A Syntax-first Approach to High-quality Morphological Analysis and Lemma Disambiguation for the TüBa-D/Z Treebank

Yannick Versley, Kathrin Beck, Erhard Hinrichs, Heike Telljohann

Seminar für Sprachwissenschaft
University of Tübingen
E-mail: versley|kbeck|eh|telljohann@sfs.uni-tuebingen.de

Abstract

Morphological analyses and lemma information are an important auxiliary resource for any treebank, especially for morphologically rich languages since such information is a useful precondition for any task that needs to link surface forms to semantic interpretation (either through wordnets or distributional measures).

In contrast to common practice in parsing, the method used in the TüBa-D/Z treebank uses syntactic information for the morphological and lemma disambiguation. We argue that this approach has an advantage in the context of treebanking since many ambiguities in morphology and lemmas can be eliminated given the syntactic context.

1 Introduction

To use lexical resources, such as wordnets (e.g., Princeton WordNet, Miller and Fellbaum, 1991; GermaNet, Kunze and Lemnitzer, 2002) in conjunction with corpora, it is necessary to map surface word forms to lemmas (or dictionary forms). Princeton WordNet offers its own lemmatizer (formulated in a dozen rules and a list of exceptions – about 6 000 in total in Princeton WordNet 3.0). For languages with richer inflection, such as German, tools for morphological analysis are considerably more complex, yet the problem of linking surface forms to the entries in lexical resources remain.

Some researchers, such as Gurevych and Niederlich (2005) solve this problem by using stemming, but remark that, contrary to their expectations, stemming delivered no better results than no morphological processing at all.

One way to relieve this problem is to annotate corpora – in particular, when they already include a multitude of annotation levels – with gold-standard lemma information, which allows researchers to perform reproducible experiments linking corpora and lexical resources, without any concerns about lemmatization errors.

In the following sections, we will describe the existing annotation in the TüBa-D/Z (section 2), the system that we used to do high-quality pre-tagging of morphology and lemma information for the manual disambiguation (section 3), and provide some statistics on both the automatic and manual annotation (section 4).

2 The TüBa-D/Z

The TüBa-D/Z treebank of German[1] is a linguistically annotated corpus based on data from the German newspaper '*die tageszeitung*' (taz). The current Release 5 comprises approximately 45 000 sentences, with a new release of the treebank consisting of more than 55 000 sentences, including lemma information, to be be released before the end of 2010.

The annotation scheme of the TüBa-D/Z treebank comprises four levels of syntactic annotation: the lexical level, the phrasal level, the level of topological fields, and the clausal level. The primary ordering principle of a clause is the inventory of topological fields, which characterize the word order regularities among different clause types of German, and which are widely accepted among descriptive linguists of German (cf. Drach, 1937; Höhle, 1986). Below this level of annotation, i.e. strictly within the bounds of topological fields, a phrase level of predicate-argument structure is applied with its own descriptive inventory based on a minimal set of assumptions that has to be captured by any syntactic theory. A set of node labels describes the syntactic categories (including topological fields and coordinations). The context-free backbone of phrase structure (i.e. proper trees without crossing branches; Telljohann et al., 2004) is combined with edge labels specifying the grammatical functions of the phrases in question as well as long-distance relations. Phrase internal dependencies are captured by a hierarchical annotation of constituent structure with head/non-head distinctions. For more details on the annotation scheme see Telljohann et al. (2009).

Over the course of the last years, the syntactic annotation has been extended in various ways. Named entity information has been added. The basic Stuttgart-Tübingen tagset (STTS: Schiller et al., 1995) labels have been enriched by relevant features of inflectional morphology. A set of anaphoric and coreference relations referring to nominal and pronominal antecedents has been incorporated to link referentially dependent noun phrases (Hinrichs et al., 2004). Current work comprises both annotating new sentences as well as adding lemmas for each word form.

(1) *Wenn alles nach Plan läuft, werden sie die ersten*
If everything to plan goes, will they the first
Umzugsbotschafter sein.
dislocation=ambassadors be.
If everything goes according to plan, they will be the first 'dislocation ambassadors'.

[1] For more information, see `http://www.sfs.uni-tuebingen.de/tuebadz.shtml`

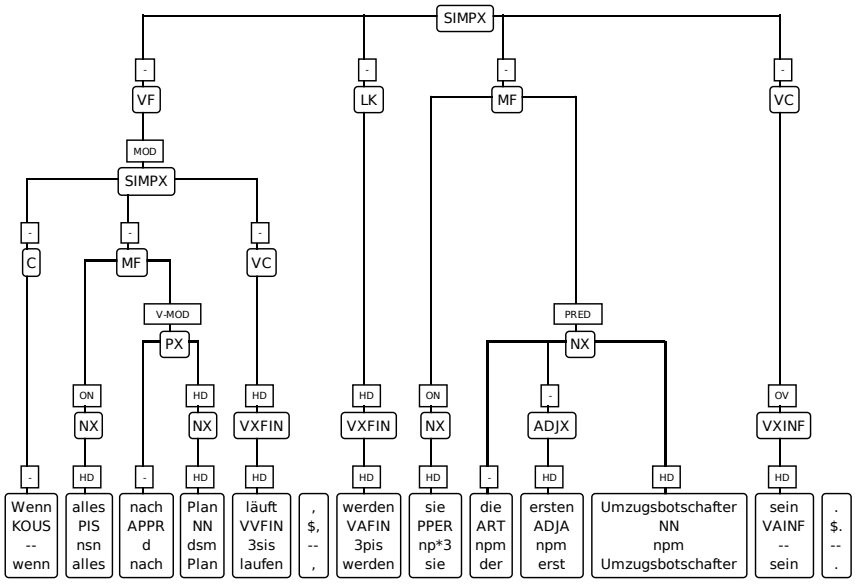

Figure 1: Example tree from the TüBa-D/Z

Figure 1 illustrates the linguistic annotation for the sentence in (1). The tree exemplifies the syntactic annotation scheme. The main clause (SIMPX) is divided into four topological fields: initial field (VF), left sentence bracket (LK), middle field (MF), and verb complex (VC). The finite verb in LK is the head (HD) of the sentence. The edge labels between the level of topological fields and the phrasal level constitute the grammatical function of the respective phrase: subject (ON), predicate (PRED), modifier (MOD), modifier of the verb (V-MOD). The modifying subordinate clause (SIMPX) in the initial field is again divided into the following fields: c-field (C), MF, and VC. The label V-MOD specifies the long-distance dependency of the prepositional phrase (PX) "*nach Plan*" on the main verb "*läuft*". Below the lexical level, the parts of speech, the morphological information, and the lemmata are annotated.

The presence of multiple layers of annotation has made it possible to use the TüBa-D/Z corpus in comparative evaluations for tasks including parsing and anaphora/coreference; the additional layers of annotation also make it possible to evaluate the impact of gold-standard versus automatic morphological information in these tasks.

3 Semi-automatic Morphological Tagging and Lemmatization

The annotation of the TüBa-D/Z corpus is carried out serially for different layers, beginning with the syntactic annotation (including part-of-speech, topological fields, and basic named entity annotation) and proceeding to morphological and lemma annotation as well as the anaphora/coreference annotation.

The annotation of syntactic structure prior to morphological annotation may appear unconventional since incremental annotation usually proceed from smaller units to larger units - thus annotation of part-of-speech and morphology typically precedes syntactic annotation. However, there are good reasons for adopting a syntax-first approach when it comes to the construction of a large treebank for a morphologically rich language.

A substantial part of the ambiguities that would occur in morphological tagging, especially with respect to case, are resolvable using syntax annotation; Furthermore, the integration of morphology only occurs after the assignment of syntactic structure, which means that we can profit from the information that has already been hand-corrected, feeding corrected syntactic information into the morphological disambiguation, and corrected morphological information into the disambiguation of lemmas.

3.1 Morphological Tagging

Morphological tags have been present in the TüBa-D/Z treebank already since the second release (Hinrichs et al., 2004); hence, the syntax-first version of the pre-tagging produces morphological tags according to the existing guidelines, but achieves greater accuracy thanks to the annotated syntactic information. Per-token analysis is based on SMOR (Schmid et al., 2004), a finite-state analyzer including derivational morphology, as well as additional heuristics that help in dealing with names and unknown words. The analyses assigned by SMOR and the heuristics are disambiguated both locally (within a noun phrase) and globally (using argument and modification information, enforcing consistent readings across coordination).

Since proper names contain morphological information (names of persons according to gender, the grammatical gender of locations is usually neutral, whereas the grammatical gender of organizations is best predicted by their organizational suffix – such as GmbH, AG, etc.), prediction of the morphology of named entities is done on the one hand by considering previous morphological tags assigned to this name string, and on the other hand by consulting gazetteer lists.

For certain classes, such as (invariant and regular) adjectives, simple suffix analysis is sufficient to predict the possible morphological tags. For nouns, which may be highly irregular, a maximally underspecified morphological tag is used so that the surrounding context may partially disambiguate them.

The first, local disambiguation step consists in disambiguating morphological tags within a single base NP: with very few exceptions, head and modifiers of a

noun phrase share the same morphological tag. Additional disambiguation is performed on strong and weak inflections of adjectives: the so-called *weak inflections* occur with definite and most indefinite articles, whereas *strong inflections* only occur when no article is present or the determiner is an ending-less form (indefinite nominative masculine/neutral determiners such as *ein*, *kein*, *sein* etc.)

The following steps in morphological disambiguation make use of the syntactic annotation that is already present in a more extensive fashion. For disambiguating case, which would be error-prone in a system based on sequence labeling, we have exact information from the grammatical functions at our disposition: Subject and copula predicates universally occur in nominative case, accusative objects in accusative case; similarly, pre- or postpositions govern the case of the noun phrase in the corresponding PP.[2] We make this information explicit by projecting down the case information across any adjunction or coordination nodes, so that the base NP with the head receives case annotation. Finally, we can also enforce number and person agreement between the inflected verb and the subject, as well as reflexives and the predicate complements of copula sentences.

Prepositions that allow both accusative and dative case (corresponding to a directional and a locational sense, similar to English *into* and *in*) are disambiguated by assuming that PPs attached to a noun are either locative (i.e., dative case) or an argument to that noun, a case for which the combination of governing noun and preposition is checked against a list acquired from a large unannotated corpus (where unambiguous combinations of a noun, a preposition, and an accusative NP are seen as indicative that the noun-preposition combination plausibly occurs with an accusative PP).

3.2 Lemmatization: open-class words

In the case of content words, the purpose of lemmatization is to map differently inflected forms of the same stem into a single form which helps to find the corresponding entries in (paper) dictionaries, wordnets, or other electronic resources as well as obtaining accurate counts to compare the relative frequency of nouns irrespective of their inflection.

There are several choices to be made with respect to lemmatization of German. For open-class words, we aimed for maximal consistency with the lemmatization in GermaNet; in particular, deadjectival nouns (such as '*Arbeitsloser*' [jobless person]: consider strong '*ein Arbeitsloser*' [*a* jobless person] versus weak '*der Arbeitslose*' [*the* jobless person]), and nouns with a corresponding inflection (such as '*Beamter*' [civil servant], which follows the weak/strong distinction normally found in adjectives and deadjectival nouns), are lemmatized to the *strong* forms.

The syntax-first strategy also provides valuable linguistic information in cases where lemmatization would be ambiguous if the lexical token is considered in isolation and not in its syntactic context: In the example below, *Summen* is ambiguous

[2] Many prepositions allow both accusative and dative case, in which case further disambiguation is necessary whenever the NP chunk is case-ambiguous.

between a singular analysis as *Summen* (humming) and a plural analysis of *Summe* (sum). Subject-verb agreement constraints in morphological disambiguation yield the necessary information to remove this ambiguity, which would still be present if only local disambiguation had been applied.

(2) *"Da hätten Summen von 165.000 Mark schon auffallen müssen".*
"There have.IRR sums of 165 000 Mark already be_conspicuous must".
"In such a situation, sums of 165 000 Mark should have been conspicious".

Verb particles are attached to the verb to which they belong syntactically. Furthermore, verbs such as *haben* (to have), *sein* (to be) and *werden* (to become) have uses as a main verb (as a verb of possession, or as copula verbs, respectively) and as an auxiliary, which are not distinguished in the part-of-speech tags according to the STTS guidelines. To help in the identification of main-verb uses of these verbs, the lemmas of word tokens used as auxiliary are suffixed with an additional tag (%passiv for passive constructions and %aux for other non-main-verb uses of auxiliaries and modals).

Again, the syntax-first strategy makes it possible to provide such a fine-grained lemma analysis of these items. This in turn allows users to perform searches for full verb uses of auxiliaries or constructions such as the passive.

The lemmatization also distinguishes between separable and inseparable verb prefixes (which can be helpful in cases where both separable and inseparable versions are possible, since the meanings of these versions are generally distinct from each other) by putting a # marker between a separable verb prefix (reattached or not) and the verb. To make this distinction in cases where SMOR returns ambiguous analyses (for example, *unter-* can be used both as a separable and as an inseparable verb prefix, as in *unter#buttern* – to ride roughshod over someone, and *untermauern* – to underpin). However, most verbs only allow, or have a strong preference for, only one of these possibilities. As a result, disambiguation is possible in most cases using frequency data[3] for unambiguous forms (in this case, the *zu*-infinitive form, which would be *unterzubuttern* in the separable case and *zu untermauern* in the inseparable case).

Reconstruction of verb and adjective lemmas from SMOR's analysis is normally possible by transforming the FST analysis. For nouns, in contrast, this is not generally possible, since the SMOR analysis is less informative than the original string and omits information about linking elements ('*Fugenelement*') in compounds, which may be semantically significant (for example, consider *Kindsmutter* – a child's mother, to *Kindermutter* – a nanny, which both get the same analysis consisting of their two stems *Kind+Mutter*).

To get around this weakness regarding linking elements, we adopt a regeneration approach similar to the one used by Versley (2007) for generating plural forms:

[3]The frequency data is extracted from the *Web 1T 5-gram, 10 European Languages Version 1* dataset produced by Google that is distributed by LDC as catalog no. LDC2009T25.

we construct an analysis string that corresponds to the desired result (i.e., nominative, singular, with weak inflection for deadjectival nouns), use SMOR in generation mode to get all possible strings (including those that SMOR overgenerates), and use a set of heuristics to predict the correct lemma out of the overgenerated set of strings. Besides similarity to the original word form (in terms of edit distance), we select for lemma candidates whose word forms are similar in frequency to the original word form, while preferring those with higher frequencies. While this approach is somewhat complicated by features of SMOR (underspecification of case for some, but not all analyses, inclusion of markers for old-standard and new-standard orthography into the analysis which need to be removed or added), we find that this approach yields high-quality lemmas for all analyzed word forms.

Finally, the lemmas of truncated items should include the understood completion. For example in the following example 3, the token *Bau-/TRUNC* should receive the lemma *Bauplanung%N* (construction planning) so that the inferred lemma and its part-of-speech are made explicit.

(3) *"Bei bedeutenden Bau- und Verkehrsplanungen müssen unsere*
"with important construction and traffic=plannings must our
Einflußmöglichkeiten gestärkt werden", fordern die
influence=possibilities strengthened become", demand the
Behindertenverbände.
disabled=associations.
"In the domain of important construction and traffic plans, our influence must become stronger", demand the associations for the disabled.

The automatic completion of truncated items comprises two parts: on the one hand, finding a corresponding item that represents the context in which the lemma is interreted; on the other hand, determining the most likely completion of the truncated item given the context item (consisting of the truncated part plus a suffix of the context item).

For the first part, we simply consider the first content word following the separating comma or conjunction token as the completing context item. The second part, determining likely completions, is done by checking concatenations of the truncated item and suffixes of the potential context item for plausibility using frequency data (from the Google n-gram dataset). Among the possible completions constructed in this way, the most frequent one is considered most likely to be correct. While this frequency-based approach works very well in most cases, there are cases which result in incorrect solutions that can only be recognized considering both coordinated parts (i.e., the proposed completion and the context item).

3.3 Lemmatization: closed-class words

While there is considerable consensus about the lemmatization of open word classes, there is substantial variation in the lemmatization guidelines for closed-class words. This is largely due to the fact that it is not always clear what the division of labour

should be between morphological tags and lemmatization. The lemmatization of definite article tokens is an example for the case in point: The TIGER treebank, for instance, uses only one lemma for each of definite and indefinite articles (mapping articles to either "*der*" for definite articles or "*ein*" for indefinite articles, corresponding to the male nominative singular form), but keeps the unmodified surface form in the case of personal pronouns.

For the TüBa-D/Z, the lemmatization guidelines prescribe that articles, possessives, and definite and indefinite pronouns are normalized to nominative singular, but keep gender and root. In cases of plurals that are unmarked for gender (e.g., *die Studierenden* the students/lit. the studying, which has only one form for both masculine and feminine), the possible strings for the determiner are all listed, separated by the diacritic 'l'. This makes it possible to find these ambiguous items when searching for either the masculine or feminine article.

4 Empirical Results

Thus far, we have focused on the empirical issues to be solved, but have not discussed the division of labour between automatic pre-tagging – both morphology and lemmas are proposed by an automatic system, either as a set of several tags to choose from in the case of morphology, or in the form of a proposed lemma – and the subsequent manual annotation. As the amount of work needed for manual correction also depends on the error rate of the automatic component, it is useful to assess the quality of our lemmatizer. To do this, we compare the hand-corrected gold standard of the upcoming Release 6 of the treebank against the lemmas proposed by the semi-automatic system on one hand, and against lemmas proposed by TreeTagger (Schmid, 1995) based on the model that comes with it.

As our lemmatization guidelines include elements that go beyond morphological analysis by itself – consider the attachment of separable verb prefixes, and the completion of truncated items – we provide evaluation figures in two different settings:

- In the *strict* setting, a lemmatizer is required to provide the exact string that is to become part of the treebank annotation.
- In the *lenient* setting, a lemmatizer is not required to mark the difference between separable and nonseparable verb prefixes; separable verb prefixes that occur as separate tokens are not required to be attached; and whenever the guidelines require a split analysis because of morphological underspecification, a result that provides only one analysis (or a subset of the analyses that make up the correct tag) is counted as correct.

For our *TreeTagger* baseline, we used the output of TreeTagger with two modifications that are common practice for lemmatization: we replaced any unknown number (`@card@`) by its surface form, and any unknown other word (TreeTagger lemma `<unknown>`) was replaced by the corresponding surface word form.

Category	TüBa-D/Z lemmatizer	TreeTagger
overall[1]	99.4	77.7
full verbs (VV...)	99.1	74.8
NN	98.3	92.5
NE	99.4	96.5
TRUNC	63.6	–
VVFIN	99.4	77.4
VVPP	98.1	69.7
VVIZU	99.6	–
VVIMP	99.0	62.1

[1]: All STTS part-of-speech categories are included in this evaluation.

Table 1: Strict evaluation (≙ necessary edits)

Category	TüBa-D/Z lemmatizer	TreeTagger
overall[2]	99.4	94.2
full verbs (VV...)	99.1	91.4
NN	98.3	92.5
NE	99.4	96.5
VVFIN	99.4	86.0
VVPP	98.2	96.2
VVIZU	99.6	96.7
VVIMP	100.0	64.1

[2]: TRUNC, pronouns, and determiners are omitted in this evaluation.

Table 2: Lenient evaluation (≙ correctness of coarse-grained information)

The *lenient* setting (cf. table 2) is better suited as a comparison to other work on German lemmatization: In the *strict* setting (table 1), TreeTagger takes an accuracy penalty for not providing the additional information required by the treebank (reattaching separable verb prefixes and/or marking them, marking auxiliary versus full verb uses, or completing truncated items) and not producing pronoun and determiners according to the guidelines of the treebank. In the *lenient* setting, pronouns, determiners and truncated words are completely left out of the evaluation, as are separated verb prefixes; and the marking of separable verb prefixes is ignored. As can be seen in table 2 we still see an error reduction of 80%, which is mostly due to nouns, with a more modest error reduction of about 44% for verbs.

Among work that uses lemmatization in treebanking for German, no accuracy figures can be found in the published literature: the mechanisms used for the TIGER treebank (Brants et al., 2002) involve interactive selection by the user (either of complete LFG parses in the LFG-to-Tiger toolset, or of morphological entries using a program named TigerMorph on which no further details are provided), whereas the Smultron parallel treebank (Volk et al., 2009) uses the GerTwoL sys-

tem (Haapalainen and Majorin, 1995) and post-corrects the predictions according to a set of guidelines that stay close to the output of the system.

In the more recent literature, Chrupala (2008) claims a lemmatization accuracy of 95% by cross-validation on the TIGER treebank using a memory-based learning approach to predict editing operations.

5 Generalizing to Unannotated Data

One important use case for treebanks (or, in general, corpora with rich linguistic annotation) is the investigation of complex phenomena that benefit from the additional annotation levels; however, in many cases the size limitations of a manually annotated treebank limit the potential usefulness. This is true for phenomena which are very rare in themselves, but also for the kind of linguistic phenomenon where multiple confounding factors make quantitative analysis a more challenging enterprise. As an example, primarily temporal discourse connectives such as *nachdem* (after/since) or *während* (while) occur relatively often in the TüBa-D/Z treebank, with more than 300 occurrences each, but a quantitative analysis that takes into account lexical and aspectual information can benefit immensely from the additional examples that would be found in larger unannotated corpora.

Leaving behind the realm of the carefully curated treebank would normally also entail rewriting most or all of the feature extraction, since neither the finer-grained lemmas nor the syntactic structure would be reproduced by a pipeline built from off-the-shelf components.

Using a parser that integrates SMOR for lexical prediction and yields the grammatical function labels necessary for the case prediction in morphology (Versley and Rehbein, 2009), however, allows us to use a syntax-first approach even for completely automatic annotation, as we have all the information that is needed for the morphological disambiguation. Remaining ambiguities (which would be left open for annotators to choose from in the case of treebank annotation) can be resolved using a simple CRF sequence tagger, as the most important global ambiguities are resolved using syntax information. Figure 2 shows an example where syntax information (phrase structure and edge labels) from the parser was automatically enriched with morphological tags and lemmas.

6 Conclusions

In this paper, we presented a lemmatization procedure devised for the TüBa-D/Z treebank, as well as a tagger that performs partial morphological disambiguation and lemmatization steps automatically, taking advantage of the existing syntactic annotation. The scope of the lemmatization guidelines incorporates some features that go beyond pure morphological analysis (such as reattaching separable verb prefixes, marking auxiliary use of verbs, and completion of truncated items), but which are squarely within the intended purpose of lemmatization as recovering the

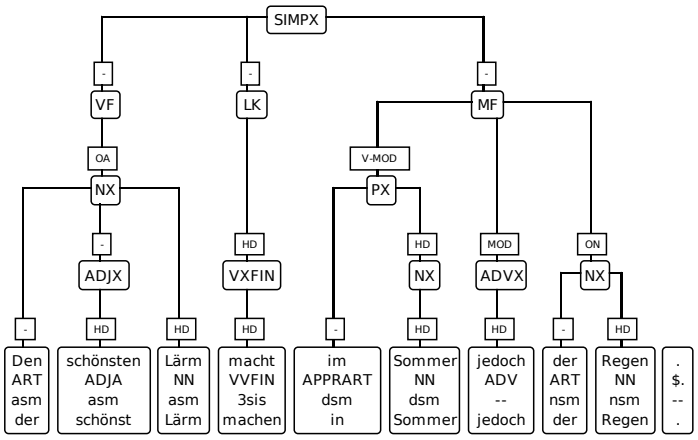

Figure 2: Example tree from the parser

terms which are mentioned in a text. The evaluation in section 4 shows that this more ambitious task definition is well within the reach of automatic tools.

Despite the high quality of fully automatic lemmatization, it is very useful to have gold-standard lemmas as part of the treebank, as this allows, in the context of more complex tasks such as coreference resolution, discourse tagging, or word sense disambiguation, to simplify the overall complexity of the task by removing the class of errors due to incorrect lemmatization. Moreover, the presence of both system-provided and gold-standard lemmatization allows to quantify the difference in the performance of these more complex applications that can be ascribed to the unsolved part of the lemmatization problem. The high quality of the automatic annotation as well as the multiple sessions of manual correction of all layers account for the gold standard quality of the TüBa-D/Z.

References

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proc. TLT 2002*.

Chrupala, G. (2008). *Towards a Machine-Learning Architecture for Lexical Functional Grammar Parsing*. PhD thesis, Dublin City University.

Drach, E. (1937). *Grundgedanken der Deutschen Satzlehre*. Diesterweg, Frankfurt/M.

Gurevych, I. and Niederlich, H. (2005). Accessing GermaNet data and computing semantic relatedness. In *ACL 2005 short papers*.

Haapalainen, M. and Majorin, A. (1995). GERTWOL und Morphologische Disambiguierung fürs Deutsche. In *NODALIDA 1995*.

Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H., and Trushkina, J. (2004). Recent developments of Linguistic Annotations of the TüBa-D/Z Treebank. In *Proceedings of TLT 2004*.

Höhle, T. N. (1986). Der Begriff 'Mittelfeld'. Anmerkungen über die Theorie der topologischen Felder. In Schöne, A., editor, *Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen*, pages 329–340.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In *Proceedings of LREC 2002*.

Miller, G. A. and Fellbaum, C. (1991). Semantic networks in English. *Cognition*, 41:197–229.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Texte mit STTS. Technical report, Univ. Stuttgart / Univ. Tübingen.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proc. ACL-SIGDAT Workshop*.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *LREC 2004*.

Telljohann, H., Hinrichs, E. W., and Kübler, S. (2004). The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proc. LREC 2004*, pages 2229–2232, Lisbon, Portugal.

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H., and Beck, K. (2009). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. University of Tübingen.

Versley, Y. (2007). Using the Web to resolve coreferent bridging in German newspaper text. In *Proceedings of GLDV-Frühjahrstagung 2007*, Tübingen. Narr.

Versley, Y. and Rehbein, I. (2009). Scalable discriminative parsing for German. In *Proc. IWPT 2009*.

Volk, M., Marek, T., and Samuelsson, Y. (2009). Smultron (version 2.0) - the Stockholm MULtilingual parallel TReebank. `http://www.cl.uzh.ch/research/paralleltreebanks_en.html`. An English-German-Swedish parallel Treebank with sub-sentential alignments.