# Disambiguation of the Semantics of German Prepositions: a Case Study

Simon Clematide, Manfred Klenner, and Lenz Furrer

Institute of Computational Linguistics
University of Zurich

**Abstract.** In this paper, we describe our experiments in preposition disambiguation based on a – compared to a previous study – revised annotation scheme and new features derived from a matrix factorization approach as used in the field of distributional semantics. We report on the annotation and Maximum Entropy modelling of the word senses of two German prepositions, *mit* ('with') and *auf* ('on'). 500 occurrences of each preposition were sampled from a treebank and annotated with syntacto-semantic classes by three annotators. Our coarse-grained classification scheme is geared towards the needs of information extraction, it relies on linguistic tests and it strives to separate semantically regular and transparent meanings from idiosyncratic meanings (i.e. of collocational constructions). We discuss our annotation scheme and the achieved inter-annotator agreement, we present descriptive statistical material e.g. on class distributions, we describe the impact of the various features on syntacto-semantic and semantic classification and focus on the contribution of semantic classes stemming from distributional semantics.

**Keywords:** Word Sense Disambiguation, Preposition, Distributional Semantics, German

## 1 Introduction

Prepositions in the sense of single word prepositions are a rather small closed lexical class with several dozen types in languages such as German, English and French. In terms of word occurrences, however, prepositions contribute a substantial amount of tokens. For instance, in the German newspaper treebank TIGER (Brants and Hansen, 2002) 12% of 768,971 word tokens (not counting punctuation tokens) are tagged as prepositions. Prepositions occurring very frequently show a high degree of ambiguity and polysemy. For 13 frequent English prepositions, Litkowski and Hargraves (2006) recorded 211 senses.

Linguistics has a long-standing tradition of sense classification of prepositional phrases used as adjuncts. Traditional dictionaries also collect detailed sense information about prepositions. In case of *mit*, the German online dictionary Duden[1] specifies 8 main senses, additionally some of them have subsenses

[1] See http://www.duden.de

resulting in a total of 12 senses. It is yet unclear which classification schemes should be used for applications that require semantic interpretation such as information extraction or questions answering – although there have been two preposition word sense disambiguation (PWSD) shared tasks for English in the past. In this paper, we want to gain experience for a larger attempt in classifying the semantic contributions of prepositions across different languages as German, English and French. Our main interest is to differentiate between semantically transparent contributions that prepositional phrases can provide in a general or productive manner on the one hand and the less transparent contributions in collocational constructions on the other hand. Additionally, many prepositions are subcategorized by verbs and the semantic contribution of the selected prepositions is weak or unspecific – a fact that is often revealed by cross-lingual comparisons of subcategorization frames.

In the Maximum Entropy model we propose, we exploit contextual and syntactic features that have proved most helpful in previous approaches on English PWSD. But we also focus on (German) language-specific features like e.g. morphological case, whichs turns out to be a strong feature for the preposition *auf* ('on'). Moreover, we have experimented with distributional semantics in order to derive semantic classes for preposition governors and for the noun phrase heads governed by the preposition. To best of our knowledge, this is the first attempt to utilize semantic knowledge derived in a corpus-driven manner in the task of PWSD.

The rest of this paper is organized as follows. Section 2 presents related work. In Section 3, we describe our syntacto-semantic classification system used in the annotation. We also present the approach borrowed from distributional semantics and used in the machine learning experiments for the automatic prediction of the classes. Section 4 contains a systematic evaluation of the different types of evidence that we have integrated in our approach.

## 2 Related Work

The meaning of a prepositional phrase (PP) depends – among others – on the meaning of its preposition and (the head of) the embedded noun phrase. Determining the functional role such a PP plays within a sentence can be regarded as semantic role labelling (SRL). Preposition word sense disambiguation, thus, is sometimes casted as a variant of SRL (e.g. O'Hara and Wiebe, 2009). For the English language, annotated data is available from the *Penn Treebank* II (Marcus et al., 1994), where thematic roles carried by prepositional phrases are marked, and *FrameNet* (Baker et al., 1998), which was annotated as part of the *Preposition Project* (Litkowski and Hargraves, 2006).

For German, the Salsa 2.0 project (Rehbein et al., 2012) made a substantial amount of FrameNet-like annotations available built on top of the TIGER corpus. About 20,000 verbs and 16,000 nouns are marked as frame-evoking concepts. In Salsa annotations, prepositional phrases appear as frame elements that are linked to the evoking target by named roles. Figure 1 shows the most fre-

59 (Message), 59 (Interlocutor_2), 52 (Partner_2), 48 (Cause), 39 (Phenomenon), 37 (Event), 37 (Response), 31 (Descriptor), 21 (Item2), 20 (Instrument), 18 (Means), 15 (Content), 13 (Goal), 13 (Side_2), 11 (Theme), 11 (Fact), 10 (Degree_of_involvement), 8 (Money), 7 (Goods), 7 (Co_Signatory), 7 (Funds), 7 (Creator), 7 (Contribution_salsa), 6 (Agent), 5 (Result), 5 (Manner), 5 (Party_2), 5 (Defendant), 5 (Outcome), 4 (State_of_affairs), 4 (Quantity), 4 (Medium), 4 (Action), 4 (Party2), 4 (Persistent_characteristic), 4 (Punishment), 4 (Award), 4 (Addressee), 3 (Specification), 3 (Effect), 3 (Body_part), 3 (Mode_of_transportation), 3 (Reason), 3 (Topic), 3 (Relation), 3 (Protagonist), 3 (Accused)

**Fig. 1.** Frequencies and names of the frame elements of the German FrameNet annotation Salsa 2.0 of PPs headed by *mit* occurring at least 3 times. In total there are 701 occurrences with 111 different frame element roles. 39 roles occur only once, 14 twice.

quent roles associated with PPs headed by *mit*. The fine-grained classification of the English FrameNet (with its larger annotation database) has been a PWSD challenge for O'Hara and Wiebe (2009). The even more fine-grained and less generalized role inventory of Salsa 2.0 makes the task of utilizing such a resource demanding.

A substantial contribution on preposition classification and disambiguation for English has been carried out in the Preposition Project (Litkowski and Hargraves, 2006) (see also the *SemEval* Task on WSD of prepositions, Litkowski and Hargraves, 2007). A fine-grained classification scheme was derived from the *Oxford Dictionary*, e.g. the preposition *on* is specified on the basis of 25 different senses. Other elaborated classification schemes can be found as part of *VerbNet* (Kipper et al., 2004) and *PrepNet* (Saint-Dizier, 2008). As can be seen from the diversity of these approaches, there is no agreed classification scheme for preposition disambiguation. Moreover, some authors argue that preposition classes are (in part) language-specific (Müller et al., 2011). They have specified an even more fine-grained and hierarchical classification scheme (compared to the Preposition Project), where German gold standard annotations are based on the traversal of manually specified preposition-specific decision trees. As a consequence of the complexity of the annotation scheme, no attempt was made so far by the authors to learn a model for preposition classification based on their semantic classes. Their approach based on logistic regression as described in Kiss et al. (2010) focuses on determiner omission in PPs.

Preposition classification is not only crucial for applications such as information extraction (see Baldwin et al., 2009, p. 134 for an application-oriented discussion), but also supports machine translation, see e.g. Shilon et al. (2012, p. 106). Although semantic information helps to tackle the translation task, the semantic class of a preposition does not perfectly determine the correct translation. As a consequence, these approaches do not strive to carry out preposition WSD, but to use semantic features in order to more directly map source prepositions to target prepositions (Li et al., 2005; Agirre et al., 2009). Turning the tables in a previous study, we used statistical machine translation for helping with WSD (Clematide and Klenner, 2013). However, using imperfect translations as a machine learning feature resulted in rather moderate improvements for only one of the prepositions in focus. Further research based on parallel corpora is needed here.

On the methodological side, preposition disambiguation with machine learning heavily relies on features derived from the surrounding context of the preposition, but also uses semantic resources such as *WordNet* (Fellbaum, 1998). The best system from the *SemEval* Task on preposition WSD, Kim and Baldwin (2007), combines collocational (surrounding words), syntactic (part of speech tags, chunks) and semantic features (semantic role tags, WordNet) in a Maximum Entropy model. They achieve an accuracy of 69.3% in the fine-grained classification task. Their conclusion is that the semantics of prepositions can be learned mostly from the surrounding context and not from syntactic or verb-related properties. O'Hara and Wiebe (2009) use an additional feature, hypernym collocations (WordNet hypernyms as collocation provider), to carry out disambiguation relative to either coarse-grained Penn Treebank functional roles or more sophisticated FrameNet roles. They achieve an accuracy of 89.3% given the six Penn Treebank annotated semantic classes. The results in the task of semantic role labelling based on preposition disambiguation are, due to the large number of frame roles (641), low, namely 23.3%.

Hovy et al. (2010) significantly improved on the results of O'Hara and Wiebe (2009); they achieved an accuracy of 91.8% (coarse-grained) and 84.8% (fine-grained using the *SemEval* data). The key to the success of their method seems to lie in the vast amount of different features ranging from suffix information to the holonyms of words. Not all of them are linguistically well motivated (e.g. the first and last two or three letters of each word, respectively). While their approach certainly sets a new standard, its utility to languages other than English is not guaranteed, since some features are geared towards English resources such as WordNet (or Roget's Thesaurus) that are not available in the same quality in other languages. Other features like capitalization are unlikely to be useful for German.

## 3 Methods

### 3.1 Resources

As mentioned in Section 2, the Penn Treebank comprises shallow semantic annotations to PPs. There, a distinction is made between several semantic classes of PPs: locative, direction, manner, purpose, temporal, and extent. Unfortunately, none of the large German treebanks (TIGER (Brants and Hansen, 2002), Tüba-D/Z (Telljohann et al., 2004)) provide such a comparable rudimentary scheme that could be a starting point for our case study. There is no resource that we could use, although one is currently being developed by another group (Müller et al., 2011), but it is not yet released. Since we believe that treebanks could benefit from such an additional annotation layer, we decided to work with the largest German treebank, the *Tübinger Baumbank* Tüba-D/Z 7.0. It comprises about 65,000 annotated sentences. Besides phrase structure, topological fields

**Table 1.** Distribution of the syntacto-semantic functions of *auf* ('on') in relation to the syntactic dependencies from the Tüba-D/Z treebank and from the ParZu parser. For Tüba-D/Z, the syntactic function "predicative" is labelled as "p", "–" is used if there is no governor available (e.g. syntactically not integrated PPs) or if there is another rare syntactic configuration.

Tüba-D/Z

| sem\syn | opp | mod | vmod | ?mod | – | p | $\sum$ |
|---|---|---|---|---|---|---|---|
| LOC | 10 | 45 | 79 | 9 | 6 | 2 | 151 |
| verbal | 119 | 2 | 1 | | 2 | | 124 |
| nominal | | 67 | | | 2 | | 69 |
| coll | 44 | 4 | 7 | | | 2 | 57 |
| DIR | 24 | 3 | 8 | | 1 | | 36 |
| MOD | 3 | 3 | 8 | 4 | 1 | | 19 |
| TLOC | | 3 | 12 | 1 | | | 16 |
| ? | 1 | 4 | 2 | | 2 | 1 | 10 |
| CAU | | | 3 | 3 | 1 | | 7 |
| TEM | | 2 | 4 | | | | 6 |
| adjectival | 1 | 3 | 1 | | | | 5 |
| $\sum$ | 202 | 136 | 125 | 17 | 15 | 5 | 500 |

ParZu

| sem\syn | v-pp | n-pp | v-objp | – | a-pp | $\sum$ |
|---|---|---|---|---|---|---|
| LOC | 92 | 29 | 8 | 17 | 5 | 151 |
| verbal | 63 | 5 | 47 | 8 | 1 | 124 |
| nominal | 5 | 62 | | 2 | | 69 |
| coll | 20 | 3 | 31 | 1 | 2 | 57 |
| DIR | 20 | 10 | 2 | 3 | 1 | 36 |
| MOD | 8 | 5 | 1 | 4 | 1 | 19 |
| TLOC | 12 | 3 | | | 1 | 16 |
| ? | 5 | 3 | 1 | 1 | | 10 |
| CAU | 6 | | | 1 | | 7 |
| TEM | 4 | 2 | | | | 6 |
| adjectival | 3 | | | | 2 | 5 |
| $\sum$ | 238 | 122 | 90 | 37 | 13 | 500 |

and grammatical functions are also specified. PPs can act as obligatory or optional (opp) complements of verbs, as NP or PP modifiers (mod), or as adjuncts (vmod).

From the ten most frequent prepositions in the Tüba-D/Z we have chosen one with a predominant local meaning (*auf* 'on') and one with a broader meaning spectrum (*mit* 'with'). We randomly sampled 500 occurrences of each.

**Dependency Parser Output** In order to have a realistic setup for our experiments, we use syntactic evidence derived from the output of a dependency parser for German, the *ParZu* (Sennrich et al., 2009). For syntactically embedded prepositional phrases, this parser applies the following dependency labels: "objp" for verb complements (analogous to the Tüba-D/Z dependency "opp") and "pp" for modifiers. In Table 1, we show the numbers for verbal ("v-"), nominal ("n-"), and adjectival heads ("a-"). There is quite a number of syntactically not embedded PPs (category "–"). This is mostly due to very complex and long sentences from the newspaper corpus where the parser cannot produce a fully connected dependency structure covering all tokens of a sentence and, therefore, emits forests of parse fragments instead of a parse tree.

**Semantics and Annotation of *auf* and *mit*** Since we envisage information extraction and question answering as an application context, a coarse-grained classification of the semantics of prepositions, tightly coupled with question words, seems appropriate.

**Table 2.** Distribution of the syntacto-semantic functions of *mit* ('with') in relation to the syntactic dependencies from the Tüba-D/Z treebank and from the ParZu parser. For Tüba-D/Z, the syntactic function "predicative" is not shown in the table because it appeared only once. For ParZu, the label "–" comprises syntactically not integrated PPs and the label "#" means cases that were not even integrated into a PP. Two dependency types occur only once and they are not shown in the table.

Tüba-D/Z

| sem\syn | vmod | mod | opp | ?mod | – | $\sum$ |
|---|---|---|---|---|---|---|
| verbal | 8 | 4 | 86 | 1 | | 99 |
| INS | 74 | 4 | 7 | | 1 | 86 |
| MOD | 54 | 4 | 4 | 9 | 2 | 73 |
| ORN | 1 | 55 | | 1 | 2 | 59 |
| nominal | 2 | 50 | | | | 52 |
| COM | 31 | 6 | 12 | 2 | 1 | 52 |
| adjectival | 1 | 8 | 9 | | | 18 |
| IDE | 7 | 1 | | 7 | | 15 |
| coll | 5 | 1 | 5 | 1 | 1 | 13 |
| SIZ | 5 | 6 | 1 | 1 | | 13 |
| ? | 2 | 3 | | 2 | 4 | 11 |
| TEM | 6 | | 1 | 1 | | 8 |
| $\sum$ | 196 | 142 | 125 | 25 | 11 | 499 |

ParZu

| sem\syn | V-pp | N-pp | – | V-objp | A-pp | # | $\sum$ |
|---|---|---|---|---|---|---|---|
| verbal | 65 | 6 | 5 | 22 | 1 | | 99 |
| INS | 69 | 7 | 8 | 1 | | 1 | 86 |
| MOD | 58 | 4 | 5 | | 4 | 2 | 73 |
| ORN | 24 | 27 | 8 | | | | 59 |
| nominal | 13 | 37 | 1 | | | 1 | 52 |
| COM | 41 | 6 | 2 | | 1 | 2 | 52 |
| adject. | 13 | | | 1 | 4 | | 18 |
| IDE | 15 | | | | | | 15 |
| coll | 11 | | 2 | | | | 13 |
| SIZ | 8 | 4 | | | 1 | | 13 |
| ? | 6 | | 3 | | | 1 | 10 |
| TEM | 8 | | | | | | 8 |
| $\sum$ | 331 | 91 | 34 | 24 | 11 | 7 | 498 |

In the case of *auf* (cf. Tab. 1), we distinguish between locative (LOC *where*), directional (DIR *where to*), temporal (TEM *when*, *how long*), modal (MOD *how*), and causal (CAU *why*) PPs. If the noun in a temporal PP is an event (e.g. *party*), then often a locative or a temporal reading is possible (e.g. *when or where did he laugh? – at his party*). We use TLOC to refer to this usage. If the PP acts as a subcategorized modifier of an adjective or noun, it is annotated with "adjectival" or "nominal"(e.g. *decision on nuclear plants*). In case that the verb governs an otherwise semantically vacuous preposition (*warten auf* 'to wait for'), the preposition is marked with "verbal". Finally, any idiomatic expression comprising a PP having a non-compositional meaning like *auf den Putz hauen* 'to kick up one's heels' is annotated as collocational ("coll"). The preposition does not contribute any semantics in these cases. Sometimes no decision was possible (e.g. given sentence fragments, missing global context, unclear semantics), and we used "?" to annotate these instances.

Table 1 shows the distribution of these classes and their syntactic analysis for the preposition *auf*, both relative to the treebank annotation (left-hand side) and the dependency labels of the parser (right-hand side). Local senses form the largest class (151), followed by the syntactic classes "verbal", "nominal" and "coll". All other senses of *auf* have lower frequencies. Syntactically, there are three groups to be distinguished: PP complements (opp, 202), NP and PP modification (mod, 136) and adjuncts (v-mod, 125).

In the case of *mit* (cf. Tab. 2), the syntactic labels "verb", "nominal" and "coll" are used as introduced above for *auf*. The prepositions *auf* and *mit* also share two semantic classes, namely TEM (temporal) and MOD (modal). The other semantic classes of *mit* are: COM for comitative use (*to watch a movie with a friend*), ORN for ornative use (*a man with humor*), SIZ indicating size or proportion (*to demonstrate with 100 people against*), INS for the instrument reading, which is a subclass of MOD (modal) (*to break with a hammer*), and IDE for identity (*with him, hope enters the room* meaning: *he represents/is identical with hope*). Note that *mit* has a more balanced distribution of semantic classes.

*Inter-Annotator Agreement* For the annotations used in the previous work (Clematide and Klenner, 2013) we have measured inter-annotator agreement in two stages. There was an initial annotation round where one annotator had created the annotation strategy and initial guidelines for one preposition based on existing sense inventories from the literature. The harmonized annotation was then built after discussing the cases where the initial annotations were different. This resulted in further clarifications and refinements of the guidelines, but we also dropped some distinctions that were difficult to apply (e.g. local meaning in a physical sense of contact versus a metaphorical sense).

**Table 3.** Inter-Annotator agreement of the annotations. We report the percentage of agreeing decisions as well as Cohen's $\kappa$.

| Annotations | *auf* | | *mit* | |
|---|---|---|---|---|
| | agreeing | $\kappa$ | agreeing | $\kappa$ |
| initial A vs. initial B | 74 | .67 | 85 | .82 |
| initial A vs. harmonized | 85 | .81 | 92 | .90 |
| initial B vs. harmonized | 86 | .83 | 92 | .90 |
| revised harm. A vs. majority | 93 | .91 | 96 | .96 |
| revised harm. B vs. majority | 92 | .90 | | |
| initial C vs. majority | 82 | .77 | 74 | .70 |

As shown in Table 3, Cohen's $\kappa$ was high for *mit* and lower, but still substantial for *auf*. There were two problems regarding this harmonized annotation: First, *auf* was missing semantic annotations for nominal and adjectival modifiers. Second, after systematically analyzing the governor lemmas we detected some global inconsistencies regarding the distinction of syntactic classes and semantic classes. As already observed by Tseng (2000), there is no dichotomous categorial distinction between subcategorized functional prepositions and semantic (also called autosemantic) ones in all cases. It is more a difference of degree. In order to give more weight to the semantics of prepositions we revised the guidelines accordingly.

Given these circumstances a third independent annotation C was mandatory. For *auf*, annotator A and B had to revise the "nominal" cases. All annotators again reviewed the cases with disagreement. The final version used in this paper was built by majority voting. Table 3 gives an overview of the agreement for the different steps of the annotations.

**Distributional Semantics: Does it help in preposition classification?** Distributional semantics (DS) is based on the assumption that similar words appear in similar contexts and that the semantic relatedness of words can be measured by a comparison of their contexts (see Erk, 2012, for an overview). Words are represented by vectors in a high-dimensional space and their "positions" can be compared e.g. by the cosine similarity measure. In order to detect the semantic dimensions underlying this huge vector space organised as a co-occurrence matrix, factorisation methods come into play, e.g. Nonnegative Matrix Factorisation (Shashanka et al., 2008). The principle of dimension reduction, which is central to these approaches, allows to cluster words into classes (hard or soft) based on their similarity in vector space.

The general idea in our experiments was to derive, by way of matrix factorisation[2] and dimension reduction, separate semantic classes of the nouns a) that govern the preposition and b) are governed by the preposition (i.e. the heads of the embedded noun phrases) – called target words, henceforth. We extracted all target words of *mit* and *auf* from the Tüba-D/Z and generated vectors based on 2000 context words. A dependency-parsed version of the DeWac corpus (90 million sentences) (Baroni et al., 2009) was used in order to detect good context words of the target words. Those context words that co-occurred most frequently with as many target words (NP heads) as possible were selected. The vectors were combined into a matrix, where rows represent target words and columns are context words, a single cell records the frequency of a context words co-occuring with the target word.

We then decomposed this matrix with Nonnegative Matrix Factorisation according to different ranks, namely 10, 20 and 50, in two different matrices, a base matrix and a coefficient matrix. The base matrix can be used to determine the class membership of the target words, the classes are produced (soft clustered) by dimension reduction according to the given ranks. We determined for each target word (governor and governed NP head, respectively) the three classes with the highest numerical impact (which determines class membership strength) and used these highest ranked classes as features. The hypothesis was that there is a correlation between these classes and the semantic classes underlying our gold standard.

### 3.2 Supervised Machine Learning Approach

In order to measure the difficulty of an automatic classification of the syntacto-semantic classes expressed by *auf* and *mit* we conducted several experiments with the Maximum Entropy Modeling tool *MegaM* (Daumé III, 2008). The *Maximum Entropy* approach for classification is also known as *Logistic Regression*

[2] We worked with the Python implementation NIMFA (Zitnik and Zupan, 2012).

and has been reported to perform very well for PWSD in Tratz and Hovy (2009). For this case study, we focused on simple features gained from the output of the ParZu dependency parser, textual data from the context, and distributional semantics. Some prepositions such as *auf* can govern two different grammatical cases depending on the semantics expressed by the PP. For instance, *auf* with dative is topological whereas *auf* with accusative case is directional. The ParZu parser does not enforce the disambiguation of grammatical case in PPs. In order to have disambiguated grammatical case for each occurrence of *auf*, we used the statistical case tagger based on Conditional Random Fields from Clematide (2013). As for the distributional semantics features, information about the governor of the PP could be provided in 74% (*auf*) and 71% (*mit*) of the samples. Information about the governed head in 78% (*auf*) and 74% (*mit*) of the samples.

In Section 4 we present and analyze the results and performance contribution of the following feature sets:

- **case** Case governed by the preposition (accusative/dative). Only for *auf*.
- **syntax** The syntactic function of the PP taken from ParZu parser output.
- **neighbor** Word, POS (part of speech), and lemma of the preceding and following token.
- **context** Word, POS, and lemma in a window of 5 preceding and following tokens (taken as a bag of words, lemmas or POS).
- **head** Word, POS, and lemma of the head word (typically a noun) of the dependent phrase of the preposition, for instance, the head of *mit Sorgfalt* is *Sorgfalt* 'care'. In case of coordinated PPs and multi-word heads, the first token was selected.
- **head n** The first 3 classes of a distributional semantics model of rank $n$ of the head.
- **governor** The lemma of the word governing the PP.
- **governor n** The first 3 classes of a distributional semantics model of rank $n$ of the governor.

## 4 Results and Discussion

The evaluations assess the performance improvement for the multi-class predictions of our 500 annotated prepositions by using different feature sets as evidence. We evaluate against a baseline system which basically predicts the majority class given the lack of any additional evidence. All results are reported as mean accuracy computed by cross-validation (stratified by classes).

### 4.1 Syntacto-Semantic Classification

We performed a 10-fold cross-validation evaluation for the scenario of predicting the full set of all syntactic and semantic class labels (cf. Tab. 1 and 2). The results of *auf* are shown in Tab. 4a. The best system uses almost all feature sets,

**Table 4.** Performance of feature sets for syntacto-semantic classification accuracy. The column "Mean" contains the average accuracy computed from the cross-validation sets. $\triangle\mathrm{rel}_{bs}$ expresses the relative performance gain. The last row contains the feature set with the best performance. Only systems beating the baseline are shown.

(a) *auf* ($N = 500$)

| System | Mean | SD | $\triangle\mathrm{rel}_{bs}$ |
|---|---|---|---|
| baseline | 30.2 | 0.6 | |
| h(ead) | 33.6 | 3.4 | +11.3 |
| h10 | 34.6 | 3.1 | +14.6 |
| g(overnor) | 35.2 | 5.3 | +16.6 |
| h50 | 37.0 | 6.9 | +22.5 |
| h20 | 37.0 | 6.7 | +22.5 |
| g10 | 38.8 | 2.1 | +28.5 |
| g20 | 39.8 | 7.4 | +31.8 |
| g50 | 43.4 | 4.9 | +43.7 |
| s(yntax) | 44.4 | 4.6 | +47.0 |
| c(ontext) | 45.4 | 8.5 | +50.3 |
| ca(se) | 52.8 | 2.5 | +74.8 |
| n(eighbor) | 53.8 | 5.8 | +78.1 |
| s/n/ca/h/g | 67.4 | 4.6 | +123.2 |
| s/n/ca/h20/g/g50 | **71.0** | 4.3 | +135.1 |

(b) *mit* ($N = 500$)

| System | Mean | SD | $\triangle\mathrm{rel}_{bs}$ |
|---|---|---|---|
| baseline | 19.8 | 0.6 | |
| g20 | 20.8 | 2.7 | +5.1 |
| h20 | 21.4 | 7.1 | +8.1 |
| g10 | 22.0 | 3.3 | +11.1 |
| g50 | 23.4 | 5.4 | +18.2 |
| h50 | 25.4 | 4.8 | +28.3 |
| s(yntax) | 26.2 | 4.5 | +32.3 |
| h(ead) | 26.2 | 4.3 | +32.3 |
| g(overnor) | 26.8 | 5.3 | +35.4 |
| c(ontext) | 33.0 | 6.1 | +66.7 |
| n(eighbor) | 35.6 | 5.2 | +79.8 |
| s/n/h/g | 42.0 | 5.7 | +112.1 |
| s/c/h/h20/g/g20 | **43.6** | 6.8 | +120.2 |
| s/n/c/h/h50/g/g50 | **43.6** | 5.7 | +120.2 |
| s/n/c/h/h20/g/g50 | **43.6** | 5.3 | +120.2 |
| s/n/c/h/h20/g/g10 | **43.6** | 4.1 | +120.2 |

"case" and "neighbor" are especially strong. The head and governor features are relatively weak, and so are their distributional equivalents. However, the distributional feature sets head 20 and governor 50 contribute to the best system. The best system without any distributional semantics shows a substantially reduced performance.

Table 4b gives the results for *mit*. The overall performance is lower. Head and governor are much stronger for *mit* compared to *auf*. The best performance is reached by rather different feature sets. The rank size, i.e. the number of distributional classes, does not have a strong influence on the results. The best system without distributional semantics performs noticeably worse.

### 4.2 Semantic Classification

In a further evaluation, we measured how well the purely semantic classes (i.e. the classes without "nominal", "verbal", "adjectival", and "coll") can be predicted. For *auf* we only have 235 cases with a defined semantic classification, for *mit* we have 306. Due to the smaller data sets we performed 5-fold cross-validation. Table 5a illustrates the problems from the skewed distribution of semantic classes in the case of *auf*: Just guessing the largest class LOC represents a strong baseline decision. Case information adds most of the improvement. However, distributional semantics of the head improves further. The best system

**Table 5.** Performance of features sets for semantic classification accuracy. The classes are LOC, DIR, MOD, TLOC, CAU, and TEM for *auf*; TEM, MOD, INS, ORN, COM, IDE, and SIZ for *mit*.

(a) *auf* ($N = 235$)

| System | Mean | SD | $_\Delta\text{rel}_{bs}$ |
|---|---|---|---|
| baseline | 64.3 | 1.0 | |
| h20 | 66.0 | 5.2 | +2.6 |
| c(ontext) | 66.0 | 3.4 | +2.6 |
| n(eighbor) | 67.2 | 7.0 | +4.6 |
| h(ead) | 67.2 | 4.7 | +4.6 |
| ca(se) | 77.0 | 2.3 | +19.9 |
| ca/s/h | 80.4 | 2.3 | +25.2 |
| ca/h20 | **81.3** | 2.8 | +26.5 |

(b) *mit* ($N = 306$)

| System | Mean | SD | $_\Delta\text{rel}_{bs}$ |
|---|---|---|---|
| baseline | 28.1 | 0.5 | |
| g(overnor) | 30.7 | 3.2 | +9.3 |
| h10 | 32.0 | 3.1 | +13.9 |
| s(yntax) | 33.7 | 4.8 | +19.9 |
| h20 | 35.3 | 5.6 | +25.6 |
| h50 | 35.6 | 4.7 | +26.7 |
| h(ead) | 37.6 | 2.4 | +33.7 |
| n(eighbor) | 42.8 | 4.2 | +52.3 |
| c(ontext) | 43.5 | 3.3 | +54.7 |
| h/s/n/c/g | 46.7 | 7.5 | +66.3 |
| s/n/c/h/h20/g20 | **51.0** | 4.3 | +81.4 |

without distributional semantics also includes syntax and performs only slightly worse than the best system.

The less skewed distribution of semantic classes in the case of *mit* allows for a significant improvement over the baseline system. Tab. 5b shows that all feature sets have a beneficial effect. For *mit*, distributional semantics with a rank of 20 increases the results considerably. It is interesting to note that for *auf* the effect of distributional semantics is strong for the syntacto-semantic classification and weak for the semantic classification. For *mit*, we have the opposite situation.

## 5 Conclusion

We have introduced a coarse-grained annotation scheme for, currently, two German prepositions, *auf* and *mit*. In our experiments with 500 annotated instances of each preposition, we did not only systematically explore the contribution of various contextual and syntactic features commonly used in the field, we also tried to work out the impact semantic information derived from distributional semantics could have on our classification tasks, the syntacto-semantic and semantic disambiguation of the two prepositions. We found that semantic classes derived by matrix factorisation do have an impact although its magnitude is not overwhelming in all cases. Further work is needed to systematically explore the contribution of these approaches. We also intent to carry out experiments with GermaNet (Kunze and Lemnitzer, 2002), the German counterpart of WordNet, in order to find out whether these distinct semantic resources interfere or rather are complementary.

The application of *Active Learning* techniques (Settles, 2012) might help to overcome another problem: the skewed distribution of semantic classes, here of *auf*. In order to relieably detect small semantic classes, more training material is needed. Active learning could be used to efficiently gather interesting new instances of such classes.

We also intend to integrate further language resources, e.g. collocation information as provided by services such as *Wortschatz Leipzig*[3] or *Digitales Wörterbuch der Deutschen Sprache*.[4] Bilingual lexicons such as dict.cc[5] might as well prove fruitful. They contain information about semantically void subcategorized prepositions, for instance *auf jdn warten* is linked to *to wait for sb*. Finally, we will continue to investigate the benefits of cross-lingual information as described in a recent paper (Clematide and Klenner, 2013).

## References


Agirre, E., Atutxa, A., Labak, G., Lersundi, M., Mayor, A., and Sarasola, K. (2009). Use of rich linguistic information to translate prepositions and grammar cases to Basque. In *Proceedings of the XIII Conference of the European Association for Machine Translation (EAMT)*, pages 58–65.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In Boitet, C. and Whitelock, P., editors, *COLING-ACL*, pages 86–90. Morgan Kaufmann Publishers / ACL.

Baldwin, T., Kordoni, V., and Villavicencio, A. (2009). Prepositions in applications: A survey and introduction to the special issue. *Computational Linguistics*, 35(2):119–149.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, pages 209–226.

Brants, S. and Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, pages 1643–1649, Las Palmas.

Clematide, S. (2013). A case study in tagging case in German: An assessment of statistical approaches. In *Systems and Frameworks for Computational Morphology: Third International Workshop (SFCM 2013)*, Communications in Computer and Information Science, pages 22–34, Berlin, Germany. Springer.

Clematide, S. and Klenner, M. (2013). A pilot study on the semantic classification of two German prepositions: Combining monolingual and multilingual evidence. In *Proceedings of Recent Advances in Natural Language Processing*, pages 148–155, Hissar, Bulgaria.

Daumé III, H. (2008). MegaM: Maximum entropy model optimization package. ACL Data and Code Repository, ADCR2008C003.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.


[3] See http://wortschatz.uni-leipzig.de

[4] See http://dwds.de

[5] See http://www.dict.cc

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Hovy, D., Tratz, S., and Hovy, E. H. (2010). What's in a preposition? Dimensions of sense disambiguation for an interesting word class. In *COLING (Posters)*, pages 454–462.

Kim, S. N. and Baldwin, T. (2007). MELB-KB: Nominal classification as noun compound interpretation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, page 231–236, Prague, Czech Republic.

Kipper, K., Snyder, B., and Palmer, M. (2004). Using prepositions to extend a verb lexicon. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, pages 23–29.

Kiss, T., Keßelmeier, K., Müller, A., Roch, C., Stadtfeld, T., and Strunk, J. (2010). A logistic regression model of determiner omission in PPs. In Huang, C.-R. and Jurafsky, D., editors, *COLING (Posters)*, pages 561–569. Chinese Information Processing Society of China.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, presentation, visualization, application. In *Proceedings of LREC 2002*, pages 1485–1491.

Li, H., Japkowicz, N., and Barrière, C. (2005). English to Chinese translation of prepositions. In *Canadian Conference on AI*, volume 3501 of *Lecture Notes in Computer Science*, pages 412–416. Springer.

Litkowski, K. and Hargraves, O. (2007). SemEval-2007 task 06: Word-sense disambiguation of prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 24–29, Prague, Czech Republic. Association for Computational Linguistics.

Litkowski, K. C. and Hargraves, O. (2006). Coverage and inheritance in The Preposition Project. In *Third ACL-SIGSEM Workshop on Prepositions*, pages 37–44.

Marcus, M. P., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn treebank: Annotating predicate argument structure. In *HLT*, pages 114–119. Morgan Kaufmann.

Müller, A., Roch, C., Stadtfeld, T., and Kiss, T. (2011). Annotating spatial interpretations of German prepositions. In *ICSC*, pages 459–466. IEEE.

O'Hara, T. and Wiebe, J. (2009). Exploiting semantic role resources for preposition disambiguation. *Computational Linguistics*, 35(2):151–184.

Rehbein, I., Ruppenhofer, J., Sporleder, C., and Pinkal, M. (2012). Adding nominal spice to SALSA – frame-semantic annotation of German nouns and verbs. In *Proceedings of KONVENS 2012, Vienna, Austria*, pages 89–97.

Saint-Dizier, P. (2008). Syntactic and semantic frames in PrepNet. In *IJCNLP*, pages 763–768. ACL.

Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A new hybrid dependency parser for German. In *Proceedings of the German Society for Computational Linguistics and Language Technology 2009 (GSCL 2009)*, pages 115–124, Potsdam, Germany.

Settles, B. (2012). *Active Learning*, volume 6 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool.

Shashanka, M., Raj, B., and Smaragdis, P. (2008). Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008. Article ID 947438.

Shilon, R., Fadida, H., and Wintner, S. (2012). Incorporating linguistic knowledge in statistical machine translation: Translating prepositions. In *Proceedings of the EACL-2012 Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 106–114.

Telljohann, H., Hinrichs, E. W., and Kübler, S. (2004). The Tüba-D/Z treebank: Annotating German with a context-free backbone. In *LREC*, pages 2229–2232. European Language Resources Association.

Tratz, S. and Hovy, D. (2009). Disambiguation of preposition sense using linguistically motivated features. In *Proceedings of NAACL-HLT 2009, Companion Volume: Student Research Workshop and Doctoral Consortium*, SRWS '09, pages 96–100, Stroudsburg, PA, USA. ACL.

Tseng, J. L. (2000). *The Representation and Selection of Prepositions*. Phd thesis, University of Edinburgh.

Zitnik, M. and Zupan, B. (2012). NIMFA: A Python library for nonnegative matrix factorization. *Journal of Machine Learning Research*, 13:849–853.