Evaluation of the Gramotron Parser for German

Franz Beil[1], Detlef Prescher[2], Helmut Schmid[3], Sabine Schulte im Walde[3]

[1]TEMIS, Paris, France
[2]DFKI GmbH, Saarbrücken, Germany
[3]IMS, Universität Stuttgart, Germany.

June 2, 2002

Overview

- Introduction
- The Gramotron Parser for German
  - Symbolic Components: Corpora and Grammar
  - Stochastic Components: Probability Model and Training
- Evaluation
  - Noun Chunks (with and without Case Labels)
  - Subcat Frames
  - Cross-Entropy
  - Error Analysis
- Conclusion

Introduction

The Gramotron Parser for German:

- Pre-conditions:
  - To be used for the **induction of a semantically annotated lexicon**.
  - Reliable parsing **results had to be aquired rapidly**.

- Grammar development:
  - The focus was on the **recognition of grammatical relations** between nouns and verbs.
  - **Development in three stages**: for (i) verb-final clauses, (ii) relative clauses, and (iii) verb-first and verb-second clauses.

Bibliography

The Gramotron Group (IMS, University of Stuttgart, 1997-2000): Franz Beil, Glenn Carroll, Marc Light, Detlef Prescher, Stefan Riezler, Mats Rooth, Helmut Schmid, and Sabine Schulte im Walde.

References:

Sabine Schulte im Walde, Helmut Schmid, Mats Rooth, Stefan Riezler, Detlef Prescher (2001). *Statistical Grammar Models and Lexicon Acquisition*. In Linguistic Form and its Computation. CSLI Publications, Stanford, CA.

Franz Beil, Glenn Carroll, Detlef Prescher, Stefan Riezler, Mats Rooth (1999). *Inside-Outside Estimation of a Lexicalized PCFG for German*. In Proceeding of ACL'99, College Park, Maryland.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, Franz Beil (1999). *Inducing a Semantically Annotated Lexicon via EM-Based Clustering*. In Proceeding of ACL'99. College Park, Maryland.

Corpora: German V-final and Relative Clauses

- Automatically extracted from a newspaper corpus of 200 million words
- V-final clauses:

	4,000,000	words
	450,000	clauses
	9.16	words per clause

- Relative clauses:

	10,000,000	words
	1,100,000	clauses
	9.12	words per clause

- No further clausal embeddings, no punctuation
- Multiply PoS tags (up to 10, collapsed morphologic information)
- Lemmas for all words (coming from a morphologic analyzer)

The Grammar

A manually developed context-free grammar for German:

- **5,000 rules** (with markers for the lexical heads)
- **350 non-terminal categories**
- **200 terminal categories**

Coverage:

- 92.4% for verb-final clauses
- 91.7% for relative clauses

Subcat Frames: Classes and Arguments

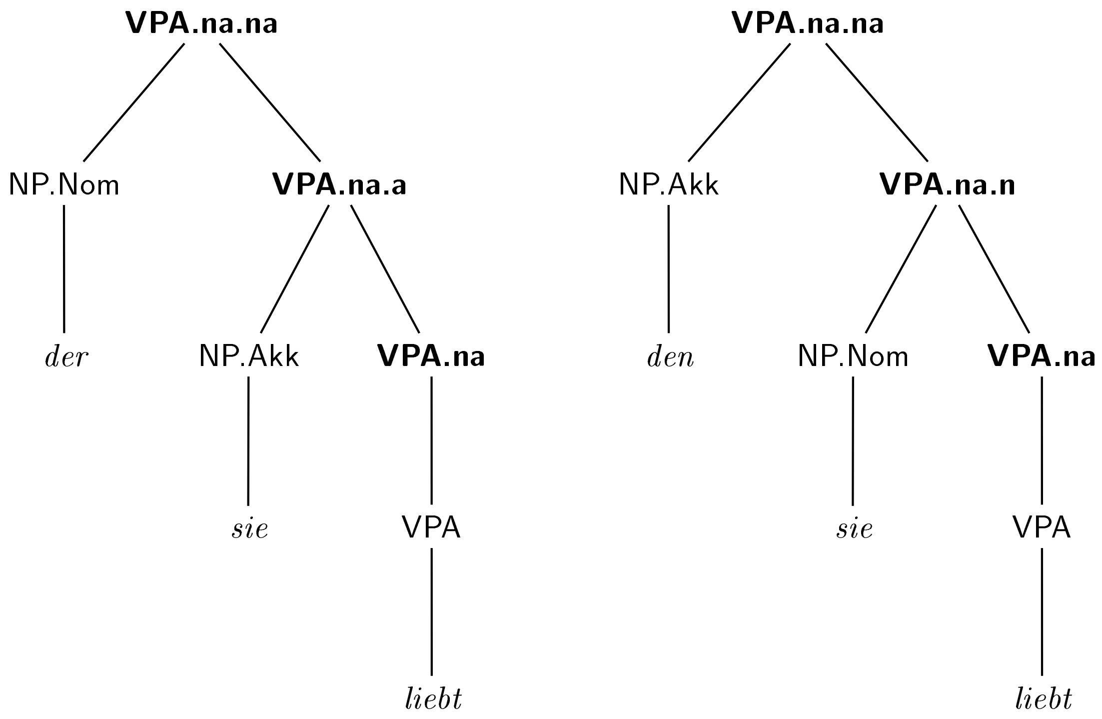
- **Four frame classes**: active (VPA), passive (VPP), non-finite (VPI) frames, and copula constructions (VPK).

- **Frame arguments:**
  nominative (n), dative (d), and accusative (a) NPs,
  reflexive pronouns (r),
  PPs (p, p-s),
  non-finite VPs (i).

- **Possible Number and types of verb frames**:

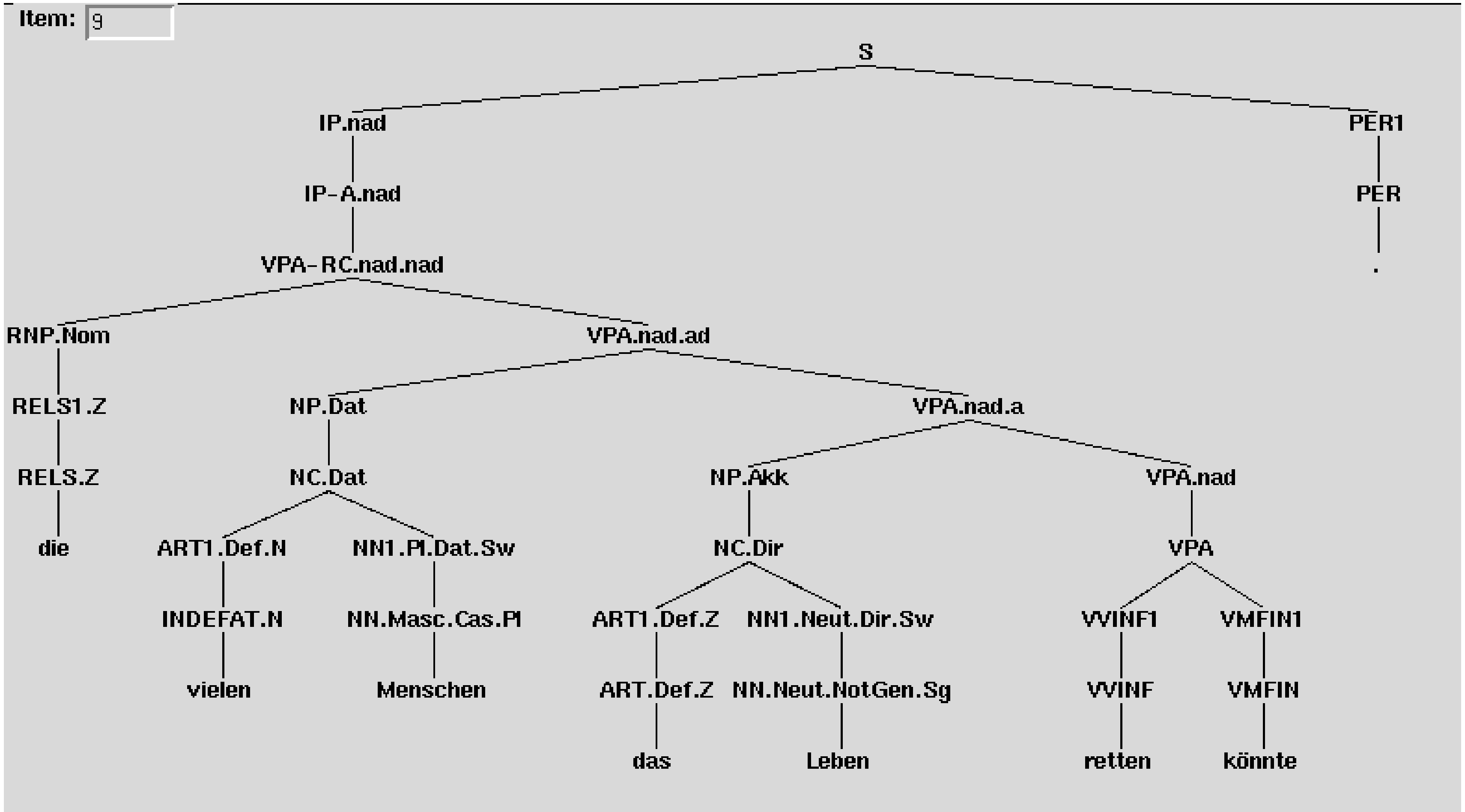
class	#	frame Types
VPA	16	n, na, nd, np, nad, nap, ndp, ni, di, nai, ndi, nr, nar, ndr, npr, nir
VPP	18	n, np-s, d, dp-s, p, pp-s, nd, ndp-s, np, npp-s, dp, dpp-s, i, ip-s, ni, nip-s, di, dip-s
VPI	8	-, a, d, p, r, ad, ap, dp, pr
VPK	2	n, i

Subcat Frames: Scrambling of VP Arguments



Srambling is reflected in the particular sequence in which the arguments of the verbs are saturated.

Example: an Analysis by the Symbolic Component



Probability Model: 4 Families of Distributions

- Root probs:

$$P_{start}(C) \text{ and } P_{start}(h|C)$$

- Rule choice probs:

$$P_{rule}(r|C,h) \text{ and } P_{rule}(\texttt{<terminal>}|C,h).$$

- Lexical choice probs:

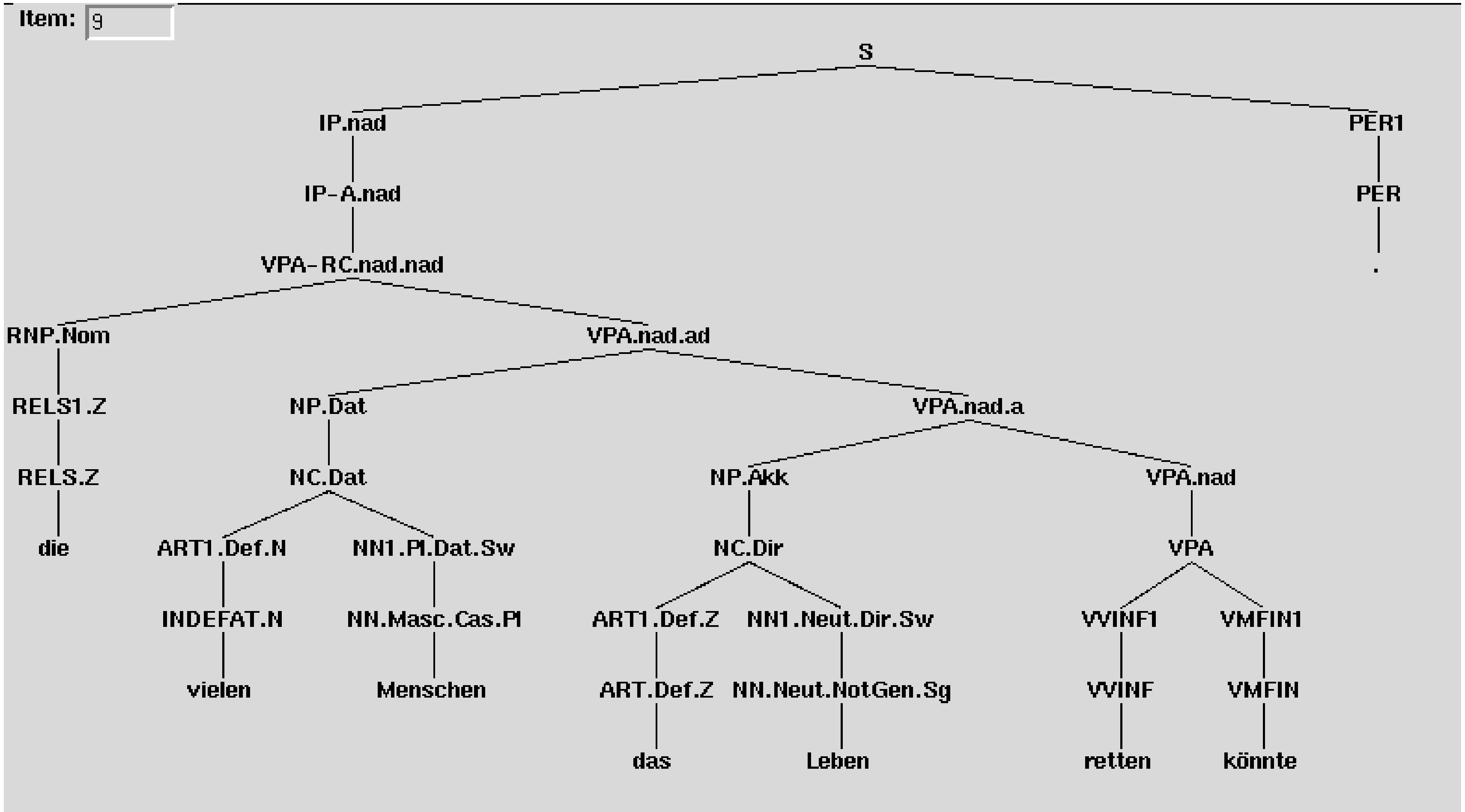
$$P_{choice}(h|C,C_p,h_p),$$

- Word probs:

$$P_{lex}(w|C,h).$$

(C = category of a node, h = lexical head of this node, C_p = category of the parent node, C_h = head of the parent node, r = a rule expanding the node, `<terminal>` = a flag indicating a terminal node, w = a word form.)

Example: The Stochastic Component



Grammar Training

The training in our main experiment was performed in the following steps:

1. **Initialization** of all CFG rules with identical frequencies.

2. **Unlexicalized training**: The training corpus was parsed once, re-estimating the frequencies twice.

3. **Lexicalization**: The unlexicalized model was turned into a lexicalized model by (i) setting the probabilities of the lexicalized rule probabilities to the values of the respective unlexicalized probabilities and (ii) initialising the lexical choice and lexicalized start probabilities uniformly.

4. **Lexicalized training**: Three training iterations were performed on the training corpus, re-estimating the frequencies after each iteration.

Linguistic Evaluation

Although an appropriate treebank is available for German, the NEGRA treebank, we did not use it for our evaluation:

- Standard parser evaluation method of measuring **precision/recall on phrase boundaries and crossing brackets is uninteresting for our objectives** (lexicon induction).

- Our initial grammar development was restricted to verb final and relative clauses while the **treebank annotates full clauses**. It turned out to be difficult to extract respective sub-treebanks.

- Two mapping problems occure for a transformation of treebank structures to a selection of ours (noun and verb chunks): (i) mapping treebank phrase spans to our chunk spans, and (ii) **finding an information-preserving mapping from our labels to treebank labels**.

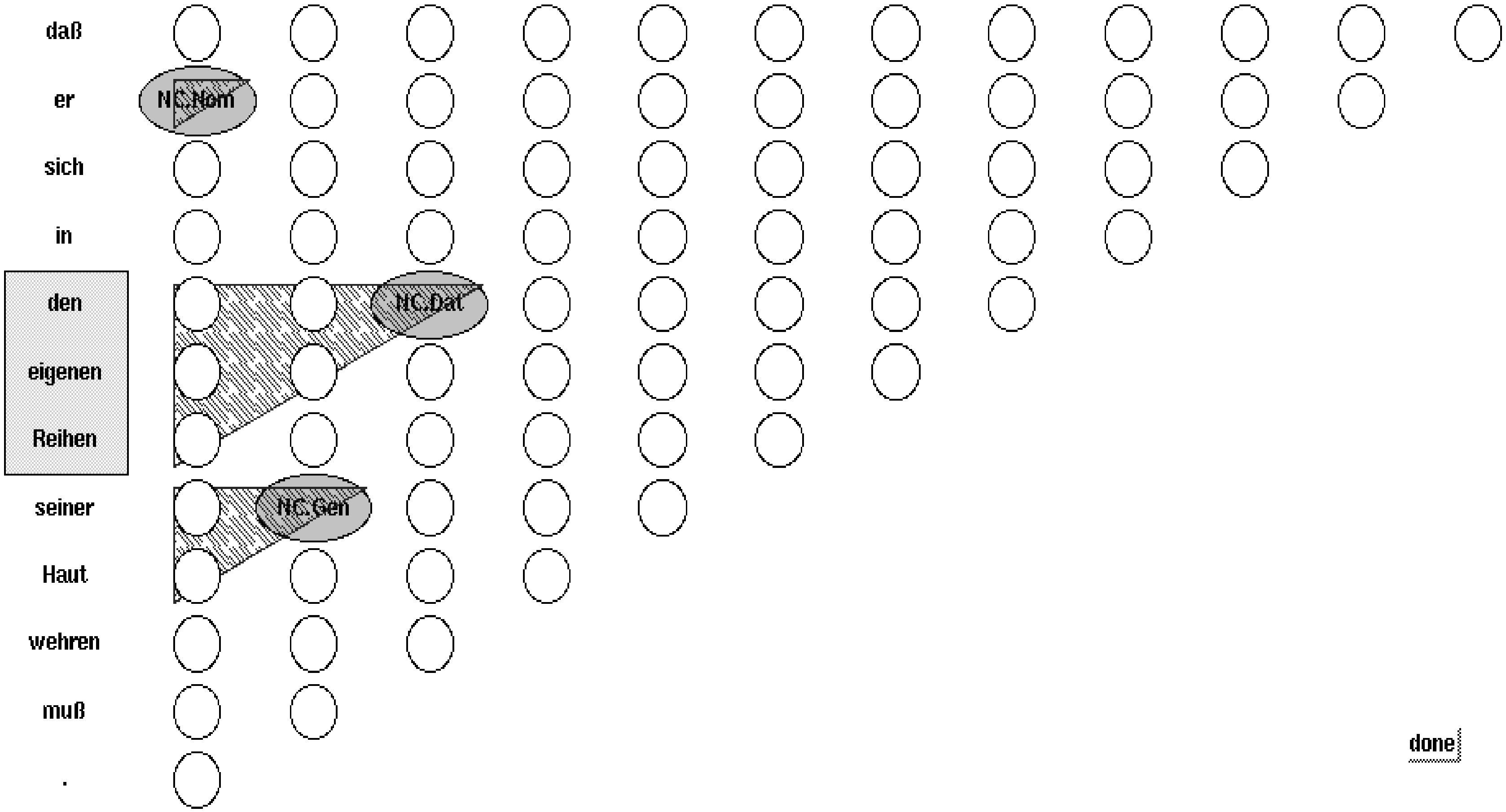
Evaluation Data

- Manual **noun-chunk annotations** for 200 randomly selected relative clauses.

- Manual **frame annotations** for 400 randomly selected verb-final and relative clauses.

- Manual **frame annotations** for 600 randomly chosen relative clauses containing six selected verbs: *beteiligen, erhalten, folgen, verbieten, versprechen, versuchen* (participate, receive, follow, forbid, promise, try).

Chart Browser for Manual Annotations



Evaluation results

Best-first consistency: Determine the Viterbi parse, extract the categories of interest, and compare them with the annotated data (i.e. compute *precision* and *recall* values).

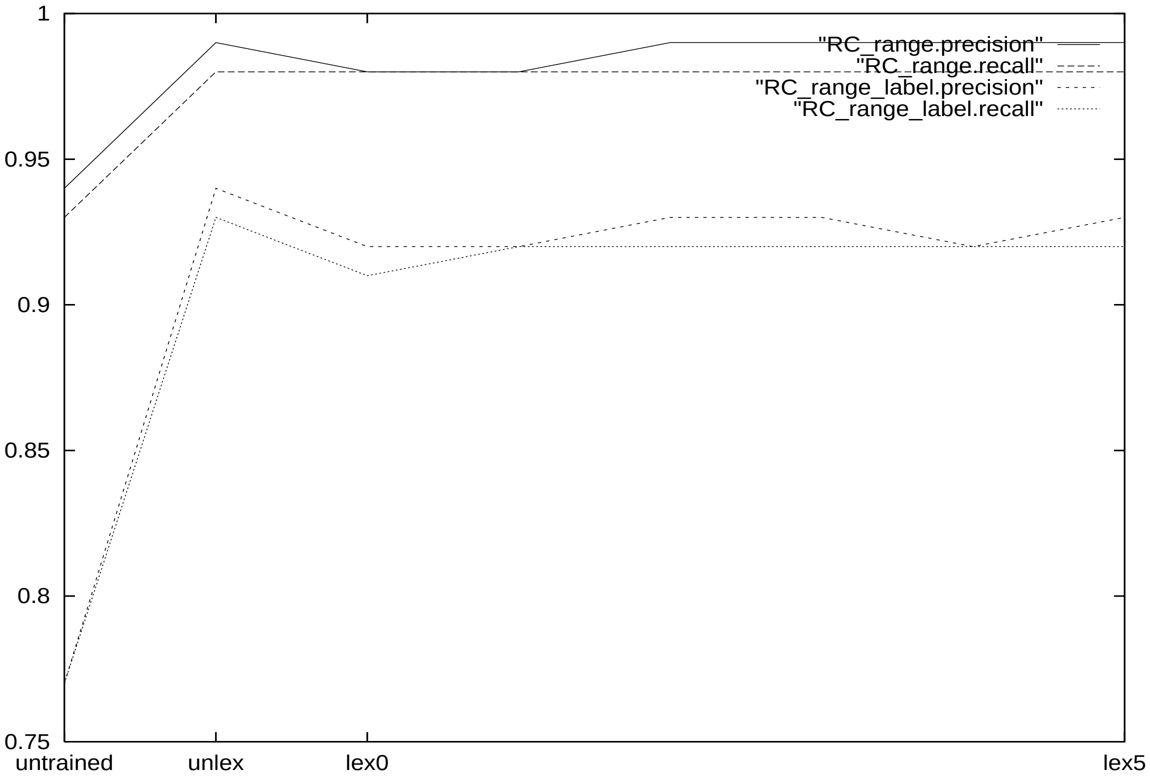
Best models (precision values):

Noun Chunks		Subcategorisation Frames on Sub-Corpora	
range	range+label	relative clauses	verb final clauses
98%	92%	63%	73%

Subcategorisation Frames on Specific Verbs					
beteiligen 'participate'	*erhalten* 'receive'	*folgen* 'follow'	*verbieten* 'forbid'	*versprechen* 'promise'	*versuchen* 'try'
48%	61%	88%	59%	80%	49%

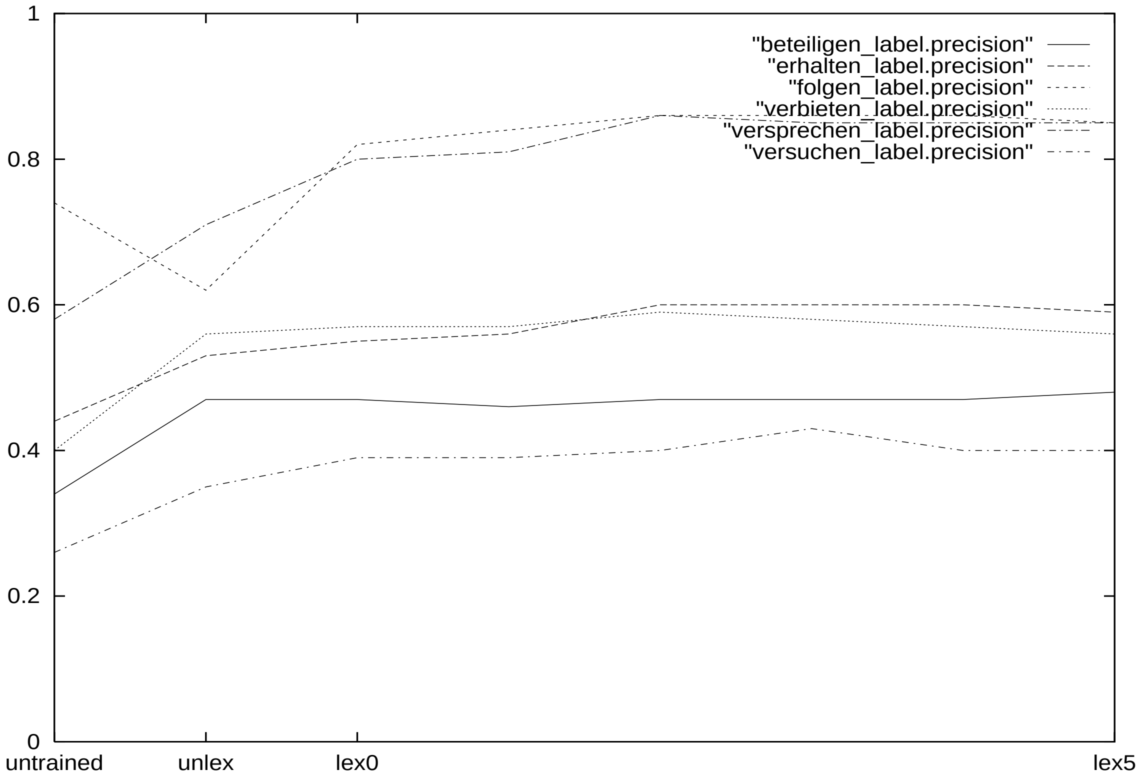
Evaluation of Training Regimes

Development of precision and recall values on **noun chunk range and label**:



Evaluation of Training Regimes

Development of precision values on **subcategorization frames for specific verbs**:



Mathematical Evaluation

Cross Entropy: Comparisons of different sizes of training and heldout data (training/heldout) for unlexicalized training in an older experiment:
(A) 50k/50k, (B) 500k/500k, (C) 4.1M/500k.

A		B		C	
1:	52.0199	1:	53.7654	1:	49.8165
2:	25.3652	2:	26.3184	2:	23.1008
3:	24.5905	3:	25.5035	3:	22.4479
⋮	⋮	⋮	⋮	⋮	⋮
15:	24.2861	57:	25.0549	90:	22.1443
16:	24.2861	58:	25.0549	95:	22.1443
17:	24.2867	59:	25.055	96:	22.1444

Selected Results of an Error Analysis

- **Highly common frames**: Slightly unlearned while lexicalized training (in contrast to less common frames).
- **Prepositional phrases**: It was not effectively learned to distinguish verbal complements and adjuncts.
- **Noun chunks belonging to a subcategorized non-finite clause**: Partly analyzed as main verb arguments.
- **Reflexive pronouns**: Confusion with accusative or dative noun chunks.

Shortcomings and Evaluation Alternatives

Shortcomings of our evaluation:

- We did not evaluate directly the **relations between lexical heads**, the main task our parsing system was designed for.
- We did not evaluate the correctness of **lexical heads of phrases**.

Alternative evaluation schemes covering these aspects:

- Evaluation of **dependency structure** (Lin, 1995).
- Evaluation of **grammatical relations** (Carroll etal., 1998)

Evaluation corpus: NEGRA, the treebank for German, might provide a basis for our future evaluation of head-head relations.

Conclusion

- Scrambling-style free-er phrase order, case morphology, and NP-internal gender, number and case agreement can be dealt within a head-lexicalized probabilistic context-free grammar.
- Inside-outside estimation appears to be benificial, however relies on a carefully built grammar and an evaluation using carefully selected linguistic criteria.
- Our evaluation focusses on regocnition of noun chunks (98% precision), noun chunks with case labels (92% precision), and subcat frames (about 70% precision).
- Two relevant aspects are still missing: direct evaluation of lexicalized grammatical relations and evaluation of lexical heads of phrases.

Corpus Annotation: Collapsing Morphological Analyses

The analyses of a computational morphological analyzer:

analyse> Bleibe
1. Bleibe+NN.Fem.Akk.Sg
2. Bleibe+NN.Fem.Dat.Sg
3. Bleibe+NN.Fem.Gen.Sg
4. Bleibe+NN.Fem.Nom.Sg
5. *bleiben+V.1.Sg.Pres.Ind
6. *bleiben+V.1.Sg.Pres.Konj
7. *bleiben+V.3.Sg.Pres.Konj

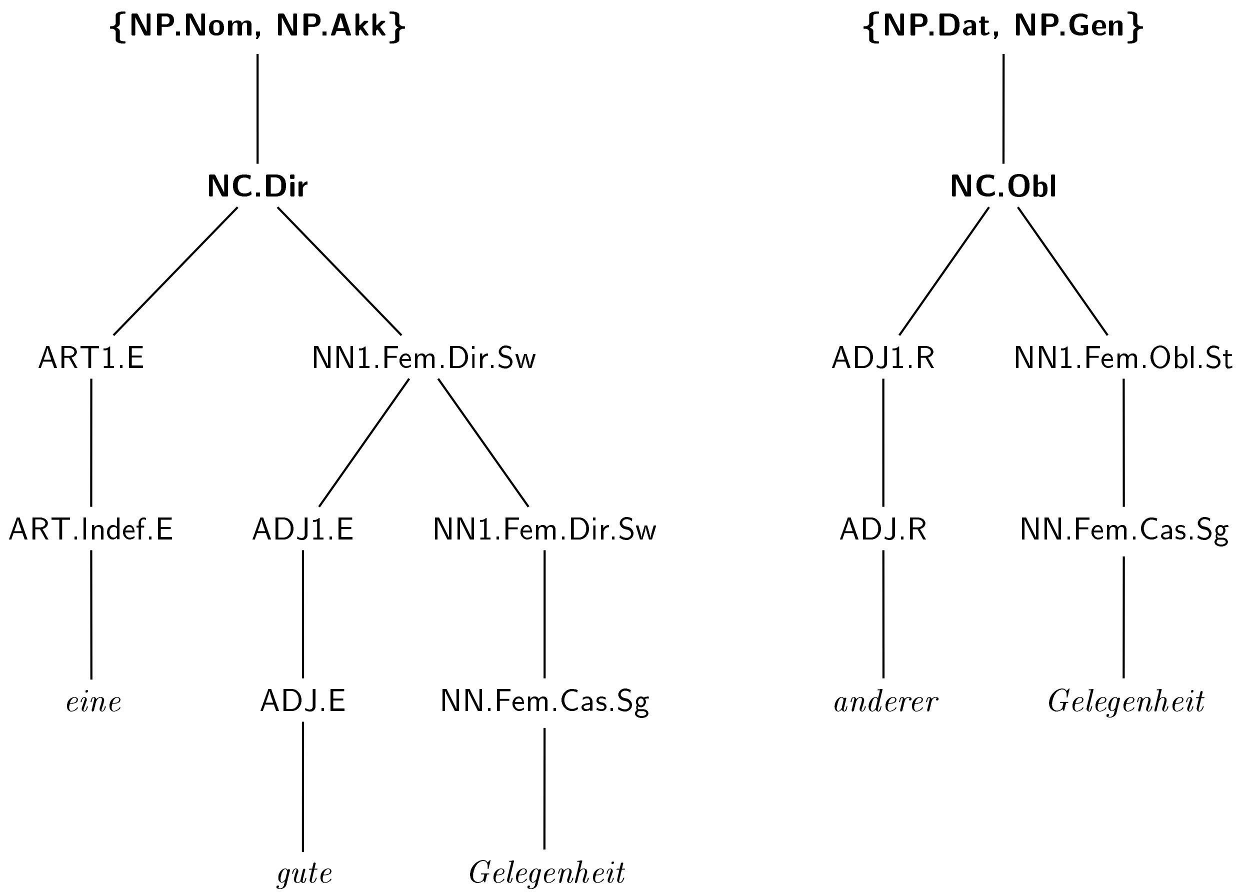
... are replaced by ambigous PoS tags:

Bleibe { NN.Fem.Cas.Sg, VVFIN }

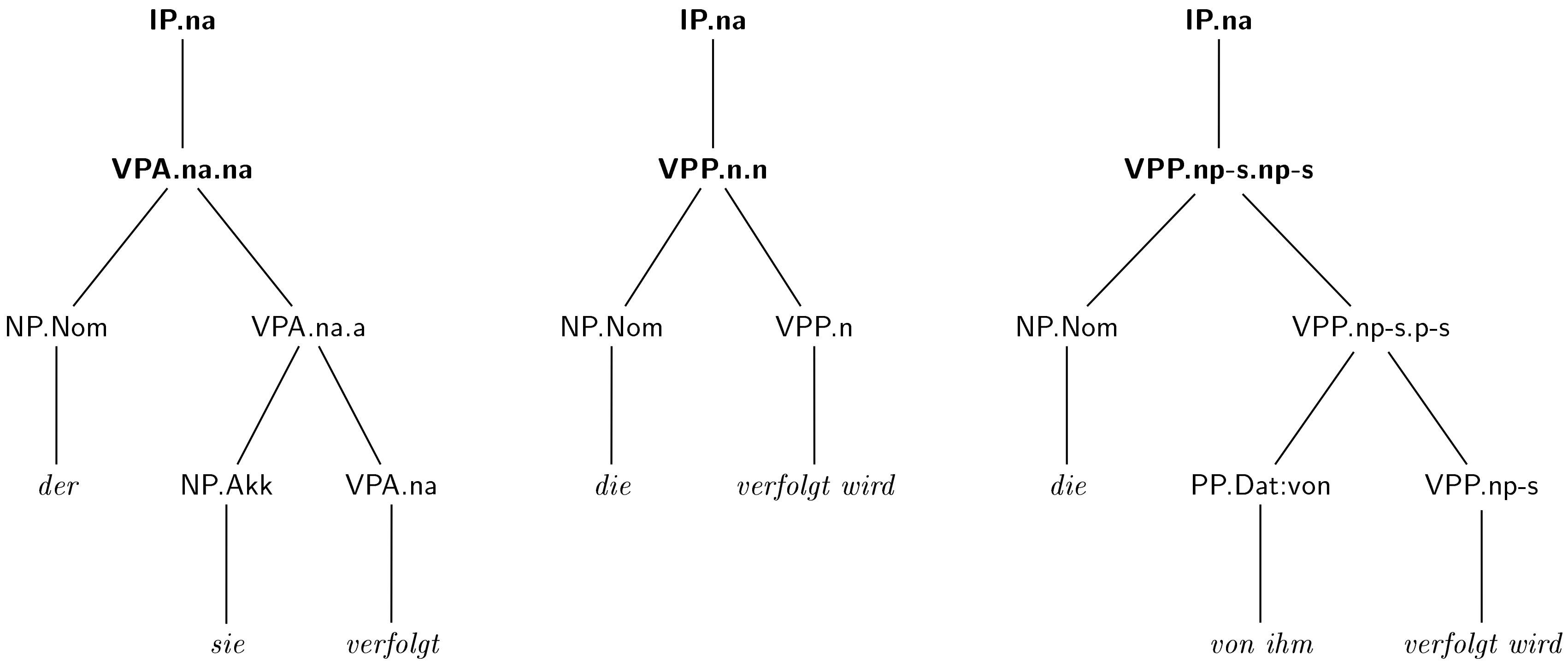
Apart from assigning morphological analyses, the tool additionally serves as **lemmatizer**.

Noun Projections: Disjunctive Category Features

Complementary to disjunctive PoS tags, we used disjunctive non-terminals:



Subcat Frames: Generalizing over the Active-Passive Alternation



Abstracting from the active and passive realization of an identical underlying deep-level syntax by defining a top-level subcat frame type.

Parameter Reduction: Parameter Pooling II

Parameter Pooling: applies to the lexical choice parameters $P_{choice}(.)$ and aims at a further reduction of parameters.

Assumption: Using uninflected lemmas rather than inflected word forms, lexical choice probabilities are unlikely to depend on inflectional features of categories. More important, **lexical choice parameters for verb frames are unlikely to depend on argument order**.

Example: Instead of assigning a lexical choice probability

$$P_{choice}(Leben|\texttt{NP.Akk}, \texttt{VPA.nad}.y,\ retten)$$

for all parameters y, the combinations are pooled to a single distribution

$$P_{choice}(Leben|\texttt{NP.Akk}, \texttt{VPA.nad}, retten)$$

Note: Active, passive and non-finite verb frames are pooled similiarily according to shared arguments.