Parsing German with Sister-head Dependencies

Amit Dubey
and **Frank Keller**
adubey@coli.uni-sb.de, keller@cogsci.ed.ac.uk

Overview

- Generalizing existing parsing models to new languages
- German: syntactic properties and annotation scheme
- Experiment 1: negative results for standard models
- Experiment 2: positive result for new model using sister-head
- Conclusions

Generalizing Existing Parsing Models

New languages:

- most parsing research has been for **English**;
- Do existing parsing models generalize to new languages? Or do they rely on the linguistic properties of English?

New annotation schemes:

- virtually all parsing research has used **one training corpus:** Penn Treebank (WSJ part);
- treebanks are now available for Bulgarian, Chinese, Czech, Dutch, German, Italian, Korean;
- some annotation schemes differ substantially from the Penn Treebank scheme.

Generalizing Existing Parsing Models

Some results are available for new languages using the Collins (1997) model:

Language	Size	LR	LP	Source
English	40,000	87.4%	88.1%	(Collins 1997)
Chinese	3,484	69.0%	74.8%	(Bikel and Chiang 2000)
Czech	19,000	—-	80.0% —-	(Collins et al. 1999)

Performance is significantly lower than for English. This might be due to the smaller training corpora.

Generalizing Existing Parsing Models

Research questions:

- Do existing models **generalize** to a new language and a new annotation scheme? Test this by applying it to German.
- Do they outperform a simple **unlexicalized** model? Test this by comparing with an unlexicalized baseline.
- Does the **size of the training** set influence the behavior of the model? Test this by computing learning curves.

Parsing German

Syntactic properties of German:

- semi-free word order (as opposed to fixed word order in English);
- order of complements (subject and objects) and adjuncts is largely free;
- order of the verb is fixed but depends on sentence type:
  - main clauses: verb in second position;
  - subordinate clauses: verb in final position;
  - questions: verb in initial position;

Word order flexibility difficult to model using CFGs.

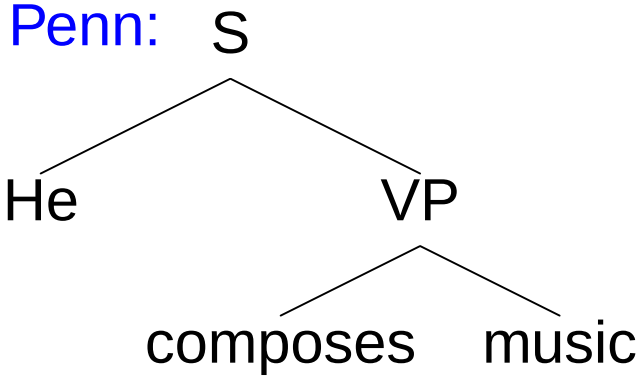
Parsing German

Training corpus: **Negra** treebank (Skut et al. 1997): 350,000 words of newspaper text, manually annotated with syntax trees.

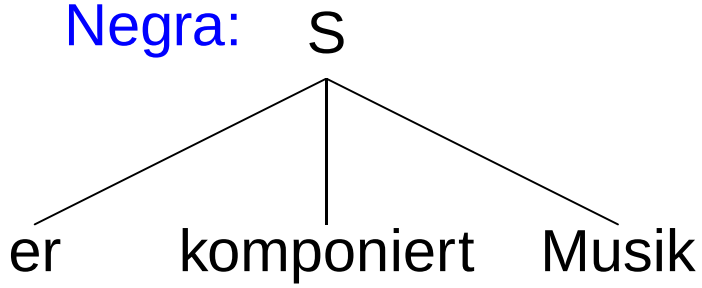
Negra annotation scheme reflects word order: much **flatter structures** than in the Penn Treebank.

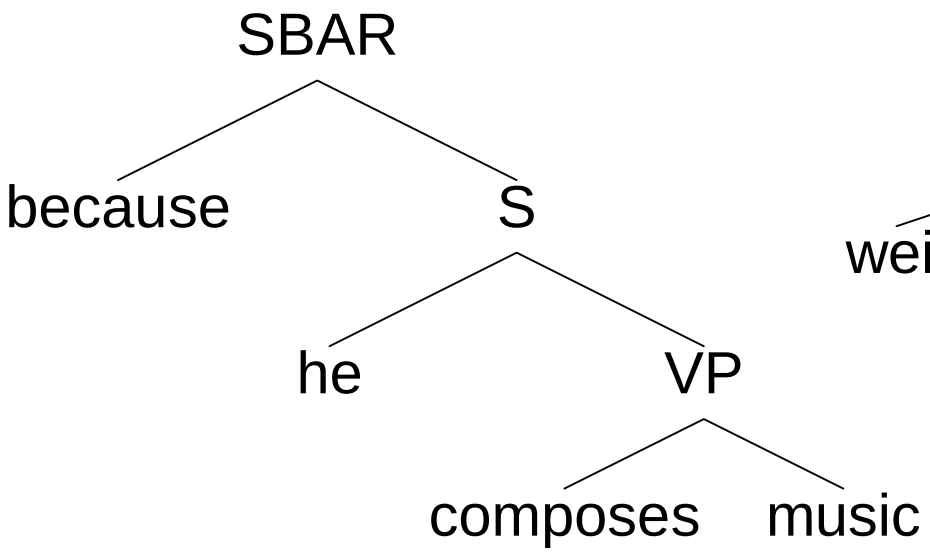
Parsing German

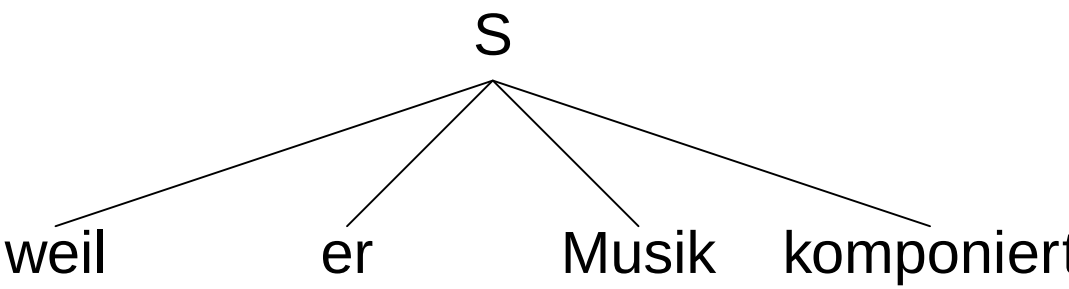
Penn:



Negra:



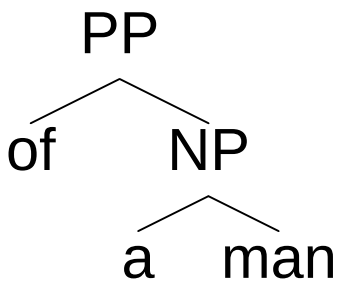




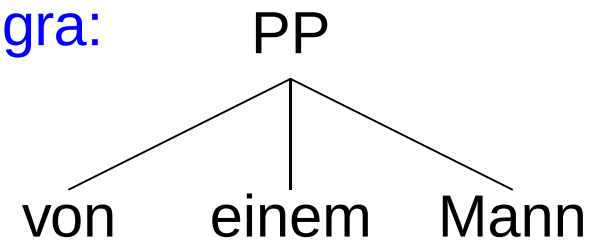
Parsing German

Also: separate categories for coordination (CNP, CPP, etc.).

Penn:



Negra:



Training on Negra

Preprocessing of Negra:

- remove traces, unary productions, sentences > 40 words (for efficiency);
- divide the resulting corpus into 90% training set, 5% development set, 5% test set.

Derive a PCFG from the training corpus:

- traverse the trees in the corpus and extract CFG rules;
- estimate rule probabilities by counting how often a given rule occurs in the corpus.

Lexicalized and Unlexicalized Models

Train three models on Negra using LoPar (Schmid 2000) and Sleepy (Dubey 2001):

- **Baseline:** standard PCFG;
- **C&R:** head-lexicalized PCFG (Carroll and Rooth 1998);
- **Collins:** Model 1 of Collins (1997); uses head-head dependencies.

Two variants:

- grammatical functions: might help with flexible word order;
- parameter pooling: might help with PPs and coordinate categories.

Results

	TnT tagging		Perfect tagging	
	LR	LP	LR	LP
Baseline	70.56	66.69	72.99	70.00
Baseline + GF	70.45	65.49	81.14*	78.37*
C&R	68.04	60.07	70.79	63.38
C&R + pool	69.07	61.41	71.74	64.73
C&R + GF	67.66	60.33	81.17*	76.83*
Collins	67.91	66.07	68.63	66.94

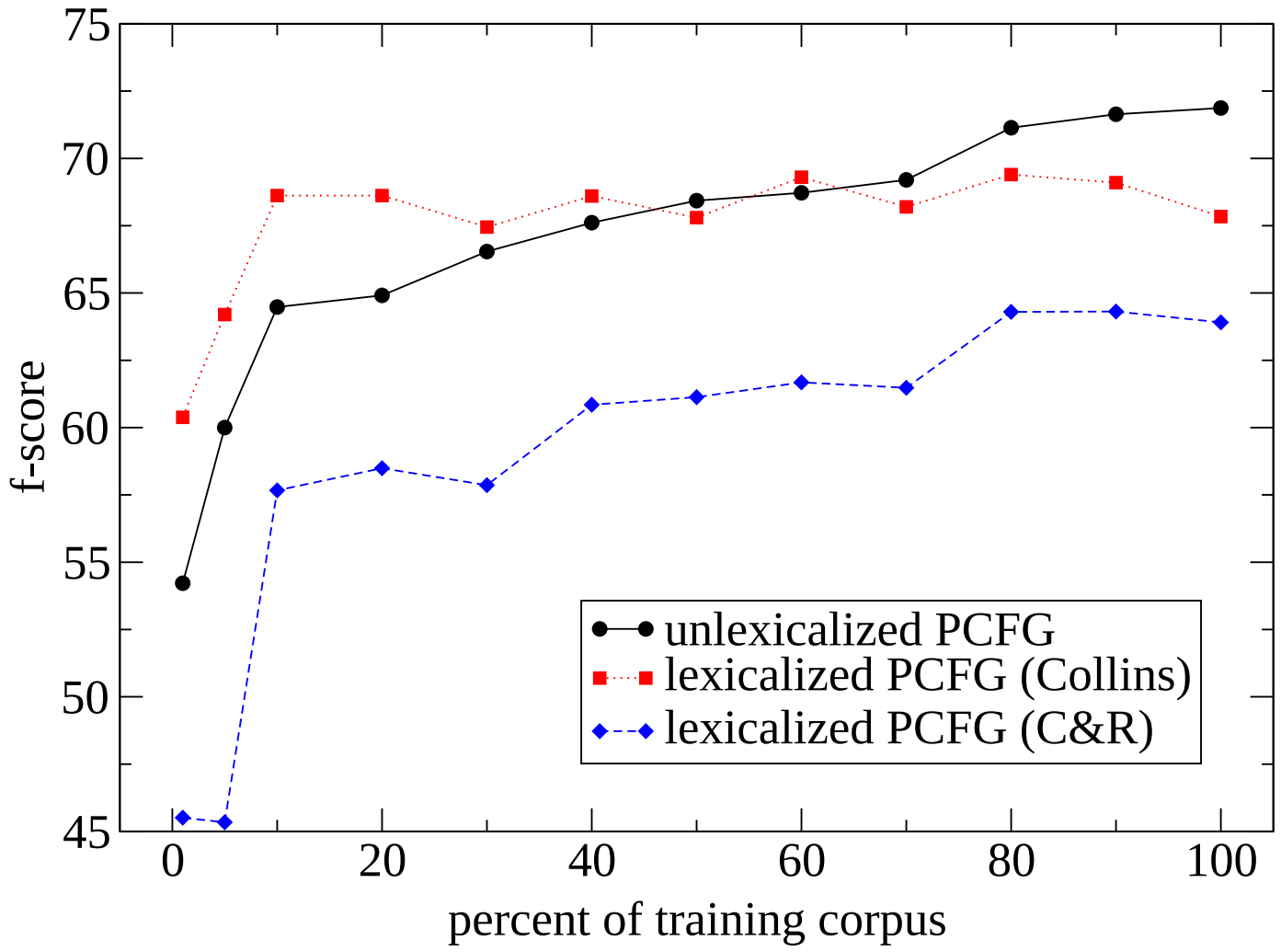
* low coverage (about 65%)

Results

Main findings:

- grammatical functions don’t improve performance;
- parameter pooling improves performance slightly;
- both lexicalized models perform worse than the baseline;
- for English, lexicalization typically improves performance by 10%;
- result is not due to sparse data, as learning curve shows.

Results


75
70
65
60
55
50
45
f-score
0
20
40
60
80
100
percent of training corpus
unlexicalized PCFG
lexicalized PCFG (Collins)
lexicalized PCFG (C&R)

Sister-Head Dependencies

Hypothesis: poor performance of lexicalized models is due to the flatness of Negra trees.

Average number of daughters in Penn and Negra:

	Penn	Negra
NP	2.20	3.08
PP	2.03	2.66

	Penn	Negra
VP	2.32	2.59
S	2.22	4.22

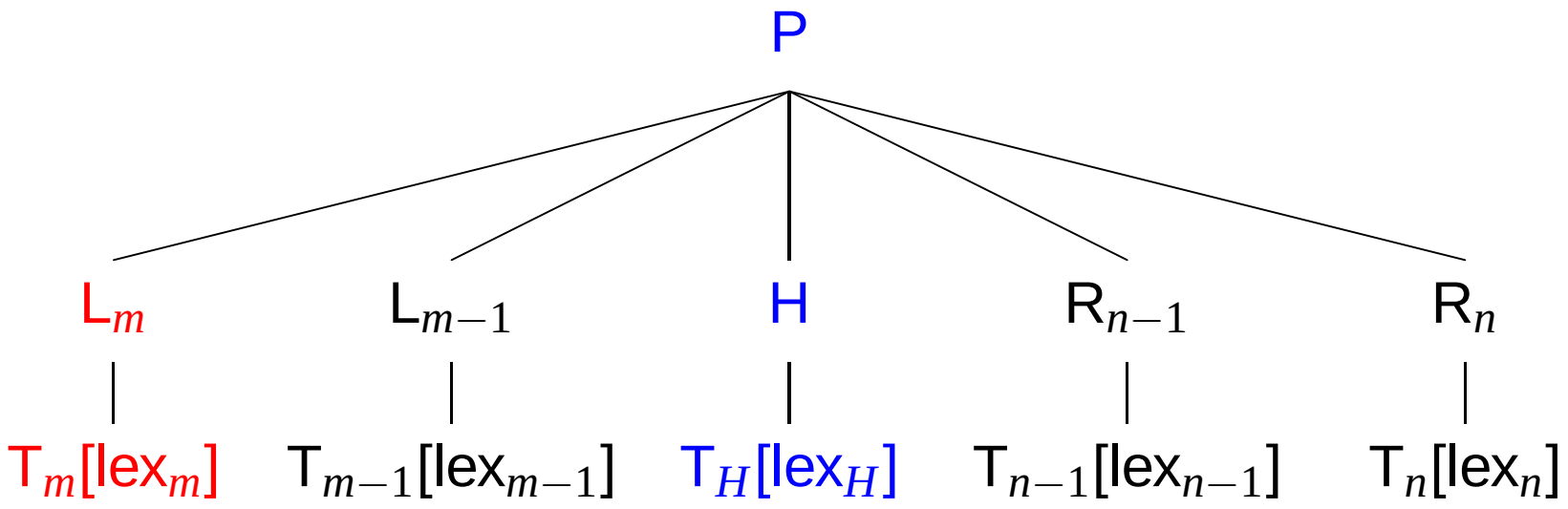
Strategy: modify the Collins model to deal with flat trees; use sister-head instead of head-head dependencies.

Sister-Head Dependencies

Standard **head-head** dependencies: in the expansion probability for the rule:

$P \rightarrow L_m \ldots L_1\ H\ R_1 \ldots R_n$

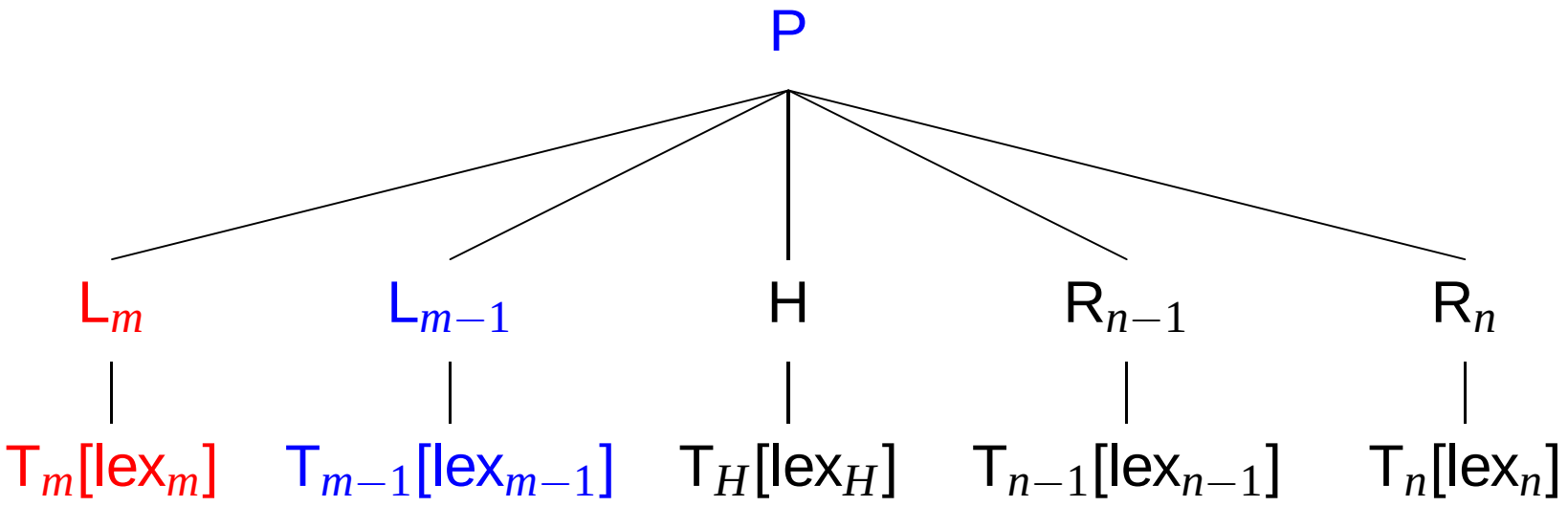
head $\langle L_m, T_m, lex_m \rangle$ is conditioned on P and head $\langle H, T_H, lex_H \rangle$:



Sister-Head Dependencies

New **sister-head** dependencies:

Head $\langle L_m, T_m, lex_m \rangle$ is conditioned on P and previous sister $\langle L_{m-1}, T_{m-1}, lex_{m-1} \rangle$:



Sister-Head Models

Test several variants of the model:

- original Collins model (all categories head-head);
- sister-head for NP, for PP, and for all categories;

Test an alternative to the sister-head model:

- split PPs (which are flat in Negra) to allow the model to generalize over NPs inside PPs;
- then test on either split or collapsed (original) PPs.

Results

	TnT tagging		Perfect tagging	
	LR	LP	LR	LP
Baseline	70.56	66.69	72.99	70.00
Unmod. Collins	67.91	66.07	68.63	66.94
Split PP	73.84	73.77	75.93	75.27
Collapsed PP	66.45	66.07	68.22	67.32
Sister-head NP	67.84	65.96	71.54	70.31
Sister-head PP	70.27	68.45	73.20	72.44
Sister-head all	71.32	70.93	73.93	74.24

Results

For which categories are sister-head probabilities most useful?

	TnT tagging		Perfect tagging	
	ΔLR	ΔLP	ΔLR	ΔLP
PP	−3.45	−1.60	−4.21	−3.35
S	−1.28	0.11	−2.23	−1.22
Coord	−1.87	−0.39	−1.54	−0.80
VP	−0.72	0.18	−0.58	−0.30
AP	−0.57	0.10	0.08	−0.07
AVP	−0.32	0.44	0.10	0.11
NP	0.06	0.78	−0.15	0.02

Results

Main findings:

- sister-head model outperforms both original Collins model and unlexicalized baseline;
- best performance with sister-head for **all** categories;
- splitting PPs doesn't improve performance (spurious improvement if testing on split PPs); same result for S (not reported here);
- explains why LP/LR figures on the Penn treebank are higher than on Negra.

Conclusion: new dependencies needed for new languages and annotation schemes.

Previous Work

Comparison with conditioning information in previous work:

	C&R	Collins	Charniak	Current
Head category	X	X	X	
Head word	X	X	X	
Head tag		X	X	
sister category	X		X	X
sister head word				X
sister head tag				X

Previous Work

Do bilexical dependencies really matter?

- Gildea (2001): **no**, as sparse data is a problem for Penn Treebank grammars;
- Hockenmaier and Steedman (2002): **yes**, it helps for a CCG grammar;
- bilexical dependencies matter for **binarized grammars;** sister-head dependencies are a way of binarizing.

Further Research

Improve the sister-head model for **German:**

- smarter use of grammatical functions: should help with word order;
- insert traces: should help with extraposition;
- integrate subcat frames: should help with attachment.

Build a **crosslinguistic model:**

- can we build a model that works well for more than one language?

Conclusions

- Previous parsing research has focused almost exclusively English, and on one training corpus;
- the results don’t generalize straightforwardly to new languages and annotation schemes;
- the standard head-head model fails to outperform an unlexicalized baseline model for German;
- this can be addressed using sister-head dependencies instead: captures flat tree structures better;
- the flat structures are motivated by linguistic properties of German (semi-free word order).

References

Bikel, Daniel M., and David Chiang. 2000. Two statistical parsing models applied to the Chinese treebank. In *Proceedings of the 2nd ACL Workshop on Chinese Language Processing*. Hong Kong.

Bod, Rens. 1993. Using an annotated language corpus as a virtual stochastic grammar. In *Proceedings of AAAI*, 778–783.

Carroll, Glenn, and Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 36–45. Granada.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on Artificial Intelligence*, 598–603. Cambridge, MA: AAAI Press.

—. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, 132–139. Seattle, WA.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational*

Linguistics, 16–23. Madrid.

Collins, Michael, Jan Hajič, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. University of Maryland, College Park.

Dubey, Amit. 2001. *Tuning Probabilisitic Parsers on Plain Text*. Master's thesis, University of Waterloo.

Gildea, Daniel. 2001. Corpus variation and parser performance. In Lillian Lee and Donna Harman, eds., *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Pittsburgh, PA.

Hockenmaier, Julia, and Mark Steedman. 2002. Generative models for statistical parsing with combinatory categorial grammar. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia.

Schmid, Helmut. 2000. LoPar: Design and implementation. Unpubl. ms., Institute for Computational Linguistics, University of Stuttgart.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing*. Washington, DC.