Determining the Degree of Compositionality of German Particle Verbs by Clustering Approaches

Natalie Kühner and Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Germany
{kuehnene,schulte}@ims.uni-stuttgart.de

Abstract

This work determines the degree of compositionality of German particle verbs by two soft clustering approaches. We assume that the more compositional a particle verb is, the more often it appears in the same cluster with its base verb, after applying a probability threshold to establish cluster membership. As German particle verbs are difficult to approach automatically at the syntax-semantics interface, because they typically change the subcategorisation behaviour in comparison to their base verbs, we explore the clustering approaches not only with respect to technical parameters such as the number of clusters, the number of iterations, etc. but in addition focus on the choice of features to describe the particle verbs.

1 Introduction

A multi-word expression (MWE) is a combination of two or more simplex words,[1] covering compounds as well as collocations. From a semantic point of view, multi-word expressions are either considered as idiosyncratic (Sag et al., 2002; Villavicencio et al., 2005; Fazly and Stevenson, 2008), i.e., non-compositional, or alternatively the MWE compositionality is assumed to be on a continuum between entirely compositional/transparent and entirely non-compositional/opaque expressions. We conform to the latter view, and consider multi-word expressions as a composition of simplex words which may or may not be entirely predictable on the basis of standard rules and lexica. This view is in line with recent work on multi-word expressions, e.g., McCarthy et al. (2003; 2007), and also theoretical considerations about compositionality, cf. Kavka (2009).

Addressing the compositionality of multi-word expressions is a crucial ingredient for lexicography (concerning the question of whether to lexicalise a MWE) and Natural Language Processing applications (to know whether the expression should be treated as a whole, or through its parts, and what the expression means). We are interested in determining the degree of compositionality of one empirically challenging class of German multi-word expressions, i.e., German particle verbs, productive compositions of a base verb and a prefix particle. The work relies on a Studienarbeit by the first author (Kühner, 2010).

We propose two clustering approaches to address the compositionality of particle verbs. The core idea is that the compositionality of the multi-word expressions is determined by the co-occurrence of the particle verbs and the respective base verbs within the same clusters. I.e., we assume that the more compositional a particle verb is, the more often it appears in the same cluster with its base verb. Note that our idea restricts the compositionality of multi-word expressions to the relationship between particle and base verb and thus for the time being ignores the contribution of the particle. As we are relying on soft clustering approaches, cluster membership is represented by a probability. We transfer the probabilistic membership into a binary membership by establishing a membership cut-off, i.e., only verbs above a certain probability threshold are considered to be cluster members.

[1] Note that the definition of multi-words is not straightforward or agreed upon Lieber and Stekauer (2009a). Our definition is one possibility among many, but has generally been adopted by computational linguistics.

German particle verbs are an empirical challenge because they are difficult to approach automatically at the syntax-semantics interface: they typically change the subcategorisation behaviour in comparison to their base verbs, cf. Section 2. Consequently, we explore the clustering approaches not only with respect to technical parameters such as the number of clusters, the number of iterations, etc. but in addition focus on the choice of features to describe the particle verbs. The compositionality scores as predicted by the clustering approaches are evaluated by comparison against human judgements, using the Spearman rank-order correlation coefficient.

The remainder of the paper is organised as follows. Section 2 introduces the reader into German particle verbs. Following an overview of the clustering approaches in Section 3, we then describe the experiments (Section 4) and the results (Section 5).

2 German Particle Verbs

German particle verbs (PVs) are productive compositions of a base verb (BV) and a prefix particle, whose part-of-speech varies between open-class nouns, adjectives, and verbs, and closed-class prepositions and adverbs. This work focuses on preposition particles.

Particle verb senses are assumed to be on a continuum between transparent (i.e. compositional) and opaque (i.e. non-compositional) with respect to their base verbs. For example, *abholen* 'fetch' is rather transparent with respect to its base verb *holen* 'fetch', *anfangen* 'begin' is quite opaque with respect to *fangen* 'catch', and *einsetzen* has both transparent (e.g. 'insert') and opaque (e.g. 'begin') verb senses with respect to *setzen* 'put/sit (down)'. Even though German particle verbs constitute a significant part of the verb lexicon, most work is devoted to theoretical investigations, such as (Stiebels, 1996; Lüdeling, 2001; Dehé et al., 2002). To our knowledge, so far only (Aldinger, 2004; Schulte im Walde, 2004; Schulte im Walde, 2005; Rehbein and van Genabith, 2006; Hartmann et al., 2008) have addressed German particle verbs from a corpus-based perspective.

This work addresses the degrees of compositionality of preposition particle verbs by clustering and then comparing the cluster memberships of the particle and base verbs. Clustering particle verbs and base verbs in turn requires the definition of empirical properties. This work relies on an automatic induction of distributional features from large-scale German corpus data, cf. Section 4.1. While inducing the distributional information is not difficult per se, German particle verbs face an empirical challenge: In general, subcategorisation properties are a powerful indicator of verb semantic relatedness and could thus point us towards the strength of relatedness between particle and base verbs (Dorr and Jones, 1996; Schulte im Walde, 2000; Korhonen et al., 2003; Schulte im Walde, 2006; Joanis et al., 2008, among others) because distributional similarity with respect to subcategorisation frames (even by themselves) corresponds to a large extent to semantic relatedness. German particle verbs are difficult to approach automatically at the syntax-semantics interface, however, because they typically change the subcategorisation behaviour in comparison to their base verbs. For example, even though *anlächeln* in example (1)[2] taken from Lüdeling (2001) is strongly compositional, its subcategorisation properties differ from those of its base verb; thus, automatic means that rely on subcategorisation cues might not recognise that *anlächeln* is semantically related to its base verb. Theoretical investigations (Stiebels, 1996) as well as corpus-based work (Aldinger, 2004) have demonstrated that such changes are quite regular, independent of whether a particle verb sense is compositional or not.

(1) Sie *lächelt*.
'She smiles.'
*Sie *lächelt* [$_{NP_{acc}}$ ihre Mutter].
'She smiles her mother.'
Sie *lächelt* [$_{NP_{acc}}$ ihre Mutter] *an*.
'She smiles her mother at.'

We believe that there are basically two strategies to address the empirically challenging class of multi-word expression from a semantic perspective: (i) avoid subcategorisation-based distributional features at the syntax-semantics interface, or (ii) incorporate the syntax-semantics subcategorisation transfer into the distributional information, cf. (Aldinger, 2004; Hartmann et al., 2008). This paper adheres to strategy (i) and basically excludes the notion of syntax from the distributional descriptions. For comparison reasons, we include an experiment that incorporates syntactic functions.

[2]Note that German particle verbs are separable, in contrast to the class of German prefix verbs that share many properties with the class of particle verbs but are inseparable (among other differences).

3 Clustering Approaches: LSC and PAC

Two soft clustering approaches were chosen to model the compositionality of German particle verbs, Latent Semantic Classes (LSC) and Predicate-Argument Clustering (PAC). Using soft clustering, each clustering object (i.e., the particle and base verbs) is assigned to each cluster with a probability between 0 and 1, and all probabilities for a certain verb over all clusters sum to 1. Cluster membership is then determined according to a probability threshold, cf. Section 4.2. In the following, we introduce the two clustering approaches.

3.1 Latent Semantic Classes

The Latent Semantic Class *(LSC)* approach is an instance of the Expectation-Maximisation (EM) algorithm (Baum, 1972) for unsupervised training on unannotated data, originally suggested by Mats Rooth (Rooth, 1998; Rooth et al., 1999). We use an implementation by Helmut Schmid. LSC cluster analyses define two-dimensional soft clusters which are able to generalise over hidden data. They model the selectional dependencies between two sets of words participating in a grammatical relationship. LSC training learns three probability distributions, one for the probabilities of the clusters, and a joint probability distribution for each lexical class participating in the grammatical relationship, with respect to cluster membership, thus the two dimensions. In our case, one dimension are the verbs (particle and base verbs), and one dimension are corpus-based features. Equation (2) provides the probability model for verb–feature pairs (v and f, respectively). Note that in our case the second dimension is crucial for the cluster analysis, but for determining the compositionality of the particle verbs, we only consider the cluster probabilities of dimension one, i.e., the particle and base verbs. Table 1 presents an example cluster that illustrates the verb and the feature dimensions, presenting the most probable verbs and direct object nouns within the cluster. The cluster is a nice example of compositional particle verbs (*verschicken, abschicken, zuschicken*) clustered together with their base verb (*schicken*).

$$p(v,f) = \sum_{c \in cluster} p(c,v,f) = \sum_{c \in cluster} p(c)\ p(v|c)\ p(f|c) \quad (2)$$

3.2 Predicate-Argument Clustering

Predicate-Argument Clustering (PAC) is an extension of the LSC approach that explicitly incorporates selectional preferences (Schulte im Walde et al., 2008). The PAC model provides a combination of the EM algorithm and the Minimum Description Length (MDL) principle (Rissanen, 1978), and refines the second dimension by explicit generalisations based on WordNet (Fellbaum, 1998) and the MDL principle. For example, instead of high probabilities of the nouns *Milch 'milk', Kaffee 'coffee', Tee 'tea'* within dimension two of a cluster, PAC might identify the generalising WordNet concept *Getränk 'beverage'*. Note that with PAC the second dimension only makes sense if WordNet provides useful generalisation information concerning that dimension, which effectively restricts the word class of the second dimension to nouns.

The PAC model is estimated through the joint probability of a verb v, a subcategorisation frame type f, and the complement realisations $n_1, ..., n_k$, cf. Equation (3). In addition to the LSC parameters in Equation (2), $p(r|c, f, i)$ is the probability that the ith complement of frame f in cluster c is realised by WordNet (wn) concept r, and $p(n|r)$ is the probability that the WordNet concept r is realised by the complement head n. Table 2 presents an example cluster where dimension two is a generalisation of WordNet concepts over PP arguments. Dimension one contains the most probable verbs in the cluster; dimension two is a selection of the most probable concepts from different hierarchical levels, plus example instances. As we are working on German data, we use the German Wordnet, i.e., *GermaNet* (Kunze, 2000).

$$p(v, f, n_1, ..., n_k) = \sum_c p(c)\ p(v|c)\ p(f|c) * \prod_{i=1}^{k} \sum_{r \in wn} p(r|c, f, i)\ p(n_i|r) \quad (3)$$

4 Clustering Experiments

To setup the clustering experiments, we need to specify the linguistic parameters (i.e., the choice of verbs and features), and the technical parameters, cf. Sections 4.1 and 4.2, respectively. The evaluation is described in Section 4.3.

dimension 1: verbs		*dimension 2: direct object nouns*	
schicken	'send'	Artikel	'article'
verschicken	'send'	Nachricht	'message'
versenden	'send'	E-Mail	'email'
nachweisen	'prove'	Brief	'letter'
überbringen	'deliver'	Kind	'child'
abonnieren	'subscribe to'	Kommentar	'comment'
zusenden	'send'	Newsletter	'newsletter'
downloaden	'download'	Bild	'picture'
bescheinigen	'attest'	Gruß	'greeting'
zustellen	'send'	Soldat	'soldier'
abschicken	'send off'	Foto	'photo'
zuschicken	'send'	Information	'information'

Table 1: Example LSC cluster.

dimension 1: verbs		*dimension 2: WN concepts over PP arguments*	
steigen	'increase'	Maßeinheit	'measuring unit'
zurückgehen	'decrease'	e.g., Jahresende	'end of year'
geben	'give'	Geldeinheit	'monetary unit'
rechnen	'calculate'	e.g., Euro	'Euro'
wachsen	'grow'	Transportmittel	'means of transportation'
ansteigen	'increase'	e.g., Fahrzeug	'automobile'
belaufen	'amount to'	Gebäudeteil	'part of building'
gehen	'go'	e.g., Dach	'roof'
zulegen	'add'	materieller Besitz	'material property'
anheben	'increase'	e.g., Haushalt	'budget'
kürzen	'reduce'	Besitzwechsel	'transfer of property'
stehen	'stagnate'	e.g., Zuschuss	'subsidy'

Table 2: Example PAC cluster.

4.1 Data

As corpus data basis, we relied on approx. 560 million words from the German web corpus *deWaC* (Baroni and Kilgarriff, 2006), after the corpus was preprocessed by the Tree Tagger (Schmid, 1994) and by a dependency parser (Schiehlen, 2003). The corpus portion contains more than 50,000 verb types (from verb-first, verb-second and verb-final clauses), which we restricted to those with a frequency above 1,000 and below 10,000, to avoid very low and very high frequent types, as they notoriously produce noise in clustering. In addition, we made sure that all verbs needed in the evaluation were covered, ending up with 2,152 verb types (comprising both particle and base verbs). The latter step, however, included some low and high frequent verbs, as many particle verbs are low frequent, and many base verbs are high frequent.

Concerning the feature choice, we relied on the main verb argument types, covering subjects, direct objects and pp objects. I.e., we used as input verb–noun pairs where the nouns were (a) subjects, or (b) objects, or (c) pp objects of the verbs. We used the information separately and also (d) merged without reference to the syntactic function, as we largely ignored syntax. The underlying assumption for this rather crude simplification refers to the observation that the selectional preferences of particle verbs overlap with those of semantically similar verbs, but not necessarily in identical syntactic functions, cf. Schulte im Walde (2004). In comparison to (d), we (e) merged the pairs, while keeping the reference to the syntactic functions. The feature choice –more specifically: comparing (d) with (e)– is based on that in Schulte im Walde (2005). We wanted to compare the individual argument types with re-

spect to their potential in addressing particle verb compositionality despite the syntax transfer hurdle. As direct objects and pp objects often remain the same function after the syntax-semantics particle–base transfer, they were supposed to provide more interesting results than subjects, which often fulfil more general roles. In addition, the syntax-unmarked input was supposed to provide better results than the syntax-marked input, because of the syntax transfer hurdle. The input variants are referred to as (a) *subj*, (b) *obj*, (c) *pp*, (d) *n-syntax*, and (e) *n+syntax*. Table 3 lists the number of input tokens and types according to the feature choices.

input	tokens	types
subj	2,316,757	368,667
obj	3,532,572	446,947
pp	4,144,588	706,377
n+syntax	9,993,917	1,346,093
n-syntax	9,993,917	1,036,282

Table 3: Input data.

4.2 Method

The data were used for both LSC and PAC, with minor formatting differences. There are basically two input dimensions (verb and argument head) as described in Section 3. When including the function markers, they were added to the (second) noun dimension, e.g., *anfangen–Job* 'begin–job' would become *anfangen–obj:Job*.

As we wanted to explore the clustering potential with respect to various parameters, we varied the number of clusters: 20, 50, 100, and 200. In addition, we varied the probability to determine cluster membership: 0.01, 0.001, 0.0005, and 0.0001, thus directly influencing precision and recall, as higher probability thresholds include less verbs per cluster. All cluster analyses were trained over 200 iterations for LSC and 100 iterations for PAC, evaluating the results after 50, 100 (and 200) iterations.

4.3 Evaluation

For the evaluation of the experiments, we relied on a gold standard created by Hartmann (2008). She had collected compositionality judgements for 99 German particle verbs across 11 different preposition particles, and across 8 frequency bands (5, 10, 18, 30, 55, 110, 300, 10,000) plus one manually chosen verb per particle (to make sure that interesting ambiguous verbs were included). The frequency bands had been determined such that there were approximately equally many particle verbs in each range.

Four independent judges had rated the compositionality of the 99 particle verbs between 1 (*completely opaque*) and 10 (*completely compositional*). The inter-rater agreement was significantly high ($W = 0.7548, \chi^2 = 274.65, df = 91, \alpha = 0.000001$), according to Kendall's coefficient of concordance. The average ratings of the judges per particle verb are considered as the gold standard scores for our experiments. Table 4 presents a selection of the average scores for particle verbs with different degrees of compositionality. Note that there are ambiguous particle verbs, whose scores are the average values of the compositionality scores for the different meanings.

particle verb		score
nachdrucken	'reprint'	9.250
aufhängen	'hang up'	8.500
ausschneiden	'cut out'	8.250
vorgehen	'go ahead' 'approach'	6.875
abwaschen	'do the dishes'	6.500
abschließen	'close' 'finalise'	6.000
nachweisen	'prove'	5.000
anklagen	'accuse'	4.500
zutrauen	'feel confident'	3.250
umbringen	'kill'	1.625

Table 4: Gold standard judgements.

The evaluation itself was performed as follows. For each cluster analysis and each probability threshold t, we calculated for each particle verb from the gold standard the proportion of how often it appeared in a cluster together with its base verb, in relation to the total number of appearances, cf. Equation (4). The ranked list of the cluster-based compositionality judgements was then compared against the ranked list of gold standard judgements, according to the Spearman rank-order correlation coefficient. This correlation is a non-parametric statistical test that measures the association between two variables that are ranked in two ordered series.

$$(4)\ comp_{pv} = \frac{\sum_c p(pv,c) \geq t \ \wedge \ p(bv,c) \geq t}{\sum_c p(pv,c) \geq t}$$

The collection of the gold standard and the evaluation procedure were performed according to a comparable evaluation task for English particle verb compositionality in McCarthy et al. (2003). The parametric tests are described in Siegel and Castellan (1988).

5 Results

The correlation scores differ substantially according to the linguistic features and the parameters of the cluster analyses. Furthermore, the probability threshold that determined cluster membership directly influenced the number of particle verbs that were included in the evaluation at all. We focus on presenting the overall best results per feature (group) in Tables 5 and 6 for LSC and PAC, respectively, and comment on the overall patterns. The tables show

- the Spearman rank-order correlation coefficient (*corr*),
- the coverage (*cov*), i.e., the proportion of gold standard verbs included in the evaluation after applying the probability threshold,
- the *f-score* (F_1) of the correlation and coverage values as usually applied to precision and recall; it indicates a compromise between the correlation and the coverage, cf. Equation (5),
- the number of clusters,
- the number of iterations, and
- the membership threshold

of the best results.

$$f\text{-}score = \frac{2 * corr * cov}{corr + cov} \quad (5)$$

5.1 Technical Parameters

The best results per feature (group) as listed in the tables are reached with different numbers of clusters (ranging from 20 to 200); with LSC, the best results are obtained after all (i.e., 200) training iterations; with PAC, the best results are obtained sometimes after 50, sometimes after 100 iterations. So in the tables (and in general), there is no clear tendency towards an optimal number of clusters with respect to our task; concerning the optimal number of training iterations, LSC seems to profit most from the largest possible number of iterations (so it might be worth testing even more training iterations than 200), and PAC does not seem to have a strong preference.

The optimal probability threshold for cluster membership is difficult to judge about, as that value strongly depends on a preference for correlation vs. coverage. The lower the threshold, the more particle verbs are included in the clusters, so the recall (coverage) increases while the precision (correlation) decreases. The tables list the best results according to the f-score, but if one wanted to use the cluster analyses within an application that incorporates particle verb compositionality values, one would have to determine a favour for precision vs. recall, to identify the appropriate threshold. The best correlation results with an acceptable coverage of 50-60% go up to .433 (LSC, obj), and .236 (PAC, n-syntax). In general, the coverage is approx. 10-30% for a threshold of 0.01, 30-60% for a threshold of 0.001, 40-70% for a threshold of 0.0005, and 50-80% for a threshold of 0.0001.

Overall, the best f-score values go up to .499 for LSC and .327 for PAC, and the PAC results are in general considerably below the LSC results. The lowest f-scores go down to zero for both clustering approaches, and sometimes even reach negative values, indicating a negative correlation. In sum, our methods reach moderate correlation values, and considering that we have worked with very simple distributional features that ignored other than some basic information at the syntax-semantics interface, we regard this a reasonable result. The dependency of the correlation scores on the clustering parameters, however, remains largely unclear.

5.2 Linguistic Parameters

Concerning the linguistic features in the clustering, the picture differs with respect to LSC vs. PAC. With LSC, direct object and pp object information is obviously valuable in comparing particle verbs with base verbs, despite the transfer at the syntax-semantics interface, while subject information is not very helpful, as expected. Comparing the unions of syntactic functions with the individual functions, LSC profits more from the individual functions, while PAC profits more from the unions. In both approaches, the unmarked *n-syntax* condition outperforms the marked *n+syntax* condition, as expected, but the difference is not impressive.

Comparing LSC and PAC, we can identify various reasons for why the PAC results are considerably below the LSC results: (i) the dependency of selectional preferences on the subcategorisation frames that represents a strength of PAC, does not play an important role in our task (rather, the ref-

input	best result corr	cov	f-score	analysis clusters	iter	membership threshold
obj	.433	.59	.499	100	200	.0005
subj	.205	.76	.323	50	200	.0001
pp	.498	.40	.444	20	200	.0005
n+syntax	.303	.54	.388	50	200	.0005
n-syntax	.336	.56	.420	100	200	.001

Table 5: LSC results.

input	best result corr	cov	f-score	analysis clusters	iter	membership threshold
obj	.100	.53	.168	100	50	.0005
subj	.783	.05	.094	20	50	.01
pp	.275	.21	.238	200	100	.01
n+syntax	.213	.61	.316	20	100	.0001
n-syntax	.236	.53	.327	200	100	.001

Table 6: PAC results.

erence to syntactic functions is supposed to have a negative influence on the prediction of compositionality, cf. Section 2); (ii) the high frequency (base) verbs included in the training data have a negative impact on cluster composition, i.e., many clusters created by PAC are dominated by few high-frequency verbs, which is sub-optimal in general but in our case has the additional effect that many compositionality predictions are 1 because it is very likely that for a specific particle verb also the base verb is in the cluster; (iii) the generalising property of PAC that would have been expected to help with the sparse data of the lexical heads, does not improve the LSC results but rather makes them worse.

Tables 7 and 8 present compositionality scores from the best LSC and the best PAC cluster analyses (cf. Tables 5 and 6), and relates them to the gold standard (gs) scores repeated from Table 4. Furthermore, the number of clusters in which the particle verb (pv), the respective base verb (bv) and both appeared, is given. While the LSC system scores are of course not perfect, we can see that there is a clear tendency towards higher overlap scores in the top half of the table, in comparison to the bottom half, even though the number of clusters the particle verbs appear in differ strongly. The only particle verb that clearly is not able to subcategorise a direct object (i.e., *vorgehen* in both of its senses) is also a clear outlier in the quality of predicting the compositionality. In comparison to the LSC results, the PAC system scores are obviously worse, the main reason being that the high frequency base verbs appear in many of the 200 clusters, especially *gehen* and *bringen*.

In sum, the optimal clustering setup to predict particle verb compositionality (with respect to the best results in the tables, but also in more general) seems to use LSC with direct object or pp object information. On the one hand, the preference for these functions is intuitive (as many particle verbs as well as their base verbs are transitive verbs, e.g., *anbauen 'build, attach', nachdrucken 'reprint', umbringen 'kill'*), but on the other hand the gold standard also includes many intransitive particle verbs (e.g., *aufatmen 'breathe', durchstarten 'touch and go', überschäumen 'foam over'*) where at least direct objects intuitively cannot help with a compositionality rating.

5.3 Comparison with Related Work

McCarthy et al. (2003) predicted the degree of compositionality of English particle verbs. Their work is probably most closely related to our approach, and we adapted their evaluation method. Their prediction relies on nearest neighbourhood, assuming that the neighbours of particle verbs should be similar to the neighbours of the respective base verbs. The definition of neighbourhood is based on Lin's thesaurus (Lin, 1998), and various statistical measures for distributional similarity. The best result they

particle verb		#clusters			score	
		pv	bv	both	gs	system
nachdrucken	'reprint'	2	5	1	9.250	0.500
aufhängen	'hang up'	4	18	4	8.500	1.000
ausschneiden	'cut out'	5	3	3	8.250	0.600
vorgehen	'go ahead' 'approach'	5	18	1	6.875	0.200
abwaschen	'do the dishes'	1	4	1	6.500	1.000
abschließen	'close' 'finalise'	2	2	1	6.000	0.500
nachweisen	'prove'	16	20	5	5.000	0.313
anklagen	'accuse'	5	8	1	4.500	0.200
zutrauen	'feel confident'	12	4	1	3.250	0.083
umbringen	'kill'	2	2	0	1.625	0.000

Table 7: LSC gold standard judgements and system scores.

particle verb		#clusters			score	
		pv	bv	both	gs	system
nachdrucken	'reprint'	0	13	0	9.250	–
aufhängen	'hang up'	3	66	3	8.500	1.000
ausschneiden	'cut out'	3	10	3	8.250	1.000
vorgehen	'go ahead' 'approach'	47	194	47	6.875	1.000
abwaschen	'do the dishes'	1	9	1	6.500	1.000
abschließen	'close' 'finalise'	63	98	48	6.000	0.762
nachweisen	'prove'	66	56	24	5.000	0.364
anklagen	'accuse'	11	35	5	4.500	0.455
zutrauen	'feel confident'	7	7	0	3.250	0.000
umbringen	'kill'	11	190	11	1.625	1.000

Table 8: PAC gold standard judgements and system scores.

achieve is a Spearman rank correlation of 0.490, which is slightly but not considerably better than our best results.

Concerning the feature choice to describe and compare German particle verbs and their base verbs (more specifically: comparing the unmarked *n-syntax* with the marked *n+syntax*), we can compare our results with previous work by Schulte im Walde (2005). Our work confirms her insight that the differences between the two versions (with vs. without reference to the syntactic functions) are visible but minimal.

6 Conclusions

This work determined the degree of compositionality of German particle verbs by two soft clustering approaches. We assumed that the more compositional a particle verb is, the more often it appears in the same cluster with its base verb, after applying a probability threshold to establish cluster membership. The overall best cluster analysis was reached by the simpler cluster approach, LSC. It could predict the degree of compositionality for 59% of the particle verbs; the correlation with the gold standard judgements was .433. Considering that we have worked with very simple distributional features that ignored other than some basic information at the syntax-semantics interface, we regard this a reasonable result. We expect that if we extended our work by incorporating the syntax-semantics transfer between particle and base verbs, we could improve on the compositionality judgements.

References

Nadine Aldinger. 2004. Towards a Dynamic Lexicon: Predicting the Syntactic Argument Structure of Complex Verbs. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Marco Baroni and Adam Kilgarriff. 2006. Large Linguistically-processed Web Corpora for Multiple Languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Leonard E. Baum. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities*, III:1–8.

Nicole Dehé, Ray Jackendoff, Andrew McIntyre, and Silke Urban, editors. 2002. *Verb-Particle Explorations*. Number 1 in Interface Explorations. Mouton de Gruyter, Berlin.

Bonnie J. Dorr and Doug Jones. 1996. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 322–327, Copenhagen, Denmark.

Afsaneh Fazly and Suzanne Stevenson. 2008. A Distributional Account of the Semantics of Multiword Expressions. *Italian Journal of Linguistics. Alessandro Lenci (guest editor): "From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science"*, 20(1).

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press.

Silvana Hartmann, Sabine Schulte im Walde, and Hans Kamp. 2008. Predicting the Degree of Compositionality of German Particle Verbs based on Empirical Syntactic and Semantic Subcategorisation Transfer Patterns. Talk at the Konvens Workshop 'Lexical-Semantic and Ontological Resources'.

Silvana Hartmann. 2008. Einfluss syntaktischer und semantischer Subkategorisierung auf die Kompositionalität von Partikelverben. Studienarbeit. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A General Feature Space for Automatic Verb Classification. *Natural Language Engineering*, 14(3):337–367.

Stanislav Kavka. 2009. Compounding and Idiomatology. In Lieber and Stekauer (2009b), chapter 2, pages 19–33.

Anna Korhonen, Yuval Krymolowski, and Zvika Marx. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Sapporo, Japan.

Natalie Kühner. 2010. Automatische Bestimmung der Kompositionalität von deutschen Partikelverben auf der Basis von Cluster-Modellen: Vergleich von LSC und PAC. Studienarbeit. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Claudia Kunze. 2000. Extension and Use of GermaNet, a Lexical-Semantic Database. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 999–1002, Athens, Greece.

Rochelle Lieber and Pavol Stekauer. 2009a. Introduction: Status and Definition of Compounding. In *The Oxford Handbook on Compounding* (Lieber and Stekauer, 2009b).

Rochelle Lieber and Pavol Stekauer, editors. 2009b. *The Oxford Handbook of Compounding*. Oxford University Press.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, Canada.

Anke Lüdeling. 2001. *On German Particle Verbs and Similar Constructions in German*. Dissertations in Linguistics. CSLI Publications.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Diana McCarthy, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting Compositionality of Verb-Object Combinations using Selectional Preferences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 369–379.

Ines Rehbein and Josef van Genabith. 2006. German Particle Verbs and Pleonastic Prepositions. In *Proceedings of the 3rd ACL-SIGSEM Workshop on Prepositions*, pages 57–64, Trento, Italy.

Jorma Rissanen. 1978. Modeling by Shortest Data Description. *Automatica*, 14:465–471.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, MD.

Mats Rooth. 1998. Two-Dimensional Clusters in Grammatical Relations. In *Inducing Lexicons with the EM Algorithm*, AIMS Report 4(3). Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico.

Michael Schiehlen. 2003. A Cascaded Finite-State Parser for German. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–166, Budapest, Hungary.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing*.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. 2008. Combining EM Training and the MDL Principle for an Automatic Verb Classification incorporating Selectional Preferences. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 496–504, Columbus, OH.

Sabine Schulte im Walde. 2000. Clustering Verbs Semantically According to their Alternation Behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 747–753, Saarbrücken, Germany.

Sabine Schulte im Walde. 2004. Identification, Quantitative Description, and Preliminary Distributional Analysis of German Particle Verbs. In *Proceedings of the COLING Workshop on Enhancing and Using Electronic Dictionaries*, pages 85–88, Geneva, Switzerland.

Sabine Schulte im Walde. 2005. Exploring Features to Identify Semantic Nearest Neighbours: A Case Study on German Particle Verbs. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 608–614, Borovets, Bulgaria.

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

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA.

Barbara Stiebels. 1996. *Lexikalische Argumente und Adjunkte. Zum semantischen Beitrag von verbalen Präfixen und Partikeln*. Akademie Verlag.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the Special Issue on Multiword Expressions: Having a Crack at a Hard Nut. *Computer Speech and Language*, 19:365–377.