Rule-based and Statistical Approaches to Morpho-syntactic Tagging of German

Erhard W. Hinrichs and Julia S. Trushkina

University of Tübingen, Seminar für Sprachwissenschaft, Wilhelmstr. 19, 72074 Tübingen, Germany

Abstract. Rule-based and statistical approaches constitute the two leading paradigms in computational linguistics. This paper applies the two types of approaches to the task of assigning morpho-syntactic categories to words in German, a language with rich inflectional morphology. The rule-based approach uses the Xerox Incremental Deep Parsing System and provides a novel constraint-based framework that integrates phrase-internal concord rules and phrase-external syntactic heuristics into one uniform architecture. The statistical approach utilizes the PCFG-parser LoPar which yields acceptable results even for moderate amounts of manually-annotated treebank training data. It is shown that tree transformations constitute a crucial step in weakening the independence assumptions inherent in probabilistic context-free grammars and in optimizing the performance for the task at hand.

1 Introduction

Over the last decade, statistical approaches and approaches based on machine learning (ML) have gained great popularity in computational linguistics. One of the tasks to which these models have been successfully applied is the assignment of morpho-syntactic categories to words, i.e., morpho-syntactic tagging. The first attempts of automatic tagging utilized rule-based methods [7]. More recently, tagging systems based on Hidden-Markov Models (HMM) [1,4] or ML techniques [5,15,17] have become the pre-dominant paradigms for this task. However, rule-based approaches are still being pursued and have proven to be competitive [2,11,22].

With a few notable exceptions [6,8,10,21], most statistical and ML approaches have focused on tagging with fairly small tagsets, distinguishing only between parts of speech (POS), but not taking into account more fine-grained distinctions of morphology, such as *case*, *number* and *gender* for nouns or *person* and *number* for verbs. At the same time, it is known that languages with rich inflectional morphology, such as Czech and German, constitute a severe problem for statistical approaches, due to the size of the tagset required for such languages and due to the non-local nature of the contextual cues which have to be taken into account for morphological disambiguation.

The purpose of this paper is to re-assess the relative strengths of rule-based and statistical approaches for one of the languages in question (German) and to compare the results of practical experiments conducted with a rule-based approach, the Xerox Incremental Deep Parsing System (XIP) [3], and with a statistical approach, the tagging mode of PCFG-parser LoPar [18].

2 Morpho-syntactic Tagging of German

For languages with (partially) free word order and with (relatively) rich inflectional morphology, case, number and gender information plays a crucial role in determining the syntactic function of noun phrases (NPs). Consider the example in (1).

(1)	Die	Politiker	gaben	verdienten	Beamten	und Lohnempfängern
	the	*politicians*	*gave*	*worthy*	*civil servants*	*and wage recipients*
	ein	höheres	Gehalt.			
	a	*higher*	*salary*			

For German, grammatical functions such as *subject*, *direct object*, and *indirect object* are strongly correlated with nominative, accusative, and dative case, respectively. Therefore, determination of case information is crucial for the correct interpretation of a sentence. However, case syncretism, i.e., the phenomenon that one and the same form may express two or more cases, is a wide-spread property across languages. German is no exception in this regard.[1] Only one NP in sentence (1) exhibits no case ambiguity: the form *Lohnempfängern* is unambiguously *dative plural*. However, all other NPs are multiply ambiguous. For example, the noun *Politiker* can be any case and number, except for *genitive singular* and *dative plural*. Thus, morphological disambiguation of this kind is a crucial prerequisite for semantic interpretation.

Morphological disambiguation may be considered as either part of syntactic parsing, i.e., as part of the task of assigning the correct constituent structure to a sentence. Alternatively, morphological disambiguation can be viewed as part of morpho-syntactic tagging, i.e., the assignment of information about word class and morphological features to individual words. In this paper we will pursue the latter strategy, partly because not all grammatical frameworks recognize syntactic constituency as an independent level of linguistic analysis.[2] In this sense, the tagging approach to morphological disambiguation provides the more general solution.

One direct consequence of incorporating morphological disambiguation into tagging is a significant expansion of the tagset. The STTS tagset [16], the widely accepted inventory of part-of-speech categories for German, comprises 54 distinct labels. When the STTS labels are enriched by morpho-syntactic

[1] Compared to other languages where case syncretism tends to conflate the same cases across different nominal paradigms, German case syncretism is particularly intricate since case syncretism is often paradigm-dependent.

[2] Dependency grammar [14,20] is a framework of this kind, treating syntactic dependencies as relations between individual words.

features such as *case*, *number*, *person*, *gender*, *tense* and *mood*, the size of the resulting tagset increases to 718 distinct tags. Such a significant expansion of the tagset has an adverse effect on the tagging accuracy of HMM taggers, since the performance of the tagger is strongly dependent on the size of the tagset and the amount of training material available.

Moreover, some essential features that supply necessary evidence for correct annotation of inflectional morphology, such as clause type, word order and subject-verb agreement, are of a non-local nature. [10] demonstrates that this non-locality of relevant syntactic properties and relations makes HMM models insufficient for the given task.

In the remainder of this paper, we will present two alternative approaches that can cope with such expanded tagsets and with the non-local nature of contextual cues for morphological disambiguation.

3 A Rule-based Approach

The rule-based approach to morpho-syntactic tagging described here is implemented as a morphological disambiguation module (MDM) based on the Xerox Incremental Deep Parsing System.

Morphological information provided by the morphological analyzer for the input string is typically highly ambiguous. In sentence (1), the only morphologically unambiguous noun phrase is the NP *Lohnempfängern* with the unique analysis `Noun+Masc+Pl+Dat`. The lexical nodes for all other NPs in the sentence are morphologically many times ambiguous. For example, as shown in (2), the NP *verdienten Beamten* has 25 analyses for the adjective and 10 analyses for the noun which include all possible values for case, number, gender and declension type features.[3]

(2)	verdienten	Adj+Fem+Sg+DatGen+Wk
	verdienten	Adj+Masc+Sg+AccGen+StWk
	verdienten	Adj+Masc+Sg+Dat+Wk
	verdienten	Adj+Neut+Sg+Gen+StWk
	verdienten	Adj+Neut+Sg+Dat+Wk
	verdienten	Adj+FMN+Pl+NomAccDatGen+Wk
	verdienten	Adj+FMN+Pl+Dat+St
	Beamten	Noun+Masc+Sg+AccGen+StWk
	Beamten	Noun+Masc+Sg+Dat+Wk
	Beamten	Noun+Masc+Pl+NomAccDatGen+Wk
	Beamten	Noun+Masc+Pl+Dat+St

[3] Beside separate morphological tags, the analyses include joint morphological tags, such as `NomAccDatGen` standing for any case, `FMN` standing for any gender, and `StWk` standing for strong or weak declension type. Each joint tag, thus, expands into several distint analyses.

In this section we will discuss two types of disambiguation rules implemented in MDM, *concord rules* and *syntactic heuristics*, which jointly provide an effective way to reduce morphological ambiguity. We will illustrate both types of rules on the example of sentence (1).

Concord rules are based on mutual agreement constraints between lexical nodes within one NP. In (2) the adjective *verdienten* has ambiguous values for all morphological features. However, in the context of the following noun *Beamten*, only masculine readings are valid since they match the gender specification of the noun. This significantly reduces the set of possible readings.

Further pruning of contextually valid readings is possible. In German, word forms for determiners, adjectives and adjectival nouns can be classified as belonging to either weak or strong declension classes.[4] If NPs do not include an overt determiner, as in (2), then only strong forms of adjective and noun are grammatical, and all weak forms can be eliminated. This results in only three analyses remaining for the NP *verdienten Beamten*.

If determiners co-occur with adjectives and nouns in the same NP, as in the NPs *ein höheres Gehalt* and *die Politiker* in example (1), adjective and noun agree in declension class, whereas the declension value of the determiner is the opposite. This regularity can be implemented in a concord rule together with mutual agreement constraints. Application of the rule to the NPs *ein höheres Gehalt* and *die Politiker* results in complete disambiguation of gender, number and declension type values. The only remaining ambiguity concerns the nominative and accusative analyses for both NPs.

More generally, concord rules need to be able to check feature agreement or feature disagreement among adjacent lexical nodes within the same phrase. In order to correctly determine the phrase boundaries and hence the domain of application, concord rules may need to take into account the local left and/or right contexts of the nodes in question.[5]

After the application of concord rules to the four NPs in sentence (1), the remaining ambiguities concern the case values of the NPs *verdienten Beamten*, *die Politiker* and *ein höheres Gehalt*. These ambiguities can be resolved by syntactic heuristics. Such heuristics impose morpho-syntactic constraints between lexical nodes which belong to different phrases.

In example sentence (1), *verdienten Beamten* is coordinated with the NP *Lohnempfängern*, which is unambiguously `Noun+Masc+Pl+Dat`. Since conjoined NPs have to agree in case, the noun in (2) also has to be `Dat`.[6] Thus,

---

[4] For a comprehensive study of the distributional properties of weak and strong forms in German NPs see [23].

[5] See [11] for a discussion of the implementation of such rules.

[6] The robust recognition of different types of conjunctions, e.g. NP conjunction, VP conjunction, and sentential conjunction, is a complex issue in itself. In the case at hand, NP conjunction is the only candidate since there is no verbal constituent following the conjunction.

the only contextually valid reading for the NP in (2) is the sequence of morphological tags `Adj+Masc+Pl+Dat+St`, `Noun+Masc+Pl+Dat+St`.

The NPs *die Politiker* and *ein höheres Gehalt* can both be either nominative or accusative case. Thus, in principle, both NPs can serve as either the subject or direct object of the finite verb. However, this ambiguity can be resolved due to subject-verb agreement. Since the finite verb is plural, only the plural NP *die Politiker* can be the subject, and the NP *ein höheres Gehalt* should be the direct object.

What this example has shown is that concord rules in conjunction with syntactic heuristics can effectively reduce the number of candidate readings. In the remainder of the section we will give a quantitative evaluation of the MDM.

At present, the MDM contains a total of 106 concord rules which aim at morphological disambiguation of German NPs. Coverage of the rules includes prenominal agreement (with determiners, adjectives, cardinals, measure phrases, participial premodifiers, etc.), head-pronoun agreement for relative clauses, case agreement with prepositions, subject-verb agreement, agreement in complex proper names and titles, as well as simple nominal coordinations. The main reason for concentrating on NP disambiguation is that it is the most crucial source of ambiguity for the assignment of grammatical functions, as discussed in section 2.

The MDM was evaluated on a corpus of 5732 lexical tokens extracted from the `taz` newspaper corpus [19]. This test corpus contains a total of 1571 NPs. The corpus was automatically annotated by the XRCE morphological analyzer for German and then manually corrected so as to provide a gold standard for the present evaluation. The corpus has an average number of 3.68 distinct readings per token. 45.18% of all tokens are morphologically unambiguous. For lexical nodes that are contained in NPs, the average number of distinct readings is significantly higher: 6.08 per token; and only 13.00% of the nodes have a unique analysis. The fact that NPs exhibit a much higher than average degree of ambiguity further attests to the priority that has to be given to morphological disambiguation of NPs.

Table 1 shows the distribution of the number of disambiguated readings for NPs after application of disambiguation rules.[7] Table 1 shows that in three out of four NPs, a unique reading can serve as input to further processing modules.

Table 2 provides numbers for quantitative measures of precision and recall for the task. By *recall* we mean the number of NPs for which the MDM provides a unique and correct morphological analysis divided by the total number of NPs in the corpus. By *precision* we mean the number of NPs for which the MDM provides a unique and correct morphological analysis

[7] Since gender ambiguities never play a role in the determination of grammatical functions, we disregard gender ambiguities in determining what counts as a unique reading.

Table 1. NP readings after application of disambiguation rules

1 reading	1211	77.08%
2 readings	226	14.39%
$\geq$ 3 readings	134	8.53%

divided by the number of NPs for which the MDM provides a unique, but not necessarily correct, morphological analysis. Since in statistical approaches evaluation is usually done on lexical nodes and not on phrases, we also provide precision and recall numbers for lexical nodes contained in NPs to make further comparison with statistical methods easier.

Table 2. Precision and recall for morphological disambiguation module

	NPs	NP lexical nodes
recall	76.26%	78.32%
precision	98.93%	99.02%

4 A Statistical Approach

As mentioned in section 2, HMM-based taggers, currently the most-widely used class of taggers for natural language processing, do not perform well for the task at hand, due to the size of the tagset and the non-local nature of the contextual cues which have to be taken into account for morphological disambiguation. [10] reports an accuracy of 80.61% as the best result they could achieve using the TnT tagger of [1] for morphological disambiguation of German. A suitable alternative is provided by probabilistic context-free grammars (PCFGs) which take more context into account and can therefore capture long-distance dependencies.

It is important to note that applying PCFGs to part-of-speech tagging constitutes a rather different task from the ordinary use of PCFGs in probabilistic parsing. For PCFG parsing, a sequence of part-of-speech tags is assumed as input, and the nature of the problem is how to project the correct syntactic structure over this given input. By contrast, when PCFGs are used for tagging, the sequence of tags is unknown, and the task is to find the model of rule probabilities that best predicts the correct part-of-speech sequence. What both tasks have in common is the need to extract those regularities from the given input that are most predictive for the structure to be found. In the case of PCFG parsing, this includes but is not limited to the identification of lexical head-head relationships. In the case of PCFG tagging, it is of primary importance to identify those regularities of syntactic structure that are highly predictive for the assignment of the correct tag sequence.

We performed a series of experiments with the PCFG-parser LoPar [18] trained on the same portion of data with different tree representations and node label encodings. LoPar provides a tagging mode that outputs the sequence of part-of-speech tags with the highest estimated frequency, i.e., the product of the inside and outside probabilities of a tag divided by the overall probability of the parse forest. The parser was trained on 51288 manually annotated tokens from the taz newspaper portion of the TüBa-D treebank and tested on 8949 tokens from the same corpus. Additional 5854 lexical tokens served as the development set. A back-up lexicon containing the set of possible morphological analyses for each unknown word was provided for the parser. A baseline accuracy of 77.65% was achieved when the parser was trained on the data that had original structure inherent to the treebank.

A weakness of PCFGs is the strong independence assumption about the distribution of words and phrases which is inherent in the context-freeness of the underlying grammar formalism. As [12] has shown, one of the ways to weaken these independence assumptions is to encode more information in node labels. This can have a dramatic effect on the performance of a PCFG parser.

The LoPar experiments show that introduction of pertinent linguistic evidence such as *case*, *number*, *grammatical function* and *finiteness* on node labels and systematic transformations in tree structure lead to a significant improvement of the results over the baseline accuracy. The optimal performance was gained by eliminating information contained in the original treebank and by percolating relevant morphological and functional information between lexical and phrasal nodes in combination with binarization of tree structure:

Elimination of information: Topological fields such as `VF`, `MF`, `NF` (see [9] for details) are present in the original treebank as an intermediate layer of syntactic structure. Elimination of all topological field information except `C` and `VC` proved advantageous since retention of only these two fields turned out to be sufficient for determining the syntactic macrostucture across different clause types.

Percolation of relevant information: Due to the context-freeness of the underlying grammar formalism, regularities in syntactic structures should whenever possible be encoded as mother-daughter relations. In the simplest case regularities among sister constituents are best encoded by percolating this information onto the mother node. In the case of more non-local dependencies information needs to be percolated to the first common ancestor. Percolation of the following morphological and functional information between lexical and phrasal nodes proved particularly significant:

1. Morphological information (case, number and gender) on phrasal categories that constitute an NP (nouns, adjectives, determiners etc.) and partial morphological information (number) on verbal categories.

2. Function labels -OA and -ON (accusative object and subject) on nodes: passing them up to SIMPX node (node of a clause) allows to capture subject-object regularities in a clause: presence of one subject in a clause and prototypicality of clauses with both subject and accusative object.
3. FIN-label of finite verbs: passing it up to SIMPX node allows to trace regularities between finiteness of a clause and morphological cases of NPs in a clause: necessary presence of subject in finite clause, absence of subject in non-finite clause. Moreover, introducing number feature to -FIN and -ON nodes makes it possible to capture subject-verb agreement.

Binarization of SIMPX subtrees: Enriching syntactic categories by morphological and functional information leads to a severe increase in the size of the rule set with an accompanying drop in recall (i.e. an increase in the number of unparsed sentences). This can be counterbalanced by binarization of tree structures, which makes the grammar more flexible by allowing for structures not present in the training data to be created in the tagging phase.

Table 3 summarizes transformations in tree structure and different node label encodings used in the experiments and the impact of the transformations on the performance of the model on the development set. The best result obtained has an accuracy of 87.69%, which amounts to significant reduction of 55.08% in the error rate of the baseline.

Table 3. LoPar PCFG experiments

	precision	# of unparsed sentences
baseline:	**77.65**	5
topological fields (except for VC and C) deleted	77.58	5
case passed up to NX and NCX	84.23	6
grammatical functions (-ON and -OA) added and passed up to SIMPX + rules binarized	84.98	5
morph. info passed up to NXs and VXFINs	87.62	7
FIN label with number passed up to SIMPX	**88.21**	9
results on test data	**87.69**	11

The present paper, thus, confirms the findings of [12,13] that treebank transformations play a crucial role in the application of PCFG models to natural language data and generalizes this result to a novel use of PCFGs for morpho-syntactic tagging of highly inflectional languages with large tagsets. It underscores the importance of careful linguistic fine-tuning of such models.

5 Comparison and Conclusion

The previous section has shown that statistical taggers based on PCFG models significantly outperform HMM taggers for the same task, using the same experimental data. The accuracy of 87.69% achieved by the LoPar model amounts to a significant reduction in the error rate of the best TnT model: a 36.51% reduction.

A comparison of the results of the rule-based and the statistical methods is not as straightforward. While both methods address the problem of morphological disambiguation, the exact specification of the task involved differs in each case: the rule-based disambiguation module relies on POS-disambiguated input and applies only to lexical tokens within NPs.[8] The PCFG model, on the other hand, applies to lexical tokens of any category and takes as input all possible morphological analyses of a given token. In both respects, the task of the PCFG model is more challenging.

Despite the differences in the tasks involved, the research results reported in this paper are sufficient for pointing out advantages and drawbacks of the two methods. Although the rule-based approach provides much more accurate results than statistical methods (precision reaches 99.02%), implementation of rules is time- and labor- consuming. High precision also comes at a price of a lower, comparatively to statistical methods results, recall. The main reason for this is that, unlike statistical methods, which always aim at assigning the most probable tag to each token, the MDM eliminates morphological readings only if there is hard linguistic evidence in terms of concord rules or syntactic heuristics. In this sense the rule-based approach is more cautious.

Which of the two approaches is more suitable will very much depend on the intended application which requires morphological disambiguation. If a unique analysis for every word in the input is required, then a statistical approach will be more appropriate. However, the more cautious rule-based approach will be preferable if partially ambiguous input is acceptable which can be further disambiguated by subsequent processing modules.

References

1. Brants, T. (2000) TnT – A Statistical Part-of-Speech Tagger. In: *Proceedings of the 6th Conference on Applied Natural Language Processing, (ANLP 2000)*.
2. Brill, E. (1992) A Simple Rule-Based Part of Speech Tagger. *Proceedings of the 3rd Conference on Applied Natural Language Processing, (ANLP 1992)*, 112–116.
3. Ait-Mokhtar, S., Chanod, J.-P., Roux, C. (2002) Robustness beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*. **8**, 121–144.
4. Church, K. (1988) A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second ACL Conference on Applied Natural Language Processing (ANLP 1998)*, 136–143.
5. Daelemans, W., Zavrel, J., Berck, P., Gillis, S. (1996) MBT: A Memory-Based Part of Speech Tagger-Generator. In: Ejerhed, E., Dagan, I. eds. *Proceedings of the Fourth Workshop on Very Large Corpora*, 14–27.

[8] Extension of the MDM to other categories as well as incorporation of the POS disambiguation module in the system is currently being investigated.

6. Dienes, P., Oravecz, C. (2000) Bottom-Up Tagset Design from Maximally Reduced Tagset. In: *International Conference on Computational Linguistics (COLING 2000)*, 42–47.
7. Greene, B., Rubin, G. (1971) Automatic Grammatical Tagging of English. Technical Report, Brown University, Providence, RI.
8. Hajič, J., Hladká, B. (1998) Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, 483–490.
9. Hinrichs, E., Bartels, J., Kawata, Y., Kordoni, V., Telljohann H. (2002) The Verbmobil Treebanks. In *5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2000)*, 107–112.
10. Hinrichs, E., Trushkina J. (2003) N-gram and PCFG Models for Morpho-Syntactic Tagging of German. In *Proceedings of the Second International Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö University Press, 81–92.
11. Hinrichs, E., Trushkina J. (to appear) Forging Agreement: Morphological Disambiguation of Noun Phrases. To appear in: *Journal of Language and Computation*.
12. Johnson, M. (1998) PCFG Models of Linguistic Tree Representation. In *Computational Linguistics 24(4)*, 613–632.
13. Klein D., Manning C. (2003) Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, 423–430.
14. Mel'čuk, I. (1987) Dependency Syntax: Theory and Practice. State University of New York Press, Albany.
15. Ratnaparkhi, A. (1996) A maximum entropy model for part-of-speech tagging. In: *Proceedings of the First Conference on Empirical Methods in Computational Linguistics (EMNLP 1996)*, 133–142.
16. Schiller, A., Teufel, S., Thielen, C. Guidelines für das Tagging deutscher Textkorpora mit STTS. Universität Stuttgart and Universität Tübingen.
17. Schmid, H. (1994) Probabilistic Part-of-speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing*.
18. Schmid, H. (2000) LoPar: Design and Implementation. *Arbeitspapiere des Sonderforschungsbereiches* 340, No. 149, IMS Stuttgart.
19. taz – die Tageszeitung (CD-ROM). (1999) September 1986 – May 1999. www.taz.de.
20. Tesnière, L. (1959) Éléments de Syntaxe Structurale. Librarie C. Klincksieck, Paris.
21. Tufiş, D. (2000) Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. In: *International Conference on Language Resources and Evaluation (LREC 2000)*, 1105–1112.
22. Voutilainen, A. (1995) Morphological Disambiguation. In: Karlsson, F., Voutilainen, A.,Heikkilä, J., Anttila, A. eds., *Constraint Grammar*, 165–285. Mouton de Gruyter. Berlin.
23. Zwicky, A. (1986) German Adjective Agreement in GPSG. *Linguistics*, 24.