Annotating topological fields and chunks – and revising POS tags at the same time

Frank Henrik Müller and **Tylman Ule**
Seminar für Sprachwissenschaft, Universität Tübingen
{fhm,ule}@sfs.uni-tuebingen.de

Abstract

Annotating a corpus of German with chunks, topological fields and clause boundaries is both a goal in itself and a step towards further syntactic annotation. Partial annotation can serve as data to test linguistic hypotheses and it can be used as a pre-structuring for further linguistic annotation steps. If, however, the underlying part-of-speech (POS) annotation is imperfect, these errors will be passed on to the subsequent levels of annotation and increase annotation errors on those levels. It is especially damaging for subsequent annotation if POS tags are incorrect which provide the framework of the German sentence by demarcating the topological fields and the clause boundaries (e.g. subordinators and verbs). This paper presents a method to automatically annotate a corpus of German with chunks, topological fields and clause boundaries, and improve tagging accuracy at the same time in order to increase the overall annotation accuracy. Tag improvement primarily relies on the linguistic knowledge encoded in the grammar for annotating the topological fields.

1 The topological field model

1.1 An outline of the model

The topological field model (cf. Höhle (1986)) is a well-established descriptive model of the constituent order in German and other Germanic languages. Topological fields describe sections in the German sentence with regard to the distributional properties of the verb (and the subordinator in subclauses). There are three different types of clauses as can be seen in Table 1:[1] verb-last clauses (VL), verb-first clauses (V1) and verb-second clauses (V2). VL clauses comprise all introduced subclauses, V1 clauses mainly comprise imperatives and yes/no questions and V2 clauses mostly comprise affirmative clauses. The topological fields CF/VCL (complementizer field/verb complex left) and VCR (verb complex right)[2] constitute the sentence bracket, relative to which the other fields can be described[3]. The section preceding the left part of the sentence bracket is called the *Vorfeld* (VF; initial field; only in V2 clauses), the section included in the sentence bracket is called the *Mittelfeld* (MF; middle field) and the section following the right part of the sentence bracket is called the *Nachfeld* (NF; final field). While the ordering of other constituents is relatively free in German, the ordering of topological fields is subject to syntactic restrictions which adhere to the unvarying pattern outlined in Table 1. Figure 1 shows that the topological field model is also capable of accounting for recursive structures. Subordinate clauses (INF, SUB, REL) may be embedded in topological fields and, with this model, it is possible to leave their attachment open until its disambiguation in further annotation steps.

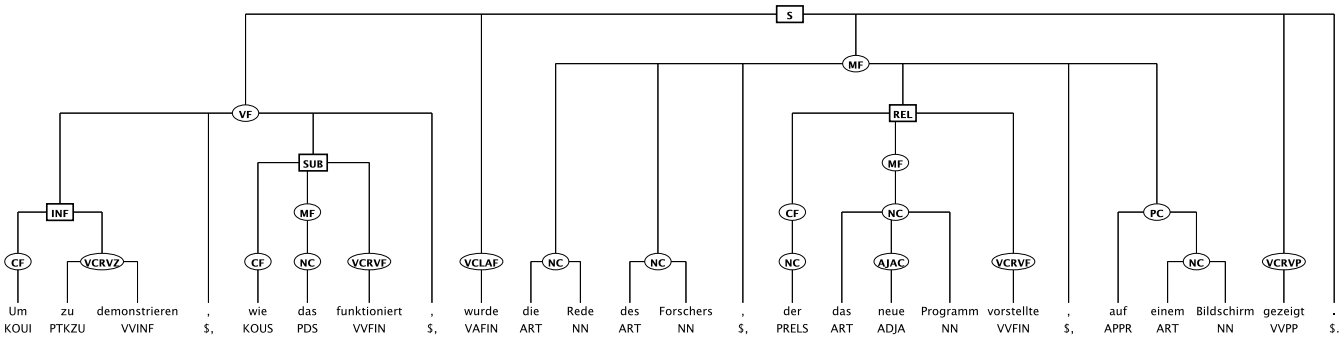
Table 1: The topological field model

clause type	topological fields						
VL:	KOORD	LV		**CF**	MF	**VCR**	NF
V1:	KOORD	LV		**VCL**	MF	VCR	NF
V2:	KOORD	LV	**VF**	**VCL**	MF	VCR	NF

[1] Cf. Figures 1 - 3 for examples for illustration.

[2] The fourth and following characters in verb complex labels abbreviate the verb POS tags (Cf. Figure 1).

[3] Obligatory fields are in bold type. The fields KOORD (coordination field) and LV (Linkversetzung, topicalization) will not be discussed in this paper. The representation in Table 1 is slightly simplified.



in order to demonstrate how that works was the talk of the researcher who the new program presented on a screen shown
'In order to demonstrate how that works, the talk of the researcher who presented the new program was shown on a screen.'

Figure 1: Sentence with shallow annotation including topological fields

1.2 The linguistic perspective

The topological field model is primarily a distributional model. It does not give any account of the verb-complement structure and it does not reveal the relation of the constituents within the topological fields, either. However, as argued by Meurers (2002), topological field annotation can "significantly help in using corpora from the perspective of theoretical linguistics". This becomes clear if one takes a look at Figure 1: The structure of topological fields along with POS tags marks the borders and the type of subclauses in a sentence. This shallow annotation thus provides the user with what one can call the skeleton of the sentence. Without the annotation of topological fields, it is by no means clear where the borders of the subclauses are and it is not clear where the potential complements of the respective verbs are. A query for a linguistic structure would therefore highly overgenerate, thus producing a high amount of false matches.

The shallow annotation as shown in Figure 1 gives the user of an annotated corpus the possibility to investigate various distributional phenomena which have to be described relative to the topological fields. The location of the embedding of subclauses is one of them, which has to be discussed in every linguistic theory and which is by no means agreed upon by linguists. Eisenberg (1999) is an adequate example: "Der Besetzung des Nachfeldes wird häufig eine kommunikativ-pragmatische Funktion zugeschrieben, [...]"[4] (p. 391). If linguists want to examine such a hypothesis, they are in need of an annotation which provides them with topological fields. A detailed account of the structure of the annotation in our system is given in Müller (2002).

1.3 The computational perspective

Concerning automatic annotation, one of the main advantages of topological fields is that they reduce the scope of ambiguity by dividing the sentence into fields and subclauses: constituents below clausal nodes can typically only be complements to verbs under the same clausal node (cf. squared nodes in Figure 1). The subclauses can now even be dealt with as single units, thus using a divide-and-conquer strategy similar to the one outlined in Peh and Ting (1996).

By first annotating topological fields, one can use a strategy termed *containment of ambiguity* by Abney (1996). This strategy proposes to annotate higher levels first if "reliable markers" are present because this considerably limits the number of possible attachment sites. Annotating topological fields first and verb-complement structure later transfers this strategy, which was originally used for chunking (i.e. basic phrase recognition), to the topological fields and clause level.[5]

[4] "A communicative-pragmatic function is very often attributed to the occupation of the Nachfeld, [...]" (our translation)

[5] A similar strategy has already been used to prestruc-

Another reason for splitting the annotation task of topological fields and chunks on the one hand and further annotation like verb-complement structure and phrase attachment on the other hand is that these linguistic phenomena adhere to two different organizing principles. While the former phenomena are mainly based on syntactic restrictions, the latter phenomena are to a large extent based on lexical selection. The syntactic pattern of topological fields in a clause is independent of the lexical entry for the respective head verb. Verb-complement structure, on the other hand, is clearly subject to lexical selection. Whether a verb takes a direct and an indirect object or not clearly depends on the lexical entry of the verb.

It is thus only logical to use two different annotation methods which fit the respective phenomena. Chunks and topological fields can be annotated with no other information than the POS tags, using a finite state transducer which works with a regular expression grammar. The rules for this grammar can be derived from linguistic knowledge. In this manner, the task of annotating a corpus is split into several tasks, thus generating a hybrid parser (cf. Hinrichs et al. (2002)).

2 POS Tagging

POS tags are annotated first in the present system. The topological field parser takes as input exclusively POS tags for syntactic annotation. Therefore, reliable POS tagging is crucial. The task of POS tagging is defined by a set of POS tags accompanied by guidelines that determine their application. In our system, the POS tags are given by the STTS German POS tagset containing 54 different tags (Schiller et al., 1995). A number of methods that combine the output of several taggers have proven to be successful in improving POS tagging (Borin, 2000; van Halteren et al., 1998).

In our system, we reduce errors of morpho-syntactic annotation along these lines by following a tagging-by-committee strategy which compares and assigns weighted probabilities to the output of several POS taggers for German, which vary in training data. For the system at hand, we use three instances of the TnT trigram tagger (Brants, 2000) trained separately on manually annotated news texts, on novels, and on all texts available.[6] Following the strategies outlined in van Halteren et al. (1998), the best POS tag is selected by simple majority voting extended by taking into account not only the number of taggers voting for each POS, but also the weights that the taggers assign to their choices. The POS tagging step results in a ranked sequence of POS tags, which is recorded in the linguistic markup for each word form token of an input text, so that later steps may access POS information in any required detail. We have tested the POS tagger by standard ten-fold cross-validation. The results are given in Table 2 (column "uncorrected").

3 Syntactic parsing and tagging errors

From what has been said in section 1 it should be clear that the correct annotation of the tokens in the sentence bracket is crucial for syntactic parsing in general and for our method in particular. The quality of a POS tagger is usually given as the overall percentage of correctly assigned tags, 97.02% in our case. This relatively high accuracy does not reveal, however, the severity of tagging errors (cf. Oliva (2001)) and which kinds of tags are mistagged at which rate. We tested the difference between the two high-frequent[7] open POS tag classes ADJA (attributive adjective) and VVFIN (finite lexical verb) according to their tagging error rate, i.e. the number of tokens which should have been assigned a certain POS tag but were not. Attributive adjectives have an error rate of 2.40% and finite lexical verbs have an error rate of 9.95%. The test has thus shown that especially the verb tags, which are of major importance for our system, have an error rate much higher than the average error rate of 2.98% and still higher than a comparable high-frequent open class. Severe tagging errors are, thus, more frequent in the POS annotation.

ture sentences for an information retrieval system (cf. Braun (1999) and Neumann et al. (2000)).

[6] The training data consist of 490.000 tokens, whereof 315.000 tokens are newspaper texts, mainly from *die tageszeitung* (*taz*), and 150.000 tokens are novels. The corpus has been compiled for the project "Deutsches Referenzkorpus" (DEREKO).

[7] The tag ADJA accounts for 5.7% of all tokens and the tag VVFIN accounts for 4.2% of all tokens.

The main reason for this is that standard taggers like the ones used in our system only take into account the immediate context of a token when searching for the correct tag. This strategy is effective for tokens where the syntactic relation is reflected in distributional proximity. This is much more the case in English than it is in German. In German, the syntactic relation between two tokens is reflected by distributional proximity mainly in noun chunks (e.g. the relation between an attributive adjective and the head noun) and prepositional chunks. The German verb complex, however, is split into two parts in the affirmative clause as explained in section 1. In the example in Figure 1 there are fourteen tokens between the two parts of the verb complex of the main clause and even one intermitting subclause.

Sentences (1) and (2) below show an example of the problem of verbs which are ambiguous as regards their POS tags. The verb *zustimmen* is an infinitive verb in (1) and a finite verb in (2). As the examples show, there is no cue in a window of eight tokens as to which POS tag is the correct one. Thus, taggers taking into account a window of only two or three tokens cannot reliably annotate verbs in such a structure. In the following section, we show how this problem can be tackled.

(1) Gestern wollten weder die
Yesterday wanted neither the
Konservativen noch die Liberalen dem
conservatives nor the liberals the
Antrag zustimmen/VVINF.
motion to accept

'Yesterday, neither the conservatives nor the liberals wanted to accept the motion.'

(2) Kommentatoren erwarten, daß weder
Commentators expect that neither
die Konservativen noch die Liberalen
the conservatives nor the liberals
dem Antrag zustimmen/VVFIN.
the motion accept

'Commentators expect that neither the conservatives nor the liberals will accept the motion.'

4 Parsing and correcting tagging errors at the same time

The parser is constructed as a cascade of transducers which use hand-crafted finite state grammars to incrementally annotate linguistic structure beginning with topological fields and continuing with subclauses before chunks are annotated. The tag correction component is one level in this cascade. Within the task of annotating topological fields, the left part of the sentence bracket (i.e. CF and VCL) and the right part of the sentence bracket (i.e. VCR) are annotated first using POS tags and local context information. The parser then tries to match the respective parts of the sentence bracket into a parsable sequence. This would typically be a sequence like in Figure 2: An initial field (VF) followed by a left part of a sentence bracket containing a finite verb (VCLM**F**) followed by a middle field (MF) followed by a right part of a sentence bracket containing a non-finite verb (VCRV**I**). If the parser fails to assign a parsable structure, it makes use of the ranked POS tag assignments recorded in the linguistic markup. The parser considers the second-best tag and tries to match a parsable sequence again. Provided that the parser succeeds, the second-best tag is promoted to be the best tag and the whole sequence is annotated. As regards Figure 2 this would mean that if the verb *zustimmen* was wrongly assigned the POS tag VVFIN for finite verb and VVINF was also in the list of possible tags, then the parser would mark VVINF as the best tag and change the label of the sentence bracket into the respective one for finite verbs.

This strategy is similar to the one described in Hirakawa et al. (2000) in that it uses linguistic knowledge already encoded in the NLP system. Apart from the language-specific characteristics, the difference between our approach and Hirakawa et al. (2000) is that they use this strategy to acquire rules for POS tag correction, while our system directly integrates the POS correction component into the parsing process. Other approaches which combine rule-based and statistical components outside the parsing process are Tapanainen and Voutilainen (1994) (for English) and Hajič et al. (2001) (for Czech). In contrast to our approach, the latter approaches both apply the linguistic rules before doing stochastic disambiguation. They first either completely disambiguate the POS of a token or reduce its ambiguity for the following stochastic process. This strategy does

not allow, however, to share the same linguistic components between the rule-based tagger and the parser. Our approach rather leaves the POS tag to a certain extent ambiguous by assigning a ranked ambiguity class and allows the parser to decide which POS tag fits into the sequence.

Figure 3 gives another example with a subclause. In this subclause, there is again the verb *zustimmen*, which this time is a finite verb. If it is incorrectly tagged VVINF (for infinitive), then there is no parsing rule to match the sequence of a CF, an MF and a non-finite VCR because this sequence does not comply with any grammatical structure. The VVINF is thus changed to VVFIN and the structure can be annotated. There are also cases in which the left part of the sentence bracket is mistagged (Cf. Figure 1 in which *um*, *wie* and *der* are ambiguous). This is the case for 3,33% of all instances of the subordinator KOUS. The tagging error rate is again higher than the average error rate in our test. This is even the more important to point out if one considers that the high-frequent German subordinators *daß/dass*, *wenn* and *weil* are unambiguous so that the error rate for ambiguous subordinators is even higher. These errors are corrected along the same lines as shown for the verbs. If a right frame containing a finite verb is lacking a left frame, then the system checks whether there is a mistagged subordinator and then tries to match the structure.

5 Results

Using rules like those described in section 4 we considerably increased the tagging accuracy for verbs. The error rate for the tag VVFIN discussed in section 3 decreased from 9.95% to 8.07%. This means that the number of clauses which were unparsable due to a VVFIN tag not being recognized has been reduced by 19%. For all verb tags, the error rate has been reduced from 6.28% to 4.78%. There has been no increase in tagging errors for other tags, because the changes of tags remained within the group of verbs. The improvement of tagging accuracy for verbs was achieved by integrating just 18 rules into our system of topological field annotation because these rules could rely on the information already annotated by our system and the information encoded in the parser. Table 2 further shows that the component for the subordinators (KOUS and KOUI) decreased the tagging error rate from 3.33% to 1.53% for KOUS and from 7.23% to 5.31% for KOUI respectively. As there have been changes across tag classes in the subordinator component, we have successfully tried to keep the precision of this component high: 96.3% of all changes of this component are true positives.

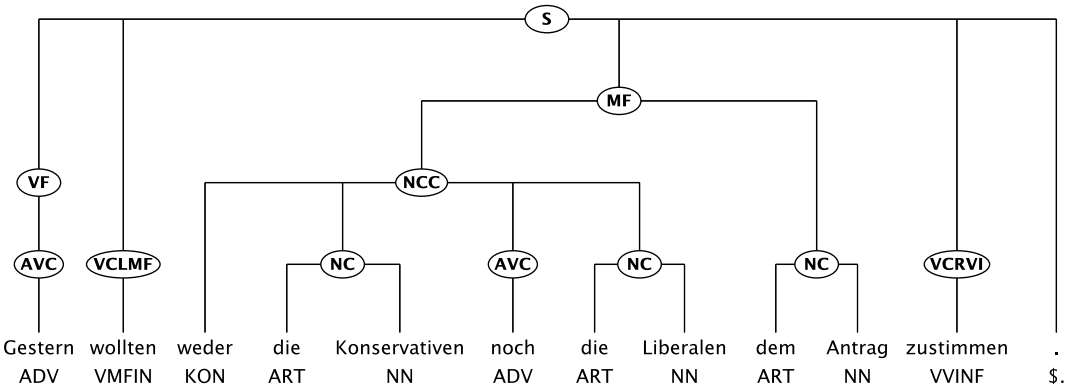
Table 2: Tagging errors for different tags

	uncorrected		corrected	
Tag	abs.	rate	abs.	rate
VVFIN	2097	9.95 %	1701	8.07 %
VAFIN	280	1.84 %	83	0.55 %
VMFIN	107	2.22 %	52	1.08 %
VVINF	605	7.55 %	469	5.85 %
VAINF	80	4.88 %	34	2.07 %
VMINF	55	17.19 %	26	8.12 %
all verbs	3914	6.28 %	2984	4.78 %
KOUS	216	3.33 %	99	1.53 %
KOUI	49	7.23 %	36	5.31 %
all sub.s	265	3.70 %	135	1.89 %
overall	14695	2.98 %	13640	2.77 %

6 Conclusion and Future Work

We have shown that it is both possible and reasonable to integrate different syntactic annotation methods into one system. The outcome of stochastic POS tagging can be considerably refined for POS tags determining the clause structure of complex sentences by taking advantage of a general-purpose topological field parser that relies on hand-crafted rules. Only a very small number of new rules have to be introduced to revise POS tags. We have also shown that by concentrating on certain classes of POS tags, we have considerably reduced the number of unparsable clauses.

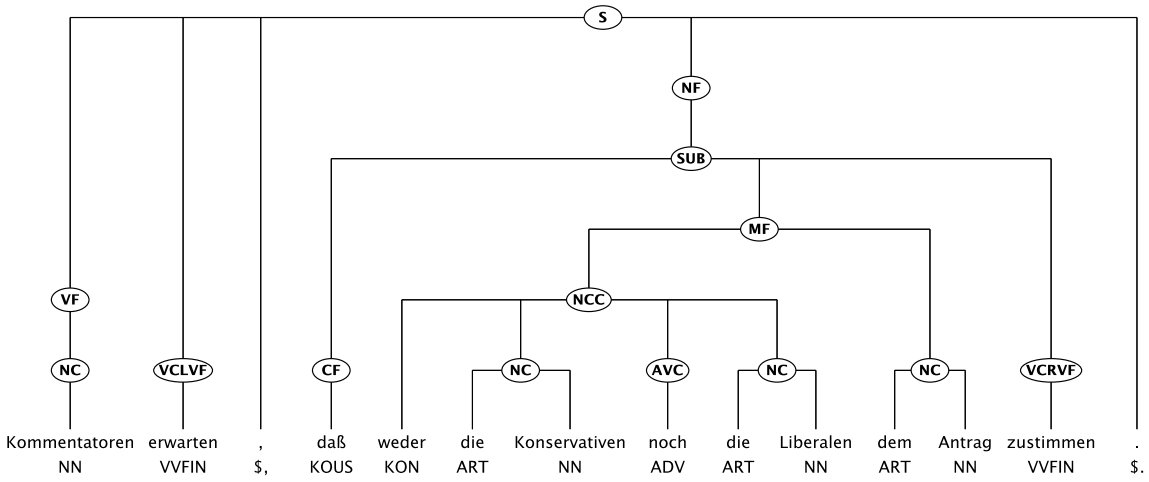
We are currently investigating the impact of tagging correction on parsing accuracy. While the parser described in the present paper seems to be competitive with machine learning methods in terms of topological field parsing (Veenstra et al., 2002), we do not know yet the influence of POS tag correction on overall parsing accuracy, which we plan to investigate in the near future.



yesterday wanted neither the conservatives nor the liberals the motion to accept

'Yesterday, neither conservatives nor the liberals wanted to accept the motion.'

Figure 2: Structurally ambiguous token *zustimmen* in affirmative sentence



commentators expect that neither the conservatives nor the liberals the motion accept

'Commentators expect that neither the conservatives nor the liberals will accept the motion.'

Figure 3: Structurally ambiguous token *zustimmen* in subclause

7 Acknowledgments

The research reported here was supported by the German Research Council (DFG) as part of the Sonderforschungsbereich 441 "Linguistische Datenstrukturen" (Linguistic Data Structures). The first author was also supported by a grant of the graduate school "Integriertes Linguistikstudium" funded by the DFG. We are grateful to the LTG Edinburgh, who made their TTT suite of tools (Grover et al., 1999) available to us. Additionally, we would like to thank Detmar Meurers for his supportive comments and the anonymous reviewers who provided us with helpful criticism.

References

Steven Abney. 1996. Partial Parsing via Finite-State Cascades. In *Proceedings of the ESSLLI-96 Workshop on "Robust Parsing"*, Prague.

Lars Borin. 2000. Something borrowed, something blue: Rule-based combination of POS taggers. In *Second International Conference on Language Resources and Evaluation (LREC 2000)*, pages 21–26, Athens.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000)*, Seattle, WA, April.

Christian Braun. 1999. Flaches und robustes Parsen deutscher Satzgefüge. Diplomarbeit, Universität des Saarlandes, Saarbrücken.

Peter Eisenberg. 1999. *Grundriß der deutschen Grammatik*, volume 2: Der Satz. Metzler, Stuttgart.

Claire Grover, Colin Matheson, and Andrei Mikheev, 1999. *TTT: Text Tokenisation Tool*. Language Technology Group, University of Edinburgh, Edinburgh.

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 of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, Toulouse, France.

Erhard W. Hinrichs, Sandra Kübler, Frank H. Müller, and Tylman Ule. 2002. A Hybrid Archictecture for Robust Parsing of German. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Gran Canaria.

Hideki Hirakawa, Kenji Ono, and Yumiko Yoshimura. 2000. Automatic Refinement of a POS Tagger Using a Reliable Parser and Plain Text Corpora. In *Proceedings of the 18th International Conference on Computational Linguistics (CoLing 2000)*, Saarbrücken.

Tilman Höhle. 1986. Der Begriff 'Mittelfeld', Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses*, pages 329–340, Göttingen.

Walt Detmar Meurers. 2002. On the use of electronic corpora for theoretical linguistics. case studies from the syntax of German. *Lingua*. Forthcoming. http://ling.osu.edu/~dm/papers/meurers-02.html.

Frank Henrik Müller. 2002. Shallow-Parsing Stylebook for German. Technical report, Universität Tübingen, Tübingen. http://www.sfs.nphil.uni-tuebingen.de/dereko/anno-doc.html.

Günter Neumann, Christian Braun, and Jakub Piskorski. 2000. A Divide-and-Conquer Strategy for Shallow Parsing of German Free Texts. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000)*, pages 239–246, Seattle, WA.

Karel Oliva. 2001. The Possibilities of Automatic Detection/Correction of Errors in Tagged Corpora: A Pilot Study on a German Corpus. In *Proceedings of the Conference 'Text, Speech and Dialogue 2001'*, Berlin.

Li-Shiuan Peh and Christopher H. Ting. 1996. A Divide-and-Conquer Strategy for Parsing. In *Proceedings of the ACL/SIGPARSE Fifth International Workshop on Parsing Technologies*, pages 57–66.

Anne Schiller, Simone Teufel, and Christine Thielen, 1995. *Guidelines für das Taggen deutscher Textcorpora mit STTS*. IMS Stuttgart und SfS Tübingen.

Pasi Tapanainen and Atro Voutilainen. 1994. Tagging accurately – Don't guess if you know. In *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP 1994)*, Stuttgart.

Hans van Halteren, Jakub Zavrel, and Walter Daelemans. 1998. Improving Data Driven Wordclass Tagging by System Combination. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 1998)*, pages 491–497.

Jorn Veenstra, Frank Henrik Müller, and Tylman Ule. 2002. Topological Fields Chunking for German. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL 2002) held in conjunction with the COLING 2002*, Taipei, Taiwan.