A Hybrid Model for Morpho-Syntactic Annotation of German with a Large Tagset

Julia TRUSHKINA and Erhard HINRICHS,
University of Tübingen,
Seminar für Sprachwissenschaft, Wilhelmstrasse 19,
72074 Tübingen, Germany,
jul@sfs.uni-tuebingen.de, eh@sfs.uni-tuebingen.de

Abstract

This paper presents a hybrid rule-based and statistical model for morpho-syntactic annotation of German, a highly ambiguous inflectional language. The proposed model makes use of a manually annotated corpus of moderate size and attains an accuracy of 92.04%, which corresponds to a 7.34% improvement of the best results reported by other researchers for the task of annotation of German with a large tagset.

1 Introduction

Morphological information, such as *case, number* and *gender*, plays an important role in parsing highly inflectional languages, such as German and Czech. It can help to resolve syntactic ambiguity in chunking and is particularly useful in dependency parsing of languages with free word-order, since it partly determines the argument structure of the sentence.

More often than not, tagsets standardly used for part-of-speech (POS) tagging do not reflect fine-grained morphological distinctions within word classes, since the inclusion of morphological information in POS labels typically results in large tagsets. For example, for Czech a tagset containing more than a thousand tags was reported in (Hajič and Hladka, 1997). This in turn raises the question as to which methods are suitable for automatic assignment of POS labels with tagsets of such size and the accompanying data sparseness problem.

A stochastic approach that aims at reducing data-sparseness for large tagsets has been described in (Tufiş, 2000). The approach is based on the idea of using a reduced tagset for an intermediate n-gram tagging step. It provides high accuracy for Romanian and Hungarian. However, as was shown in (Hinrichs and Trushkina, 2003), the approach has a rather low performance on German data due to the higher ambiguity rate of German (cf. statistics in Table 1).

For Czech, a hybrid model with rule-based and statistical modules (Hajič et al., 2001) has yielded excellent results. A similar hybrid tagging system was successfully designed for English (Tapanainen and Voutilainen, 1994). Given the previous success of hybrid models for other languages, the purpose of this paper is to apply such models for German, which, due to its high ambiguity rates, provides a particularly challenging case for the task at hand.

2 Architecture of the model

The model has a layered and sequential architecture consisting of morphological analysis, rule-based disambiguation and statistical tagging. The order of these modules reflects the relative strengths of the rule-based and statistical methods involved.

The morphological analyzer provides all possible analyses for a given sequence of tokens. This highly ambiguous output is then fed to the rule-based module. Its task is to reduce the candidate analyses to be considered by the statistical module. If used cautiously, the rule-based method will rule out only those candidates for which it has sufficient evidence and will retain all those that are contextually plausible. The task of the statistical module is to disambiguate the remaining cases of ambiguity. Statistical disambiguation is made considerably easier by the rule-based pre-filtering module, since the remaining set of hypotheses is greatly reduced. This reduction in search space corresponds to a gain in precision compared to purely statistical disambiguation.

3 Data

The `taz` newspaper portion of the Tübingen Treebank of German (TüBa-D/Z) (Telljohann et al., 2003) provides the basis for the experiments reported in this paper. The treebank is manually annotated with morphological information, constituent structures and argument

language and source of the statistics		average # analyses	ambiguous tokens	tagset size
German	(current paper)	7.10	68.87%	718
Czech	(Hajič and Hladka, 1997)	3.65	not avail.	1171
	(Hajič and Hladka, 1997)	2.36	not avail.	882
Turkish	(Oflazer and Tür, 1996)	1.83	50.66%	not avail.
English	(Tapanainen and Voutilainen, 1994)	1.77	not avail.	139
German (STTS)	(current paper)	1.77	39.57%	54
Romanian	(Tufiş, 2000)	1.71	38.17%	410
Hungarian	(Tufiş et al., 2000)	1.33	31.90%	> 1265

Table 1: Ambiguity of German data in comparison to other languages

function information. The treebank tagset is based on the Stuttgart-Tübingen tagset (STTS) (Schiller et al., 1995), the widely accepted inventory of POS categories for German, which contains 54 distinct POS labels. The STTS tagset is enriched by morpho-syntactic features such as *case*, *number*, *person*, *gender*, *tense* and *mood*. The resulting tagset distinguishes 718 tags. The treebank tagset is used in the statistical and combined model experiments, as well as in the evaluation of all modules. 11 361 tokens from the corpus were set apart for test data and 5 891 tokens for development data. 104 049 tokens were used as training data. The statistical component also uses 115 098 tokens from the second part of the treebank for weakly supervised training. These tokens are annotated only with constituent structures and argument functions and do not have morphological information.

4 Data ambiguity

Table 1 presents ambiguity rates for German estimated on the test data. Analyses for the tokens in the test data were generated by the Xerox morphological analyzer[1] and then mapped into the treebank tagset. The average number of analyses is counted as the ratio between the number of analyses assigned to the tokens in the text and the total number of tokens in the text. The percentage of ambiguous tokens in the data is provided in column 3.

For comparison, ambiguity rates reported in the literature for other languages, as well as for pure POS tagging of German, are included in the table. What this comparison shows is that morpho-syntactic annotation of German constitutes a much harder task than the same problem for other languages or for pure POS tagging of German. The average number of analyses is double that of Czech and at least by factor of 4 higher than for the other languages and for POS tagging of German. This is reflected in the percentage of ambiguous tokens, which is almost twice as big compared to Romanian and to POS tagging of German and more than twice compared to Hungarian.

5 Rule-based disambiguation module

The initial set of analyses for the rule-based disambiguation module is provided by the Xerox morphological analyzer, whose tagset makes even more fine-grained distinctions than the treebank tagset. These finer distinctions are in some cases useful in order to precisely delineate the applicability of highly specific disambiguation rules. After rule-based disambiguation the Xerox tagset is mapped into the treebank tagset, which is used by the statistical module.

The rule-based disambiguation module was developed in the Xerox Incremental Parsing System (XIP) (Aït-Mokhtar et al., 2002). It consists of a POS disambiguation submodule and a subsequent morphological disambiguation submodule. An earlier version of the morphological disambiguation submodule which applied to noun phrases only was described in (Hinrichs and Trushkina, 2002). This submodule has now been generalized to other parts of speech and has been combined with POS disambiguation submodule.

5.1 Disambiguation rules

Two types of disambiguation rules are used in the XIP system: syntactic heuristics and concord rules. They jointly provide an effective way to reduce morpho-syntactic ambiguity.

[1] Consult www.xrce.xerox.com/competencies/content-analysis/demos/german.de.html for more information.

Concord rules are based on mutual agreement constraints between lexical nodes within one phrase, e.g. between articles and nouns within one noun phrase. They are, therefore, best suited for morphological disambiguation of lexical nodes that make up phrasal categories. Syntactic heuristics rely on constraints that a surrounding context imposes on the set of possible analyses for a given token and can, therefore, be used for both POS and morphological disambiguation.

In the remainder of the subsection we will illustrate the use of the two types of rules with sentence (1).[2]

(1) Der/ART/PRELS/PDS Fahrer/NN konnte/VMFIN nicht/PTKNEG mehr/ADV/VVIMP/PIS/PIAT bremsen/VVFIN/VVINF.
The driver could not break anymore.

The POS-disambiguation module applies its rules in sequential order, eliminating readings that violate constraints stated in the rules. First, the relative pronoun reading (PRELS) can be eliminated in sentence-initial position. The demonstrative pronoun reading (PDS) is also ungrammatical in sentence-initial position followed by an unambiguous noun (NN) and a finite verb (VMFIN), since it cannot construct a phrase with the noun. This constraint relies on the theory of topological fields (Höhle, 1985), according to which only one element or phrase can occupy a Vorfeld position (i.e., the position between a clause boundary and a finite verb). A finite verb reading (VVFIN) of *bremsen* in clause-final position can be eliminated since there is a preceding unambiguous finite verb. An imperative verb reading (VVIMP) is ungrammatical in a non-clause-initial position and can also be deleted.

All the constraints mentioned above eliminate ungrammatical readings. Another possible operation of the disambiguation rules is to identify the correct analysis among the set of legitimate readings and to delete all the others. A heuristic of this type chooses an adverbial reading for *mehr* if the immediate left context contains a negation particle (PTKNEG).

After the application of POS disambiguation rules the sentence is further processed by the morphological disambiguation module.

[2]For expository purposes, (1) shows only POS ambiguity but ignores morphological features.

Table 2 demonstrates the remaining morphological ambiguity for each token of the sentence. Morphological information is encoded in the tags after the fullstop separator. For articles and nouns this information contains *case* (`a`ccusative, `d`ative, `g`enitive or `n`ominative), *number* (`s`ingular or `p`lural) and *gender* (`f`eminine, `m`asculine, `n`eutral or underspecified (`*`)) values; verbal morphology includes *person* (`1`st, `2`nd or `3`rd), *number* (`s`ingular or `p`lural), *mood* (`i`ndicative or subjunctive (`k`)) and *tense* (pas`t` or presen`t`) values.

(2)	Der	ART.dsf, ART.gsf, ART.nsm ART.gp*, ART.gpf, ART.gpn, ART.gpm
	Fahrer	NN.asm, NN.dsm, NN.nsm NN.apm, NN.gpm, NN.npm
	konnte	VVFIN.3sit, VVFIN.1sit
	nicht	PTKNEG
	mehr	ADV
	bremsen	VVINF
	.	$.

Concord rules in the morphological disambiguation module rely on the fact that lexical nodes within the same NP mutually constrain each other as to the set of possible readings. Application of such rules leads to elimination of all non-shared readings on tokens *Der Fahrer*, reducing the set of possible analyses to `.gpm` and `.nsm` for both tokens. Further disambiguation of the tokens *Der Fahrer* is performed by a syntactic heuristic that restricts the use of genitive NPs to positions preceded by a preposition or another NP. Resolving the *person* ambiguity of the finite verb is based on the absence of a nominative pronoun with first person value in the clause, which allows to eliminate the first person reading of the verb.

This example has shown that concord rules in conjunction with syntactic heuristics can effectively reduce the number of candidate readings.

5.2 Evaluation of the rule-based disambiguation module

Table 2 provides the results of the experiments with the rule-based disambiguation module. The first line gives a baseline for the module performance as the performance of the morphological analyzer. The next two lines stand for the POS and morphological disambiguation modules, respectively.

The table contains numbers for precision, recall and f-measure, as well as the percentage of

module	precision	recall	F-measure	LE	DE	ambiguity tokens	rate
morph. analyzer	13.61%	96.64%	23.86%	100%	0%	68.76%	9.87
POS disamb.	19.93%	96.11%	33.01%	86.01%	13.99%	59.79%	7.39
morph. disamb.	42.53%	94.86%	58.73%	64.51%	35.49%	31.05%	4.96
+ adding analyses	46.93%	95.64%	62.97%	60.12%	39.88%	30.13%	4.44

Table 2: Evaluation of the rule-based disambiguation module

ambiguous tokens in the test data together with the ambiguity rate for ambiguous tokens. To simplify comparison with the results obtained by other researchers, the formulas described in the earlier literature (Hajič et al., 2001) are used:

$$Precision = \frac{\#Tokens\ with\ a\ correct\ tag}{\#Analyses\ generated}$$

$$Recall = \frac{\#Tokens\ with\ a\ correct\ tag}{\#Tokens\ in\ data}$$

$$F-measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Following (Volk and Schneider, 1998) we split the errors made by the module into lexical errors (LE; column 5) and disambiguation errors (DE; column 6). Lexical errors are caused by the morphological analyzer: the correct analysis is not present among the set of analyses assigned by it. Disambiguation errors are proper errors of the module: the correct analysis was deleted during rule application.

As column 5 (LE) of Table 2 shows, the majority of the errors are lexical errors which are due to the deficiency of the morphological analyzer. A big part of such errors concerns foreign material and proper names – these lexemes do not contain necessary morphological clues for successful identification of correct analyses and are often confused with other parts of speech: common nouns, adjectives and verbs.

An attempt was made to decrease the initial error rate caused by the morphological analyzer. Unknown words were looked up in the external list of foreign material words and proper names extracted from the training data, as well as in the list of personal names extracted from the online newspapers of the period of 2001-2003.[3] If the lists contained the unknown word, the analyses provided by the morphological analyzer were replaced with the analyses from the lists.[4] Performance of the disambiguation module run on such an extended input is shown in the last line of Table 2.

Another major source of lexical errors is the confusion between adverbs and predicative adjectives, the annotation of which is often guided by semantic criteria and constitutes a difficult case for the morphological analyzer. Together with errors caused by rule over-application, lexical errors lead to a decreased recall.

The low precision (46.93%) of the module is due to the remaining ambiguity on the tokens. The disambiguation rules are mostly eliminative in nature, and in many cases surrounding context does not provide enough evidence for deleting an analysis as ungrammatical.

While for some applications the ambiguity on 30% of tokens and high accuracy may be a reasonable starting point, for dependency parsing, which relies on unique morphological analyses of verbs and their arguments and modifiers, such a high rate of ambiguity constitutes a major obstacle in reliably identifying the dependency structure. Therefore, further means for decreasing the ambiguity rate are needed. This additional disambiguation is performed by the statistical module.

6 Statistical module

(Hinrichs and Trushkina, 2003) have shown that a tagger based on probabilistic phrase structure grammars (PCFGs) is better suited for the task of morpho-syntactic annotation of German than n-gram taggers, which currently are the most-widely used class of taggers for natural language processing. The reason for this lies in the fact that PCFGs are able to incorporate more global structural information and can therefore capture long-distance dependencies between tokens that have to be taken into account for correct

[3] Thanks to Christian Biemann for making this list available for us.

[4] Analyses were added to 1.62% of tokens in test data.

module	precision	recall	F-measure	no tag	LE	DE
statistical	89.20%	88.10%	88.68%	1.23%	11.55%	88.45%

Table 3: Evaluation of the statistical module

assignment of morpho-syntactic categories.[5]

A PCFG tagger described in (Hinrichs and Trushkina, 2003) was used as a statistical module of the combined system presented in the current paper. The tagger was retrained on a larger set of training data. The initial grammar and lexicon of the tagger were also extended. The extension involved inclusion of new rules in the grammar and of new tokens and their analyses in the lexicon.

6.1 PCFG tagger

The PCFG parser LoPar (Schmid, 2000) was used in the statistical module experiments. LoPar provides a tagging mode which defines the best tag sequence as the sequence of those tags that yield the maximal product of the inside and outside probabilities among the candidate tags for a given token. Thus, the tagger bases its decision about tag assignment on both the probability of a tag, given the token, and the probability of a tag, given the full surrounding context of the token. However, due to the independence assumption about the distribution of words and phrases which is inherent in the PCFGs, the influence of the surrounding context tends to decrease the further it is removed from the focus token. To weaken this independence assumption, systematic transformations of the tree structures and enrichment of the node label set are introduced so as to boost those contextual features that are external to individual local trees.

The tagger was trained on the 104 048 tokens from the Tüba-D/Z treebank with transformed tree representations. In line with techniques and insights of (Pereira and Schabes, 1992), 115 098 tokens from the second part of the treebank are used for weakly supervised training of the statistical module. These tokens, although tagged and labeled only partially, help to reduce data sparseness problem and lead to improved performance. Further improvement is due to the inclusion of 60 901 tokens from the Negra corpus into the tagger lexicon. The lexicon is also extended by a list of possible morphological analyses for each unknown word.[6] The evaluation of the statistical module is given in the next subsection.

6.2 Evaluation of the statistical module

Table 3 summarizes the performance of the statistical module. Precision, recall and f-measure are calculated in the same way as for the rule-based module. The additional metric `no tag` represents the percentage of tokens for which no tag was assigned. The tagger is unable to assign a tag to a token if the PCFG cannot provide any parse for a sentence due to data sparseness.[7]

In case of the statistical module, errors are considered lexical if the correct tag is not present in the lexicon of the tagger. Otherwise they are viewed as disambiguation errors.[8] It is important to note that the set of analyses provided for a token in the input of the rule-based module may be different from the set of analyses contained in the tagger lexicon for the same token, since the tagger lexicon is compiled from the training data and augmented by the analyses from the morphological analyzer for unknown tokens, whereas the input for the rule-based module consists almost exclusively of the analyses provided by the morphological analyzer.[9] Therefore, the tagger lexicon may contain correct analyses for words which are unknown to the morphological analyzer. Given that the test and the training data come from the same source of newspaper articles, such improvement of the tagger lexicon over the morphological analyzer often happens for proper names and foreign material tokens. This fact,

[5] We are aware of the fact that the term *long-distance dependencies* in linguistics usually refers to unbound dependencies between fillers and arguments. Here we are using the term more loosely to refer to dependencies between tokens that more often than not involve a large context window. Well-known cases of this sort in German are dependencies between verbs and separable verbal particles (the latter can be confused with adverbs, pre- or postpositions) and between finite and non-finite verbs in complex verb forms.

[6] Such a list was generated by the Xerox morphological analyzer.

[7] This adverse effect on recall is clearly due to the modest size of the training data and should be alleviated as more training data are added.

[8] Distribution of errors is given only for tagged tokens.

[9] Only for some unknown tokens analyses were taken from the external lists.

experiments	precision	recall	F-measure	no tag	LE	RBE	SE
full input	90.23%	86.29%	88.22%	4.37%	0%	42.98%	57.02%
partial input	90.59%	87.67%	89.11%	3.23%	8.12%	19.54%	72.34%

Table 4: Evaluation of the combined model

together with the fact that the statistical module has to assign only one tag, explains the opposite distribution of lexical and disambiguation errors for the statistical module.

The combined model aims at bringing together the strengths of the rule-based and the statistical approaches: the rule-based module accurately narrows down the set of possible analyses for input tokens and provides unique analyses for almost 70% of the tokens. The statistical module resolves the remaining ambiguity, selecting for every token the most probable reading, given the context.

7 Combined model

Two experiments were performed with the combined model. In the first experiment all analyses left after application of the rule-based module are provided as input to the statistical module. The search space of the statistical module is thus restricted to the readings that the rule-based component considers grammatical. In case there is a single analysis for a token available, it remains unchanged. The main advantage of such a strategy is the elimination of ungrammatical readings prior to probabilistic processing. This helps to avoid errors made in cases that are traditionally hard for statistical models and are comparatively easy for rule-based approaches (such as long distance dependencies among tags). The drawback of the strategy, however, consists in carrying all the errors of the rule-based component into the final model performance, which mainly concerns unknown words: the rule-based module will produce an error if the lexicon or the guesser of the morphological analyzer provides a set of hypotheses that does not include the correct analysis.

In the second experiment the input to the statistical model was limited to the categories that are most reliably tagged by the rule-based module. This provides the intended division of labor between the two modules according to their strengths: the rule-based component eliminates analyses in sure cases and performs disambiguation based on long-distance relations, whereas the statistical module resolves remaining ambiguity and assigns tags to unknown tokens.

The input to the statistical model in the second experiment was prepared in the following way: all unambiguous analyses produced by the rule-based module are included in the input, except for the analyses of predicative adjectives (the morphological analyzer often mistakens them for adverbs), of imperative verbs (these are often erroneous analyses that are actually foreign material), and of proper nouns without morphology. In addition, all analyses that have unambiguous POS readings of article, attributive adjective, finite verb, demonstrative, relative or personal pronoun are included in the input for the statistical module, since these parts of speech in most cases are tagged correctly by the rule-based module and since resolution of remaining morphological ambiguity for them does not usually constitute a problem for the statistical module. Due to the unreliable treatment of unknown words by the rule-based component, the statistical module is used as its own pre-processor to identify categories that most often correspond to unknown words. These categories include foreign material (FM), special symbols (XY) and pronominal adverbs (PROP). The tags for these categories are then included into the input to the combined model, replacing the analyses of the rule-based module.

The results achieved in the experiments described above are given in the next subsection.

7.1 Evaluation of the combined model

Table 4 presents the performance of the combined model. The first line provides statistics for the model that takes full input from the rule-based module. The second line demonstrates performance of the model that takes partial input of most reliable categories from the rule-based component. To reflect the impact of the errors by the rule-based component on the error rate of the model, the disambiguation errors are split into errors of the rule-based module (RBE; column 7) and errors of the statistical module (SE; column 8).

errors	POS	case	number	gender	person	tense	mood
SE	27.54%	40.51%	2.01%	10.03%	0.94%	0.00%	0.80%
RBE	62.87%	20.79%	1.49%	2.48%	0.00%	0.00%	0.00%
LE	70.24%	5.95%	3.57%	1.19%	0.00%	0.00%	0.00%
all	37.91%	33.85%	2.03%	7.83%	0.68%	0.00%	0.58%

Table 5: Error analysis for the combined model with partial input

experiments	precision	recall	F-measure	no tag	LE	RBE	SE
full input	92.04%	88.35%	90.16%	4.00%	0%	22.93%	77.07%
partial input	91.82%	89.02%	90.40%	3.05%	0%	10.21%	89.79%

Table 6: Evaluation of the combined model given the perfect lexicon

7.2 Error analysis for the best model

Table 5 demonstrates the distribution of errors among the morpho-syntactic features. In the second column (POS) the percentage of errors involving POS categories is presented. Columns 3-6 provide the error distribution that occur (only) in one of the morphological feature values, while the values of the other morphological features and of the POS category are correct. Errors involving more than one morphological feature are ignored, since in such cases it is unclear which feature is ultimately responsible for the error. Isolating errors of morphological features in this way clearly demonstrates that case, together with POS, is the hardest category to disambiguate correctly.

8 Using a perfect lexicon

The error analysis of the model performance demonstrates a high rate of lexical errors, i.e. errors caused by the deficiency of the morphological analyzer: lexical errors amount to 11.55% in the statistical module (see Table 3) and to 60.12% in the rule-based module (see Table 2). An additional experiment which aims at the evaluation of the proper model and disregards the initial lexical error rate was designed. In this experiment, a set of possible analyses for every token was augmented by a correct analysis in case it was not originally provided by the morphological analyzer. The data were further processed as described in the experiments above. The experiment facilitates comparison of the current model to the models described in the literature which usually assume a perfect lexicon available.[10]

[10]Compare, for example, the recall of 100% reported for the morphological analyzer in (Hajič et al., 2001).

The performance of the model given the perfect lexicon is shown in Table 6. The improvement over the basic model performance amounts to 1.29% in f-measure and 1.45% in precision.

9 Conclusion

The combined model outperforms the rule-based and statistical modules applied in isolation. The best result of the model attains an accuracy of 92.04%, which corresponds to a 7.34% improvement of the best results reported by other researchers for the same task for German (Lezius et al., 1996).

By comparison, the best results for standard POS tagging of German achieved an accuracy of 96.70% (Brants, 1998). The difference in performance reflects the relative difficulty of the task. Morpho-syntactic annotation with a large tagset is much harder than standard POS tagging described in (Brants, 1998), since it employs a tagset which is 13 times larger than the standard POS tagset STTS used in (Brants, 1998) and since it has to deal with much more severe data ambiguity (cf. statistics for "German" and "German (STTS)" in Table 1).

Expansion of the tagset, on the other hand, allows for a much higher range of applications for the tagger. Among the possible applications is the use of the tagger in dependency parsing. Identification of the case value on noun phrases plays a crucial role in correct dependency assignment for languages with free word-order, such as German. To assess the utility of the model for this task, an additional evaluation on only the case value (disregarding POS and all other morphological features) was performed, resulting in a precision of 92.26%. This evaluation demonstrates that the model provides a sound basis for the correct assignment of dependencies and highlights the applicability of the model.

References

Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. 2002. Robustness beyond shallowness: Incremental deep parsing. *Journal of Natural Language Engineering*, 8:121–144.

Thorsten Brants, 1998. *TnT–A Statistical Part-of-Speech Tagger.* Universität des Saarlandes, Computational Linguistics, Saarbrücken, Germany.

Jan Hajič and Barbora Hladka. 1997. Probabilistic and rule-based tagger of an inflective language - a comparison. In *Proceedings of ANLP'97*, pages 111–118, Washington, D.C., USA.

Jan Hajič, Pavel Krbec, Pavel Květoň, Karel Oliva, and Vladimír Petkevič. 2001. Serial combination of rules and statistics: A case study in Czech tagging. In *Proceedings ACL'2001*, pages 260–267, Toulouse, France.

Erhard W. Hinrichs and Julia Trushkina. 2002. Forging agreement: Morphological disambiguation of noun phrases. In *Proceedings of the First Workshop on Treebanks and Linguistic Theory (TLT'2002)*, pages 78–95, Sozopol, Bulgaria.

Erhard W. Hinrichs and Julia Trushkina. 2003. N-gram and PCFG models for morpho-syntactic tagging of German. In *Proceedings of Second Workshop on Treebanks and Linguistic Theories (TLT'2003)*, pages 177–188, Växjö, Sweden.

Tilman Höhle. 1985. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340, Göttingen, Germany.

Wolfgang Lezius, Reinhard Rapp, and Manfred Wettler. 1996. A morphology-system and part-of-speech tagger for German. In *Proceedings of KONVENS'96*, Bielefeld, Germany.

Kemal Oflazer and Gökhan Tür. 1996. Combining hand-crafted rules and unsupervised learning in constraint-based morphological disambiguation. In *Proceedings of the EMNLP'96*, Philadelphia, PA, USA.

Fernando Pereira and Yves Schabes. 1992. Inside-outside reestimation for partially bracketed corpora. In *Proceedings of ACL'92*, pages 128–135, Newark, DE, USA.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen, September.

Helmut Schmid. 2000. Lopar: Design and implementation. Technical Report 149, IMS Stuttgart. Arbeitspapiere des Sonderforschungsbereiches 340.

Pasi Tapanainen and Atro Voutilainen. 1994. Tagging accurately – Don't guess if you know. In *Proceedings of ANLP'94*, Stuttgart, Germany.

Heike Telljohann, Erhard W. Hinrichs, and Sandra Kübler, 2003. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).* Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.

Dan Tufiş, Péter Dienes, Csaba Oravecz, and Tamás Váradi. 2000. Principled hidden tagset design for Tiered Tagging of Hungarian. In *Proceedings of LREC'2000*, pages 1421–1426, Athens, Greece.

Dan Tufiş. 2000. Using a large set of EAGLES-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In *Proceedings of LREC'2000*, pages 1105–1112, Athens, Greece.

Martin Volk and Gerold Schneider. 1998. Comparing a statistical and a rule-based tagger for German. In *Proceedings of KONVENS'98*, Bonn, Germany.