Distinguishing Literal and Non-Literal Usage of German Particle Verbs

Maximilian Köper **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{maximilian.koeper,schulte}@ims.uni-stuttgart.de

Abstract

This paper provides a binary, token-based classification of German particle verbs (PVs) into literal vs. non-literal usage. A random forest improving standard features (e.g., bag-of-words; affective ratings) with PV-specific information and abstraction over common nouns significantly outperforms the majority baseline. In addition, PV-specific classification experiments demonstrate the role of shared particle semantics and semantically related base verbs in PV meaning shifts.

1 Introduction

Automatic detection of non-literal expressions (including metaphors and idioms) is critical for many natural language processing (NLP) tasks such as information extraction, machine translation, and sentiment analysis. For this reason, the last decade has seen an increase in research on identifying literal vs. non-literal meaning (Birke and Sarkar, 2006; Birke and Sarkar, 2007; Li and Sporleder, 2009; Sporleder and Li, 2009; Turney et al., 2011; Shutova et al., 2013; Tsvetkov et al., 2014), as well as the establishment of workshops on metaphorical language in NLP.[1]

In this paper, we explore the prediction of literal vs. non-literal language usage of a computationally challenging class of multiword expressions: German particle verbs (PVs) such as *anlachen* (laugh at) are compositions of a base verb (BV) such as *lachen* (smile/laugh) and a verb particle such as *an*. German PVs are highly productive (Springorum et al., 2013b; Springorum et al., 2013a), and the particles are notoriously ambiguous (Lechler and Roßdeutscher, 2009; Haselbach, 2011; Springorum, 2011). Furthermore, the particles often trigger (regular) meaning shifts when they combine with base verbs (Springorum et al., 2013b), so the resulting PVs represent frequent cases of non-literal meaning. The contributions of this paper are as follows:

1. We present a random forest classifier that correctly identifies 86.8% of literal vs. non-literal language usage within a novel dataset of 6 436 annotated sentences, in comparison to a majority baseline of 64.9%.

2. We successfully incorporate salient PV-specific features and noun clusters in addition to standard bag-of-words features and affective ratings.

3. We demonstrate that PVs with semantically similar particles and semantically similar base verbs can predict each others' literal vs. non-literal language usage.

4. We illustrate the potential and the limits of the most salient classification features in predicting PV non-literal language usage.

In the remainder of this paper we describe previous work on non-literal language identification and computational models of German particle verbs (Section 2), before we introduce our dataset

[1] sites.google.com/site/metaphorinnlp2016/

Proceedings of NAACL-HLT 2016, pages 353–362,
San Diego, California, June 12-17, 2016.

on German particle verbs (Section 3), the particle verb features (Section 4), and the experiments, results and analyses (Section 5).

2 Related Work

Previous work relevant to this paper includes research on identifying non-literal language usage, and computational work on (German) particle verb meaning.

Identification of non-literal language usage: Birke and Sarkar (2006), Birke and Sarkar (2007), Li and Sporleder (2009) and Sporleder and Li (2009) performed binary token-based classifications for English datasets, relying on various contextual indicators. Birke & Sarkar exploited seed sets of literal vs. non-literal sentences, and used distributional similarity to classify English verbs. Li & Sporleder defined two models of text cohesion (a cohesion chain and a cohesion graph) to classify V+NP and V+PP combinations. Shutova et al. (2013) performed both metaphor identification and interpretion (by paraphrasing), focusing on English verbs. She relied on a seed set of annotated metaphors and standard verb and noun clustering, to classify literal vs. metaphorical verb senses. Gedigian et al. (2006) also predicted metaphorical meanings of English verb tokens, however heavily relying on manual rather than unsupervised data (i.e. labeled sentences and PropBank annotation) and a maximum entropy classifier. Turney et al. (2011) assumed that metaphorical word usage is correlated with the degree of abstractness of the word's context, and classified word senses in a given context as either literal or metaphorical. Their targets were adjective–noun combinations and verbs. Tsvetkov et al. (2014) presented a language-independent approach to metaphor identification. They used affective ratings, WordNet categories and vector-space word representations to train a metaphor-detecting classifier on English samples, and then applied it to a different target language using bilingual dictionaries.

Computational research on particle verbs was initially concerned with the automatic acquisition of particle verbs from corpora (Baldwin and Villavicencio, 2002; Baldwin, 2005; Villavicencio, 2005). Afterwards, the main focus has been on modelling the degree of compositionality of particle verbs as based on distributional features (McCarthy et al., 2003; Baldwin et al., 2003; Bannard, 2005). All these approaches were type-based, and predicting the compositionality was mainly concerned with PV–BV similarity, not taking the contribution of the particle into account. In cases where the particle semantics was respected (such as Bannard (2005)), the results were disappointing because modelling particle senses is still an unsolved problem.

Regarding German particle verbs, there has also been a focus on modelling PV compositionality (Kühner and Schulte im Walde, 2010; Bott and Schulte im Walde, 2014; Bott and Schulte im Walde, 2015). As in English, the approaches were all type-based and mainly concerned with PV–BV similarity. Another line of research categorized particle meanings by relating formal semantic definitions to automatic classifications (Rüd, 2012; Springorum et al., 2012). Furthermore, Springorum et al. (2013b) recently provided a corpus-based study on regular meaning shift conditions for German particle verbs.

3 Particle Verb Dataset

We selected 165 particle verbs across 10 particles, based on previous experiments and datasets that incorporated German particle verbs with regular meaning shifts, various degrees of ambiguity, and across frequency ranges (Springorum et al., 2013b; Springorum et al., 2013a; Bott and Schulte im Walde, 2015). For the 165 PVs, we randomly extracted 50 sentences from *DECOW14AX*, a German web corpus containing 12 billion tokens (Schäfer and Bildhauer, 2012; Schäfer, 2015). The sentences were morphologically annotated and parsed using *SMOR* (Faaß et al., 2010), *MarMoT* (Müller et al., 2013) and the MATE dependency parser (Bohnet, 2010). Combining part-of-speech and dependency information, we were able to reliably sample both separated and non-separated PV occurrences ("Der␣Ast␣**bricht**␣**ab**" vs. "Der␣Ast␣ist␣**abgebrochen**").

Three German native speakers with a linguistic background annotated each of the 8128 sen-

tences[2] on a 6-point scale [0,5], ranging from clearly literal (0) to clearly non-literal (5) usage. The total agreement of the annotators on all six categories was 43%, Fleiss' $\kappa = 0.35$. Dividing the scale into two disjunctive ranges with three categories each ([0,2] and [3,5]), the total agreement of the annotators on the two categories was 79%, Fleiss' $\kappa = 0.70$. In the experiments we used the binary-class distinction, and disregarded all cases of disagreement. This final dataset comprises 6436 sentences: 4174 literal and 2262 non-literal uses across 159 particle verbs and 10 particles.[3] Figure 1 shows the distribution of literal and non-literal sentences across the particles.

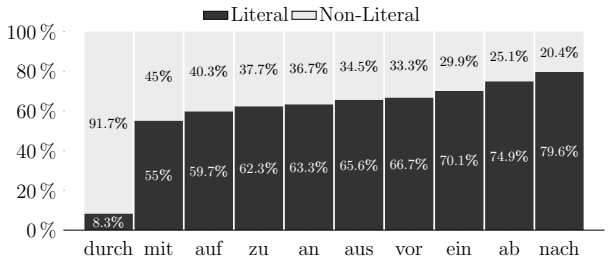


Figure 1: Lit/Non-lit distribution across particles.

4 Particle Verb Features

Our feature space includes standard features to detect non-literal language uses (bags-of-words and affective ratings) as well as PV-specific features and abstraction over common nouns.

4.1 Unigrams

As a standard feature in vector space models, we used all words in the particle verb sentences, i.e., a bag-of-words model relying on unigrams. We expected this standard information to be useful, because some words such as the abstract noun *Hoffnung* (hope) and the concrete noun *Geld* (money) frequently occur with non-literal rather than literal language usage:

1. (non-lit.) "Die Hoffnung **keimte** früh **auf**." *That hope **arose (lit: sprouted)** early.*
2. (non-lit.) "Er versucht das Geld **abzugraben**." *He tries to **demand (lit: dig off)** the money.*

[2]Some PVs appeared < 50 times in the corpus.

[3]The dataset is accessible from `http://www.ims.uni-stuttgart.de/data/pv_nonlit`.

To overcome data sparseness, we did not use the unigrams as individual features ($|V|$ = feature space), but implemented this feature as the output of a text-classifier. We relied on the Multinomial Naive Bayes (MNB) classifier by McCallum and Nigam (1998). While the classifier was designed for document classification, we considered a sentence as a document and the possible class outcomes were literal and non-literal.

Noun Clusters Because of the severe data sparseness in our PV feature sets, we performed noun generalization and applied the generalized information to all nouns in our PV contexts. Using all approx. 430000 nouns that appeared >100 times in the *DECOW14AX* corpus, we applied k-Means clustering with $k \in [2, 10000]$. As an alternative to the standard unigrams, we then replaced every noun in the PV sentences with its corresponding cluster tag.

4.2 Affective Ratings

Previous work on detecting non-literal language often makes use of psycholinguistic attributes, namely abstractness and concreteness ratings (Turney et al., 2011), and imageability ratings (Tsvetkov et al., 2014). Words with high abstractness ratings refer to entities that cannot be perceived with our senses; a large subset of which are non-visual (i.e., receive low imageability). It has been shown that non-literal expressions tend to occur with abstract words (*dark humor* versus *dark hair*). We thus expected affective ratings to be useful for particle verbs as well:

1. (lit.) "Den Lippenstift kannst du dir **abschminken**." *You can **remove** the lipstick.*
2. (non-lit.) "'Den Job kannst du dir **abschminken**." *You can **forget about** the job.*

We reimplemented the algorithm from (Turney and Littman, 2003) to create large-scale abstractness and imageability ratings for German (Köper and Schulte im Walde, 2016). Based on these ratings, we defined the following (partially redundant) features for the PV sentential contexts:

1. Rating of the PV subject
2. Rating of the PV object

3. Average rating of all nouns (excluding proper names)
4. Average rating of all proper names
5. Average rating of all verbs, excluding the PV
6. Average rating of all adjectives
7. Average rating of all adverbs

While features 3–7 have been adopted from (Turney et al., 2011), features 1–2 represent additional, PV-specific features.

4.3 Distributional Fit of PV, BV and Context

Particle verbs with a meaning shift are non-compositional regarding their base verbs. We thus implemented a PV-specific feature that measures the distributional fit of PVs and their BVs in the PV contexts. For example, looking at the following two PV sentences containing the BV *klingen* (to sound), the context of the first, literal sentence fits well to the BV meaning, but the context of the second, non-literal sentence does not. The distributional fit of the BV in the literal context should therefore be high, but the distributional fit of the BV in the non-literal context should be low.

1. (lit.) "Der Ton der Gitarre **klingt aus**."
   *The tone of the guitar **fades**.*

2. (non-lit.) "Den Abend lassen wir mit Wein **ausklingen**." *We **end** the evening with wine.*

To measure the distributional fit of PVs and BVs to PV contexts, we created 400-dimensional word representations using the *hyperwords* toolkit (Levy et al., 2015) and the *DECOW14AX* corpus. We relied on a symmetrical window of size 3 and applied positive pointwise mutual (PPMI) feature weighting together with singular value decomposition (SVD). Based on the word representations, we calculated cosine similarities between the PVs and their contexts, and likewise between the respective BVs and the PV contexts. The contexts we used were the same seven dimensions we used for the affective ratings (cf. Section 4.2). For example, regarding the sentence "Die Katze **springt** auf den Tisch" (*The cat **jumps** on the table*), we calculated the distributional similarity between the PV "aufspringen" and the subject "Katze", and the distributional similarity between the BV "springen" and the subject "Katze", etc. Each PV–context and each BV–context dimension represents an individual feature.

5 Classification Experiments

In this section, we present a series of binary classification experiments to distinguish literal and non-literal PV usage. Section 5.1 presents the main experiments comparing our features in a global classification setup, and Section 5.2 presents PV-specific additional experiments that zoom into the role of particle types and into the role of semantically related PVs and BVs. Section 5.3 provides a qualitative analysis of the features.

5.1 Main Experiments

We used a random forest with multiple (in our case 100) random decision trees,[4] with each tree voting for an overall classification result. The unigram information was represented by stacking the output of a multinomial naive bayes text classifier as a single feature into the random forest. For all machine learning algorithms we relied on the WEKA toolkit (Witten et al., 2011).

The experiments were performed in two modes, (a) without knowledge of the particle (i.e., the individual particle was not provided as a feature), and (b) with explicit knowledge of the particle. In this way, we could identify the contribution of the particle.

The classification results are shown in Table 1. We report on the feature type, and on the size[5] of the feature set f. We further present literal and non-literal f-scores F_1, and accuracy with and without particle knowledge. We compare against the majority baseline (literal). The right-most columns indicate whether the differences in performance are statistically significant, using the χ^2

[4] Experiments with other classification methods showed similar but inferior performance. Simple Logistic Regression performed 2nd best.

[5] Remember from Section 4.1 that the unigram information is based on all tokens (12 427) but we implemented the unigrams as a single feature (using the output of a classifier), thus the combined setting is only based on 22 features.

	Feature Type	$\|f\|$	Lit. F_1	Non-Lit. F_1	*Acc.*	*Acc.* + *P*
1	Majority Baseline	0	78.7	0.0	64.9	-
2	Unigram	12 427	83.2	55.5	75.6	76.5
3	Unigram + NN Clusters	6 305	81.6	66.7	76.3	79.3
4	AC Ratings	7	81.3	60.7	74.7	76.3
5	IMG Ratings	7	77.5	48.1	68.6	71.6
6	Distributional Fit	14	83.0	61.8	76.5	80.2
7	Comb. (2+4+6)	22	88.6	77.1	84.8	86.6
8	Comb. (3+4+6) + NN Clusters	22	88.8	77.3	85.0	86.8

(a) Results across feature types and their combinations.

	1	2	3	4	5	6	7	8
1								
2	*/*							
3	*/*	-/∗						
4	*/*	-/-	∘/*					
5	*/*	*/*	*/*	*/*				
6	*/*	*/*	-/-	∘/*	*/*			
7	*/*	*/*	*/*	*/*	*/*	*/*		
8	*/*	*/*	*/*	*/*	*/*	*/*	-/-	

(b) Statistical significance of differences *Acc*/*Acc* + *P*.

Table 1: Main classification results.

test and $*$ for $p < 0.001$ and $\circ$ for $p < 0.05$ to mark significance.

The results demonstrate that the classification results across all feature types are significantly better than the majority baseline. The single best performing feature type (cf. lines 1–6) is the unigram information; in combination with the particle information (+*P*), the distributional PV/BV–context fit is best. Combining the best feature types (2+4+6) once more improves the results, and ditto when adding noun cluster information.[6] We can also see that abstractness (AC) ratings outperform imageability (IMG) ratings.

So overall, the best performing feature set successfully combines unigrams that incorporate clusters for noun generalization; abstractness ratings; and PV-specific information regarding the distributional PV/BV–context fit and knowledge about the particle. This setup correctly classifies literal sentences with an f-score of 88.8 and non-literal sentences with an f-score of 77.3; overall accuracy is 86.8 over a baseline of 64.9.

It is difficult to compare our results against previous approaches on different datasets and in different languages. Regarding the closest approaches to our work, Tsvetkov et al. (2014) report an accuracy score of 82.0 using 10-fold cross-validation on a training dataset with a majority baseline of 59.2, combining multiple lexical semantic features on a dataset of 1 609 English subject–verb–object triples. Birke and Sarkar (2007) trained a single classifier for each of twenty-five verbs in the English TroFi verb dataset and reported only an average f-score: 64.9 against a majority baseline of 62.9. Turney et al. (2011) obtained an average f-score of 63.9 and additionally report an accuracy score of 73.4 on the same dataset, using abstractness ratings.

In contrast to our work, the two approaches by Birke and Sarkar (2007) and Turney et al. (2011) treated each group of sentences for a given target verb as a separate learning problem, while we learn one classifier across different verbs. Our method 4 (AC Ratings) can be considered a German re-implementation of the approach by Turney et al. (2011). In comparison to the results of previous work, our approach can safely be considered state-of-the-art.

5.2 PV-Specific Experiments

5.2.1 Incorporating Standard Measures of Multiword Idiomaticity

One traditional line of research to identify type-based multiword collocations or idiomatic expressions relies on the association strength between the multiword parts (Evert and Krenn, 2001; Krenn and Evert, 2001; Stevenson et al., 2004): The stronger the association between the parts of a multiword expression (as determined by raw frequency, some variant of mutual information, etc.), the stronger the collocation/idiomaticity of the combination of the parts. Based on this assumption, we calculated the association strength between PVs and their contextual subjects/objects, using *local mutual information (LMI)*, cf. Evert (2005). The LMI scores were based on type-based frequency counts in the *DECOW14AX* corpus and added as features to the respective contexts, assuming that large LMI scores indicate non-literal PV usage.

[6] The best cluster analysis in our experiments contained 750 noun clusters.

Adding the LMI values to the overall best feature set from the main experiments decreased accuracy from 86.8 → 86.0. Using the LMI association strength values of the PV–subject and PV–object pairs by themselves provided slightly but non-significantly better results in comparison to the majority baseline: 65.9 > 64.9.[7] Manual investigations revealed that verb–noun pairs with high LMI scores represent collocations in many cases, but the collocations are not only used in non-literal language but also in literal language, e.g., "Sendung ausstrahlen" ("broadcast a program").

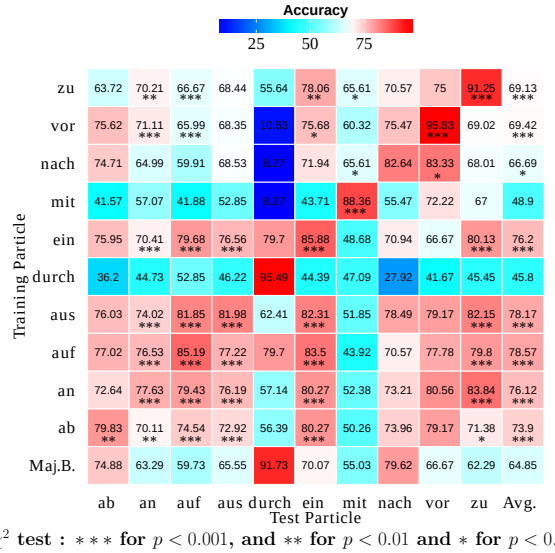
5.2.2 Non-Literality across Particles

In order to explore the predicability of literal vs. non-literal uses with respect to specific particles, we trained the best classifier from the main experiments on all particle verbs with particle *X* and applied the classifier to all particle verbs with particle *Y*. Our hypothesis was that pairs of particles with similar ambiguities might predict each other better than pairs with different particle meanings.

This PV-specific setup could also be applied within a PV group with the same particle: We trained the classifier on all PVs with particle *X* except for one, and then applied the trained classifier to the missing PV with particle *X*. The setup was repeated for all PVs with particle *X*, and the average accuracy was calculated.

Figure 2 provides the results as a heat map, with red indicating high and blue indicating low accuracy scores. The vertical particles on the left correspond to the training particles, and the horizontal particles at the bottom correspond to the test particles. The bottom line shows the majority baseline. For example, training a classifier on "ein" PVs and evaluating it on "aus" PVs results in an accuracy of 76.56, which is significantly better ($***$ for $p < 0.001$) than the baseline for "aus" (65.55).

The diagonal in the heat map (showing the within-particle setup) provides particularly high accuracy scores, so the PVs with the same particle predict (non-)literality within the group very well. This demonstrates that the meanings and the meaning shifts across PVs with the same particle (e.g., *aufdecken* and *auftischen*) are quite regular. A comparably strong prediction is found between "vor" (before/in front of) and "nach" (after/behind), with both particles carrying highly similar temporal and local senses. Other examples of strongly related antonymous particle pairs are "auf"/"zu", "ein"/"aus", and "aus"/"an". Examples of strongly related synonymous particle pairs are "an"/"ein", and "aus"/"zu". "durch" correlates poorly with all other particles, which is probably due to the few sentences we collected from the corpus. "mit" also correlates poorly with all other particles, because it is the only particle with little ambiguity. So overall, the heat map corresponds to intuitions about semantic relatedness across particle pairs.



χ^2 test : $***$ for $p < 0.001$, and $**$ for $p < 0.01$ and $*$ for $p < 0.05$.

Figure 2: Train a classifier on PVs with particle *X* and test it on PVs with particle *Y*.

5.2.3 Non-Literality across Particle Verbs

An even more fine-grained experiment setting explored the predictability of a specific particle verb based on the classifier trained on a different particle verb. Our hypothesis was that pairs of PVs that predict each other particularly well share some meaning aspects, either (i) because the training and the test verb share the same BV (**SameBV**: ab*graben*:auf*graben*), or (ii) the PVs are synonymous according to the German *Duden*[8]

[7] We also experimented with the other five contextual feature dimensions, but the results were even worse.

[8] http://www.duden.de

dictionary (**PVSyn**: *auftragen*:*auftischen*), or (iii) because the BVs of two PVs with identical particles are synonymous according to the *Duden* (**BVSyn**: auf*reissen*:auf*platzen*).

Figure 3 shows the f-scores for predicting literality and non-literality across the three settings, in comparison to the main experiments ("All"). The number of PV pairs in the settings and the majority accuracy for these PV pairs are also provided, because the experiment sets differ in size. We can see that PVs with the same BV (**SameBV**) predict each other's classifications well regarding literal but not regarding non-literal sentences. This behaviour illustrates the contribution of the particle to the PV meaning: The same BVs with different particles potentially differ strongly, if the particles do not agree on one or more senses. Synonymous PVs (**PVSyn**) predict each other as well in literal as in non-literal cases. Since the PVs in all cases are supposed to have the same meaning, this behaviour is also reasonable. An increase in both literal and non-literal F1 is reached for PV pairs with the same particle and synonymous BVs (**BVSyn**), because the BVs are supposed to carry the same meaning, and the identical particles trigger similar meaning shifts. Overall, the experiment demonstrates that synonymous verbs undergo similar meaning shifts, and that a particle initiates similar meaning shifts when applied to synonymous BVs.

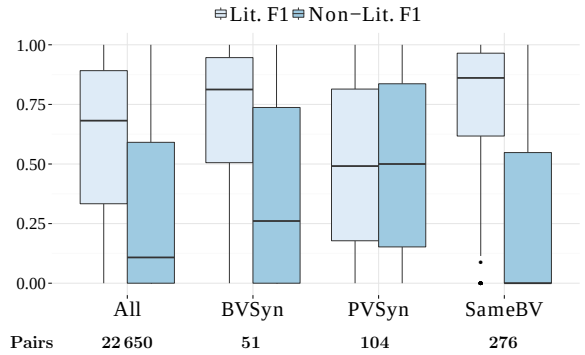


Figure 3: Prediction for semantically related PVs.

5.3 Indicators of Non-Literality

In the final part of the paper, we perform a qualitative analysis of the most salient features.

5.3.1 Information Gain

First of all, we looked into the feature space by computing the information gain within the best random forest classifier. The information gain (*I-Gain*) provides the improvement in information entropy regarding our feature space and the class labels, as defined by equation (1).

$$I\text{-}Gain(Class,Feat) = H(Class) - H(Class|Feat) \quad (1)$$

The information gain does not take feature interaction into account, but determines the importance of the individual features. Applying this method reveals the three most salient features: unigrams (0.31), abstractness ratings of the context nouns (0.17), and distributional fit of the base verbs (0.11). The information gain therefore confirms our results from the main experiments, where these three features worked best.

In addition, we noticed that for all features higher weights were given to dimensions that depend on nouns (such as the common nouns in the PV contexts, and the subject and object nouns), in comparison to proper names, verbs, adjectives and adverbs. For example, the abstractness ratings of the adverbs were ranked second lowest with a score of 0.005, and the distributional fit between BVs and adjectives was ranked last with a zero score, indicating that this feature provides no additional information for our dataset.

5.3.2 Distributional Fit

We now take a look at the distributional fit feature, which was the best performing feature in the main experiments, when combined with particle knowledge. Figure 4 focusing on the distributional fit between BVs and common nouns (as determined third best by the information gain) confirms that the feature is helpful in distinguishing literal vs. non-literal PV sentences across particles: The medians in the boxplots for literal sentences are clearly above those for non-literal sentences. The plots confirm that BVs can be exploited to identify compositional uses of PVs (which in turn refer to literal usage).

Looking into individual PVs confirms that this feature distinguishes well between the literal and non-literal sentences. On the other hand, we also

find PVs where this feature is not able to identify non-literal language use. Figure 5 presents the boxplots with cosine values for *aufblühen* (blossom out) and *auflodern* (burn up), where the feature works well, in comparison to *absaufen* (drown), where the feature cannot distinguish (non-)literal language usage.

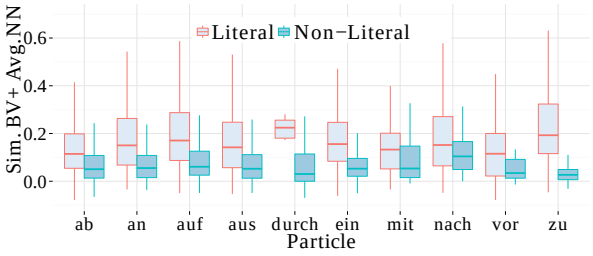


Figure 4: Distributional fit of BVs and context nouns in (non-)literal sentences across particles.

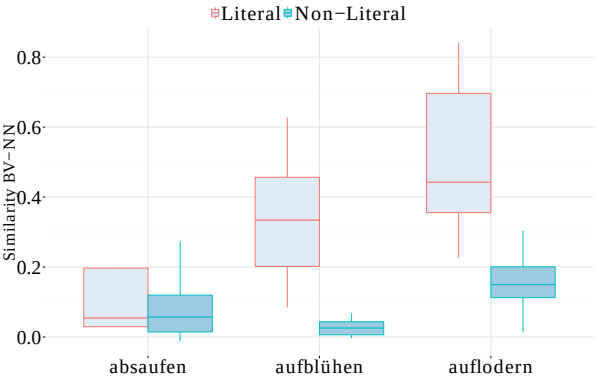


Figure 5: Example PVs and their distributional fit of BVs and context nouns in (non-)literal use.

5.3.3 Abstractness of Contexts

Finally, we take a look at the abstractness feature, which was also among the best performing features in the main experiments, and which is generally assumed to represent a salient indicator of non-literal language usage. Figure 6 focusing on the abstractness of common nouns in the PV sentences[9] (as determined second best by the information gain) confirms that the feature is also helpful in distinguishing literal vs. non-literal PV sentences across particles: Again, the medians in the boxplots for literal sentences are clearly above those for non-literal sentences. The plots confirm that contextual abstractness is a salient indicator of non-literal language usage.

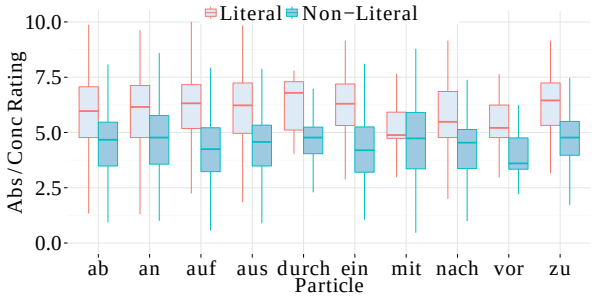


Figure 6: Average abstractness ratings of context nouns in (non-)literal sentences across particles.

Looking into individual PVs again confirms that this feature distinguishes well between the literal and non-literal sentences but also that there are PVs where this feature is not able to identify non-literal language use. Figure 7 presents the boxplots with abstractness ratings for *anstauen* (accumulate) and *durchsickern* (leak through), where the feature works well, in comparison to *antanzen* (waltz in) and especially *ausklingen* (fade/finish), where the feature cannot distinguish (non-)literal language usage.

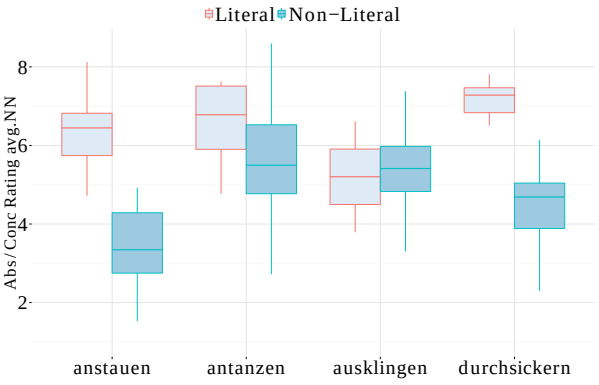


Figure 7: Example PVs and their average abstractness ratings of context nouns in (non-)literal use.

Two example sentences where the abstractness feature goes wrong for a good reason are as follows. In (1) "Aber wir sollten doch um fünf zum Essen **antanzen**." (*But we should* ***show up (lit: waltz in)*** *for dinner at five*), the context nouns are concrete (*we; dinner*) but the language usage is non-literal. In contrast, in (2) "Ich liebe Emotionen, deshalb **summen** alle **mit**." (*I love emotions, there-*

[9] High values indicate concreteness.

*fore everyone **hums along***), the object noun in the sentence is highly abstract (*emotion*), but the language usage is literal. These examples illustrate that contextual abstractness is not a perfect indicator of non-literal language usage.

6 Conclusion

We presented a classifier that predicts literal vs. non-literal language usage for German particle verbs, a semantically challenging type of multiword expressions. The classifier significantly outperformed the baseline by improving standard features with noun clusters and a PV-specific distributional fit feature. PV-specific experiments indicated that PVs whose particles share aspects of ambiguity and which incorporate semantically related BVs seem to undergo similar meaning shifts.

7 Acknowledgements

The research was supported by the DFG Collaborative Research Centre SFB 732 (Maximilian Köper) and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde).

References

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the Unextractable: A Case Study on Verb Particles. In *Proceedings of the 6th 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 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.

Julia Birke and Anoop Sarkar. 2006. A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, pages 329–336, Trento, Italy.

Julia Birke and Anoop Sarkar. 2007. Active Learning for the Identification of Nonliteral Language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28, Rochester, NY.

Bernd Bohnet. 2010. Top 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, Iceland.

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 Conference on Computational Semantics*, pages 34–39, London, UK.

Stefan Evert and Brigitte Krenn. 2001. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France.

Stefan Evert. 2005. *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 803–810, Valletta, Malta.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching Metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, pages 41–48, New York City, NY.

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.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatically generated affective norms of abstractness, arousal, imageability and valence for 350000 german lemmas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than Frequency? A Case Study on Extracting PP-Verb Collocations. In *Proceedings of the ACL Workshop on Collocations*, pages 39–46, Toulouse, France.

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.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity with Lessons learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Linlin Li and Caroline Sporleder. 2009. Classifier Combination for Contextual Idiom Detection Without Labelled Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 315–323, Singapore.

Andrew McCallum and Kamal Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of the AAAI Workshop on Learning for Text Categorization*.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, WA, USA.

Stefan Rüd. 2012. Untersuchung der distributionellen Eigenschaften der Lesarten der Partikel *'auf'* mittels Clustering-Methoden. Master's thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

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.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, editors, *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28 – 34.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical Metaphor Processing. *Computational Linguistics*, 39(2):301–353.

Caroline Sporleder and Linlin Li. 2009. Unsupervised Recognition of Literal and Non-Literal Use of Idiomatic Expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 754–762, Athens, Greece.

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. 2013a. Sentence Generation and Compositionality of Systematic Neologisms of German Particle Verbs. Talk at the Conference on Quantitative Investigations in Theoretical Linguistics.

Sylvia Springorum, Jason Utt, and Sabine Schulte im Walde. 2013b. Regular Meaning Shifts in German Particle Verbs: A Case Study. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 228–239, Potsdam, Germany.

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

Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2004. Statistical Measures of the Semi-Productivity of Light Verb Constructions. In *Proceedings of the 2nd Workshop on Multiword Expressions: Integrating Processing*, pages 1–8, Barcelona, Spain.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258.

Peter D. Turney and Michael L. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21(4):315–346.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK.

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

Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.