

# A Statistical Grammar as Empirical Resource for Inducing Semantic Phenomena

---

Sabine Schulte im Walde  
Universität des Saarlandes

Heike Zinsmeister  
Universität Tübingen

DGfS Annual Meeting 2006, February 22-24

*AG 13: Lexikalisch-semantische Ressourcen zur Sprachdokumentation  
und maschinellen Sprachverarbeitung*

# Overview

---

1. HeadLex PCFG framework and properties
2. Implementation of the German HeadLex PCFG
3. Lexical acquisition from the German HeadLex PCFG:
  - » semantic verb classes
  - » (non-)compositionality of noun compounds
4. Discussion of design and implementation

# CFG → PCFG → HeadLex PCFG

---

## CFG

[rule]

VP → V

VP → V NP

VP → V NP NP

## PCFG

[relevance]

407,220

VP → V

706,754

VP → V NP

167,101

VP → V NP NP

## HeadLex PCFG

[lexicalisation]

1

$VP_{[give]} \rightarrow V'_{[give]}$

84

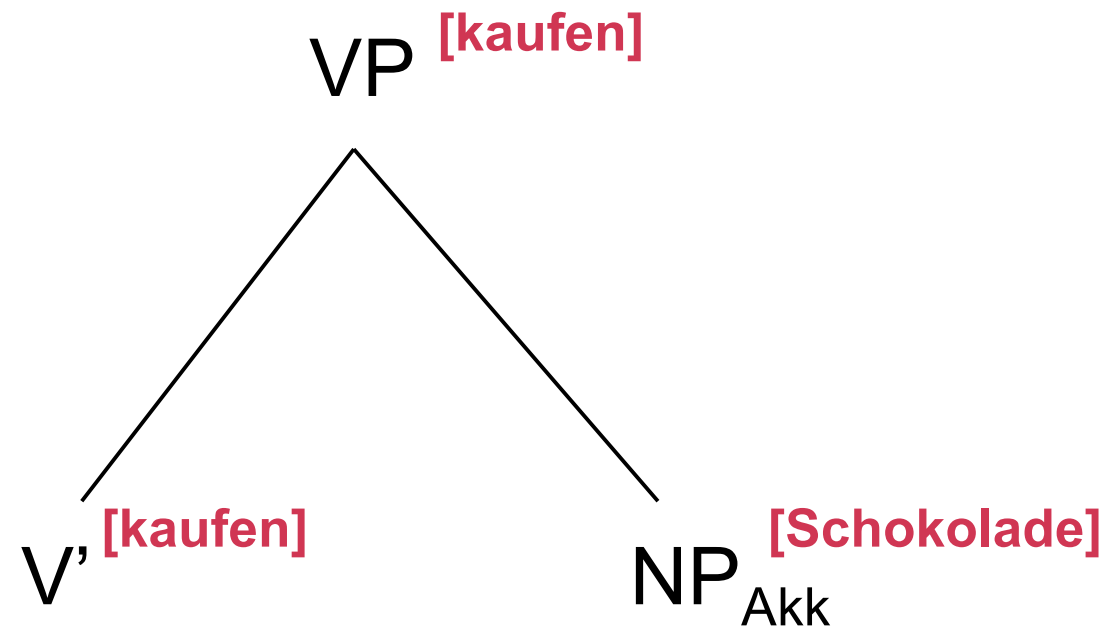
$VP_{[give]} \rightarrow V'_{[give]} NP_{[present]}$

242

$VP_{[give]} \rightarrow V'_{[give]} NP_{[friend]} NP_{[present]}$

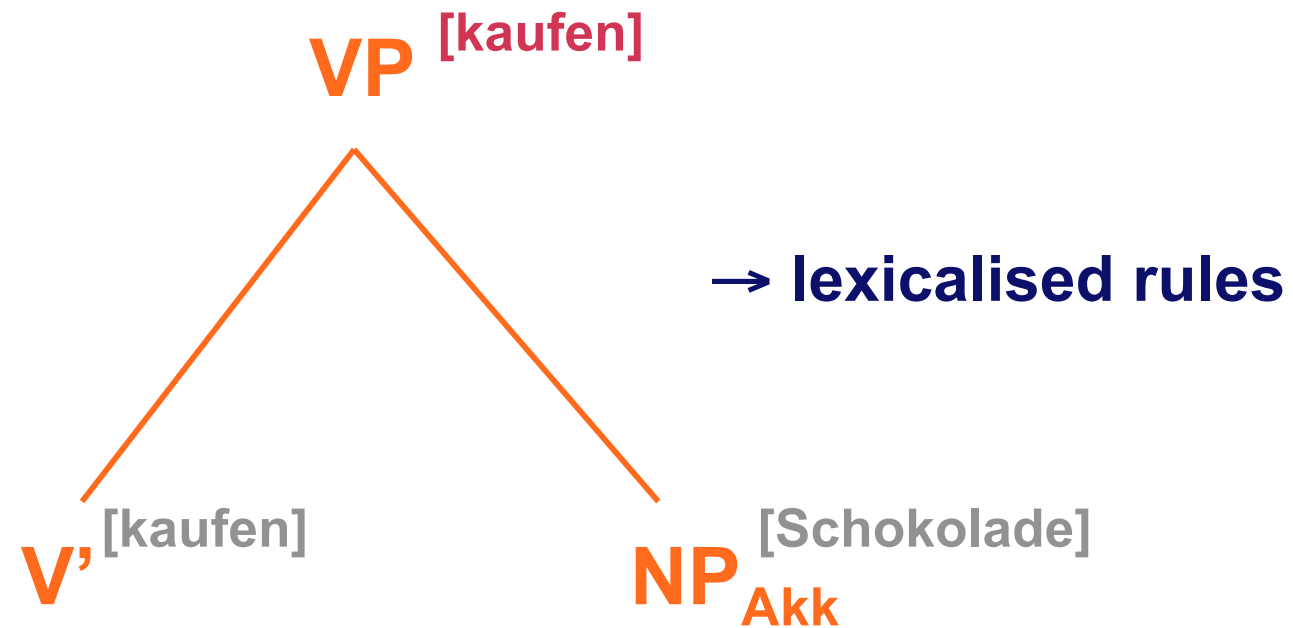
# HeadLex PCFG: Tree Parameters

---



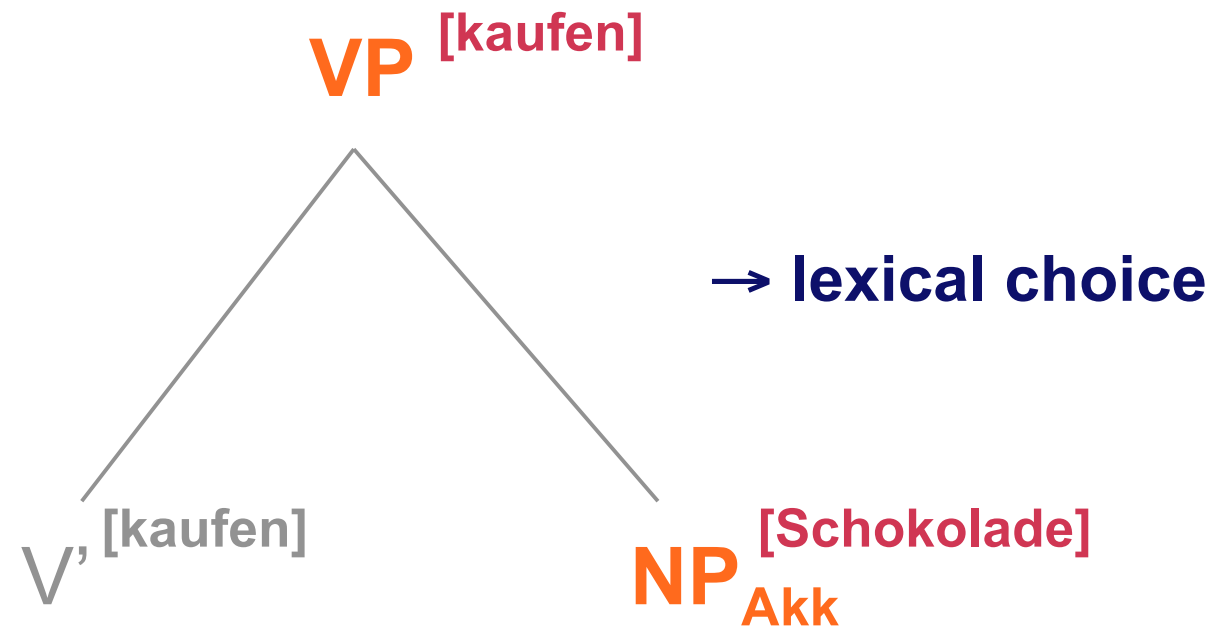
# HeadLex PCFG: Tree Parameters

---



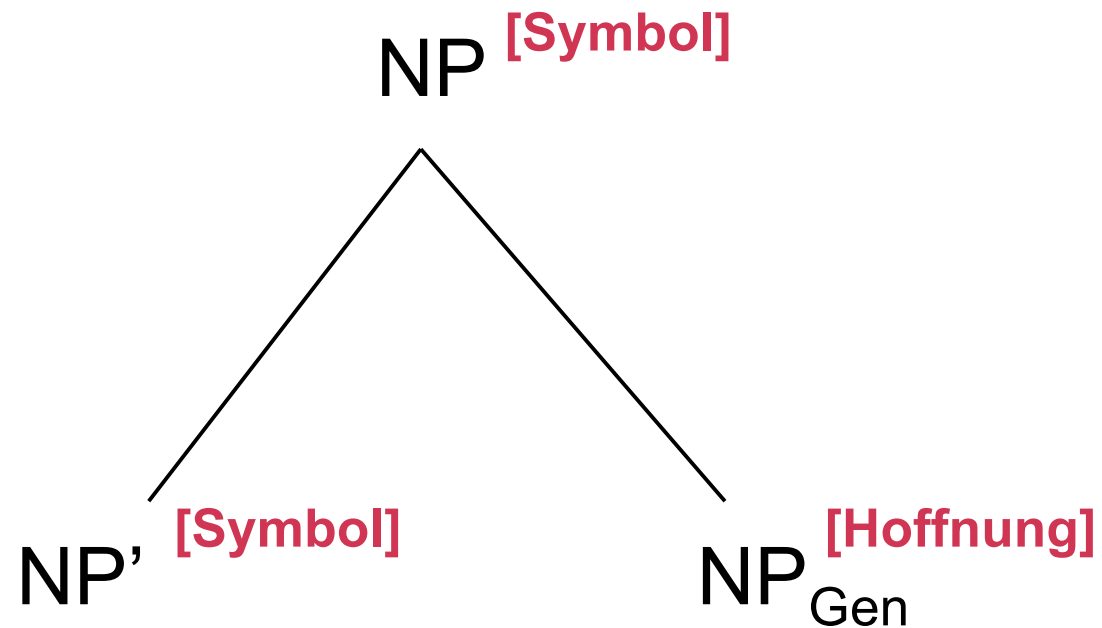
# HeadLex PCFG: Tree Parameters

---



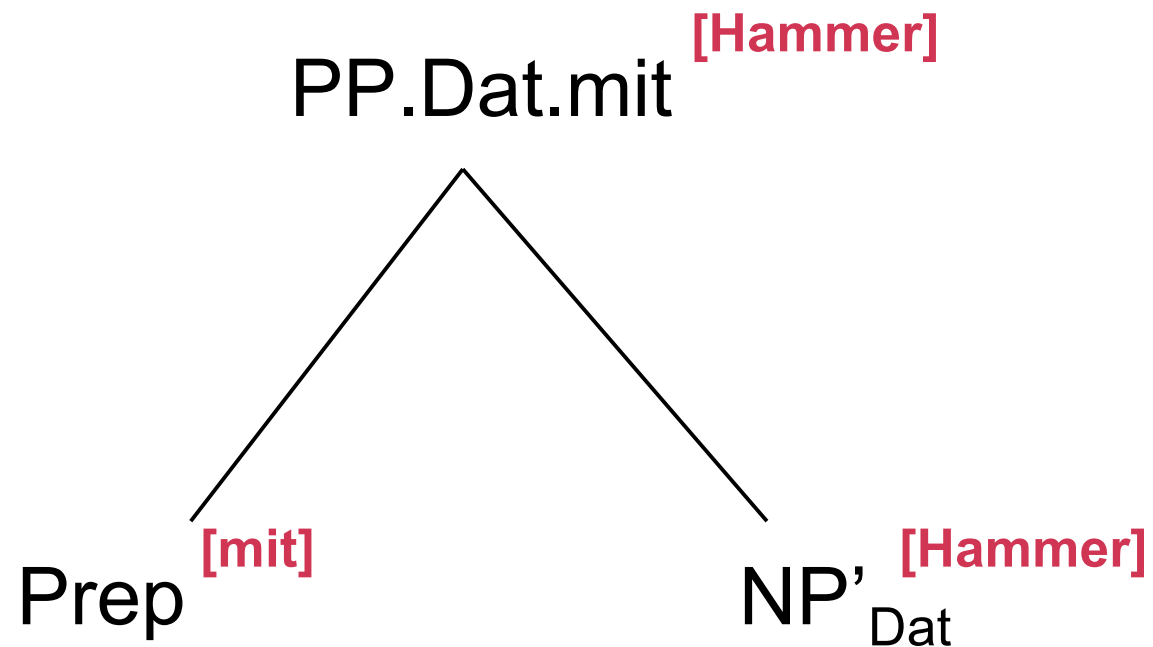
# HeadLex PCFG: Tree Parameters

---



# HeadLex PCFG: Tree Parameters

---



# HeadLex PCFG: Parameter Estimation

---

- **Unsupervised training** based on an unannotated corpus
- **Expectation-Maximisation** (EM) algorithm (Baum, 1972)
- Alternation between assessing frequencies and estimating probabilities
- **E-step = estimation** → calculation of expected values  
evaluates probability distribution for data given the model parameters from the previous iteration
- **M-step = maximisation** → calculation of max. likelihood  
finds the new parameter set that maximises the distribution

# Grammar Development and Training Strategy

---

1. Definition of **CFG rules** with head-specification
2. Assigning uniform frequencies to CFG rules  
(extension of **CFG to PCFG**)
3. **Unlexicalised training** of PCFG:  
one iteration on 82 million words
4. **Lexicalisation** of PCFG (bootstrapping HeadLex PCFG)  
on 19 million words
5. **Lexicalised training** of HeadLex PCFG:  
three iterations on 35 million words

# Grammar Design Decisions

---

- Focus on verb subcategorisation:  
detailed rules, additional category levels
- Modeling mass phenomena, disregarding selected irregularities, e.g. adjectival and genitive arguments
- Disregarding rule-inflating structures, e.g. coordination
- Disregarding fine-grained features, e.g. number, gender
- Concentration on binary rules ensures step-by-step analysis of verb phrase saturation → deep tree structure
- Overgeneralisation: high degree of ambiguity

# Number of CFG Grammar Rules

---

|              |        |
|--------------|--------|
| total        | 35,821 |
| verb-related | 33,671 |
| other        | 2,150  |

# Lexical Semantic Phenomena

---

- Selectional preferences
- Collocations and idiomatic expressions
- Compositional phenomena,  
e.g. noun compounds, particle verbs
- Semantic classes for verbs, nouns, adverbs, etc.
- Distinction of PP arguments vs. adjuncts
- Structural categorial preferences,  
e.g. predicative adverbs, clause adverbs
- Word order

# Lexical Semantic Phenomena

---

- Selectional preferences
- Collocations and idiomatic expressions
- **Compositional phenomena**,  
e.g. noun compounds, particle verbs
- **Semantic classes for verbs**, nouns, adverbs, etc.
- Distinction of PP arguments vs. adjuncts
- Structural categorial preferences,  
e.g. predicative adverbs, clause adverbs
- Word order

# Semantic Verb Classes

---

- Hypothesis:  
**verb behaviour ↔ verb meaning aspects**
- Distributional verb descriptions:  
**syntactic frames, PPs, selectional preferences**
- Clustering with k-Means algorithm
- Result: semantic verb classes
- Different feature description for particle verbs

# Subcategorisation Lexicon

---

- **C-<type> → S-<type>.<frame>**

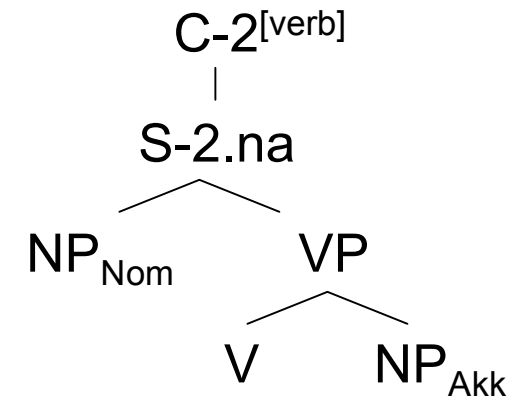
- Lexicalised Preferences:

$p_1$     C-2<sup>[verb]</sup> → S-2.frame<sub>1</sub>

$p_2$     C-2<sup>[verb]</sup> → S-2.frame<sub>2</sub>

...

$p_n$     C-2<sup>[verb]</sup> → S-2.frame<sub>n</sub>



# Subcategorisation Frames

---

|               |                                |
|---------------|--------------------------------|
| <b>n</b>      | noun phrase (case: nominative) |
| <b>a</b>      | noun phrase (case: accusative) |
| <b>d</b>      | noun phrase (case: dative)     |
| <b>r</b>      | reflexive pronoun              |
| <b>p</b>      | prepositional phrase           |
| <b>x</b>      | expletive <i>es</i>            |
| <b>i</b>      | non-finite clause              |
| <b>s-2</b>    | finite verb second clause      |
| <b>s-dass</b> | finite <i>dass</i> -clause     |
| <b>s-ob</b>   | finite <i>ob</i> -clause       |
| <b>s-w</b>    | indirect <i>wh</i> -question   |
| <b>k</b>      | copula construction            |

*Examples:*

- *na*
- *np*
- *npr*
- *nds-dass*

# Subcategorisation Frame Distribution

---

*glauben*

`to think, to believe`

| Frame Type     | Freq  |
|----------------|-------|
| <b>ns-dass</b> | 1,929 |
| <b>ns-2</b>    | 1,888 |
| <b>np</b>      | 687   |
| <b>n</b>       | 608   |
| <b>na</b>      | 555   |
| <b>ni</b>      | 346   |
| <b>nd</b>      | 234   |
| <b>nad</b>     | 160   |
| <b>nds-2</b>   | 70    |
| <b>nai</b>     | 62    |

# Clustering Example: Random Input

---

- konsumieren kriegen vermuten
- anfangen
- ahnen bekanntgeben bestehen **fahren fliegen** liegen nieseln pochen
- aufhören **bekommen erhalten** essen insistieren regnen segeln vermitteln
- beginnen freuen interpretieren
- rudern saufen schneien ärgern
- eröffnen folgen glauben
- zustellen
- charakterisieren dämmern stehen
- blitzen verkünden wissen
- beschreiben **dienen** donnern schließen **unterstützen**
- beenden darstellen **liegen sitzen**
- ankündigen denken enden lesen schicken öffnen
- beharren bringen erlangen helfen trinken

# Clustering Example: Output

---

- **ahnen vermuten wissen** - *Propositional Attitude*
- **denken glauben** - *Propositional Attitude*
- **anfangen aufhören beginnen beharren enden insistieren rudern** - *Aspect*
- **liegen sitzen stehen** - *Position*
- **dienen folgen helfen** - *Support*
- **nieseln regnen schneien** - *Weather*
- **dämmern**
- **blitzen donnern segeln** - *Weather*
- **bestehen fahren fliegen pochen** - *Insistence, Manner of Motion*
- **freuen ärgern** - *Emotion*
- **essen konsumieren saufen trinken verkünden** - *Consumption*
- **bringen eröffnen lesen liefern schicken schließen vermitteln öffnen** - *Supply*
- **ankündigen beenden bekanntgeben bekommen beschreiben charakterisieren darstellen erhalten erlangen interpretieren kriegen unterstützen** - *Description, Obtain*
- **zustellen**

# Noun Compounds

---

- **Goal:**
  - » automatic identification of lexicalised German noun+noun compounds
  - (1) Rücktrittsangebot – Angebot  
‘offer of resignation’ – ‘offer’
  - (2) Wahlkampf – Kampf  
‘election campaign’ – ‘fight’
- **Method:**
  - » study of noun+verb collocations
  - » compare collocation preferences of compound and of head noun

# Experiment

---

- Candidate extraction from trained HeadLex PCFG
  - » (Compound-)Noun+Verb
- Morphological analysis by DMOR (Schiller 1995)
  - » Bergwerks → Berg=Werk
- Head nouns with high number of compound types
  - » 85 head nouns → 7,518 compound types
- Identification of lexicalised compounds
  - » frequency-based extraction
  - » log-likelihood scores

# Shared Collocations

---

- *Fest* ('party'): compositional compounds
- 94 compound types:  
*Sommerfest* ('summer party'), *Straßenfest* ('street party'),  
*Starkbierfest* ('beer feast'), etc.

| <u>verb</u>  | <u>freq(Fest,verb)</u> | <u>co-occurring compounds</u> |
|--------------|------------------------|-------------------------------|
| feiern       | 88                     | 52.13%                        |
| eröffnen     | 21                     |                               |
| planen       | 13                     |                               |
| veranstalten | 12                     |                               |
| machen       | 11                     |                               |
| organisieren | 6                      |                               |
| besuchen     | 6                      | 24.87%                        |

→ 77% of all compound occurrences

# No Shared Collocations

---

- 40 candidates, extraction and categorisation:

**$f(\text{compound}, \text{verb}_i) > 5$  &  $f(\text{head noun}, \text{verb}_i) = 0$**

» **idiomatic collocations** (29):

Alarmanlage, Anhaltspunkt, Autobahn, Besatzungsmitglied, Bußgeld, Eigentumsverhältnis(se), Fahrbahn, Feindbild, Feuerwerk, Gangart, Größenordnung, Handwerk, ...

» **mixed but rather idiomatic** (9):

Arbeitskampf, Arbeitskraft, Autofahrer, Grenzwert, Kopfgeld, Motorradfahrer, Ozonwert, Sozialhilfe, Wahlkampf

» **sharing of collocations** (2):

Mißtrauensantrag, Pressekonferenz

# Extraction of Significant Pairs

---

- Association measure:  
**log-likelihood ratio** (Dunning, 1993)
  - » compare observed frequency with co-occurrence by chance
  - » high deviance → significant combination
  - » threshold for frequency and log-likelihood value

# Results: Log-Likelihood Scores

---

| <b>Head</b> | <b>#c-types</b> | <b>lexicalised c-types</b>  |
|-------------|-----------------|---|
| Abend       | 65              | Elternabend, Feierabend, Lebensabend  |
| Art         | 35              | Eigenart, Gangart, Mundart, Sportart, Spielart, Tonart  |
| Fest        | 94              | none  |
| Kampf       | 64              | Wahlkampf   |
| Stellung    | 46              | Fragestellung, Hilfestellung, Problemstellung, Schlüsselstellung, Themenstellung                                |
| Werk        | 192             | Bauwerk, Bergwerk, Fachwerk, Feuerwerk, Handwerk, Kraftwerk, Laufwerk, Mundwerk, Netzwerk, Triebwerk, Schuhwerk |

# Results: Log-Likelihood Scores

---

| <b>Head</b> | <b>#c-types</b> | <b>lexicalised c-types</b>  |
|-------------|-----------------|---|
| Abend       | 65              | Elternabend, <b>Feierabend</b> , <b>Lebensabend</b>   |
| Art         | 35              | <b>Eigenart</b> , <b>Gangart</b> , Mundart, <b>Sportart</b> , <b>Spielart</b> , Tonart  |
| Fest        | 94              | none  |
| Kampf       | 64              | <b>Wahlkampf</b>  |
| Stellung    | 46              | Fragestellung, <b>Hilfestellung</b> ,<br>Problemstellung, Schlüsselstellung,<br>Themenstellung  |
| Werk        | 192             | <b>Bauwerk</b> , Bergwerk, Fachwerk,<br><b>Feuerwerk</b> , <b>Handwerk</b> , <b>Kraftwerk</b> ,<br>Laufwerk, Mundwerk, <b>Netzwerk</b> ,<br><b>Triebwerk</b> , <b>Schuhwerk</b> |

# Discussion: Design and Implementation

---

- Idiosyncratic design and exploitation
- Design criteria (examples):
  - » development and training effort
  - » lemmatisation (sparse data vs. lexeme features)
  - » mass vs. detailed phenomena description
  - » deep vs. flat analysis structures
  - » ambiguity in grammar rules
- Comparison:
  - » lexical acquisition from treebanks / Viterbi parses
- No contexts or examples for lexical phenomena