Syntactic Transfer Patterns of German Particle Verbs and their Impact on Lexical Semantics

Stefan Bott **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverabeitung
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{stefan.bott,schulte}@ims.uni-stuttgart.de

Abstract

German particle verbs, like *anblicken* (*to gaze at*) combine a base verb (*blicken*) with a particle (*an*) to form a special kind of Multi Word Expression. Particle verbs may share the semantics of the base verb and the particle to a variable degree. However, while syntactic subcategorization frames tend to be good predictor for the semantics of verbs in general (verbs that are similar in meaning also tend to have similar subcategorization frames and selectional preferences), there are regular changes in subcategorization frames by particle verbs with regard to the corresponding base verbs. This paper demonstrates that the syntactic behavior of particle verbs and base verbs together (modeling regular changes in subcategorization frames by particle verbs and corresponding base verbs) and applying clustering techniques allows us to distinguish particle verb meaning and shows the tight connection between transfer patterns and the semantic classes of particle verbs.

1 Introduction

In German, particle verbs (PVs), like *anblicken* in (1), are a highly productive class. PVs present challenges for a both theoretical analysis and their computational treatment. One of the central problems is the prediction of their meaning from their constituent parts: the base verb (BV, e.g. *blicken* in (1)) and the particle (e.g. *an*). Many PVs derive their meaning from the corresponding BVs – with a varying degree of transparency. It is often not clear, however, how to interpret the semantics of the particles and their contribution to the meaning of the PVs. Since particles never occur isolated, without the context of the verb, it is difficult to assign them a lexical semantic entry on their own. Even more, German particles are a notoriously ambiguous word class.

(1) Das Kind blickt seine Mutter an.
The child gazes his-acc mother PRT.
The child looks at his mother.

One way to approximate the meaning of particles is to group together the particle verbs which share the same particle into semantic groups (such as *anblicken, anstarren, anschauen* 'to stare/look at'), such that both the meaning of the PV and the meaning of the BV is similar in each group. This allows us to make inferences like "taking a BV from semantic group α and particle β, we will derive a PV from semantic group δ". Such groups can be established and they represent productive paradigms. Springorum et al. (2013) have shown in a generation experiment setup that subjects are able to associate a meaning to artificially created, previously unattested PVs and to construct example sentences for them.[1] Different subjects also agree to a large degree on the meaning they attribute to the newly formed lexical items.

But this approach also rises a series of questions, especially concerning the way in which such groups can be distinguished, both from a theoretical and a corpus-based perspective. For example, which kinds of linguistic features allow us to discriminate such semantic classes? In this paper we investigate the influence of syntax, which represents one of the possible feature sources. Syn-

[1] For example for the neologism *anlauschen*, referring to a partitive meaning of the particle, senentence like the following could be found: *Er hatte an der Wand angelauscht und wusste Bescheid.* ('He had listened at the wall and knew everything.')

*Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 182–192,
Dublin, Ireland, August 23-24 2014.

tactic subcategorization frames tend to be good predictors for the semantics of verbs in general: verbs that are similar in meaning also tend to have similar subcategorization frames and selectional preferences (Schulte im Walde, 2000; Merlo and Stevenson, 2001; Korhonen et al., 2003; Schulte im Walde, 2006a; Joanis et al., 2008). But, as we will show below, PV-BV pairs tend to have a special behavior with respect to their subcategorization, even if their meanings are closely related. Because we are interested in pairs of PVs and their BVs, we thus have to look at pairs of subcategorization preferences, and rely on the concept of *syntactic transfer*. We use *syntactic transfer* as a technical term here, which we define as regular changes in subcategorization frames by PVs and corresponding BVs, e.g., the incorporation or addition of complements of PVs in comparison to their BVs (Stiebels, 1996; Lüdeling, 2001; Fleischer and Barz, 2012a). We claim that the syntactic behavior of PVs and BVs together allows us to distinguish semantic classes.

A better understanding of the nature of the connection between syntactic transfer patterns and semantic classes may be beneficial for both theoretical and computational linguistics. On the theoretical side we can hope to find new arguments to guide and justify lexical semantic classifications. We may also shed light on what particles actually mean, a topic which is not trivial by itself. In computational semantics, a better understanding of syntactic transfer patterns can potentially contribute to a better treatment of PVs in meaning-related areas, such as machine translation and information retrieval.

In sum, this paper makes the following contributions:

- We show that the meaning of verb particles can be modeled as classes of pairs of PVs and their corresponding BVs, where both PVs and BVs in each class are closely related in meaning. In addition, the PV-BV pairs in each class undergo the same syntactic transfers, i.e. the selectional preferences of PV-BV pairs within each class tend to be very similar, even if the subcategorization preferences may be different between PVs and BVs.
- We show that automatic clustering can replicate a gold standard classification of PV-BV pairs to a large degree when clustering only relies on syntax and the gold standard reflects semantic regularities.

The rest of this paper is organized as follows: In section 2 we describe the task and our goals. Here we also define the term *syntactic transfer pattern*, which is central to our discussion. Section 3 is dedicated to related work relevant for our study. In section 4 we describe the experimental setup, while sections 5 and 6 present the experiment results and discuss them.

2 Goal and Motivation

The work we describe here centers around the concept of semantic classes and syntactic transfer patterns. As concerning the semantic side, the PVs which share the same particle may be grouped into different classes according to their meaning. For example, among the PVs incorporating the particle *an* we find a group of verbs whose meanings center around the concept of "*to look at someone/something in manner X*", "*to attach something somewhere in manner X*", "*to make an unpleasant sound towards someone in a manner X*" and "*to start an action X on something which starts consuming it*", as exemplified in (2) a-d.

(2)
a. A blickt/schaut/starrt/stiert/ B an.
A looks/stares/gazes B PRT.
A looks/stares/gazes at B.

b. A klebt/heftet/schraubt/nagelt B
A glues/affixes/screws B
an C an.
at/onto C PRT.
A glues/affixes/screws B onto C.

c. A brüllt/faucht/bellt/meckert B an.
A roars/hisses/bleats B PRT.
A brawls/hisses/scolds at B.

d. A schneidet/bricht/reißt B an.
A cuts/breaks/tears B PRT.
A cuts/breaks/tears the first slice/piece of B.

Such semantic classes are not easy to define and they are also difficult to induce automatically. Although there is general agreement in the theoretical literature that such semantic classes for PVs exist (cf. Lechler and Roßdeutscher (2009), Kliche (2011) and Springorum (2011)) the agreement on the number and nature of such classes is not very high. For example, Springorum (2011) (who develops her analysis within Discourse Rep-

resentation Theory (Kamp and Reyle, 1993)) distinguishes between 11 classes of PVs with the particle *an*, while Fleischer and Barz (2012b) only distinguish 3 major de-verbal classes, based on their *aktionsart*, which can be divided into some 9 minor classes.[2] It should be noted that all the PVs and BVs in (2) a-d are not only quite homogeneous in their *semantics*; they also form coherent *syntactic* classes. The PVs and BVs of these examples are quite similar in the way they typically select their syntactic complements. For example, the BVs of (2-a) typically take a PP argument that expresses the direction of gaze using a prepositional phrases with one of the prepositions *auf*, *zu*, *nach* or *in* subcategorizing a dative noun phrase. The corresponding PVs, however, typically express this semantic role by an accusative object. The type of change from the typical frame of a BV to the typical frame of a PV is an example of what we mean by a *syntactic transfer pattern*.

So, while similar syntactic behavior of two verbs in general may indicate that the verbs are also semantically similar, this is typically *not* the case for PV-BV pairs. Compare (1) to (3), which are nearly synonymous but (3) uses the BV *blicken* instead of the PV *anblicken* in (1). We can only induce the similarity of the PV and the BV if we take the syntactic transfer into consideration.

(3) a. Das Kind blickt zu seiner Mutter.
The child looks at his-dat mother.

b. Das Kind stiert/starrt/schaut zu
The child stares/stares/looks at
seiner Mutter.
his-dat Mother.

Looking at the class to which this PV belongs, all the variants of (3-b) are semantically very similar to (3-a). This also corresponds to a syntactic similarity: all the verbs of this group share the same preferred syntactic subcategorization frames. The dominant frame of theses verbs is "NPnom+PP-dat" (the head preposition of the PP may vary, but within well-defined limits). But this is not the case for the PV *anblicken* in (1). (1) is nearly synonymous to (3-a), but the PV in this example has a totally different frame, namely the simple transitive "NPnom+NP-acc". It may not come as a surprise that all of the verbs in (3-b) have PV counterparts (*anstieren*, *anstarren*, etc.), which all behave syntactically like *anblicken*.

In sum, we part from the hypothesis that there is a tight connection between transfer patterns and the semantic classes of PVs. There is only one more point to make: the classes shown in (2), could actually be seen as reflecting different meanings of the particle *an* itself.

3 Related Work

Particle verbs have been studied from the theoretical perspective and, to a more limited extend, from the aspect of the computational predictability of the degree of semantic compositionality (the transparency of their meaning with respect to the meaning of the base verb and the particle) and the semantic classifiabilty of PVs.

For English, there is work on the automatic extraction of PVs from corpora (Baldwin and Villavicencio, 2002; Baldwin, 2005; Villavicencio, 2005) and the determination of compositionality (McCarthy et al., 2003; Baldwin et al., 2003; Bannard, 2005).

To the best of our knowledge Aldinger (2004) is the first work that studies German PVs from a corpus based perspective, with an emphasis on the syntactic behavior and syntactic change. Schulte im Walde (2004), Schulte im Walde (2005) and Schulte im Walde (2006b) present preliminary distributional studies to explore salient features at the syntax-semantics interface that determine the semantic nearest neighbours of German PVs. Relying on the insights of those studies, Schulte im Walde (2006b) and Hartmann (2008) describe experiments which model the subcategorization transfer of German PVs with respect to their BVs in order to strengthen PV-BV distributional similarity. The main goal for them is to use transfer information in order to predict the degree of semantic compositionality of PVs. Kühner and Schulte im Walde (2010) use clustering to determine the degree of compositionality of German PVs, via common PV-BV cluster membership. They are, again, mainly interested in the assessment of compositionality, which is done on the basis of lexical information. They use syntactic information, but only as a filter and for lexical heads as cooccurrence features in order to limit the selected argument slots to certain syntactic functions. They conclude that the best results can be obtained with information stemming from direct objects and PP-objects. The incorporation of syntactic informa-

[2]The subdivision is, however not fully spelled out and only implicit in their description.

tion in the form of dependency arc labels (concatenated with the head nouns) does not yield satisfactory results, putting the syntactic transfer problem in evidence, again. They conclude that an incorporation of syntactic transfer information between BVs and PVs could possibly improve the results.

Based on a theoretical study (Springorum, 2011), which explains particle meanings in terms of Discourse Representation Theory (Kamp and Reyle, 1993), Springorum et al. (2012) show that four classes of PVs with the particle *an* can be classified automatically. They take a supervised approach using decision trees. The use of decision trees also allows them to manually inspect and analyze the decisions made by the classifier. As predictive features they use the head nouns of objects, generalized classes of these nouns and PP types.

The approach we take here is not fully comparable to any of the former approaches, since we try to derive a semantic classification BV-PP *pairs* in an unsupervised manner and we only use syntactic features, stemming from corpus instances of both the BVs and the PVs. In other words, we do not attempt to classify PVs, but we try to classify syntactic transfers and, by doing so, we identify syntactic transfer patterns which we hypothesize to have a close relation to semantic PV classes and the semantics of the particles.

4 Experimental Setup

4.1 Gold Standard Classification

For testing our hypothesis, we created a gold standard of 32 PVs, including 14 with the particle *an* and 18 with the particle *auf*. We concentrated on two particles here in order to have a small and controlled test bed which allows us to study the syntactic transfers.

We based the creation of the gold standard on the classification by Fleischer and Barz (2012b), but we further distinguished the classes based on the meanings of the BVs. For example, we grouped all the BVs with the meaning of '*looking in a manner X*' or '*tying X to Y in a manner Z*'. From these classes we selected those which had a clear subcategorization pattern for both the BVs and the PVs. We discarded such PVs where either the PV itself or its underlying BV was clearly ambiguous. The full gold standard can be seen in table 2. The table also lists the expected dominant subcategorization frames for the BVs and PVs of each category.

While the gold standard was based on theoretic considerations, we expected it to correlate with human intuitions. To test this, we presented the gold standard verbs to 6 human raters. These raters were all German native speakers with working practice in various areas of linguistics or language didactics. The raters were not directly asked to group PVs into categories. Instead the PVs were presented in pairs[3] and raters had to make a decision on whether or not the pairs belong to the same semantic category (even if they could not think of a name or description of that category). No pre-defined categories were given, nor were raters asked to provide a name or description for these categories. The annotators were asked to take the similarity of the BVs and the similarity of the PVs into consideration for their judgements. In order to avoid possible bias, the verbs were presented without given context. What is important here is that we did *not* ask them to take any syntactic criterion into consideration, the criterion we used for the initial compilation of the gold standard.

The inter-annotator agreement was substantial with a Fleiss' Kappa score of 0.68 (Fleiss, 1971).[4] As a measure of agreement between raters and the previously created gold standard, we performed pair-wise calculations between the ratings of each annotator and the gold standard. For the comparison, the gold standard was transformed into PV pairs and the value *true* was assigned if the two verbs of a pairs belonged to the same category, and *false* otherwise. We calculated the Kappa scores for each annotator and took the average of the agreement scores. Table 1 resumes the comparison. Values are given for the parts of the gold standard corresponding to PVs with *an* and *auf* separately and also for the gold standard as a whole.

It can clearly be seen that humans agreement with the gold standard is as high as the agreement among different annotators. This shows that the gold standard used here is a valid representation of human language intuition. Most importantly, the annotators did not use syntactic criteria

[3] All possible PV combinations were generated, but the PVs with *an* were kept separate from those with *auf* in order to avoid an unnecessary explosion of the number of pairs to be rated.

[4] One of the 6 raters showed less agreement with the other raters. If we eliminate this rater from the calculation of agreement, we achieve an even higher Kappa score of 0.76 and also agreement scores with the gold standard improved. Two of the annotators even achieved Kappa scores of over 0.80 when compared to the gold standard.

and still validated a gold standard whose creation was explicitly based on syntactic subcategorization frames. In other words: there is an apparent tight interrelation between syntax and semantics for PVs, at least in the sense that semantic distinctions can be used to predict different syntactic behaviour. The inverse case - predicting semantic classes from syntactic information - will be discussed below.

4.2 Corpus Data

We used a lemmatized and tagged version of the SdeWaC corpus (Faaß and Eckart, 2013), a web corpus of 880 million words. For linguistic preprocessing we used the MATE parser (Bohnet, 2010), which allowed us to extract syntactic subcategorization frames.

4.3 Feature Selection

For each PV-BV pair we extracted two parallel sets of features, one pertaining to the BV and one for the PV. This allows us to model the syntactic transfer. For example, we expected that an ideal transfer from a group of transitive BVs to a group of intransitive PVs should be reflected in high values for the features BV:transitive and PV:intransitive[5] and, in turn, low values for BV:intransitive and PV:transitive.

We had two ways of selecting the feature types: manually and automatically. For the manual feature selection we extracted only those features from the parsed frames which we already used in the creation of the gold standard and which are listed in table 2. This resulted in a small feature set of 30 features (15 features for PVs and BVs, respectively). For the automatic feature selection we simply used the *n* most frequent frames which could be observed in the corpus for the set of verbs of the gold standard.

From the syntactic dependency representation provided by the parser, we excluded subjects and modifiers (except for PP-modifiers) in the representation of subcat frames. We did not use information on subjects, because in German all verbs have subjects, which may be implicit in the case of subordinate clauses. We found that for this reason that with the representation of subjects in the extracted features no relevant information was gained, but some distortion was introduced. Modifiers in the MATE parser represent information which is too general to be good predictors. Based on theoretical considerations on the best lexicographic representation of verbs, we included PP-modifiers, however, because quantitative information on PP-adjuncts has proven successful next to that of PP-arguments (Schulte im Walde, 2006a; Joanis et al., 2008), and in addition the parser often distinguishes poorly between PP-modifiers and PP-arguments.

In order to create an idealized artificial upper bound, we also created a set of idealized "lexicographic" descriptions in the form of manually instantiated feature vectors and feature values, using the manually selected feature configuration we just described (and ultimately based on the gold standard description represented by table 2). These idealized vectors were also used for clustering experiments in order to estimate an upper bound.

4.4 Clustering Methods

For the clustering experiments we used two different clustering algorithms: K-means and Latent Semantic Classes (LSC). K-means is a standard flat, hard-clustering algorithm; we used the Weka implementation (Witten and Frank, 2005). LSC (Rooth, 1998; Rooth et al., 1999) is a two-dimensional soft-clustering algorithm which learns three probability distributions: one for the clusters, and one for the output probabilities of each element and for each feature type with regard to a cluster. The latter two (elements and features) correspond to the two dimensions of the clustering. In our case the elements are the PV-BV pairs, and the features are normalized counts of the subcategorization frames.

4.5 Evaluation

Our feature vectors are a combination of the feature vector for the BV and the feature vector for the PV of each PV-BV pair. Since the length of each vector depends on the base frequency of each verb we need to apply a feature normalization: we simply reduce each feature to its unit vector of length 1. Because the frequency ratio between BV and PV may vary strongly, we need to normalize PV vectors and BV vectors separately before they can be combined.

The vector combination for each PV-BV pair is done by simply adding the dimensions (and not the

[5]Note that *transitive* and *transitive* are only convenient abreviations for the labels *NPnom* and *NPnom+NPacc*, which are used in table2.

	an	auf	an+auf
Inter-annotator agreement	0.79	0.64	0.70
Average agreement between annotators and gold standard	0.73	0.74	0.73

Table 1: Inter-annotator agreement and comparison of the gold standard to the ratings of 6 human annotators (Fleiss' Kappa Scores).

Particle	Typical frames for the BV	Typical frames for the PV	Semantic Class	Verbs in Class	
an	NPnom +NPacc +PP-an	NPnom +NPacc +PP-an	locative/ relational tying	an|binden an|ketten	to tie at to chain at
	NPnom +PP-zu/in/ nach/auf	NPnom +NPacc	locative/ relational gaze	an|blicken an|gucken an|starren	to glance at to look at to stare at
	NPnom +NPacc +PP-mit	NPnom +NPacc +PP-mit	ingressive consump- tion	an|brechen an|reißen an|schneiden	start to break start to tear start to cut
	NPnom	NPnom +NPacc	locative/ relational sound	an|brüllen an|fauchen an|meckern	to roar at to hiss at to bleat at
	NPnom +NPacc +PP-an	NPnom +NPacc	locative/ relational fixation	an|heften an|kleben an|schrauben	to stick at to glue at to screw at
auf	NPnom	NPnom	locative blaze- bubble	auf|brodeln auf|flammen auf|lodern auf|spudeln	to bubble up to light up to blaze up to bubble up
	NPnom +PP-zu/in/ nach/auf	NPnom	locative gaze	auf|blicken auf|schauen auf|sehen	to glance up to look up to look up
	NPnom +NPacc	NPnom +NPacc	locative/ dimensional instigate	auf|hetzen auf|scheuchen	to instigate to rouse
	NPnom +NPacc +PP-auf	NPnom +NPacc	locative/ relational fixation	auf|heften auf|kleben auf|pressen	to staple on to glue on to press on
	NPnom	NPnom	ingressive sound	auf|brüllen auf|heulen auf|klingen auf|kreischen auf|schluchzen auf|stöhnen	suddenly roar suddenly howl suddenly sound suddenly scream suddenly sob suddenly moan

Table 2: The gold standard classes for the experiments, with subcategorization patterns.

		an			auf			an+auf		
		Purity	RI	ARI	Purity	RI	ARI	Purity	RI	ARI
	Human ratings		0.93			0.92			0.92	
K-means	idealized features (manually set)	0.83	0.91	0.70	0.88	0.92	0.72	0.93	0.97	8.2
	selected features (extracted)	**0.67**	**0.82**	**0.29**	0.75	**0.87**	0.52	0.46	0.88	0.32
	20 feat	0.58	0.74	0.18	0.69	0.69	0.40	0.43	0.88	0.14
	50 feat	**0.67**	0.80	0.20	0.75	0.83	0.38	0.43	0.90	0.19
	100 feat	**0.67**	0.79	0.18	0.75	0.83	0.40	0.49	0.90	*0.21*
	200 feat	0.58	0.74	0.13	**0.81**	0.86	0.52	0.43	0.88	0.18
LSC	selected features (extracted) Cutoff: 0.1	0.63	0.78	0.22	0.80	0.85	**0.55**	**0.85**	**0.92**	**0.59**

Table 3: Comparison of the results from different clustering methods and feature configurations.

dimension extensions) of the two vectors. In this way, each subcategorization frame is represented separately for the BV and the PV. For example, the vectors for the intransitive frame will be represented as *BV:intransitive* and *PV:intransitive*.

We evaluated the clusterings in terms of Purity (Manning et al., 2008), Rand Index and Adjusted Rand Index (Rand, 1971; Hubert and Arabie, 1985). Purity is a measure with values between 0 and 1 which captures the *purity* of individual clusters in terms of the ratio between the number of elements of the majority class in each cluster and the total of elements in the cluster. A perfect clustering will have a purity of 1. What Purity does not capture is the amount of clusters over which each target class is distributed. That means that also non-perfect clusters may achieve a Purity of 1 if there are more clusters than target classes. As long as the number of clusters is constant, however, purity is a good and intuitive approximation to clustering evaluation.

The Rand Index (RI) looks at pairs of elements and assesses whether they have been correctly placed in the same cluster (which is correct if they pertain to the same target class) or in different clusters (correct if they belong to different target classes). RI is sensitive to the number of non-empty clusters and can capture both the quality of individual clusters and the amount to which elements of target categories have been grouped together. RI looks as pair-wise decisions, which makes it also applicable to the human ratings described in section 4.1. The Adjusted Rand Index (ARI) is a version of RI which is corrected for chance. While RI has values between 0 and 1, ARI can have negative values; 1 still represents a perfect clustering.

The Adjusted Rand Index (ARI) is a version of RI which is corrected for chance. While RI has values between 0 and 1, ARI can have negative values; 1 still represents a perfect clustering.

We evaluated the clustering of the verbs with the particles *an* and *auf* separately from each other, since we have to expect that there is a different set of semantic classes for each verb particle. We also ran the same experiments for the gold standard as a whole (*an+auf*), in order to test if we could find some tendencies across clusters.

We set the number of clusters equal to the number of target categories from the gold standard. This gave us 5 clusters for both the *an*-set and the *auf*-set and 10 clusters for the classification of the whole gold standard.

Note that LSC is a soft clustering algorithm. For the evaluation of LSC clusters with respect to purity and RI and ARI, a conversion to hard clustering must be done. We did this conversion by simply applying a cutoff value for the output probabilities for cluster membership. We tried out various cut-off levels and found that for the sets of *an* and *auf* PVs the value of 0.1 gave a good trade-off between coverage (the total number of elements retained in all clusters) and ARI (cf also Table 4 below). This value is also the one used in Kühner and Schulte im Walde (2010).

5 Results

The comparison of the results from different methods can be seen in table 3. The strongest automatically obtained results are printed in bold face. The human rating scores are given in the first row and allow for a direct comparison between automatic clustering and human decisions.[6] The second row shows the artificial upper bound represented by the manually set feature vectors as lexicographic entries. Note that this is an *artificial* upper bound and not an experimental result, even if obtained by clustering.

The third row corresponds to the evaluation results for the manually selected corpus-based feature configuration used within K-means. They are to be compared with the following rows concerning the results based on automatically selected n most frequent features. The last row shows the results obtained with the LSC soft clustering algorithm, applying a cutoff of 0.1 output probability for cluster membership, again for the manually selected feature configuration. This result is not fully comparable to the rows above, which are obained with K-means or human ratings. Since LSC is a soft clustering algorithm, there is a tradeoff between coverage and accuracy which depends on the cutoff point selected for the conversion into hard clusters.

Note that the Purity values are comparable among each other since the number of clusters was held constant. We always chose a number of clusters equal to the number of target categories (5 categories for *an*, 5 for *auf* and 10 for *an+auf*).

Table 4 shows the results for LSC clustering in more detail. The soft clusterings have to be converted to hard clusterings. Because of this the cut-off point within the conversion becomes an important parameter. We chose here cut-off points which correspond to the output probability of cluster-elements (e.g. PV-BV pairs) with regard to each cluster. The table shows a clear tendency towards better ARI scores when higher cutoff points are chosen. But this is counterbalanced by the fact that for higher cutoff points less elements are retained. Below a certain cutoff-point the total number of elements retaind is smaller than the target set of verbs in the gold standard.

[6] RI is a measure which is based on pair-wise clustering decisions, we were able to calculate these scores for the human ratings described in section 4.1. Since purity is not based on a pair-wise decision, it was not applicable to the human ratings. For the same reason ARI was also not adaptable to the human rating scenario.

6 Discussion

It is not surprising that the manually defined feature configuration in our "lexicographic" setting perform best. These results are also similar to those obtained by the human validation of the gold standard. They do not get perfect scores of 1 because of small lexicographic differences concerning individual entries. The automatic clustering results relying on corpus-based features are worse, as expected, but they still represent a very strong tendency to group together PV-BV pairs into semantic classes. We can achieve relatively high purity scores, thus demonstrating that our approach is generally valid.

Concerning the feature selection for the corpus-based data, the manually selected set seems to perform slightly better than the automatic feature selection settings. Moreover, the manual selection represents a more stable setting since automatic selection seems to vary with the number n of features. There appears to be no optimal setting for n which gives the best results for all sets. For the *an* set the local maximum is reached with the selection of the 50 or 100 most frequent subcat frames. The selection of more or less features leads to worse evaluation scores. For the *auf* set this local maximum is reach with much higher values for n. The manually created feature set, on the other hand, always results in a relatively good performance. This is also an expected result since the feature selection already contains human linguistic knowledge on which syntactic arguments represent the core set of the semantic roles which the verbs can realize.

It is apparently surprising that for the joint gold standard set *an+auf* LSC performs much better than K-means. But this high ARI value comes at the cost of a very low coverage. If we compare this value to table 4, it can be seen that the cutoff point of 0.1, which works very well for sets of *an* and *auf* is inadequate for the set *an+auf*: only 20 verbs are retained in the converted clusters while the target size is 32. While we can observe the general tendency of LSC to perform on a roughly comparable level to K-means, an exact comparison is hard to obtain with the used evaluation metrics. There are, nevertheless, possible problem settings where soft clusters are more adequate, which justifies to include LSC in this comparison.

	an		auf		an+auf	
Cutoff	ARI	n_{clust}	ARI	n_{clust}	ARI	n_{clust}
0.07	0.17	25	0.39	22	0.31	40
0.08	0.18	23	0.55	20	0.39	32
0.09	0.19	21	0.55	20	0.56	23
0.10	0.22	19	0.55	20	0.59	20
0.11	0.30	16	0.5	19	0.48	17
0.12	0.30	16	0.41	16	0.56	16
$n_{classes}$		14		18		32

Table 4: Evaluation with LSC using extracted selected features for different cutoff points (probabilities of class membership) when creating hard clusters from soft clusters. ($n_{classes}$ refers to the number of elements across target classes, n_{clust} refers to the number of elements across hard clusters.)

The class of *anketten/anbinden* tends to end up in singleton clusters, especially *anketten*. We first suspected that this is due to the fact that *anketten* is a relatively infrequent verb and is represented by a sparse vector. But a comparison to the human ratings reveals that human raters show a similar and quite consistent disagreement with the gold standad with respect to this the locative relational *tying* and *fixation* classes. All 6 raters judged *anheften* (a fixation verb) and *anbinden* (a tying verb) as pertaining to the same category, contrary to the gold standard. Interestingly, this fixation-tying distinction is the only one, where a majority of raters deviated in their judgements from the gold standard at the same point. On the other hand some of the raters were confused by the fact the class of *aufbrodeln* combines two different elements: water and fire. This did not affect the majority of raters, nor was the disagreement consistent, but it is reflected in the somewhat lower inter-annotator agreement for the *auf* set (cf. table 1). These findings strongly suggest that the problem should be located in the gold standard rather than in the clustering method.

Finally, is interesting to compare the automatic clustering results to the human ratings from section 4.1. The human annotation task was complementary to the automatic clustering because clustering was done on the basis of corpus-based purely syntactic features while for the human rating the annotators focused on purely semantic information. Apart from the expectably worse performance of an automatic clustering it can be concluded that both information from the semantic and the syntactic perspectives ultimately lead to the creation of quite similar clusters, which is probably the most important conclusion we can draw from the experiment.

7 Conclusion

In this paper we have shown that a pairwise clustering of particle verbs in combination with their base verbs can be done with success if syntactic subcategorization frames for PVs and BVs are taken as features separately. By combining the extracted subcategorization frame count from base verbs and particle verbs as separate dimensions in a common vector space, we are able to model syntactic transfer patterns. We can also show that within our setting we are able to replicate a gold standard classification with a reasonable degree of success when we apply various clustering algorithms. The gold standard by itself can be validated by human judgements to a high degree. Human judges based their annotations on semantic factors and still they converge largely with an automatic clustering which is purely based on syntactic subcategorization.

In future work we plan to adress the problem of finding correspondences between the syntactic subcategorization slots, hence model the syntactic transfer proper, and to investigate if the syntactic transfer information can be used to predict the degree of semantic compositionality of PVs.

Acknowledgements

This work was funded by the DFG Research Project "Distributional Approaches to Semantic Relatedness" (Stefan Bott, Sabine Schulte im Walde), and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde). We would also like to thank the participants of the human rating experiment.

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.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the Unextractable: A Case Study on Verb Particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning*, pages 98–104, Taipei, Taiwan.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Timothy Baldwin. 2005. Deep Lexical Acquisition of Verb–Particle Constructions. *Computer Speech and Language*, 19:398–414.

Collin Bannard. 2005. Learning about the Meaning of Verb–Particle Constructions from Corpora. *Computer Speech and Language*, 19:467–478.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – a Corpus of Parsable Sentences from the Web. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, Darmstadt, Germany. To appear.

Wolfgang Fleischer and Irmhild Barz. 2012a. *Wortbildung der deutschen Gegenwartssprache*. de Gruyter.

Wolfgang Fleischer and Irmhild Barz. 2012b. *Wortbildung der deutschen Gegenwartssprache*. Walter de Gruyter, 4th edition.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Silvana Hartmann. 2008. Einfluss syntaktischer und semantischer Subkategorisierung auf die Kompositionalität von Partikelverben. Studienarbeit. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Supervision: Sabine Schulte im Walde and Hans Kamp.

Lawrence Hubert and Phipps Arabie. 1985. Comparing Partitions. *Journal of Classification*, 2:193–218.

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

Hans Kamp and Uwe Reyle. 1993. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Number 42. Springer.

Fritz Kliche. 2011. Semantic Variants of German Particle Verbs with *"ab"*. *Leuvense Bijdragen*, 97:3–27.

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 and Sabine Schulte im Walde. 2010. Determining the Degree of Compositionality of German Particle Verbs by Clustering Approaches. In *Proceedings of the 10th Conference on Natural Language Processing*, pages 47–56, Saarbrücken, Germany.

Andrea Lechler and Antje Roßdeutscher. 2009. German Particle Verbs with *auf*. Reconstructing their Composition in a DRT-based Framework. *Linguistische Berichte*, 220.

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

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

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.

Paola Merlo and Suzanne Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408.

William M. Rand. 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carr oll, and Franz Beil. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proceedings of the 37th Annual Meeting of the Association for Co mputational 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.

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

Sabine Schulte im Walde. 2006b. The Syntax-Semantics Interface of German Particle Verbs. Panel discussion at the 3rd ACL-SIGSEM Workshop on Prepositions at the 11th Conference of the European Chapter of the Association for Computational Linguistics.

Sylvia Springorum, Sabine Schulte im Walde, and Antje Roßdeutscher. 2012. Automatic Classification of German *an* Particle Verbs. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 73–80, Istanbul, Turkey.

Sylvia Springorum, Sabine Schulte im Walde, and Antje Roßdeutscher. 2013. Sentence Generation and Compositionality of Systematic Neologisms of German Particle Verbs. Talk at the 5th Conference on Quantitative Investigations in Theoretical Linguistics.

Sylvia Springorum. 2011. DRT-based Analysis of the German Verb Particle *"an"*. *Leuvense Bijdragen*, 97:80–105.

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

Aline Villavicencio. 2005. The Availability of Verb-Particle Constructions in Lexical Resources: How much is enough? *Computer Speech & Language*, 19(4):415–432.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques wi th Java Implementations*. Morgan Kaufmann.