Latent Semantic Clustering of German Verbs with Treebank Data

Holger Wunsch and Erhard W. Hinrichs
SfS-CL, University of Tübingen
Wilhelmstr. 19
72074 Tübingen, Germany
{wunsch,eh}@sfs.uni-tuebingen.de

1 Introduction

Treebank data have been utilized as data sources for a wide range of tasks in computational linguistics, including statistical parsing, anaphora resolution, induction of valence lexica, etc. More recently, researchers have experimented with extracting semantic information from syntactically annotated data. Here, treebank data have been used for the purposes of identifying selectional preferences of verbs and for the purposes of clustering verb classes (most notably using *latent semantic clustering*, or LSC for short).

The present paper follows this recent tradition of extracting semantic information from syntactically annotated data. The goal of this work is to determine verb classes for German verbs by means of latent semantic clustering. The ultimate goal of this research is task-oriented. We would like to investigate whether verb clusters obtained by the LSC method can be used as semantic knowledge for the purposes of anaphora resolution. In this sense, the current paper is a preparatory study and awaits a task-oriented evaluation in future work.

We will present experiments with two treebanks, TüBa-D/Z (Telljohann et al., 2003) and TüPP-D/Z (Müller, 2004b) that are both based on German newspaper text from the daily newspaper *die tageszeitung* (taz). The two resources differ significantly along the following dimensions:

1. **method of annotation:** The TüBa-D/Z treebank was manually annotated with the help of the tool *annotate* (Brants and Plaehn, 2000) and checked for consistency of annotation in a post-editing phase. The TüPP-D/Z was automatically annotated with the help of the KaRoPars parser described in Müller and Ule (2002) and not checked for errors of annotation in any way.

Hajič, J. and Nivre, J. (eds.): Proceedings of the TLT 2006, pp. 151–162.

However, as Müller (2004a) has shown, the quality of annotation produced by KaRoPars is quite competitive with the best results of other parsers of German for the categories that are annotated in TüPP-D/Z. The TüPP-D/Z experiments described in this paper corroborate this finding.

2. **granularity of annotation:** Both treebanks contain annotations about clause structure, topological fields, and grammatical functions of major constituents. However, at the clausal level, the depth of annotation differs considerably. In TüPP-D/Z only chunks in the sense of Abney (1991) are annotated below the clause level, and attachments of chunks to other chunks is not provided. The TüBa-D/Z annotation, on the other hand, contains ordinary phrases (as opposed to chunks), and attachment among phrases is fully specified.

3. **size:** The version of the TüBa-D/Z treebank that was used in the experiments contains 27,125 sentences and 473,747 lexical tokens, while the TüPP-D/Z corpus is much larger in size: appr. 11.5 million sentences and 204,661,513 lexical tokens.

It turns out that the TüBa-D/Z data source is not sufficient in size for inducing good-quality clusters by the LSC method. Rather, the LSC experiments show that much larger resources such as TüPP-D/Z are needed to overcome the data sparseness issues that arise with smaller resources such as TüBa-D/Z. At the same time, automatic annotation of partial syntactic structure in combination with annotation of grammatical functions as in TüPP-D/Z suffices for LSC methods, as long as the annotation is sufficiently accurate and contains relevant information about clause structure.

2 The TüBa-D/Z treebank of German

Due to their fine grained syntactic annotation, the TüBa-D/Z treebank data are ideally suited as a basis for extracting the type of information relevant for LSC experiments, i.e. syntactic and semantic properties of verbs and their complements.

The TüBa-D/Z annotation scheme distinguishes four levels of syntactic constituency: 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. e.g. Höhle (1986)). The TüBa-D/Z annotation relies on a context-free backbone (i.e. proper trees without crossing branches) of phrase structure combined with edge labels that specify the grammatical function of the phrase in question.

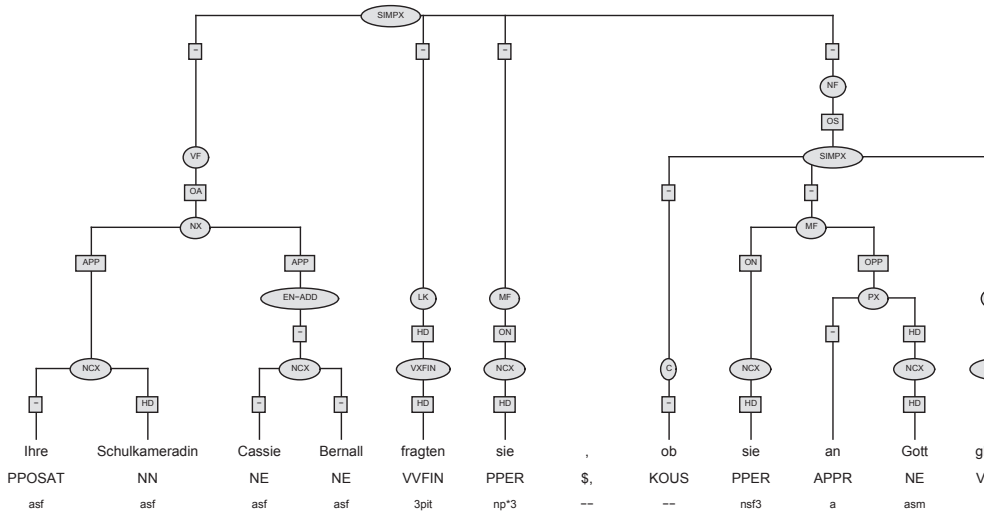

Figure 1: A sample tree from the TüBa/D-Z treebank.

Figure 1 shows an example tree from the TüBa-D/Z treebank for sentence (1). The sentence is divided into two clauses (SIMPX), and each clause is subdivided into topological fields. The main clause is made up of the following fields: VF (mnemonic for: *Vorfeld* – 'initial field') contains the sentence-initial, topicalized constituent. LK (for: *linke Satzklammer* – 'left sentence bracket') is occupied by the finite verb. MF (for: *Mittelfeld* – 'middle field') contains adjuncts and complements of the main verb. NF (for: *Nachfeld* – 'final field') contains extraposed material – in this case an indirect yes/no question. The subordinate clause is again divided into three topological fields: C (for: *Komplementierer* – 'complementizer'), MF, and VC (for: *Verbalkomplex* – verbal complex). Edge labels are rendered in boxes and indicate grammatical functions. The sentence-initial NX (for: *noun phrase*) is marked as OA (for: *accusative complement*), the pronouns *sie* in the main and subordinate clause as ON (for: *nominative complement*).

(1) Ihre Schulkameradin Cassie Bernall fragten sie , ob sie
Their fellow student Cassie Bernall asked they[subj] , whether she[subj]
an Gott glaube.
in God believes.
'They asked their fellow student Cassie Bernall whether she believed in God.'

Topological field information and grammatical function information are crucial for the extraction of verbs and their complements. Topological fields provide the

regions for grouping the right complements with the right verbs, and grammatical function labelling provides the necessary information for identifying the role of each complement.

3 The TüPP-D/Z treebank of German

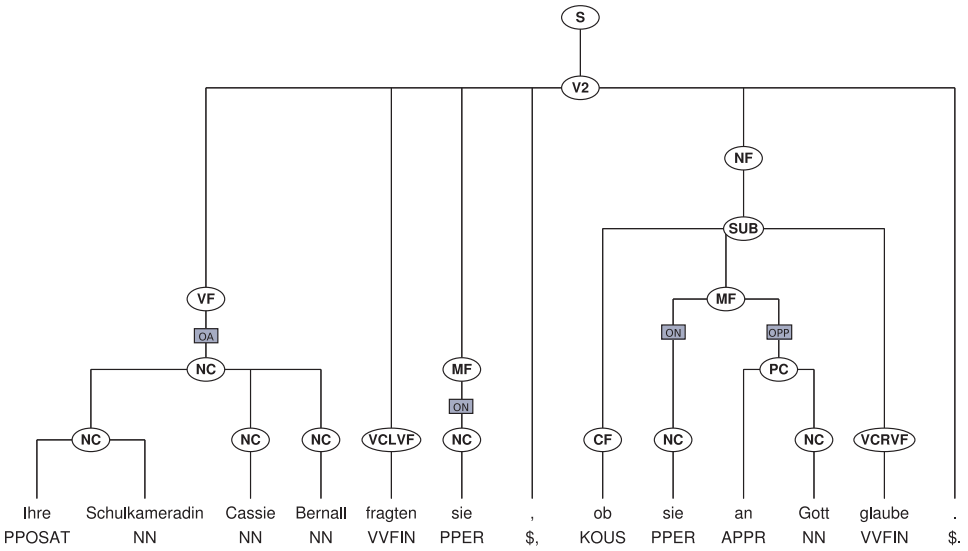

Figure 2: A sample from the automatically annotated TüPP-D/Z treebank.

TüPP-D/Z (Müller, 2004b) has been automatically annotated using the cascaded finite state parser KaRoPars. Four levels of syntactic constituency are annotated: the lexical level, the chunk level (in this respect, TüPP-D/Z differs from TüBa-D/Z), the level of topological fields, and the clausal level. Unlike TüBa-D/Z, which assumes a relatively deep syntactic structure, trees are quite flat in TüPP-D/Z. Due to limitations of the finite state parsing model, the attachment of chunks remains underspecified. Major constituents are annotated with grammatical functions. Figure 2 shows the example sentence (1) from section 2 in TüPP-D/Z annotation style. The automatic variant is fairly close to the manual annotation. There are differences in the annotation of the complex noun phrase *"Ihre Schulkameradin Cassie Bernall"*, where the additional grouping of the proper name *Cassie Bernall* is missing from TüPP-D/Z. The categories indicating left and right sentence brackets are merged with the categories of verb chunks.

Although the annotation of TüPP-D/Z provides less syntactic structure, the rel-

evant information for the extraction of verb-object pairs, most importantly the annotation of topological fields and of noun chunks with grammatical functions, is present with sufficient accuracy.

4 Latent Semantic Clustering

The kinds of entities that can occur as complements (i.e. subjects and objects) of a verb are strongly determined by the verb's meaning. For the same reason, nouns preferably co-occur with certain classes of verbs. For example, nouns denoting types of *food* typically occur as objects of verbs like *cook* and *eat*, while verbs in the semantic field of *hear* may select objects like *music*, *opinion*, or *word*. It is extremely unlikely that *words* are *cooked*, and *cucumbers* are *heard*.

Although a verb's selectional preferences are immediately clear to a speaker or hearer in most cases, it is more difficult to find out about the semantic properties of a verb for the purpose of automatic processing. Given sufficient amounts of corpus data however, it is possible to conclude a verb's selectional preferences by considering pairs or n-tuples of co-occurring verbs and objects. Nouns that do belong to the preferred semantic field of a verb will occur with significantly higher frequency in a corpus together with that verb than nouns that do not. Combining verbs and nouns that co-occur with high frequency will result in groups that reflect classes of verbs with similar selectional preferences and the entities they prefer in their argument slots.

Latent Semantic Clustering (LSC) (Rooth, 1998) is a method for the automatic extraction of selectional preferences from large corpora. Given samples of tuples of a verb and its objects, the algorithm arranges verbs and nouns in clusters. No human intervention is required, so LSC is an unsupervised approach. Unlike other clustering methods that allow a tuple to become element of only one cluster (so-called *hard-clustering methods*), LSC puts the tuple in all clusters. Tuples are assigned probabilities in a cluster, where most tuples will receive very low probabilities. This is called *soft clustering*. Soft-clustering methods are especially well suited to capture semantic properties: The meaning of a word is never clear-cut, but rather a blend of multiple semantic fields, some more typical than others. Hard clustering methods that allow a word to occur only in one cluster put unnatural restrictions on the distribution of selectional preferences, while soft clusters provide much more fine-grained representations.

LSC employs three structures: Sets of verbs, sets of nouns, and sets of selectional types (i.e. clusters) (Rooth, 1998). It assumes probability distributions for all structures: The probability p^τ of a selectional type τ with respect to all other selectional types, the probability of any verb to be member of a selectional type

Cluster 31 (0.0401107)			
Feature 0		Feature 1	
sein 'be'	0.500682	entscheidung 'decision'	0.0217412
lassen 'let'	0.197448	schröder	0.0181574
fallen 'fall'	0.132553	krieg 'war'	0.0104872
feststellen 'determine'	0.0179019	bombe 'bomb'	0.00966531
antworten 'answer'	0.0165691	ergebnis 'result'	0.00675502
beenden 'finish'	0.0155959	polizei 'police'	0.00639594
formulieren 'formulate'	0.0128891	mann 'man'	0.00606199
festhalten 'hold onto'	0.0107859	zeit 'time'	0.00552692
erfassen 'capture'	0.00741737	demonstrant 'demonstrator'	0.00552304
lenken 'steer'	0.00549608	rede 'speech'	0.00552304

Figure 3: Top-ranked subject-verb clusters extracted from TüBa-D/Z.

(p_v^τ), and the probability of any noun to be member of a selectional type (p_n^τ). For any type, LSC constructs a probability distribution which gives the probability of a pair[1] of a verb and a noun being member of a selectional type:

$$p_{\tau,v,n} = p^\tau p_v^\tau p_n^\tau$$

LSC iteratively estimates these probabilities by employing an expectation-maximization (EM) strategy.

5 Latent Semantic Clustering on TüBa-D/Z

The first set of experiments uses TüBa-D/Z as its data source. From the treebank, two sets of pairs were extracted. The first set of pairs comprises the lemmatized main verb and the lemmatized head of the subject noun phrase (grammatical function ON). The second set of pairs again consists of the main verb but this time the head of the accusative object (grammatical function OA) as the second element. For both sets, pair frequencies were calculated. The set of verbs and subjects contains 16,846 pairs, where the most frequent pair occurs 11 times (*sterben – Mensch* / *die – human being*). The set of verbs and accusative object contains 8,160 pairs. There, the most frequent pair occurs 35 times (*spielen – Rolle* / *play – role*).

The results were used as the input to the *lsc* program (Schmid, 2006) which performed the actual soft clustering. *lsc* requires both the number of clusters and

[1] We assume pairs of verbs and nouns. For n-tuples, this generalizes in the obvious ways.

Cluster 30 (0.0737073)			
Feature 0		Feature 1	
geben 'give'	0.909575	alternative 'alternative'	0.0191835
starten 'start'	0.0236105	antwort 'answer'	0.014756
ankündigen 'announce'	0.0162279	mühe 'effort'	0.0103296
unterrichten 'teach'	0.00737828	auskunft 'information'	0.0103296
plazieren 'place'	0.00464185	meinung 'opinion'	0.00885393
überreichen 'hand over'	0.00442697	position 'position'	0.0075398
aktivieren 'activate'	0.00442697	absprache 'agreement'	0.00737828
durchspielen 'run through'	0.00317725	möglichkeit 'possibility'	0.00737828
vernachlässigen 'neglect'	0.00295131	licht 'light'	0.00737827
leihen 'lend/borrow'	0.00212336	krieg 'war'	0.00600035

Figure 4: Top-ranked verb-object clusters extracted from TüBa-D/Z.

the number of iterations for the model estimation to be specified. A value of 40 was chosen for the number of clusters, a number which turned out to be optimal in previous work (Wagner, 2005; Schulte im Walde, 2003). 30 was chosen for the number of iterations. Altering this number does not noticeably change the clustering results.

Figures 3 and 4 show the top-ranked verb-subject and verb-object clusters as calculated by *lsc*. Each cluster consists of two features. Feature 0 contains the verbs and the corresponding probabilities that a verb has a selectional preference that is represented by the cluster. This probability corresponds to p_v^τ described above. Feature 1 contains the nouns and corresponding values for p_n^τ. Note that for each feature, only the ten most probable words are shown. Due to the nature of the soft-clustering algorithm, all verbs and all nouns are in fact members of each cluster, but the ones not shown received very low probabilities.

It is obvious from manual inspection of the clusters that the LSC algorithm is not able to produce semantically coherent clusters with this input data. Consider the top-ranked verb-subject cluster in figure 3. The figure shows the ten most prototypical (measured in terms of relative frequency) verbs (under feature 0) and nouns (under feature 1) for this cluster. Neither the verbs nor the nouns exhibit natural lexical fields. In particular the nouns are scattered among different ontological categories such as abstract entities (e.g. *decision* 'Entscheidung' and *Krieg* 'war'), humans (e.g. *Polizei* 'police' and *Mann* 'man') as well as inanimate objects (e.g. *Bombe* 'bomb'). Likewise, in figure 4 the verbs are almost equally divided between two disparate lexical fields: change of possession verbs (e.g. *geben* 'give'

Cluster 19 (0.0456084)			
Feature 0		Feature 1	
wollen 'want'	0.199787	Regierung 'government'	0.0258177
beschließen 'decide'	0.0383597	Senat 'senate'	0.0207326
ablehnen 'reject'	0.032485	SPD	0.0184641
aussprechen 'articulate'	0.0215241	CDU	0.010199
ankündigen 'announce'	0.0204867	Bundesregierung 'federal government'	0.00844105
zustimmen 'agree'	0.018803	USA	0.00816486
einigen 'agree on'	0.0180794	Parlament 'parliament'	0.00801274
fordern 'demand'	0.0169813	Präsident 'president'	0.00780153
aufrufen 'call on'	0.0164636	Grünen 'green party'	0.00759958
verabschieden 'pass (law)'	0.0158671	Prozent 'percent'	0.00577177

Figure 5: Top-ranked subject-verb clusters extracted from TüPP-D/Z.

and *leihen* 'lend') and verbs of mental action (e.g. *vernachlässigen* 'neglect' and *durchspielen* 'run through'). Moreover, the nominal objects that have been clustered for these verbs are only appropriate for the change of possession verbs, but do not represent realistic candidates for the verbs of mental action included in the verb cluster.

The two clusters are but two examples of the general picture that emerges from the LSC clusters obtained for the TüBa-D/Z data. Their lack of cohesion must be attributed to the relatively small size of the input data presented to the clusterer. With most of the pairs occurring only once, and the highest number of occurrences being below 40, the samples are nearly uniformly distributed, which means that the clustering algorithm cannot rely on much more information than random choice.

6 Latent Semantic Clustering on TüPP-D/Z

The second set of experiments uses TüPP-D/Z as its data source. Sets of lemmatized verbs and subjects or accusative objects are extracted from the automatically parsed corpus and presented to the *lsc* clusterer in the same fashion as for the TüBa-D/Z experiments described in section 5. The size of the data sets extracted from TüPP-D/Z however exceeds the TüBa-D/Z data by several orders of magnitude. The set of verbs and subjects contains 4,309,330 pairs. The most frequent pair occurs 7,240 times (*Prozent – sein* / *percent – to be*). The set of verbs and accusative objects comprises 5,315,778 different pairs. The most frequent pair

Cluster 16 (0.0372036)			
Feature 0		Feature 1	
sagen 'say'	0.04944	Menschen 'people'	0.0364247
verletzen 'injure'	0.0297649	Frau 'woman'	0.013469
töten 'kill'	0.0245558	Mann 'man'	0.0125814
glauben 'believe'	0.0172522	Leute 'people'	0.012347
erschießen 'shoot'	0.0139877	Kinder 'children'	0.0112188
fragen 'ask'	0.0133666	Frauen 'women'	0.0110965
meinen 'believe'	0.0102231	Personen 'persons'	0.00700736
ermorden 'murder'	0.00950939	Männer 'men'	0.00679796
angreifen 'attack'	0.00945653	Soldaten 'soldiers'	0.00544463
festnehmen 'arrest'	0.00792689	Opfer 'victim'	0.00472603

Figure 6: Top-ranked verb-object clusters extracted from TüPP-D/Z.

occurs 9,205 times (*spielen – Rolle* / *play – role*).

Figures 5 and 6 show two clusters that were generated by *lsc* in this experiment and that are representative of the overall quality obtained.[2] Manual inspection of the results shows that the increased size of the input data clearly improves the quality of the clusters. Especially the elements of the verb-object clusters yield intuitive selectional preferences. For example, the nouns in cluster 16 are all about people, and the verbs deal with actions that can be done to or with people. The verbs *verletzen* ('injure'), *töten* ('kill') , or *erschießen* ('shoot') belong to a more restricted domain of war, with corresponding nouns like *Soldaten* ('soldiers') or *Opfer* ('victim'). Likewise the subject-verb cluster in figure 5 also exhibits natural semantic classes of both verbs and nouns. The verbs are all members of the semantic field of communication verbs, with the subject nouns representing prototypical agents for this verb class.

7 Comparison with other work and conclusion

With the exception of Schulte im Walde (2003), Schulte im Walde (2004b), and Schulte im Walde (2006) we are not aware of any data-driven studies of German verb classifications. To the best of our knowledge, the present paper is the first study to employ LSC for soft-clustering of German verb classes. Schulte im Walde employs hard clustering algorithms for generating verb classes and limits herself to a detailed study of 168 German verbs. The main goal of her work is to see whether

[2] Auxiliary verbs were removed from the clusters.

clustering techniques can yield empirically adequate results for the set of verbs that she considers. By comparison, the present work does not limit itself to a preselected number of verbs and uses soft clustering. Another interesting difference between Schulte im Walde's work and ours concerns the way in which she generalizes over the nominal complements obtained for a particular cluster. In Schulte im Walde (2004a), all nominal heads are projected to 15 most general concepts superimposed on the GermaNet hypernym hierarchy of nouns, the German version of WordNet (Hamp and Feldweg, 1997), so that selectional preferences of verbs can be expressed by very general ontological categories such as *situation, concrete object, abstract object*, etc.

An aspect that is currently missing from this research is an objective way of evaluating clusters. One possibility for evaluation of quality would be to map the elements of the clusters to their corresponding GermaNet concepts and then to search for hypernyms in the GermaNet hierarchy. If it is possible to find a restricted number of hypernyms that cover most, if not all, of the nouns in a cluster, this is an indication for the coherence of the cluster. An alternative way of measuring the quality of a cluster could be to consult lists of word associations as described in Dennis (2003) for English. Unfortunately, resources of this kind and of suitable size are not available for German. Yet another evaluation strategy could be to employ other techniques of corpus-based inference of semantic properties, such as Latent Semantic Analysis (Landauer and Dumais, 1997). However, the results of such a comparison would certainly have to be taken with a grain of salt since LSA is a much more general technique for measuring semantic relatedness.

We conclude with some brief remarks about two additional directions for future research. Wagner (2005) has shown for English how selectional preferences can be obtained by data abstraction on nominal argument positions of verb classes that are obtained by LSC. Wagner's approach differs from Schulte im Walde's in that the latter always generalizes to a set of very general ontological categories while Wagner tries to generalize up the hypernym hierarchy only as high as is supported by the data. This has the effect that ceteris paribus the selectional preferences that Wagner's approach produces are more specific than those obtained by other abstraction methods. This in turn leads to crisper selectional preferences. Another direction of future research concerns a task-based evaluation of the LSC clustering results. In the present paper we have limited ourselves to a purely manual inspection of the LSC clusters for the two treebanks we have considered. While this seems adequate for comparing the relative quality of clusters obtained by the two treebanks, it remains to be seen whether the clusters obtained from the TüPP/D-Z treebank are of sufficient quality to be used in NLP applications for which selectional preferences of verbs can play an important role.

Automatic pronoun resolution seems to be a good candidate for such a task-based evaluation since it has often been argued that selectional preferences can provide an important source of knowledge for this task.

References

Abney, S. (1991). Parsing by chunks. In R. Berwick, S. Abney, and C. Tenny (Eds.), *Principle-based Parsing*. Boston: Kluwer Academic Publishers.

Brants, T. and O. Plaehn (2000). Interactive corpus annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

Dennis, S. (2003). A Comparison of Statistical Models for the Extraction of Lexical Information from Text Corpora. In *Proceedings of the Twenty Fifth Conference of the Cognitive Science Society*, pp. 330–335.

Hamp, B. and H. Feldweg (1997). GermaNet – a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications)*, Madrid, Spain.

Höhle, T. (1986). Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, Göttingen, Germany, pp. 329–340.

Landauer, T. K. and S. T. Dumais (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review 105*, 221–240.

Müller, F. H. (2004a). *A Finite State Approach to Shallow Parsing and Grammatical Functions Annotation of German*. Ph. D. thesis, University of Tübingen.

Müller, F. H. (2004b). Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z).

Müller, F. H. and T. Ule (2002). Annotating topological fields and chunks – and revising POS tags at the same time. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan, pp. 679–701.

Rooth, M. (1998). Two-dimensional clusters in grammatical relations. In M. Rooth, S. Riezler, D. Prescher, S. Schulte im Walde, G. Carroll, and F. Beil

(Eds.), *Inducing Lexicons with the EM Algorithm*, Volume 4 of *AIMS*, pp. 7–24. Universität Stuttgart.

Schmid, H. (2006). LSC. Institut für maschinelle Sprachverarbeitung, University of Stuttgart, `http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/LSC.html`.

Schulte im Walde, S. (2003). *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph. D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Published as AIMS Report 9(2).

Schulte im Walde, S. (2004a). GermaNet Synsets as Selectional Preferences in Semantic Verb Clustering. *LDV-Forum – Zeitschrift für Computerlinguistik und Sprachtechnologie 19*(1/2), 69–79. also published in Proceedings of the GermaNet Workshop, 2003.

Schulte im Walde, S. (2004b). Induction of Semantic Classes for German Verbs. In S. Langer and D. Schnorbusch (Eds.), *Semantik im Lexikon*, Volume 479 of *Tübinger Beiträge zur Linguistik*, pp. 59–86. Tübingen: Gunter Narr Verlag.

Schulte im Walde, S. (2006). Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics 32*(2), 159–194.

Telljohann, H., E. W. Hinrichs, and S. Kübler (2003). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

Wagner, A. (2005). *Learning Thematic Role Relations for Lexical Semantic Nets*. Ph. D. thesis, University of Tübingen.