# Topological Field Chunking in German

Jorn Veenstra, Frank H. Müller, Tylman Ule
[veenstra,fhm,ule]@sfs.uni-tuebingen.de

ESSLLI Summerschool

Workshop on Machine Learning Approaches in Computational Linguistics

August 7, 2002

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# German $\neq$ English

- Sentence structure in English lends itself to chunk analysis, followed by grammatical role assignment.
- In other Germanic languages the sentence structure can be different. NPs structure is different, VPs are less clustered.
- $[_{NP}$ The strategy $]$ $[_{VP}$ proposed $]$ by $[_{NP}$ the USA $]$ $[_{VP}$ was accepted $]$ by $[_{NP}$ the NATO $]$ yesterday.
- $[_{NP}$ Die von den USA vorgeschlagene Strategie $]$ $[_{VP}$ wurde $]$ gestern von $[_{NP}$ der NATO $]$ $[_{VP}$ übernommen $]$.

# Topological Field Theory (TopF)

- Topological Field Theory gives a handle to divide the sentence into labeled verbal and non-verbal parts:
  - Non-Verbal parts: VF: Vorfeld, MF: Mittelfeld, NF: Nachfeld
  - Verbal parts: CF: Complementizer, LK: Linke Klammer, RK: Rechte Klammer
- $[_{VF}$ `Die von den USA vorgeschlagene Strategie` $]$ $[_{LK}$ `wurde` $]$ $[_{MF}$ `gestern von der NATO` $]$ $[_{RK}$ `übernommen` $]$`.`

# Overview

- Sentence Structure in German
- Topological Field Theory
- The Data: TüBa-D/Z
- Three TopFChunkers: FSA, PCFG, MBL
- Results
- Future Research

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# Sentence Structure in German I

- Verbs can be far apart in German sentences
- In so-called V2 sentences the finite verb is always in second position. The other verbal elements is further up in the sentence.
- Ich $[_{VP}$ habe $]$ gestern mit meinen Freunden Spaghetti $[_{VP}$ gegessen $]$.
- I have yesterday with my friends spaghetti eaten.

# Sentence Structure in German II

- The finite verb will stay in second position:
- Mit meinen Freunden $[_{VP}$ habe $]$ ich gestern Spaghetti $[_{VP}$ gegessen $]$.
- With my friends have I yesterday spaghetti eaten.
- “Mit meinen Freunden” is in first position now, forcing the subject “Ich” after the verb.

# Sentence Structure in German III

- In the Verb-Last constructions, verbs are clustered at the end of the sentence. This construction occurs e.g. in relative clauses:
- Er glaubt, dass ich gestern mit meinen Freunden Spaghetti $[_{VP}$ gegessen habe $]$.
- He thinks that I yesterday with my friends spaghetti eaten have.

# Sentence Structure in German IV

- In yes/no-questions the finite verb is in first position, the so-called V1 sentences:
- $[_{VP}$ Hast $]$ du gestern mit deinen Freunden Spaghetti $[_{VP}$ gegessen $]$?
- Have you yesterday with your friends spaghetti eaten?

# Sentence Structure in German V

- There can be more after the last verbal chunk:
- Ich $[_{VP}$ habe $]$ gestern Spaghetti $[_{VP}$ gegessen $]$ in einer Pizzeria.
- I have yesterday spaghetti eaten in a pizzeria.

# Sentence Structure in German VI

- The finite verb forms its own constituent.
- However, the non-finite verbal constituent can contain many verbs, the so-called verb cluster.
- Das $[_{VP}$ sind $]$ allerdings Gelder, die vom Schatzmeister $[_{VP}$ hintergezogen worden sein sollen $]$ .
- These are, however, funds which by the treasurer defrauded been have should.

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

# Topological Fields

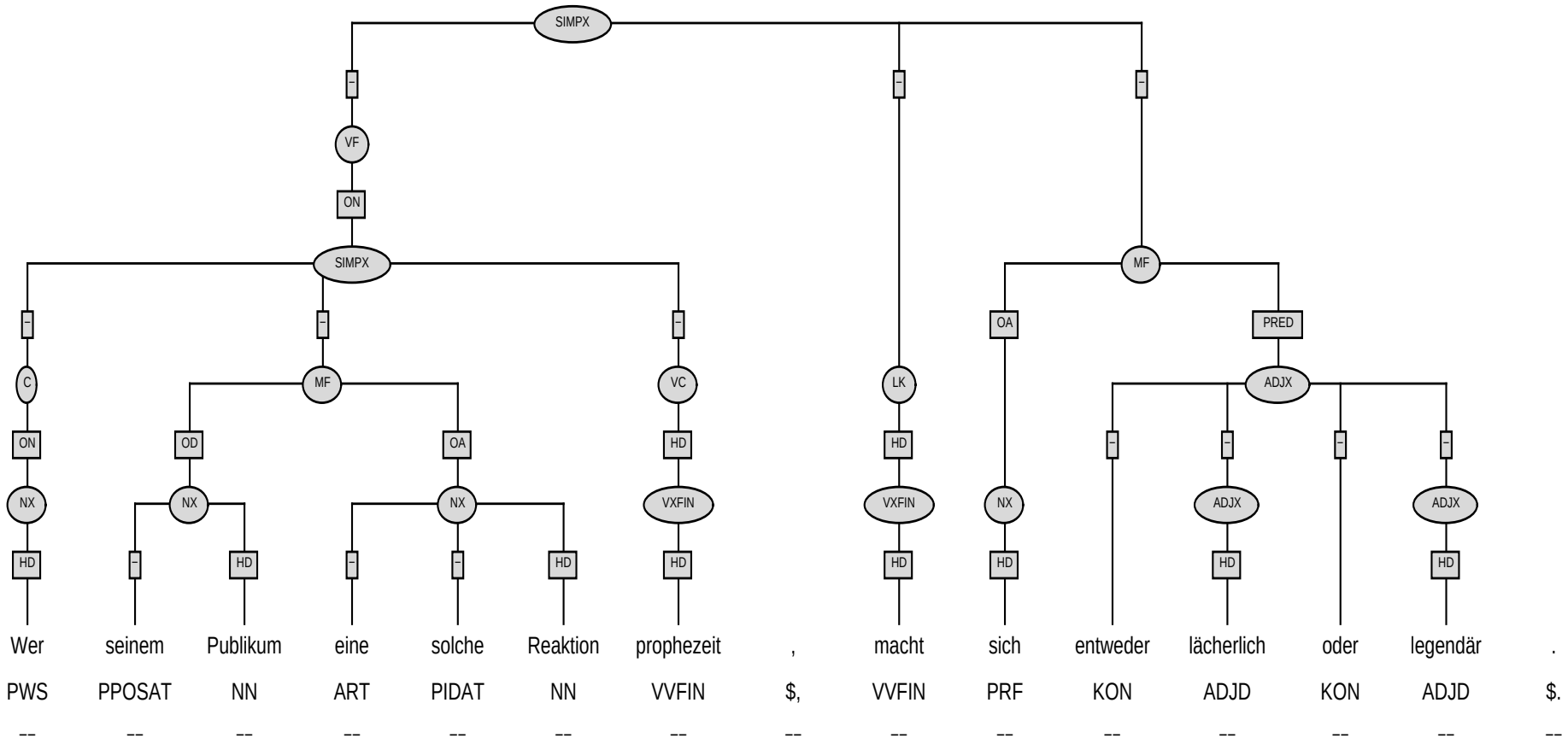
- Descriptive model of the constituent order in German(ic), the TopF structure gives the skeleton of the sentence.
- Topological fields describe sections with respect to the distributional properties of the verbs in German sentences.
- CF: complementizer, LK: left bracket, RK: right bracket, VF: initial field, MF: middle field, NF: final field
- The TopFs CF, LK and RK form the verbal frame relative to which the other fields are defined.
- The VF, MF and NF are defined relative to the verbal fields.

# Sentence Structure in German

| type | topological fields | | | | | |
|---|---|---|---|---|---|---|
| (VL) | | **compl. field (CF)** | (MF) | **verb complex (RK)** | (NF) |
| (V1) | | **finite verb (LK)** | (MF) | verb complex (RK) | (NF) |
| (V2) | **(VF)** | **finite verb (LK)** | (MF) | verb complex (RK) | (NF) |
| | | left part | | right part | |

Figure 1: VL=Verb-Last; V1=Verb-First; V2= Verb-Second; VF=Initial Field; MF=Middle Field; NF=Final Field

# An example of an analysed sentence




Who his audience a such reaction predicts, makes himself or ridiculous or legendary.

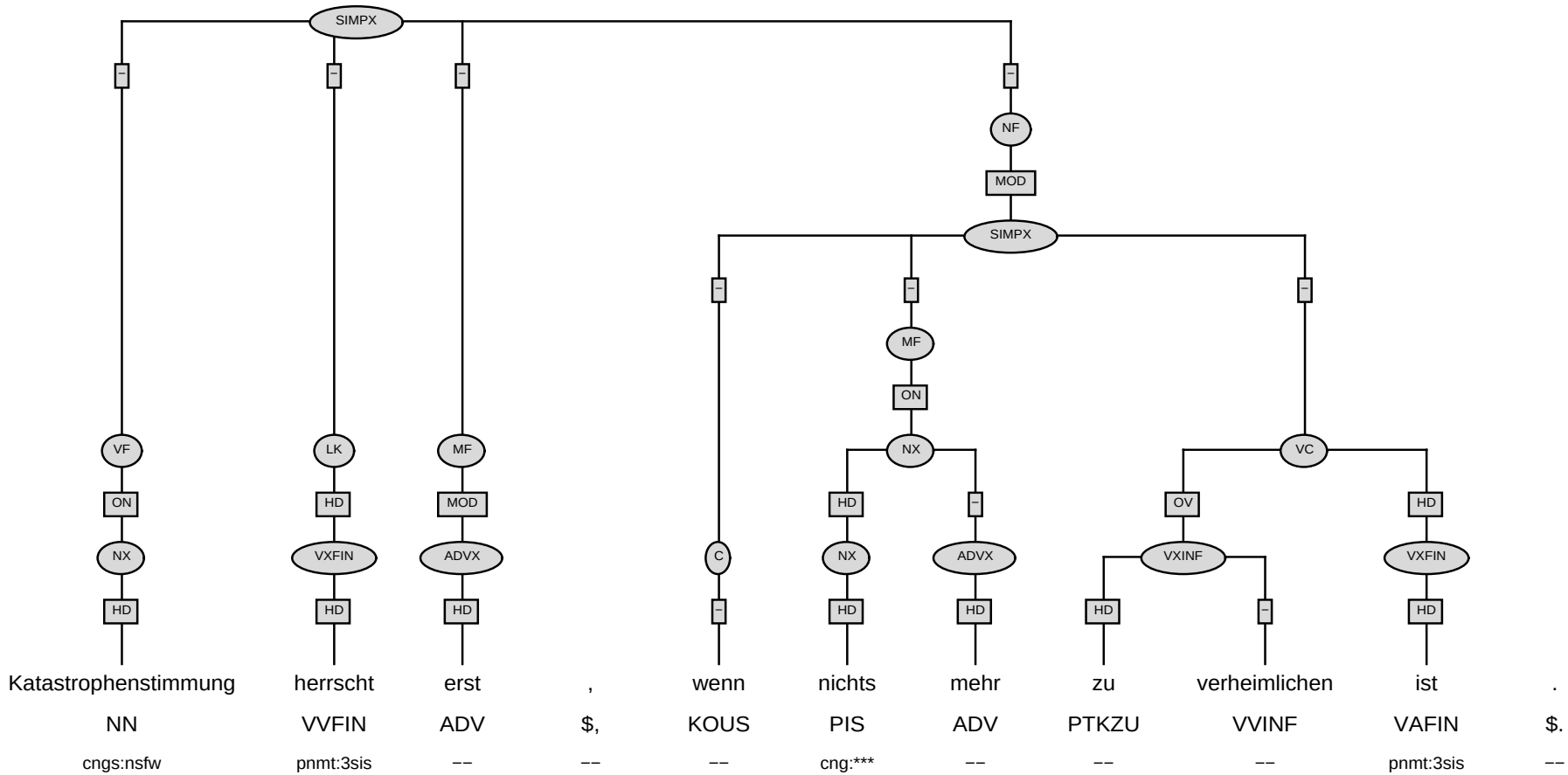
# Other Languages, e.g. Dutch

- Ik $[_{VP}$ had $]$ gisteren met mijn vrienden pasta $[_{VP}$ willen eten $]$ bij de pizzeria.
- I had yesterday with my friends pasta want eat at the pizzeria.
- $[_{VF}$ Ik $]$ $[_{LK}$ had $]$ $[_{MF}$ gisteren met mijn vrienden pasta $]$ $[_{RK}$ willen eten $]$ $[_{NF}$ bij de pizzeria $]$.

# The Corpus: TüBa-D/Z

- German newspaper text from the Tageszeitung (taz).
- Annotation style comparable to VERBMOBIL
- about 7300 sentences
- annotated and POS-tagged
- with Topological Field structure

# An example of an annotated sentence




Catastrophe-mood prevails only, when nothing anymore to hide is

# The Task: TopF chunking

- Find a task with high accuracy to find the basic structure of German sentences.
- TopFChunking finds the most common verbal parts of the Topological Field structure: CF (compl), LK (left) and RK (right).
- This gives us a handle to find the basic structure of the sentence.
- TopFChunking is clearly defined, and can be expected to be performed with a high accuracy.

# TopF chunking: some examples

- Die von den USA vorgeschlagene Strategie $[_{LK}$ wurde $]$ gestern von der NATO $[_{RK}$ übernommen $]$.
- Ich $[_{LK}$ habe $]$ gestern mit meinen Freunden Spaghetti $[_{RK}$ gegessen $]$.
- Das $[_{LK}$ sind $]$ allerdings Gelder, die vom Schatzmeister $[_{RK}$ hintergezogen worden sein sollen $]$ .

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# Three Chunkers

- Rule-based hand-crafted Finite State Automata (FSA), by Frank H. Müller.
- Rule-Based corpus trained PCFG: Probabilistic Context Free Grammars, by Tylman Ule
- Machine Learning Approach: Memory-Based Natural Language Processing, by Jorn Veenstra

# Finite-State Automata I

- topological fields structure can be captured by regular expression grammar,
- thus, it can be translated into finite state automaton (FSA)
- FSA = automaton which accepts input symbols according to transition function
- FSA can act as a transducer
- POS tags are input of transducer and topological fields and POS tags are output
- transition function is described by hand-written rules

# FSA II, a simplified example

- simple examples:
- $(APPR) - PRELS - (APPO) \rightarrow CF$
- ‘eine Tatsache, [mit der] man leben kann’
- ‘ein Mensch, [dem] man vertrauen kann’
- $VVPP - VAINF - VMFIN \rightarrow RK$
- ‘eine Frage, die [aufgeworfen werden könnte]’
- recursion can be covered by iteration of FSAs (cascade)

# FSA III, some POS tags

- APPR=preposition,
- PRELS=relative pronoun,
- APPO=postposition,
- VVPP=past participle,
- VAINF=auxiliary verb infinitive,
- VAFIN=auxiliary verb finite,
- VMFIN=modal verb finite

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# FSA IV, cannot be too simple

- The seemingly simple and effective FSA: $V+ \rightarrow RK$ is not correct.
- e.g. `Ich [bin] [gegangen]`
- `I have gone`
- or: `Mit meinem Bruder [gespielt] [hab] ich nie.`
- `with my brother played have I never`

# Probabilistic Context Free Grammars

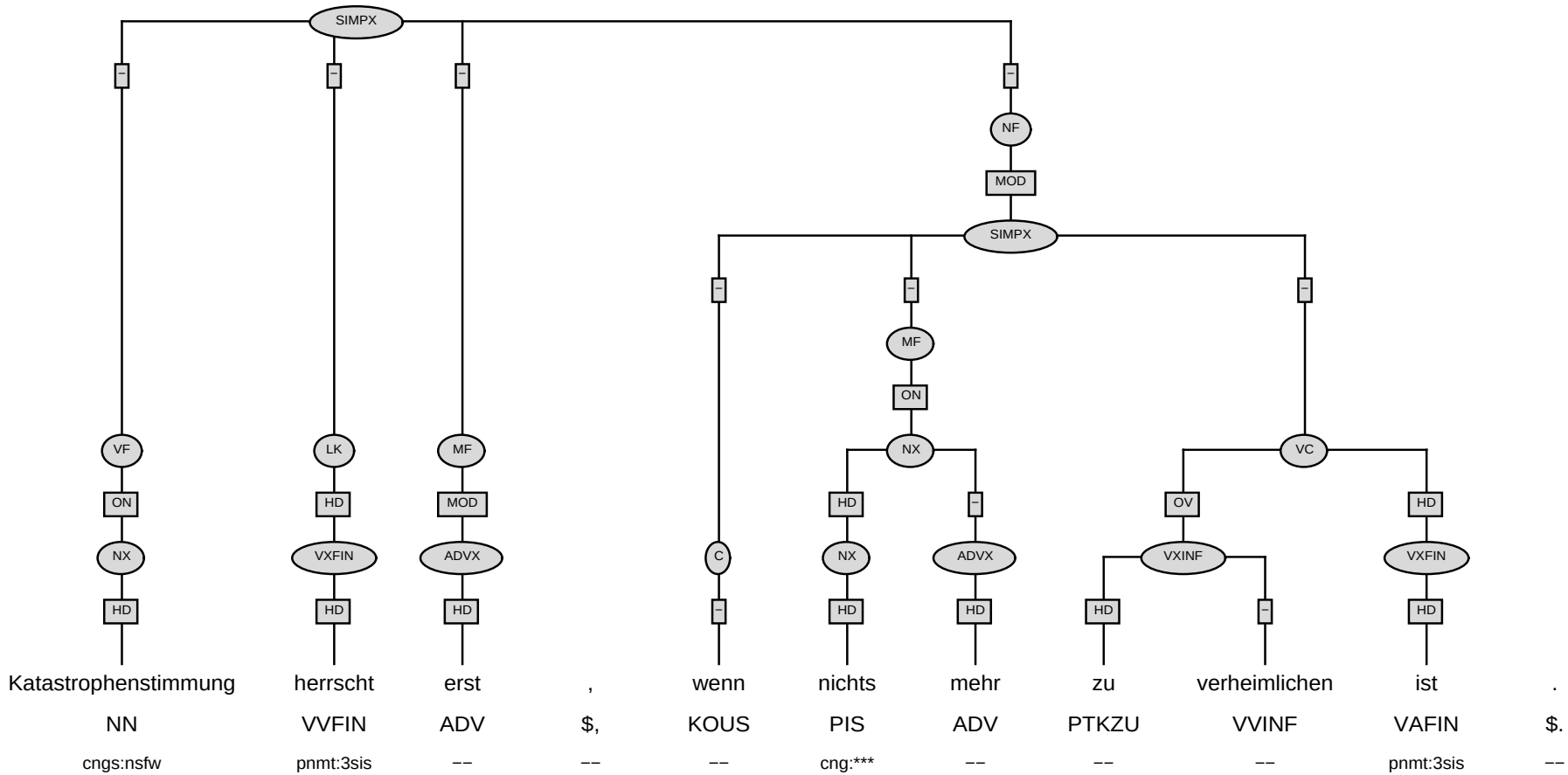
- The FSA approach uses hand-crafted non-probabilistic rules to recognise the TopFChunks.
- The PCFG approach extracts context-free, probabilistic rules from the corpus.
- The PCFG chunker uses POS information only.
- Used the LOPAR (Schmidt, 2000).

# PCFG II

- Design rules over POS tags to predict TopF chunks.
- Extract CF rules from the corpus and count their frequencies.
- Assign probabilities to rules according to these frequencies.
- For parsing new sentences parse bottom up, prune unprobable parses and work out the most probable routes.

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# PCFG III: an example




Catastrophe-mood prevails only, when nothing anymore to hide is

# PCFG IV: an example

- $SIMPX \rightarrow CF - MF - RK$
- $C \rightarrow KOUS$
- $ADVX \rightarrow ADV$
- $NX \rightarrow PIS$
- $MF \rightarrow NX$
- $NX \rightarrow NX - ADVX$
- $VC \rightarrow VXFIN - VXINF$
- $RK \rightarrow VXFIN$
- $VXINF \rightarrow PTKZU - VVINF$
- $VXFIN \rightarrow VVFIN$

# Memory-Based Learning

- MBL is a machine learning approach to NLP
- Learning Stage is Memory-Based
- Classification stage is similarity-based
- We have used TiMBL, an MBL software tool.
- Approach TopFChunking as a word-by-word classification task, comparable to POS tagging and NP chunking.

# Memory-Based Learning

- Assign to each word a tag: inside or outside one of the TopF chunks, or between two chunks.
- This gives us three tags (I, O & B) per TopF type (CF, LK & RK).
- Each word in the corpus data is annotated with one of these nine IOB tags.
- The MBL chunker is first trained on the corpus annotated with these IOB tags.
- We have used a context window of two words and POS tags to the left and to the right, and information gain to weight the feature relevance.

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

# Memory-Based Learning

- Der Mann $[_{LK}$ **hat** $]$ gestern $[_{RK}$ gelacht $]$ .
- Der$_O$ Mann$_O$ hat$_{I-LK}$ gestern$_O$ gelacht$_{I-RK}$ .$_O$
- der DET Mann NN **hat** V gestern TEMP gelacht V --- **I_LK**

# Baseline

- As a baseline we have computer the most probable TopF IOB tag for each POS tag in the train set.
- We have used these TopT tags to chunk the test set.

# Experimental setup

- first 5000 sentences as training data
- last 2300 sentences as test data
- FSA developer was not allowed to see the test set.
- Since the FSA and PCFG chunkers predicted too much, their output was converted to the TopF chunk format
- As the TüBa corpus is still under development, some changes were made during the development of the chunkers, this might have harmed the performance of the chunkers.

# Results

| | ALL | LK | VC | C |
|---|---|---|---|---|
| FSA TnT | 94.1 | 96.2 | 92.0 | 93.8 |
| FSA gold | 98.4 | 98.8 | 98.3 | 97.5 |
| PCFG TnT | 94.4 | 97.0 | 92.2 | 92.3 |
| PCFG gold | 98.1 | 98.9 | 98.1 | 96.1 |
| MBL TnT | 93.3 | 96.0 | 90.0 | 91.6 |
| MBL gold | 97.2 | 98.0 | 96.7 | 96.6 |
| baseline TnT | 75.5 | 75.2 | 72.3 | 83.2 |
| baseline gold | 77.9 | 75.0 | 79.1 | 86.2 |

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

# Conclusions

- $F_{\beta=1}$ scores are high for all three approaches
- Many errors are caused by POS tag errors
- Since TopFChunking is a well defined task, rule-based and hand-crafted approaches work relatively well.
- Memory-Based Approaches seem to suffer more from sparse data than FSA and PCFG.
- Such a high performance gives a solid basis for parsing and other text analysis tasks, such as information extraction and grammatical function assignment.

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

# Future Work

- Combining chunkers to optimise performance
- Annotating the full TopF structure of a sentence
- Sentence type classifier
- NP chunking with the German idiosyncrasies
- Grammatical function assignment

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen