G_hoSt-PV: A Representative Gold Standard of German Particle Verbs

Stefan Bott, Nana Khvtisavrishvili, Max Kisselew, Sabine Schulte im Walde
Institute for Natural Language Processing
University of Stuttgart, Germany
{stefan.bott,nana.khvtisavrishvili,max.kisselew,schulte}@ims.uni-stuttgart.de

Abstract

German particle verbs represent a frequent type of multi-word-expression that forms a highly productive paradigm in the lexicon. Similarly to other multi-word expressions, particle verbs exhibit various levels of compositionality. One of the major obstacles for the study of compositionality is the lack of representative gold standards of human ratings. In order to address this bottleneck, this paper presents such a gold standard data set containing 400 randomly selected German particle verbs. It is balanced across several particle types and three frequency bands, and accomplished by human ratings on the degree of semantic compositionality.

1 Introduction

German particle verbs (PVs), such as *auf|schauen* (*to look up*) represent a type of multi-word expression composed of a particle and a base verb (BV). As example (1) shows, they may be written together or syntactically separated but they always form one semantic unit.

(1) a. Das Kind **sah** seine Mutter **an**.
The child looked his/her mother an-PRT.
'The child looked at his/her mother.'

b. .... dass das Kind seine Mutter **an**|**sah**.
... that the child his/her mother looked|an-PRT.
'... that the child looked at his mother.'

In German, PVs are particularly frequent and form a highly productive paradigm in the lexicon, which often leads to neologisms and is subject to creative language use in puns and word plays (Springorum et al., 2013). Like many other multi-word expressions, PVs differ with respect to their compositionality. Some PVs can be deduced entirely from the meaning of the BV but others have meanings which are totally distinct. Most PVs fall on a continuum in between the two extremes. Some examples are the following:

FULLY COMPOSITIONAL: *an|leuchten* (*to illuminate*); the BV *leuchten* means *to shine*, and *an* expresses directionality (among other senses), cf. (2-a).

SEMI-COMPOSITIONAL: *ab|segnen* means *to approve*; literally, *segnen* means *to bless*. The two verb meanings are related, but a meaning shift occurred (cf. (2-b)). Semi-compositional PVs are usually part of a productive paradigm. In our case, *ab|segnen* patterns with verbs like *ab|nicken* (also meaning *to approve*, where the BV means *to nod*), and *ab|zeichnen* (*to give the approval signature*, where *zeichnen* means *to sign*).

NON-COMPOSITIONAL: *nach|schlagen* means *to look up (e.g. a reference)* or *to consult (e.g. a dictionary)*; the BV *schlagen* means *to beat* (cf. (2-c)).

(2) a. Peter **leuchtete** das Bild mit der Taschenlampe **an**.
Peter shined the picture with the flashlight an-PRT.
'Peter illuminated the picture with the flashlight.'

b. Der Chef **segnete** die Pläne **ab**.
The boss blessed the plans ab-PRT.
'The boss approved the plans.'

c. Stella **schlug** das Wort im Wörterbuch **nach**.
Stella beat the word in-the dictionary nach-PRT.
'Stella looked up the word in the dictionary.'

The compositionality of PVs has received some attention in Computational Linguistics. For example, the assessment of compositionality grades has been studied for English (Baldwin et al., 2003; McCarthy et al., 2003; Bannard et al., 2003; Bannard, 2005; Reddy et al., 2011; Salehi and Cook, 2013; Salehi et al., 2014) and German (Hartmann et al., 2008; Kühner and Schulte im Walde, 2010; Bott and Schulte im Walde, 2014), mostly with the use of methods from distributional semantics. A central requirement for such studies is the availability of gold standards of human ratings which can serve as the basis for evaluation.

Only few gold standards of this kind are available (cf. section 2), and they tend to require a high amount of human work to create. While humans have relatively clear intuitions on the grade of compositionality of PVs, the ambiguity of PVs often represents a problem both for the elicitation of ratings and automatic assessment. Most of the studies that are dedicated to PV compositionality have created their own gold standards, but both the workload and the issue of comparability among studies make larger, public ally available data sets highly desirable. In addition, the availability of standard resources is a prerequisite for inter-study comparability. In this paper we present such a resource containing 400 German PVs. The gold standard was designed as a target selection which is balanced over different types of particles and various ranges of corpus frequency. A subset of the gold standard has already been used in Bott and Schulte im Walde (2015) for the assessment of PV-compositionality. The data set has been created in a larger project which also produced G$_h$ost-NN (Schulte im Walde et al., 2016), a gold standard for German noun-noun Compounds with a similar design and a similar rating collection process. The resource is available to the research community under a Creative Commons License.[1]

In the remainder of this paper, section 2 discusses the availability of similar existing resources. In section 3 we describe the criteria which were important for the design of the new resource. In sections 4 and 5 we describe the creation and the properties of the gold standard.

2 Previously Existing Data

The only comparable previously existing data set which contains human ratings on German PVs can be found in Hartmann (2008). This data set is balanced over 8 frequency bands and rated by 3 expert raters, but it only contains 99 PVs, corresponding to 11 particles. The collection of this data set considered polysemy by asking raters to indicate ambiguities and, if they noticed any, to disambiguate them in their own words. The ambiguity was not a criterion for the selection of the PVs in that data set, and the compositionality ratings did not distinguish between different word senses. This inability to distinguish between word senses for annotation is a problem with no obvious solution, as we will argue in section 4.4 below. We found that this resource, even if highly valuable, was too small for many purposes, especially because statistic significance depends highly on the size of the sample.

Also for English particle verbs, a limited number of data sets do exist. Bannard et al. (2003) present a corpus-based approach to the semantics of particle verb constructions in English. To this end they collected a gold standard containing 40 randomly selected phrasal verbs which were rated by 26 annotators. This gold standard contains ratings on compositionality for each particle verb construction with respect to both the BV and the particle. Ratings were given regarding only three levels: *yes*, *no* and *don't know*. For our new gold standard we wanted to avoid a simple binary classification, cf. the discussion in the previous section (2).

Somewhat related to our topic is the data set created by Cook and Stevenson (2006) for the evaluation of the prediction of particle senses. This gold standard consists of a list of of 389 English particle verb constructions with *up* balanced over three different frequency bands. Each of the PVs was annotated by two annotators for four different particle senses. The focus of their research was, however, not the study of compositionality, but the classification of particle meanings, and specified for one particle type.

3 Considerations for the Creation of the Gold Standard

For the creation of the gold standard we defined a series of properties which we wanted to find reflected in the data set, based on theoretical considerations and previous experiences.

- *Scalar judgments on compositionality:* As we already argued, the degree of compositionality falls on a continuum from fully compositional and non-compositional. For this reason we wanted scalar compositionality judgments.
- *Random selection:* In order to avoid bias we wanted to obtain a random sample from all existing PVs, but we also wanted different PV properties reflected in our selection, such as frequency and ambiguity levels.
- *Balanced over frequency bands:* From earlier studies (Bott and Schulte im Walde, 2014) we know that both very frequent and very sparse PVs tend to present special problems in comparison to mid-frequency PVs: high-frequency items tend to be strongly lexicalized and ambiguous, while low-frequency items are often

[1] http://www.ims.uni-stuttgart.de/data/ghost-pv

subject to problems that can be attributed to data sparseness. So we were faced with an inherent conflict between a strict balancedness of the GS –which would require us to represent PVs from the extreme ends of the frequency spectrum proportionally– and the goal to select PVs with prototypical behavior –which is contradicted by the fact that we know a priori that extremely frequent and extremely infrequent PVs tend to behave idiosyncratically.

- *Different ambiguity levels:* Polysemy is a factor which influences both human ratings and automatic computational assessment. We thus wanted semantic ambiguity levels to represent a feature in the data set. Ideally, we wanted compositionality ratings which correspond to different word senses. In section 4.4 below we discuss the complications this point brings about.

- *Selection of particles:* We were interested in de-prepositional particles which are semantically ambiguous and abstract (Lechler and Roßdeutscher, 2009; Haselbach, 2011; Kliche, 2011; Springorum, 2011). We chose to sample PVs corresponding to 11 verb particles, which were already used in (Hartmann et al., 2008): *an*, *auf*, *aus*, *nach*, *ab*, *zu*, *ein*, *über*, *unter*, *um*, *durch*. These particles are all de-prepositional, and their semantics are all highly ambiguous and show a high proportion of abstract readings.

4 Creation of the Gold Standard

The creation of the gold standard involved a number of steps: We collected a list of all PVs across particle types, as found in a large corpus. From this list a random selection was created automatically, which was balanced over three different frequency ranges. This initial list was manually filtered and finally this data set was annotated by human raters for PV compositionality. In the following, we describe these steps in some detail.

G$_h$ost-PV was designed with similar goals and similar desired properties in mind as G$_h$ost-NN (Schulte im Walde et al., 2016), a gold standard of German noun-noun compounds which was compiled within the same research project and in a very similar crowdsourcing process. Both PVs and noun-noun compounds are multi-word-expressions, but their different nature required also some different design-decisions which makes the two gold standards comparable, but not not entirely parallel.

4.1 Compilation of a Complete List of Existing PVs

We wanted to select PVs out of a list of all PVs that could be attested in German corpora. This required the compilation of a full corpus-extracted list of PVs. We only targeted PVs which are built with one of the particles we mentioned in section 3. An automatic detection of adequate candidate lemmas is not entirely trivial for three reasons.

1. If the lemma of a PV starts with the string that coincides with one of the particles, this can produce false positive PVs because also non-PVs start with the same string. For example, the simplex verb *zupfen* (*to pluck/pick*) happens to start with the character sequence that is idiomorphic to the particle *zu*.

2. Lemmatizers and parsers tend to produce errors in the detection and treatment of PVs, especially in the case of syntactically separate occurrences. This is problematic since prepositions may be wrongly interpreted as syntactically separated particles.

3. Some particles have counterparts which act as verb prefixes, so prefix verbs may be confounded with PVs. Some complex verbs are even ambiguous between a prefix verb and a particle verb, e.g. the verb *umfahren* in (3), which can be a PV which means *to drive over* or a prefix verb with the meaning of *to drive around*. Prefix verbs resemble particle verbs, but behave syntactically very different because they are never separated from the BV, as exemplified in example (3-b).

   (3) a. Er **fuhr** den Baum **um**.
   He drove the tree um-PRT.
   'He knocked over the tree (with a car).'

   b. Er **umfuhr** den Baum.
   He over-drove the tree.
   'He drove around the tree.'

   In order to exclude prefix verbs, we looked for combinations of verbs and particles which occurred both syntactically separated and written together as one word, relying on a dependency-parsed version of the *SdeWaC* corpus (Bohnet, 2010; Faaß and Eckart, 2013).

4.2 Selection of the Particle Verbs

Since our goal was to create a random but balanced selection of PVs, we automatically selected 938 PVs from the list obtained in the previous step. We aimed for a selection of 990 PVs (11 particles, 3 frequency bands and 30 PVs per combination), but for one particle (*unter*) the corpus only contained 38 PVs. We sampled from three different frequency ranges: Frequency tertiles were used to determine the three frequency bands: L(ow), M(id) and H(igh). Since the frequencies of PVs are not independent from the particles they correspond to, the tertiles were computed for each particle separately.

The frequencies were obtained as the harmonic mean of frequencies obtained from four different corpora: *SdeWaC* (Faaß and Eckart, 2013), *DECOW12* (Schäfer and Bildhauer, 2012), *HGC* (Fitschen, 2004) and the German *Wikipedia* (dump `dewiki-20110410`). The calculation of word frequency over different corpora was done to balance out known and suspected deficits in the balancedness of each corpus.

4.3 Cleaning of the Gold Standard

Since the original list of PVs was created randomly, the gold standard of 938 PVs still contained a certain amount of noisy entries. To remedy this problem we created a reduced gold standard which eliminated problematic entries. The most noticeable problem was the fact that some of the listed verbs were either ambiguous between homophone versions as a prefix verb and a particle verb (cf. example (3)) or only existed as prefix verbs. This means that we had to eliminate such verbs which were not detected by the filters described in section 4.1.

A second problem was that the automatically harvested PVs often contained wrong entries which were produced by parsing or lemmatization errors. We eliminated all verbs for which no consensus among the authors could be obtained on the basis of their graphic form whether they are existing PVs or not. In the same process also PVs were deleted which could be attested, but only for a very specific and limited domain, such as the verb *ab|teufen* (*to sink*), which could be attested for highly technical domains, but was not known by all authors.

Finally we considered all highly frequent and highly infrequent PVs as not desirable for practical experiments, as we found out in earlier experiments (Bott and Schulte im Walde, 2014). For this reason we excluded the 20 PVs with the highest and lowest frequency for each particle type. As a result of the manual filtering, the balance over frequency bands changed, as the number of mid-frequency PVs in the final gold standard is now higher than the number of low-frequency and high-frequency PVs. The distribution across particle types was however kept similar, because we removed the same number of PVs from the gold standard across particle types. We consider the manual cleaning more beneficial than harming since it excludes problematic entries while it retains those which are most prototypical and especially interesting for the task of compositionality assessment. The three parts of Table 1 present the final numbers of PV elements for each particle, frequency band and ambiguity level.

Particle	an	auf	aus	nach	ab	zu	über	unter	ein	um	durch
	47	45	48	45	47	37	9	12	45	37	28

Frequency Level	H	M	L
	88	238	74

Ambiguity Level	A1	A2	A3	AG3
	141	143	56	60

Table 1: Number of items per particle, frequency band and ambiguity band (A1 refers to one PV sense (i.e., monosemy); ambiguity of >3 is coded as AG3) after the manual selection process.

4.4 Collection of Compositionality Ratings

We collected compositionality ratings via Amazon Mechanical Turk (AMT)[2], allowing only for German native speakers as raters. Raters were asked to evaluate in how far the meaning of the PV is related to the meaning of its base verb. Each item was rated by 7 to 31 raters, with an average of 16.14 raters per item. Rating was done on a scale from 1 to 6, with 6 representing the maximum rating for compositionality. Raters with an insufficient level of German were detected by the inclusion of non-existing verbs which had to be detected in the rating process. If participants did not recognize the fake words, all of their ratings were rejected.

One problematic aspect of the collection of ratings on compositionality is the treatment of polysemy. It is evident that different readings of PVs correspond to different ambiguity levels. For example, the PV *zu|schlagen* has at least two meanings: *to strike* and *to take advantage of a good offer/bargain*. In addition, it can mean *to slam a door* and *to hit quickly and hard*. The BV *schlagen* means *to hit*. It is evident that the *strike* meaning is closely related

[2] `https://www.mturk.com`

to the meaning of the BV, and even more so the meaning of *hit-quickly*, while the meaning of *take advantage* is less compositional. But how many readings are there exactly? Is *striking* and *hitting* one sense or two? Which sense is predominant, and does the predominant sense exist in terms of frequency or in terms of some cognitive aspect? We found that these aspects are extremely hard to assess and even more so in a data collection based on crowd-sourcing. For this reason, we tried not to bias the raters choices by providing them contextual information, or any other information to disambiguate the target PV. We are aware of the fact that this is problematic, but we considered any other alternative even more problematic. Items were thus presented without context, and the rated word sense was assumed to be the predominant word sense as perceived by the raters. Our ongoing and future work explores alternative methods of data collection which addresses this problem, but which is necessarily more costly and more limited in scope.

5 Properties of the Gold Standard

The resulting gold standard data set contains 400 PVs accomplished by the following information:

- PV lemma
- Harmonic mean of PV corpus frequencies across four corpora
- The PV frequency band (low, mid, high)
- The PV level of ambiguity (ambiguities of 1, 2, 3 or greater than 3)
- The number of human ratings for the PV
- The mean compositionality rating for each PV
- The standard deviation of ratings among raters, as a measure of agreement
- The proportions of syntactically separated and syntactically non-separated appearances of the PV

The degree of semantic ambiguities is represented as the average grade of semantic ambiguity according to four lexical resources: GermaNet, Duden online, DictCC and Wiktionary. Any resource shows cases of a) spurious sense distinctions and b) under-representation of word senses. We tried to overcome definition problems by combining different lexical resources. In practice it is of course still very difficult to find an optimal representative listing of the number of word senses. Table 2 shows some sample entries of PVs from different frequency and ambiguity bands.

PV	PV freq	freq band	ambig. band	no. raters	mean rating	std dev	prop. synt. sep.	prop. synt. non-sep.
abkratzen	39.80	M	AG3	14	5.29	2.52	0.16	0.84
absegnen	23.38	H	A1	14	4.07	1.90	0.09	0.91
anleuchten	6.37	L	A1	20	5.95	1.50	0.62	0.38
anstiften	7.92	M	A2	15	1.80	0.86	0.17	0.83
aufhorchen	74.58	H	A1	29	4.55	1.97	0.16	0.84
aufschneiden	43.31	H	AG3	14	6.07	1.73	0.32	0.68
ausreizen	19.35	M	A2	29	3.62	2.13	0.07	0.93
durchrosten	9.66	M	A1	14	6.29	0.73	0.31	0.69
einstampfen	33.34	H	A1	14	4.07	2.06	0.15	0.85
nachschicken	22.81	H	A1	15	6.00	1.07	0.29	0.71
nachtragen	3.97	L	A2	15	4.47	2.03	0.21	0.79
umplanen	14.44	M	A2	15	4.93	1.83	0.10	0.90
zukneifen	8.53	M	A2	14	4.71	1.77	0.33	0.67
zulegen	4.00	L	AG3	14	3.86	2.07	0.29	0.71

Table 2: Sample entries from the gold standard.

The data collection scenario via Amazon Mechanical Turk makes it difficult to calculate inter-annotator agreement. Items were annotated by a varying number of annotators and each annotator annotated a different set of items. In the gold standard we include the standard deviation per item as a measure of agreement for each PV. The average standard deviation per rating target was 1.82 points on a 6-point scale. In Figure 1, the distribution of standard deviation over items can be seen in the form of a histogram. The x-axis shows the standard deviation per PV, where PVs were binned into intervals of 0.2 points of standard deviation. The y-axis shows the count of PVs per bin.

The plot shows that the highest peak is reached at a standard deviation of approximately 2.3 on a 6-point scale. This reflects the difficulty of the annotation task but is also to a certain extent a consequence of the crowd-sourcing

approach, which on the one hand allows for a larger collection of data, but on the other hand provides less control on the background of the raters. There is a strong tendency of PVs with a low deviation in rating –the ones that represent the tail to the left in Figure 1– to be monosemous, like *nach|reisen* (*to follow s.o. or s.th by traveling*) and *durch|rosten* (*to rust through*), and the ones with strongly deviating ratings to be polysemous. A good example for this is the PV *ab|kratzen*, which can either mean *to scratch off* or *to die*. Among the latter group we also find PVs which are clearly monosemous, like *nach|denken* (*to meditate on s.th.*) or *durch|rechnen* (*to thoroughly calculate*). The strong variation in the ratings for such cases is surprising. A final, more expected, tendency that can be observed is that PVs with strong deviation in ratings also tend be the least compositional ones like *unter|jubeln* (*to plant s.th. on s.o.*).

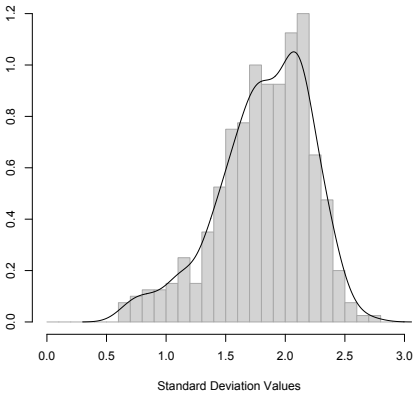

Figure 1: Histogram of the distribution and approximate density of standard deviation values for compositionality ratings across PVs. Standard deviation is provided to approximate inter-annotator agreement per item.

Figures 2 and 3 show the distribution of the obtained ratings and log-transformed word frequencies in relation to the different particles. The plots confirm some of the already known facts about the particles in question. For example, the particle *über* is predominantly locative and nearly always occurs in PVs which express some kind of movement or state (*über|streifen*, *to pull over*), even if it may be implicit (*über|schäumen*, *to foam over*). These PVs are always highly compositional, but not highly frequent. PV with *ab*, *an* and *ein* are much more varied in their semantics. Consequently, the corresponding PVs show a wider distribution in both frequency and compositionality. In general, the variation among particles is expected and thus confirmed by the gold standard.

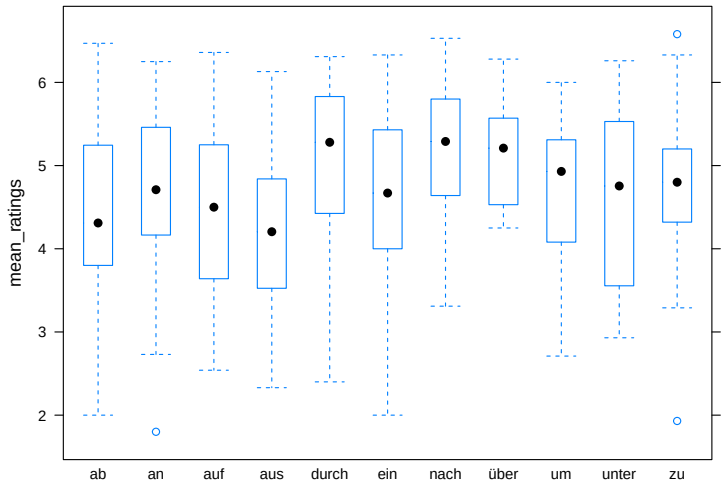

Figure 2: Mean ratings of particle verbs across particles types.

Figures 4 and 5 show the variation of ratings over frequency bands and ambiguity levels. We can observe little variation, which is good, since we intended the gold standard to be balanced. The ratings are quite evenly distributed over the different frequency bands. The mean value of the ratings is 4.67, which shows that PVs with a

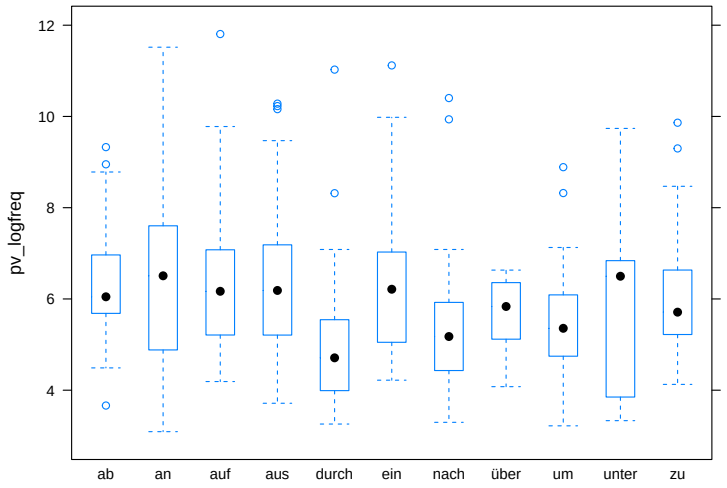

Figure 3: Log frequencies of particle verbs across particle types.

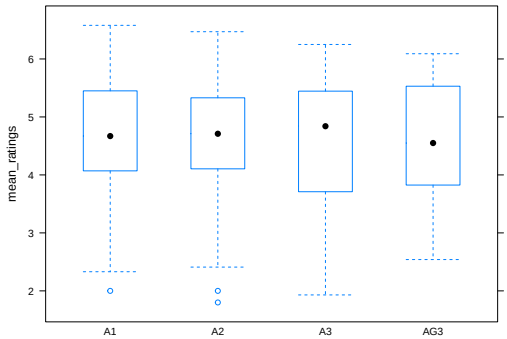

Figure 4: Mean compositionality ratings across frequency bands.

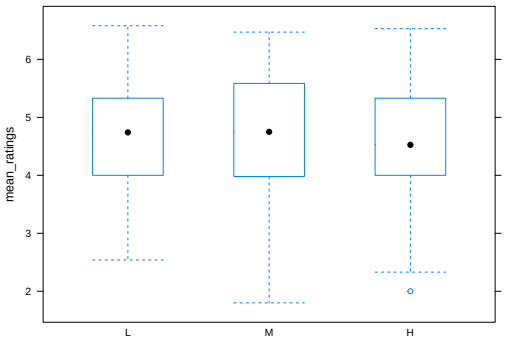

Figure 5: Mean compositionality ratings across ambiguity levels.

higher compositionality are slightly more dominant than those with low compositionality. Since the selection of PVs was done randomly we can assume that this reflects a general tendency of PVs to be compositional. Figure 4 shows the distribution of compositionality ratings for different ambiguity levels. The PVs with the highest polysemy (number of senses greater than 3) show a slight tendency to be rated in the medium range of compositionality. Highly ambiguous PVs tend to have senses with different levels of compositionality. They tend to mix word senses with different compositionality level, which should result in less PVs in the very high and the very low range. We expected this effect to be more pronounced than we could finally observe. We did not find a straightforward explanation for this, except for the already known fact that information on the grade of ambiguity extracted from lexical resources are never fully reliable, which might have caused the observed behavior.

6 Conclusion

This paper introduced a new gold standard for the evaluation of predicting German particle verb compositionality. The selection of particle verbs for this data set was carefully designed, especially in compiling a random selection of PVs which are balanced over different frequency bands. We provided some descriptive statistics which show that the data set is balanced in the distribution of PV compositionality across frequency and the grade of polysemy. The gold standard is available for research and education.

One of the problems which we could not resolve in a fully satisfactory way is the fact that the compositionality ratings per particle verb do not distinguish between different word senses. We have argued that this is a problem which is difficult to solve in a crowdsourcing approach for various reasons. Ongoing and future work addresses this specific aspect, but is necessarily limited to smaller amounts of target verbs and a smaller number of ratings.

References

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.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A Statistical Approach to the Semantics of Verb-Particles. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan.

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

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, Beijing, China.

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.

Stefan Bott and Sabine Schulte im Walde. 2015. Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs. In *Proceedings of the 11th International Conference on Computational Semantics*, page 34–39, London, UK.

Paul Cook and Suzanne Stevenson. 2006. Classifying Particle Semantics in English Verb-Particle Constructions. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 45–53, Sydney, Australia.

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*, pages 61–68, Darmstadt, Germany.

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

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. In *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.

Boris Haselbach. 2011. Deconstructing the Meaning of the German Temporal Verb Particle *'nach'* at the Syntax-Semantics Interface. In *Proceedings of Generative Grammar in Geneva*, pages 71–92, Geneva, Switzerland.

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

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):439–478.

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*, pages 73–80, Sapporo, Japan.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Bahar Salehi and Paul Cook. 2013. Predicting the Compositionality of Multiword Expressions Using Translations in Multiple Languages. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 266–275, Atlanta, GA.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using Distributional Similarity of Multi-way Translations to Predict Multiword Expression Compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden.

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, Istanbul, Turkey.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. G$_h$ost-NN: A Representative Gold Standard of German Noun-Noun Compounds. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2285–2292, Portoroz, Slovenia.

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.