Wie oft schreibt man das zusammen? The Puzzle of Why some Separable Verbs in German are More Separable than Others

Nana Khvtisavrishvili **Stefan Bott** **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
{khvtisna,bottsn,schulte}@ims.uni-stuttgart.de

Abstract

In this work we address the question why different German particle verbs tend to occur with different frequency proportions in syntactically separated vs. non-separated forms. The problem has been studied from a theoretical point of view and the syntactic conditions that determine particle verb realization in separated and non-separated paradigms are quite clear. But, to the best of our knowledge, the question of why there is a variation among particle verbs with respect to how often they appear in different paradigms has never been addressed empirically so far. In this paper we present a corpus-based study which tackles this question. We formulate various morphological, semantic and pragmatic hypotheses which might explain the variation and we test them with clustering and linear regression techniques.

1 Introduction

German particle verbs (PVs) may occur in different syntactic paradigms, depending on the type of clause and the finite/infinite status of the base verb. One of their best known characteristics is that of syntactic separability. PVs may be written together as one word or appear syntactically separated, as illustrated by (1) and (2). Finite PVs occur obligatorily separated in verb final clauses and syntactically non-separated in verb first and verb second clauses. This is the reason why they are also often called separable verbs.

(1) Die Praxis *sieht* meist noch ganz
The practice *looks* mostly still entirely
anders *aus*.
different *PRT*.
"The practice usually looks entirely different"

(2) Es sind keine Softkorallen, obwohl sie
They are no soft corrals, even they
so *aus*|*sehen*.
so *PRT*|*look*.
They are no soft corrals, even if they look like them.

The case of syntactically separated PVs is a quite cumbersome issue for NLP applications, especially in parsing and machine translation. The linear distance between verb and particle can be quite large, which makes it difficult to detect the syntactic dependency between them. Additionally, many verb particles, especially the most frequent, are homophonous to prepositions and other function words.

Even if PVs occur syntactically non-separated, this case is not homogeneous because PVs can also be separated morphologically by a functional morpheme, such as *-ge-* or *-zu-*. Non-separated uses of particle verbs can correspond to one of the following cases, as illustrated in Figure 1.

- Finite verbs in subordinate clauses (FIN): e.g. *... dass er sie an*|*lächelt.* (*... that he smiles at her.*)
- Infinitive, e.g. in combination with a modal verb (INF): e.g. *Ich kann da nur an*|*schließen.* (*I simply have to subscribe to that.*)
- Participle perfect (PP): e.g. *Er hat sie ein*|*ge*|*laden.* (*He has invited her.*)
- Infinitive with "zu" (IZU): e.g. *Die Sitzung war auf*|*zu*|*zeichnen.* (*The session had to be recorded.*)

Non-separated instances of PVs can occur in subordinate clauses (FIN) or appear in certain grammatical constructions which involve auxiliary verbs (PP, IZU & INF). The syntactically sep-

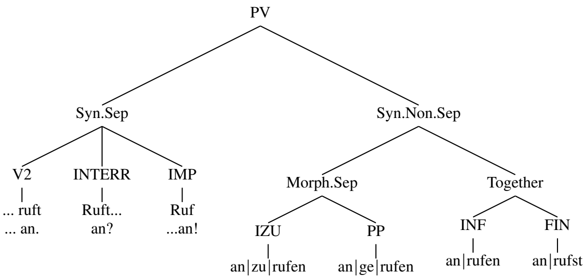

Figure 1: Different syntactic paradigms of the use of particle verbs

arated paradigm (SEP)[1] is much easier to define as a coherent class, since it consists of an inflected main verb and a clause final verb particle, as exemplified in (1). As Figure 1 shows, PVs may occur in indicative (V2), interrogative and imperative root clauses. However, in this work we do not distinguish between interrogative and indicative and we do not consider the imperative because of its low corpus frequency. From this discussion it should be clear that in German the realization of the PV as either separated or non-separated is fully determined by the clause type and the finite/infinitive distinction.

The syntactic and morphological aspects of the separated/non-separated dichotomy have been described adequately in traditional grammars and research literature (Lüdeling, 2001; Jacobs, 2005; Fuhrhop, 2007), but there is one aspect which has never been investigated, namely the proportions or relative frequencies with which different PVs occur in the different syntactic paradigms. To illustrate this, consider the verb *an|sehen* (*to watch/to resemble*) in (3) in contrast to *aus|sehen* (*to appear/to look like*) in (1)/(2).

(3) . . . ein Millionenpublikum das sich
. . . a million audience that REFL
Schrott an|sieht.
trash PRT|looks.
. . . an audience of millions that watches rubbish.

Neither *an|sehen* nor *aus|sehen* appear to be marked for a certain genre or register, they are both ambiguous and they have a similar corpus frequency (114 and 126 per million tokens, respectively). Nevertheless, they behave quite differently with respect to the proportions in which they occur in the syntactically separated paradigm: 20.5% vs 64.7%. This is surprising and, based on the relevant literature, we could not find an indication of why we observe such differences among PVs. Figure 2 shows the distribution of the proportion of separated occurrences over PVs as observed in the SdeWaC-Corpus (Faaß and Eckart, 2013). The x-axis represents relative frequency bands of syntactically separated occurrences of different PVs, the y-axis represents the count of PVs which falls into each relative frequency band. The PV *an|sehen* from example (3) would fall into the relative frequency band 20%-25%, which is the most densely populated one. The black curve represents the approximate density function, a smoothed representation of the histogram. It can be clearly seen that there is quite an amount of variation for which there is no straightforward explanation. Most notably, there is a long tail to the right, which means that a small number of PVs have a high tendency to occur syntactically separated.

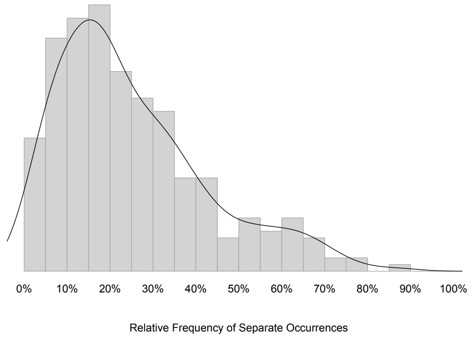

Figure 2: Histogram (bandwidth = 0.5) and density for the distribution of PVs according to their proportion of syntactically separated occurrence

In this paper we attempt to find reasons behind the variation we just described. We formulate various hypotheses on different syntactic, semantic and pragmatic factors which might influence the proportional distribution of PVs over the different syntactic paradigms. The current work mainly has a theoretic interest. It discovers a topic which has, so far, not received attention. It also represents a first attempt to solve the puzzle. Since PVs are very frequent in German, the question may have important implications for practical NLP applications. Parser performance is often poor in cases of separated PVs, in part because of the long lin-

[1] We use the term *paradigm* in a wide sense since the different PV realizations listed in Figure 1 are mutually exclusive. We do not intent to make a statement, however, on the exact theoretical status of this relation.

ear distance between verb and particle. Shedding some light on the issue of PV separability might improve parsing and all subsequent NLP tasks that depend on a reliable detection of the syntactic dependencies between the base verbs and the verb particles of PVs. Challenges that arise in the process of handling PVs for different NLP tasks show, in turn, the importance of investigating PVs to understand the problems associated with them.

2 Related Work

A number of studies have already investigated the topic of German PVs from both a theoretical as well as a corpus-based perspective. German PVs were extensively studied from the theoretical perspective in works of Lüdeling (2001) and Stiebels (1996); some other works have focused on a single particle such as Springorum (2009), dealing with the semantic of PVs with *an*; Lechler and Roßdeutscher (2009) studied PVs with the particle *auf*; Kliche (2009) looked at PVs with the particle *ab*.

The theoretical studies of PV separability have so far mostly dealt with German PVs with respect to their idiosyncratic behavior. Lüdeling (2001) investigated whether PVs are morphological objects or phrasal constructions and how they can be distinguished from secondary predicate constructions or adverbial constructions. She revealed a series of theoretical problems and analyzed PVs as lexicalized phrasal constructions, considering separability the strongest argument for this analysis. Müller (2001; 2003), in turn, argued for a syntactic analysis of PVs.

Jacobs (2005) studied PVs as one of several cases that pose problems for the determination of word boundaries. This affects the question of separability and orthographic separation. Also Fuhrhop (2007) was concerned with the morphological and orthographic aspect of the separability of German PVs. In contrast to Lüdeling's analysis of PVs as lexicalized phrasal constructions, Fuhrhop analyzed them as graphemic words.

Corpus-based, empirical investigations of PVs have received less attention. Schulte im Walde (2004) used statistical grammars to identify German PVs and provided quantitative description and a preliminary distributional analysis of German PVs. Schulte im Walde (2005) addressed the issue of feature selection to identify semantically nearest neighbors.

Some other works aimed at determining the degree of semantic compositionality of PVs. Bott and Schulte im Walde (2014) predicted the degree of PV compositionality relying solely on word window information. In their approach only lexical distributional distance between a PV and its corresponding BV was considered to be a predictor for compositionality. They were the first to automatically correct PV lemmas which occur in the syntactically separated paradigm, where they are consistently listed as the lemma of the base verb. They reported on problems with automatically parsed data in this respect.

As for the statistical study of variation of particle placement, Gries (2001; 2002; 2011) analyzed the variation of particle placement in English. Since in English the placement of verb particles is subject to relatively free variation (*John picked up the book* vs *John picked the book up*) and in German the realization of PVs as separated or unseparated is tied to the clause type, Gries' work cannot be directly replicated for German data.

To our knowledge no work comparable to what we propose here has been performed so far. In this study, we want to explore the behavior of German PVs with respect to the relative frequency distribution over different syntactic paradigms. In other words, we want to assess the empirical distribution of proportions corresponding to these paradigms and, by doing so, learn something about the nature of PVs.

3 Experiments and Data Analysis

Since we found that it is hard to understand why different PVs tend to occur in different syntactic paradigms in different proportions, our goal was to find factors which might explain the variation we could observe. For our experiments we used clustering techniques and simple correlation analysis based on least squared error regression. Clustering produces a partitioning of the data into classes which are derived in an unsupervised manner. This has two advantages; the first is that the classification is overt and the derived clusters can be inspected directly. The second advantage is that, by virtue of being an unsupervised technique, clustering classifies data points without the need of a previously given classification scheme. The clusters derived in one clustering procedure can be matched against various gold standards.

3.1 Hypotheses

We started out from the basic hypothesis H_b that the variation of the proportion with which different PVs can be observed in different paradigms is not a random factor but must be governed by some underlying reason. We therefor formulated a series of hypotheses which are elaborations of H_b. When formulating our hypotheses we considered two factors: 1) It must be possible to evaluate each hypothesis in an empirical way and 2) it should be interpretable in grammatical terms.

If we turn again to *aus|sehen* and *an|sehen*, the pair of verbs in examples (1) - (3), we already saw that these verbs share a number of common features, even if they occur in the syntactically separated paradigm in different proportions: they correspond to the same base verb, they have a similar corpus frequency and they are both ambiguous. The most evident difference they have is that they appear with different verb particles. It might also be argued that *aus|sehen* is ambiguous to a higher degree than *an|sehen*. Of course different verbs may also differ in the register and genre in which they tend to be used. Since the use of the syntactically separated paradigm is mainly tied to main clauses and different genres/registers may use more or less subordinate clauses, there may also be an indirect relation between genre/register and the tendency of verbs to occur in the non-separated paradigm. Genre and register are, however, often difficult to assess in corpora since genre-specific corpora tend to be much smaller than mixed-genre corpora. Nevertheless we can assume that average sentence length is a rough indicator for such difference.

Based on these considerations we formulated the following hypotheses:

- H_b: The variation of PVs with respect to the proportions that correspond to different syntactic paradigms is not a random factor. It is governed by some other underlying phenomenon.
- H1: Particles: Different particles influence the use of a corresponding PV in different syntactic paradigms.
- H2: Corpus Frequencies: The total corpus frequency has an impact on the proportional distribution over different paradigms.
- H3: Ambiguity: The degree of ambiguity of individual PVs can explain the behavior of PVs with respect to its proportional distribution over different paradigms.
- H4: Sentence Length: We take sentence length as a rough indicator for differences in text genre and register and hypothesize that there are correlations between average sentence length per PV and the proportion with which the PV occurs in different paradigms.

3.2 Data

For the extraction of features we used the SdeWaC corpus (Faaß and Eckart, 2013), a cleaned version of the deWaC corpus, which was compiled by the WaCky initiative (Baroni et al., 2009). SdeWaC consists of sentences from German web pages which contains syntactically well-formed and parseable sentences (880 million tokens). The corpus was tokenized with Schmid's tokenizer (Schmid, 2000), POS-tagged and lemmatized with Schmid's tree-tagger (Schmid, 1994) and parsed with Bohnet's MATE dependency parser (Bohnet, 2010). The parses provide information about dependency relations between components of a sentence. Dependency information is important in cases in which a PV occurs in a separated form, because it shows dependency between a particle and a corresponding base verb. The format of the corpus further provides lemma and part-of-speech annotation. For annotation the STTS tagset (Thielen et al., 1999) was used.

For the selection of the PVs in our data set, three additional corpora were used: HGC (Fitschen, 2004), DECOW12 (Schäfer and Bildhauer, 2012) and the German Wikipedia.[2]

3.3 Selection of Particle Verbs and building of the Data Set

For our experiments we created a data set of PVs which was balanced over the corpus frequencies of PVs and the particles to which they correspond. For this, we selected PVs randomly from three frequency bands - high, mid and low frequency - to investigate the behavior of particle verbs from different frequency bands. Occurrence frequencies are calculated as the harmonic mean of four different frequencies gained from the following corpora: SdeWaC, HGC, DECOW12 and Wikipedia.

Frequency bands are determined for each particle separately, i.e. thresholds for determining fre-

[2] Wikipedia dump (dewiki-20110410). `http://en.wikipedia.org/wiki/Wikipedia:Database_download`

PV	Sep.	PP	IZU	INF	FIN	Non-sep.
aussehen	0.5801	0.0207	0.0123	0.1886	0.1982	0.4198
anblicken	0.7994	0.0252	0	0.0466	0.1288	0.2005
ansehen	0.2025	0.3389	0.1907	0.1659	0.1019	0.7975
zuhören	0.3946	0.0569	0.0019	0.3136	0.2329	0.6054

Table 1: Feature Vectors

quency areas for different particles are different. The thresholds for the frequency bands were calculated by dividing the PVs with the same particle into equally large sets according to their overall corpus frequency (tertiles). In our work we investigateed PVs with the following 11 prepositional particles: *an*, *auf*, *ein*, *aus*, *zu*, *um*, *ab*, *unter*, *durch*, *über* and *nach*. We randomly selected 30 PVs for each particle from three frequency areas. The resulting list contained 938 PVs. Each particle was represented through 90 PVs (30 of low frequency, 30 of mid frequency and 30 of high frequency). The particle *unter* had only 38 corresponding PVs.

One of the problems that arose in the creation of the data set was the fact that PVs may be easily confounded with prefix verbs. Prefix verbs are not separable at all and have a quite different syntactic behaviour. For example, the verb *umarmen* (*to hug*), has the prefix *um-* which has a homophonic verb particle. There are also verbs which are ambiguous between a prefix verb and a particle verb interpretation: *übersetzen* may be a PV (meaning *to cross a body of water*) or a prefix verb (meaning *to translate*). Four of the verb particles we used - *um*, *unter*, *über* and *durch* - are ambiguous between prefix and particle use. Since the data set with 938 PVs was generated automatically, there was a number of verbs which were ambiguous between particle verb and prefix verb interpretation. In order to make sure that no prefix verbs were included in our data set, we manually edited the list of 938 PVs by excluding such cases. PVs whose verbal base correspond to a prefix verb were excluded as well (e.g. *ausverkaufen/to sell off*).

We know from previous experiments that PVs with very low and very high frequencies tend to be problematic for automatic assessment: low frequency items are likely to present data-sparseness problems and high-frequency items tend to be highly lexicalized and very idiosyncratic in their behavior. For this reason we excluded the top 20 frequent PVs and the 20 PV with the lowest frequency. This revision of the original list (938 PVs) resulted in a new list of 400 PVs. The data set for our experiments contains 400 PVs (targets). Each PV is represented through a six-dimensional feature vector. Features correspond to the different syntactic paradigms a PV can occur in plus the syntactically non-separated use of a PV, which is a sum of the paradigms: PP, IZU, INF and FIN. The values are normalized (relative) frequencies over the total frequency of a PV. For feature extraction only counts from the SdeWaC corpus were used. Table 1 shows a sample of vectors.

3.4 Clustering Experiments

In order to assess our first three hypotheses (H1-H3) we carried our clustering experiments (see section 3.5 for H4). The goal of clustering algorithms is to partition a set of objects in groups (clusters), so that the objects within one group are similar to each other and dissimilar to those in other groups. Object are compared based on particular features. To perform the task of analyzing German PVs empirically, we use a hard clustering method, namely the simple K-means algorithm.[3] On the account that PVs are represented in terms of feature vectors, similarity (or dissimilarity) between two objects is defined as the euclidean distance between the corresponding vectors. The greater the distance, the more dissimilar the objects are; they are then assigned to different clusters.

One of the challenges of the K-means algorithm is to find the optimal K, which must be specified in advance so that the structure of the data can be revealed. The experiments were carried out with different K values : K = 3, 5, 7, 11, 15, 20 for H1 (particles) and H2 (frequency). In addition to these values K = 4 clustering experiments were performed for the H3 (ambiguity), because for this hypothesis we used a reference set with 4 classes of different ambiguity levels (cf. 3.4.1 below).

3.4.1 Reference Sets

For evaluation we used a series of reference sets against which the clusterings were compared and the evaluation metrics were computed. Each reference set was built to represent the information corresponding to our hypotheses listed in section 3.1. For the partitioning of data into ambiguity classes, for example, the degree of ambiguity of a verb was determined by the information gained from different dictionaries. Due to the inconsistency in the

[3] We use the WEKA implementation (Witten and Frank, 2005)

degree of ambiguity a verb has in different dictionaries - GermaNet (Hamp and Feldweg, 1997), Wiktionary,[4] Duden[5] and DictCC[6] - mean ambiguity was calculated and a verb was assigned to a certain class according to its mean ambiguity value. In sum, we used the following reference sets:

- RS1 corresponds to H1, the particle hypothesis. RS1 contains 11 classes which correspond to the 11 particles described in section 3.3. For example *an|sehen* belongs to the class *an*. In this case class affiliation can be defined unambiguously for each verb.
- RS2 models the corpus frequency of PVs (H2): The PVs of RS2 are divided in three classes: H(high), M(mid) and L(low). Because we discarded the 20 most frequent and the 20 PVs with the lowest frequency from the original list of 938 PVs we also had to randomly reduce the mid frequency class in order to obtain a balanced representation of each class. This led to a selection of 88 high-frequency, 80 mid-frequency and 74 low-frequency PVs.
- RS3 captures the ambiguity of PVs. Ambiguity of each PVs is determined by computing the rounded mean ambiguity out of four ambiguities gained from the four sources mentioned above: GermaNet, Wiktionary, Duden and DicctCC. The RS has four classes for unambiguous PVs (A1), and PVs which have two, three or more than three readings (A2, A3 and AG3). *Nach|zahlen*, for example, is unambiguous (A1) and means *to pay later*; *an|sehen* from example (3) has more than three meanings (AG3) and may mean *to look at*, *to watch*, *to have the look of something*, *to consider someone as*.

Note that there is no reference set for Hypothesis 4 (average sentence length), because we do not test this hypothesis with clustering techniques and use corpus counts of sentence lengths (on a continuous scale) instead.

3.4.2 Evaluation

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 at pair-wise decisions, which makes it also applicable to comparison with reference data which lists pairwise class membership decisions, but does not necessarily define closed sets of reference classes.

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 again represents a perfect clustering. An ARI of 0 indicates a clustering which is close to the random level. While ARI is corrected for chance, the two metrics require a baseline for comparison. For this purpose we use random clustering, where each PV is assigned to a random cluster. In our case we averaged the values over 100 random clusterings.

3.5 Correlations

In order to tackle hypothesis 4 we used corpus-extracted counts of sentence lengths. In this case we deviate from the clustering approach because sentence length is a feature which is easily extracted from the corpus and there appears to be no natural way to bin sentence length into reference classes. In order to test H4, we matched the average sentence length with the percentage of verb realizations in the separated paradigm per PV. Each PV is thus matched to a point in a two-dimensional

[4] `http://wiktionary.org`

[5] We consulted the online edition: `http://www.duden.de`

[6] `http://www.dict.cc`

	K	RS1:Particle			RS2:Frequency			RS3:Ambiguity		
		Purity	ARI	RI	Purity	ARI	RI	Purity	ARI	RI
K-Means	3	0.16	0.0168	0.62	0.42	0.0174	0.56	0.42	0.00422	0.56
	4							0.40	-0.00127	0.60
	5	0.17	0.0150	0.74	0.41	0.0101	0.59	0.40	-0.00727	0.61
	7	0.18	0.0110	0.77	0.44	0.0066	0.62	0.40	-0.00354	0.65
	11	0.23	0.0173	0.83	0.47	0.0098	0.64	0.41	0.00002	0.67
	15	0.23	0.0101	0.84	0.52	0.1579	0.65	0.43	0.00512	0.68
	20	0.25	0.0108	0.85	0.53	0.0132	0.65	0.44	0.00221	0.68
Random Clustering	3	0.15	-0.0000	0.63	0.39	0.0001	0.56	0.38	0.0005	0.57
	4							0.38	0.0002	0.60
	5	0.16	0.0006	0.74	0.40	-0.0002	0.60	0.39	0.0002	0.62
	7	0.17	0.0000	0.78	0.42	0.0004	0.62	0.39	-0.0002	0.65
	11	0.19	-0.0000	0.83	0.44	0.0007	0.64	0.41	-0.0002	0.67
	15	0.21	0.00000	0.85	0.46	-0.0019	0.65	0.42	0.0001	0.68
	20	0.22	-0.00000	0.86	0.48	0.0001	0.65	0.43	0.0007	0.69

Table 2: Results of the clustering experiments for hypotheses 1, 2 and 3

space. Then a simple least squared error regression is applied, using the *lm* function of the R language (R Development Core Team, 2008).

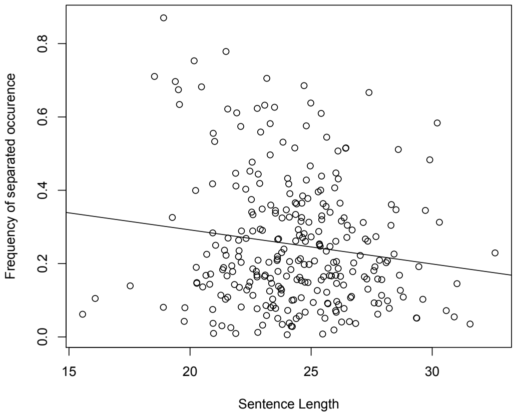

Figure 3: The relation between proportion of the syntactically separated paradigm and average sentence length

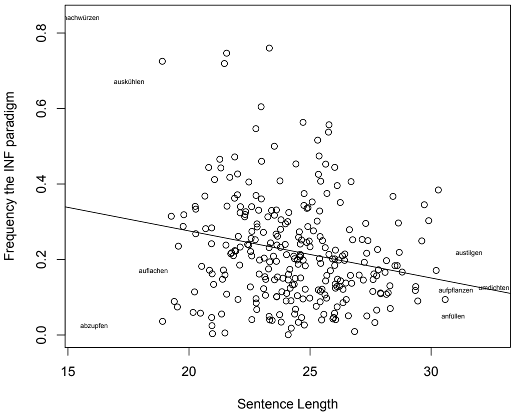

Figure 4: The relation between proportion of the infinite paradigm and average sentence length

4 Results and Discussion

Table 2 shows the results of the clustering experiments for hypotheses H1 to H3. The top part lists the results obtained with K-means while the lower part lists the results of the baseline random clustering. It can be seen that the results are nearly consistently above the baseline, but the difference is not significant. The factors we capture by the reference sets which correspond to these hypotheses seem to have a certain effect on the proportion of PVs with which they occur syntactically separated, but by themselves they do not explain the observed variation with which PVs occur in different syntactic paradigms.

Figure 3 plots the distribution of sentence lengths (H4) against the distribution of relative frequencies (proportions) for the separated paradigm per PV. The black line represents the regression line for the syntactically separated paradigm depending on the average sentence length per PV. The scatterplot suggests that there is a direct relation between the two factors - there is a tendency that PVs that occur very often separately also tend to occur in shorter sentences - but it is not very strong. This first impression is corroborated by a simple regression analysis which only examines the correlation between average sentence lengths and the relative frequency of the syntactically separated paradigm: the correlation between sentence length and the frequency of the separated paradigm reaches significance, but not with a very high confidence (p=0.018).

This analysis models the branching of the top-node of the tree in Figure 1, the distinction between syntactically separated vs non-separated. In order to check possible correlations between sentence length and *all* syntactic paradigms we carried out a multivariate regression analysis with sentence length as the dependent variable and each of the syntactic paradigms as independent variables (PP, INF, IZU, FIN, SEP), but in this analysis none of the independent variables showed a significant correlation with sentence length.

Just for the sake of error analysis and visual data exploration we also examined the relation of individual syntactic paradigms to sentence length. The most notable correlation we found is the one between the INF paradigm and sentence length. The corresponding scatterplot can be seen in Figure 4. This scatterplot resembles Figure 3, but also shows some differences. The most interesting observation which can be made here concerns the outliers on the x-axis, which is the reason why they are plotted out as PV lemmas (the dots correspond to the rest of all the PVs). The outliers in the left upper corner, the PVs that occur in short sentences and tend to occur very frequently in the INF paradigm all appear in the cooking domain, such as *nach|würzen* (*to add additional spice*), *aus|kühlen* (*to cool down*) or *auf|kochen* (*to reboil*). This hints at an influence of the text domain.

We have made some observations which are worth a closer investigation. We have noticed that a number of PVs sharing the same BV tend to be assigned to the same cluster. What was remarkable was the behavior of PVs with the BV *bauen* (*to build*), i.e. *auf|bauen* (*to build up*), *ab|bauen* (*to dismantle/reduce/mine*), *nach|bauen* (*to reversely engineer*), *aus|bauen* (*to enlarge/equip*), *ein|bauen* (*to install/integrate/build in*), which were very often found in the same cluster across different clusterings. This behavior was observed also in synonym clustering: some synonym pairs which share the same BV were repeatedly found in the same cluster.

Some verbs tend to appear in more formal register and hence have other preferences for syntactic paradigms. To give an example: *zu|senden* and *zu|schicken* (both meaning *to sent to*) can be used in different registers. *Zu|senden* is predominantly used in formal style, whereas *zu|schicken* tends to occur in informal style. This, again, highlights the influence of pragmatic factors, such as register and genre.

Finally we found that much noise was introduced into our data by errors stemming from the linguistic preprocessing. We found errors in the POS tags, most notably verb particles which were tagged as preposition and vice versa. This also means that the syntactic dependency between base verb and particle is not identified correctly. Often the lemmas of PVs were predicted incorrectly, incorporating functional morphemes into the lemma (e.g. *auf|zumachen* instead of *auf|machen*). This shows again that a better treatment of PVs in linguistic processors would be very desirable. A better understanding of empirical aspects of PVs could contribute to an improvement.

5 Conclusions

In this paper we described the empirical distribution of the proportions of German particle verbs with respect to their occurrence in different syntactic paradigms. We were able to show that there is observable variation in the frequencies in which PVs occur in different syntactic paradigms. We could find no explanation for this variation in the relevant literature. We parted from the basic hypothesis that there must be underlying factors which influence the behaviour of PVs in this respect. Building on this assumption, we formulated and tested a series of syntactic, semantic and pragmatic hypotheses about the source of the variation.

We could not provide a definitive answer to our initial question of what factors determine the proportional distributions of PVs over the different syntactic paradigms, but our findings suggest that pragmatic factors, such as genre and register, play an important role. We consider the problem well worth further study, considering that a better understanding of the behaviour of PVs has a high potential to improve the treatment of PVs in NLP tasks such as parsing and machine translation. In future work we plan to take pragmatic factors more strongly into account.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: a Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is Not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Stroudsburg, PA, USA.

Stefan Bott and Sabine Schulte im Walde. 2014. Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 509–516, Reykjavik, Island.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – A Corpus of Parsable Sentences from the Web. In *Language Processing and Knowledge in the Web*, pages 61–68. Springer.

Arne Fitschen. 2004. Ein computerlinguistisches Lexikon als komplexes System. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*.

Nanna Fuhrhop. 2007. *Zwischen Wort und Syntagma: Zur grammatischen Fundierung der Getrennt-und Zusammenschreibung*. Walter de Gruyter.

Stefan Gries. 2001. A Multifactorial Analysis of Syntactic Variation: Particle Movement Revisited. *Journal of Quantitative Linguistics*, 8(1):33–50.

Stefan Gries. 2002. The Influence of Processing on Syntactic Variation: Particle Placement in English. *Verb-particle Explorations*, 1:269–288.

Stefan Gries. 2011. Acquiring Particle Placement in English: A Corpus-Based erspective. *Morphosyntactic Alternations in English: Functional and Cognitive Perspectives. London/Oakville, CT: Equinox*, pages 235–263.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – A Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Lawrence Hubert and Phipps Arabie. 1985. Comparing Partitions. *Journal of Classification*, 2(1):193–218.

Joachim Jacobs. 2005. *Spatien: Zum System der Getrennt-und Zusammenschreibung im heutigen Deutsch*, volume 8. Walter de Gruyter.

Fritz Kliche. 2009. Zur semantik der partikelverben auf '*ab*'. eine studie im rahmen der diskursepräentationstheorie. *Masters thesis, Universität Tübingen*.

Andrea Lechler and Antje Roßdeutscher. 2009. German Particle Verbs with *auf*. Reconstructing their Composition in a DRT-based Framework. *Linguistische Berichte*, (220):439–478.

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

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Stefan Müller. 2001. Syntax or Morphology: German Particle Verbs Revisited. In Nicole Dehé, Ray Jackendoff, Andrew McIntyre, and Silke Urban, editors, *Verb-Particle Explorations*, Interface Explorations. Mouton de Gruyter, Berlin, New York.

Stefan Müller. 2003. Solving the Bracketing Paradox: an Analysis of the Morphology of German Particle Verbs. *Journal of Linguistics*, 39(2):275–325.

R Development Core Team, 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

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

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12, pages 44–49. Manchester, UK.

Helmut Schmid. 2000. Unsupervised Learning of Period Disambiguation for Tokenisation. Technical report, Universität Stuttgart.

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

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.

Sylvia Springorum. 2009. Zur Semantik der Partikelverben mit *an*. Eine Studie zur Konstruktion ihrer Bedeutung im Rahmen der Diskursrepräsentationstheorie. *Studienarbeit. Universität Stuttgart*.

Barbara Stiebels. 1996. *Lexikalische Argumente und Adjunkte. Zum semantischen Beitrag von verbalen Präfixen und Partikeln*. Number 39. Akademie Verlag.

Christine Thielen, Anne Schiller, Simone Teufel, and Christine Stöckert. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universitiät Stuttgart and Universität Tübingen.

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