

Training and Test Data

Training corpora: Constructed by parsing German newspaper text with an LFG Grammar, implemented in the Xerox Linguistic Environment (XLE).

- **Incomplete data corpus:** Restriction to low-ambiguity data enabled by highly constrained grammar: 36,000 sentences with at most 20 parses, in sum 250,000 parses, average sentence length 10.5,
- **Automatically constructed parsebank:** 4,000 unambiguously parsed sentences, average sentence length 7.5.

Test corpora:

- 550 manually disambiguated examples from foreign language learner's grammar (Helbig-Buscha), average ambiguity 5.4, average sentence length 7.5,
- 375 manually disambiguated examples held-out from newspaper corpus, average ambiguity 25, average sentence length 11.2.

Evaluation Tasks and Measures

Tasks:

- Exact match task = exact correspondence of full c/f-structure pair of most probable parse and manually selected correct parse.
- Frame match task = exact correspondence only of subcategorization frame of main verb of most probable parse and correct parse.

Measures:

$$\text{Precision} = \frac{\#\text{correct}}{\#\text{correct}+\#\text{incorrect}},$$

$$\text{Effectiveness} = \frac{\#\text{correct}}{\#\text{correct}+\#\text{incorrect}+\#\text{don't know}}.$$

"Correct" and "incorrect" specifies a success/failure on the respective evaluation tasks; "don't know" cases are cases where the system is unable to make a decision.

Exact-Match Evaluation

Results on Helbig-Buscha corpus with random baseline of about 33%:

exact match evaluation	basic model	lexicalized model	selected + lexicalized model
complete-data estimation	P: 68 E: 59.6	P: 73.9 E: 71.6	P: 74.3 E: 71.8
incomplete-data estimation	P: 73 E: 65.4	P: 86 E: 85.2	P: 86.1 E: 85.4

Best results on newspaper corpus with random baseline of about 17%:
60.1% precision, 57.9% effectiveness.

Frame-Match Evaluation

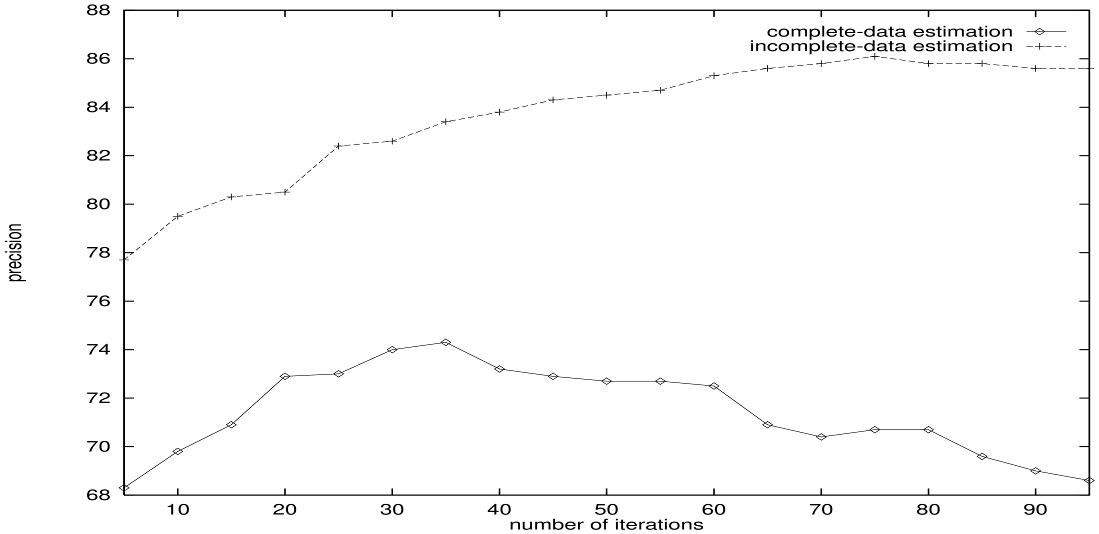
Results on newspaper corpus with random baseline of about 70%:

frame match evaluation	basic model	lexicalized model	selected + lexicalized model
complete-data estimation	P: 80.6 E: 70.4	P: 82.7 E: 76.4	P: 83.4 E: 76
incomplete-data estimation	P: 84.5 E: 73.1	P: 88.5 E: 84.9	P: 90 E: 86.3

Best results on Helbig-Buscha corpus with random baseline of about 80%:
97.6% precision, 96.7% effectiveness (Caution: Possibly high number of examples which are unambiguous in subcat frames!)

Controlling Hidden Parameters

Optimal iteration numbers were decided by repeated evaluation at every fifth iteration. Precision of lexicalized and selected models on exaxt match task plotted against number of IM iterations:



Comparison of starting values: Initialization with uniform model shows an increase in precision of 10-40% over models with randomly chosen starting values.

Conclusion

- **Estimation from large corpora of unannotated data does help statistical disambiguation with SUBGs**, e.g., using about 10 times more sentences and about 100 times more parses for unsupervised training than for training from unambiguous sentences can improve precision by 10%.

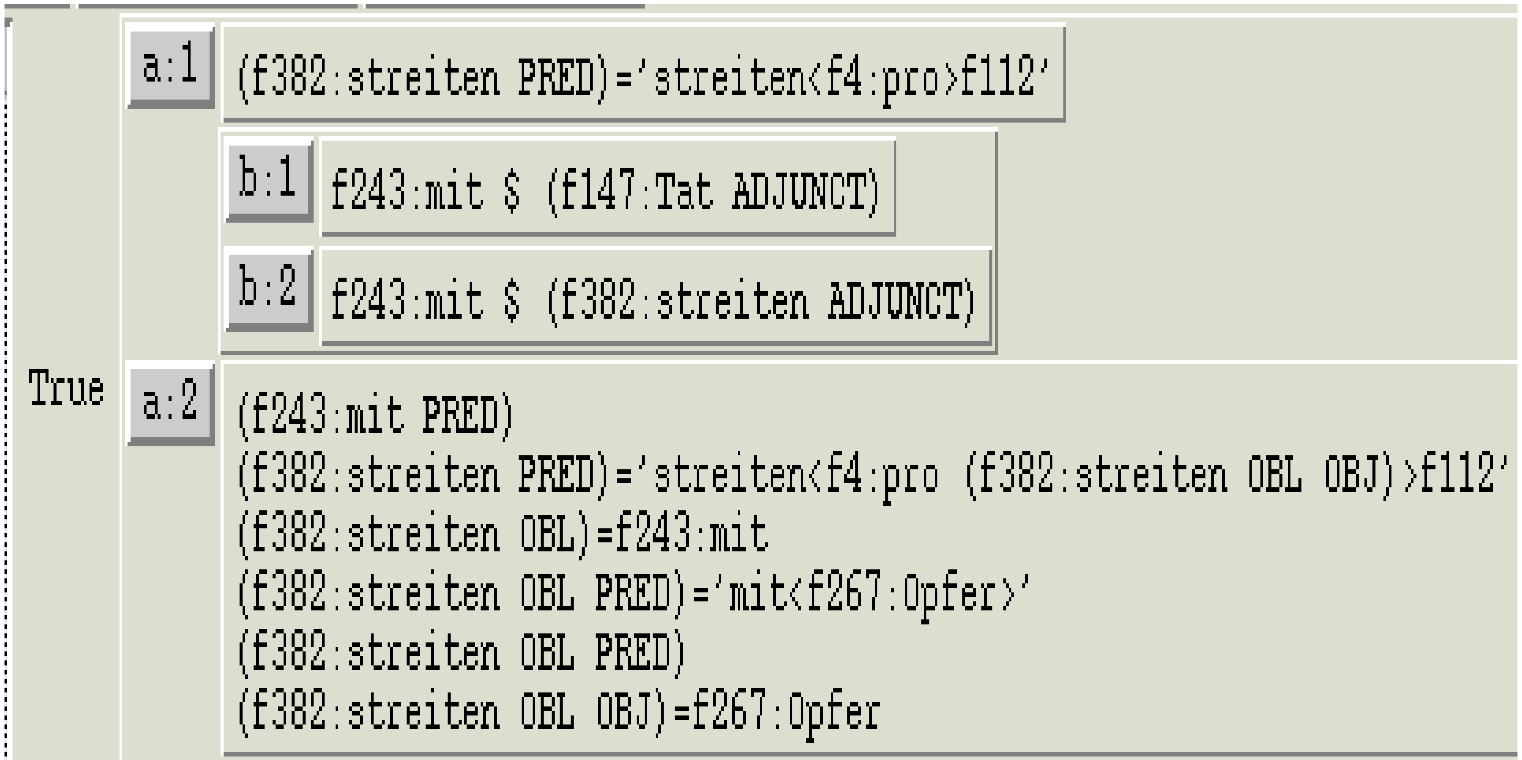
- **Unsupervised class-based lexicalization** can yield a **further gain** in precision of 10%.

- An **important prerequisite** for precise statistical disambiguation with SUBGs is the **low ambiguity rate of hand-crafted UBGs**, i.e., the parse space to be searched in estimation is only "**mildly incomplete**"!

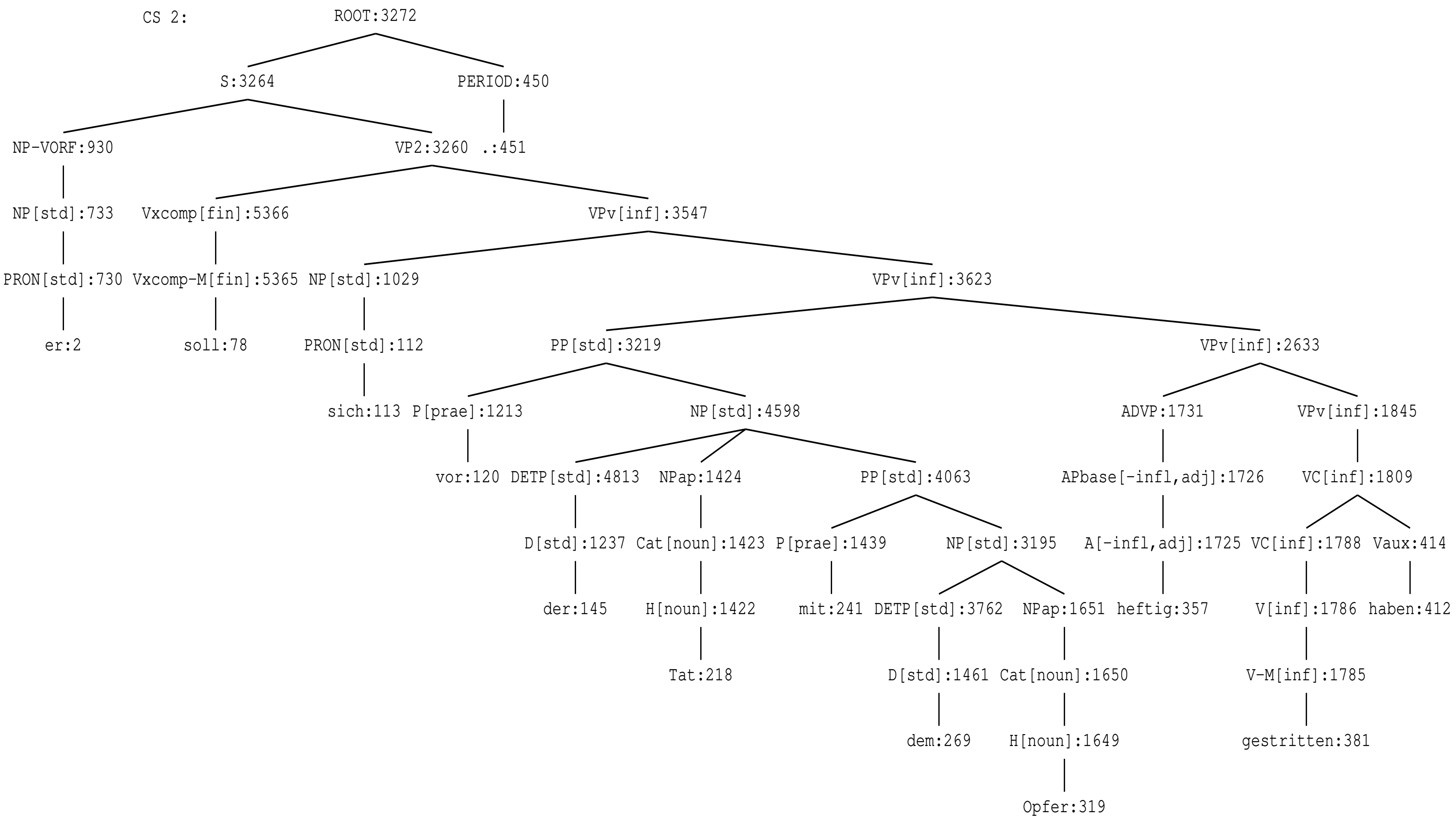
A UBG Example: Lexical-Functional Grammar

“Er	soll	sich	vor	der	Tat	mit	dem	Opfer	heftig	gestritten	haben.”
He	is said		before	the	deed	with	the	victim	heavily	quarreled	have.

He is said to have quarreled heavily with the victim before the deed.



First Analysis (c-structure)

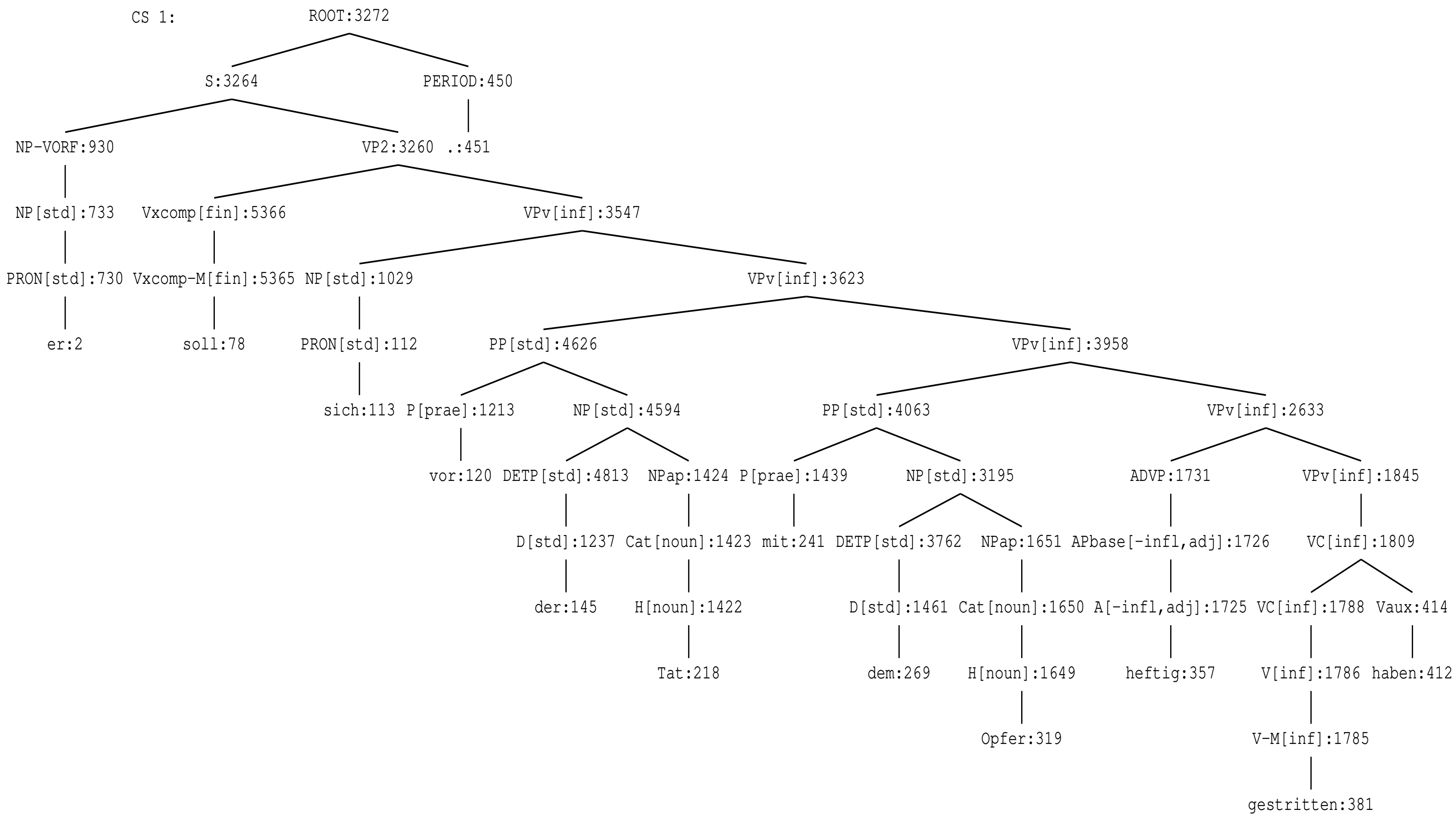


First Analysis (partial f-structure)

"Er soll sich vor der Tat mit dem Opfer heftig gestritten haben".

```
[PRED      'sollen<[381:streiten]>[2:pro]'
 SUBJ   2    [PRED 'pro']
        730
        733
        930
 XCOMP  412  [PRED    'streiten<[2:pro]>[112]'
        414   SUBJ    [2:pro]
        381   ADJUNCT {  120 [PRED  'vor<[145:Tat]>'
        1785            1213  OBJ   218 [PRED    'Tat'
        1786            3219       1422  ADJUNCT {  241 [PRED  'mit<[269:Opfer]>'
        1788                       1423            1439  OBJ   319 [PRED 'Opfer']
        1809                       1424            4063       1649
        1845                        145                       1650
        2633                       1237                       1651
        3623                       4813                        269
        3547                       4598                       1461
                                                              3762
                                                              3195 ] } ] ]
                          357 [PRED 'heftig']
                         1725
                         1726
                         1731 } ]
451
450
78
5365
5366
3260
3264
3272 ]
```

Second Analysis (c-structure)

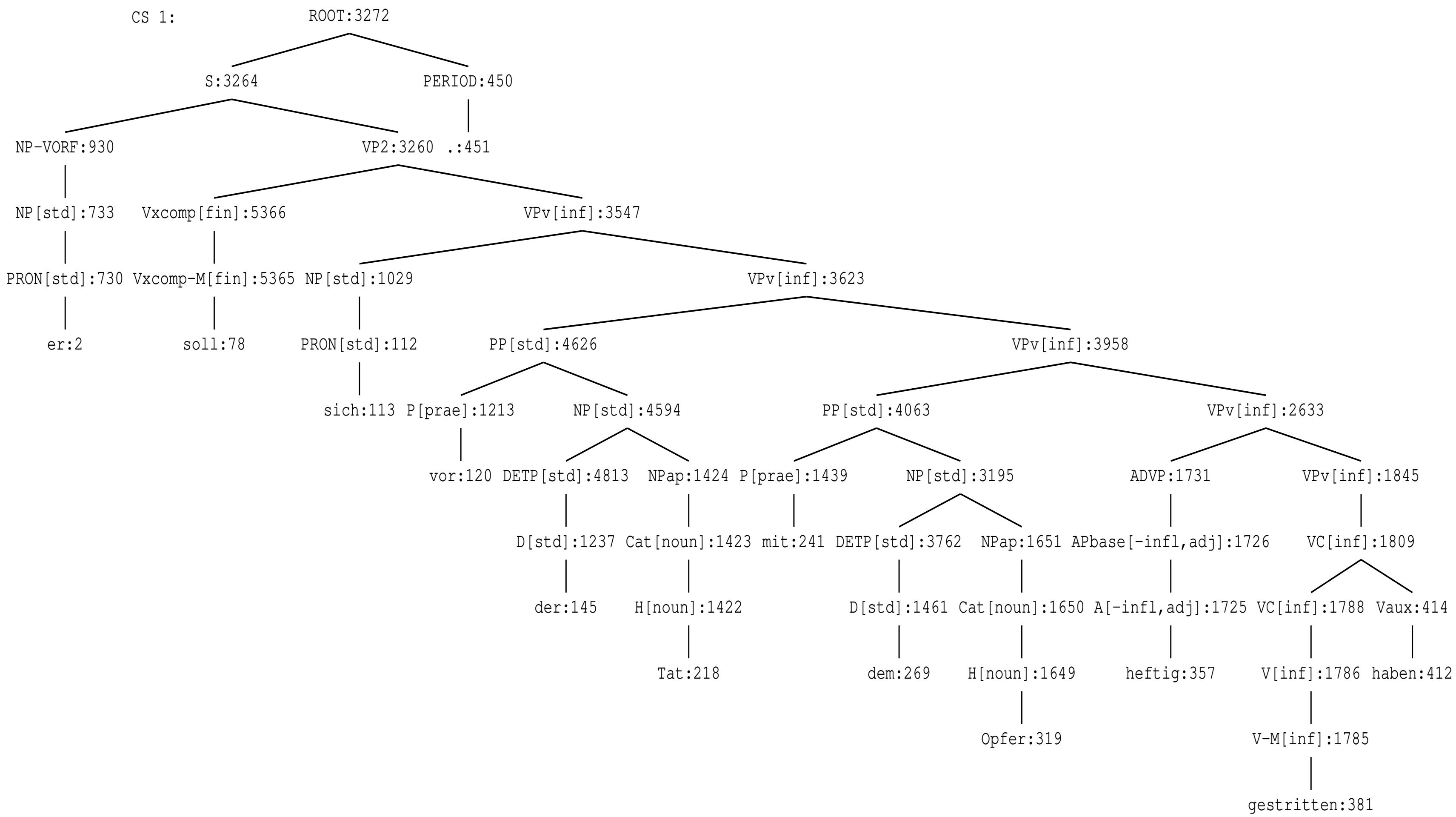


Second Analysis (partial f-structure)

"Er soll sich vor der Tat mit dem Opfer heftig gestritten haben"

```
     [PRED     'sollen<[381:streiten]>[2:pro]'                              ]
     [               2[PRED 'pro']                                           ]
     [SUBJ         730[          ]                                           ]
     [             733[          ]                                           ]
     [             930[          ]                                           ]
     [                [PRED     'streiten<[2:pro]>[112]'                    ]]
     [                [SUBJ     [2:pro]                                     ]]
     [                [         {     [PRED    'vor<[145:Tat]>'  ]          }]]
     [                [         {     [          218[PRED 'Tat']  ]          }]]
     [                [         {     [         1422[           ]  ]          }]]
     [                [         {     [         1423[           ]  ]          }]]
     [                [         {     [OBJ      1424[           ]  ]          }]]
     [                [         {     [          145[           ]  ]          }]]
     [                [         {  120[         1237[           ]  ]          }]]
     [                [         { 1213[         4813[           ]  ]          }]]
     [                [         { 4626[         4594[           ]  ]          }]]
     [                [         {     [PRED    'mit<[269:Opfer]>' ]          }]]
     [XCOMP           [         {     [          319[PRED 'Opfer']]          }]]
     [                [ADJUNCT  {     [         1649[             ]]          }]]
     [             412[         {     [         1650[             ]]          }]]
     [             414[         {     [         1651[             ]]          }]]
     [             381[         {     [OBJ       269[             ]]          }]]
     [            1785[         {     [         1461[             ]]          }]]
 451 [            1786[         {  241[         3762[             ]]          }]]
 450 [            1788[         { 1439[         3195[             ]]          }]]
  78 [            1809[         { 4063[                            ]          }]]
5365 [            1845[         {  357[PRED 'heftig']                        }]]
5366 [            2633[         { 1725[              ]                       }]]
3260 [            3958[         { 1726[              ]                       }]]
3264 [            3623[         { 1731[              ]                       }]]
3272 [            3547[                                                      ]]
```

Third Analysis (c-structure)



Third Analysis (partial f-structure)

"Er soll sich vor der Tat mit dem Opfer heftig gestritten haben"

```
      [PRED      'sollen<[381:streiten]>[2:pro]'                                              ]
      [          2[PRED 'pro']                                                                ]
      [SUBJ    730[          ]                                                                ]
      [        733[          ]                                                                ]
      [        930[          ]                                                                ]
      [           [PRED     'streiten<[2:pro], [269:Opfer]>[112]'                            ]]
      [           [SUBJ     [2:pro]                                                          ]]
      [           [         [PRED     'mit<[269:Opfer]>'        ]                            ]]
      [           [         [     319[PRED 'Opfer']             ]                            ]]
      [           [         [    1649[             ]            ]                            ]]
      [           [         [    1650[             ]            ]                            ]]
      [           [OBL      [OBJ 1651[             ]            ]                            ]]
      [           [         [     269[             ]            ]                            ]]
      [           [      241[    1461[             ]            ]                            ]]
      [           [     1439[    3762[             ]            ]                            ]]
      [           [     4063[    3195[             ]            ]                            ]]
      [           [         {    [PRED   'vor<[145:Tat]>'       ]}                           ]]
      [XCOMP      [         {    [    218[PRED 'Tat']           ]}                           ]]
      [        412[         {    [   1422[           ]          ]}                           ]]
      [        414[         {    [   1423[           ]          ]}                           ]]
      [        381[         {    [OBJ1424[           ]          ]}                           ]]
      [       1785[         {    [    145[           ]          ]}                           ]]
  451 [       1786[ADJUNCT  { 120[   1237[           ]          ]}                           ]]
  450 [       1788[         {1213[   4813[           ]          ]}                           ]]
   78 [       1809[         {4626[   4594[           ]          ]}                           ]]
 5365 [       1845[         {                                    }                           ]]
 5366 [       2633[         { 357[PRED 'heftig']                 }                           ]]
 3260 [       3958[         {1725[              ]                }                           ]]
 3264 [       3623[         {1726[              ]                }                           ]]
 3272 [       3547[         {1731[              ]                }                           ]]
```

Class 23: “Financial Transactions”

Class 23 PROB 0.0202		0.1309	0.0899	0.0414	0.0341	0.0304	0.0182	0.0143	0.0104	0.0097	0.0088	0.0085	0.0081	0.0079	0.0074	0.0069	0.0059	0.0057	0.0057	0.0053	0.0050	0.0049	0.0049	0.0046	0.0045	0.0044	0.0040	0.0040	0.0036	0.0036	0.0033
		Prozent	Mark	Million	Geld	DM	Milliarde	Dollar	Kosten	Teil	Gewinn	mehr	Miete	Haelfte	Gebiet	Dividende	Jahr	Hoehe	Drittel	Preis	Niveau	Umsatz	Anteil	Zins	Betrag	Pfennig	Summe	Steuer	Schaetzung	Kost	Platz
0.0378	erreichen.a.a	●	●	●		●	●	●		●		●		●	●		●	●	●		●	●	●		●		●				●
0.0369	kosten.a.a	●	●	●	●	●	●	●		●		●	●	●					●	●				●		●	●				●
0.0255	betragen.a.a	●	●	●		●	●	●	●			●	●	●		●	●	●	●				●			●	●				
0.0203	ausmachen.a.a	●	●	●			●	●		●	●	●	●	●	●		●		●				●			●					
0.0196	liegen.a.p-d-adj	●	●	●		●	●	●			●	●	●	●	●		●			●	●			●	●		●		●	●	
0.0171	zahlen.a.a	●	●	●	●	●	●	●		●		●	●	●		●	●		●	●			●	●	●	●	●	●		●	
0.0141	betragen.a.n	●	●	●		●	●	●	●		●	●	●			●	●	●		●		●	●	●				●		●	
0.0135	stammen.a.p-d-obl	●								●		●		●	●		●	●							●						
0.0120	bezahlen.a.a	●	●	●	●		●	●		●			●	●	●	●			●	●				●		●	●	●		●	
0.0115	erzielen.a.a	●	●				●				●	●		●						●		●									●
0.0098	beteiligen.a.p-d-adj	●	●	●		●	●		●								●					●							●		
0.0097	erhalten.a.a	●	●	●	●	●	●	●		●		●		●		●	●		●	●			●	●	●	●	●				●
0.0096	halten.a.a	●	●		●	●				●	●	●	●	●	●		●	●	●	●	●	●	●				●	●		●	●
0.0094	vorsehen.a.a	●	●	●	●		●	●		●		●		●													●				
0.0090	liegen.a.p-d-obl	●	●			●	●	●		●		●	●	●	●		●	●			●				●				●		●
0.0084	verdienen.a.a	●	●	●	●		●	●				●	●	●			●		●	●				●		●	●				●