Analogies in Complex Verb Meaning Shifts:  
The Effect of Affect in Semantic Similarity Models

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

Abstract

We present a computational model to detect and distinguish analogies in meaning shifts between German base and complex verbs. In contrast to previous corpus-based studies, a novel dataset demonstrates that “regular” shifts represent the smallest class. Classification experiments relying on a standard similarity model successfully distinguish between four types of shifts, with verb classes boosting the performance, and affective features for abstractness, emotion and sentiment representing the most salient indicators.

1 Introduction

German particle verbs are complex verb structures such as anstrahlen ‘to beam/smile at’ that combine a prefix particle (an) with a base verb (strahlen ‘to beam’). They are highly ambiguous, and they often trigger meaning shifts of the base verbs (Springorum et al., 2013; Köper and Schulte im Walde, 2016). More specifically, Springorum et al. (2013) presented a manual corpus exploration suggesting regular mechanisms in meaning shifts from base verbs (BVs) to particle verbs (PVs) that apply across a semantically coherent set of BVs. For example, the two sound BVs brummen ‘to hum’ and donnern ‘to rumble’ both describe a displeasing loud noise. Combining them with the particle auf, the PVs aufbrummen and aufdonnern are near-synonyms in one of their senses, roughly meaning ‘to forcefully assign a task’. In a similar vein, Morgan (1997) used schematic diagrams to illustrate meaning shifts of English complex verbs with the particle out.

The goal of this work is to provide a computational model of meaning shifts for German particle verbs. We define our task from the perspective of an analogy, comparing a BV pair with a PV pair, cf. Figure 1. A BV–PV model of regular meaning shifts expects (i) semantic coherence $\text{sim}(BV_1, BV_2)$ between the two BVs (i.e., overlap in a selected set of semantically salient features), (ii) strong semantic similarity $\text{sim}(PV_1, PV_2)$ between the PVs, and (iii) low semantic similarity $\text{sim}(BV_i, PV_i)$ between the corresponding BV–PV pairs, where the shifts take place.

Figure 1: Analogy model applied to BV–PV shifts.

In a similar vein, a rich tradition on computational work on analogies focuses on finding a relational analogy in multiple choices as required by the SAT Scholastic Aptitude Test (Turney, 2006, 2012; Speer et al., 2017). While the SAT questions provide a limited set of possible answers, more recent attention has been spent on open vocabulary tasks of the form $A:B::C:?$ (Mikolov et al., 2013; Levy and Goldberg, 2014).

The contribution of our analogy model is two-fold: (i) it makes a step forward from hand-selected manual datasets of meaning shifts to larger-scale automatic classification; and (ii) it aims to deepen the linguistic insights into complex verb meaning shifts. While we focus on German particle verbs, we expect our explorations to be applicable also to other types of meaning shifts or languages. Most importantly, we show that (a) there are variants of (ir)regular meaning shifts that go beyond what was found in corpus-based explorations; (b) generalisation via classification boosts the strengths of salient verb features; and (c) affective features (i.e., abstractness, emotion and sentiment) play the predominant role in similarity models of meaning shifts.
2 A Collection of BV–PV Analogies

As to our knowledge, no datasets of human-annotated complex verb meaning shifts are available, apart from small-scale case studies (Springorum et al., 2013). We therefore collected human judgements for analogy combinations of BV–PV pairs of the form

$$BV_1 :: PV_1 :: BV_2 :: PV_2$$

such as klappern:abklappern::klopfen:abklopfen. We aimed for \( \approx 200 \) analogies per particle type, focusing on the four highly frequent particle types ab, an, auf, aus. The target selection was restricted to \( PV_1/PV_2 \) combinations with identical particles, and where the two PVs were deemed (near-)synonyms according to the German standard dictionary DUDEN\(^1\) or the German Wiktionary\(^2\), as we were interested in BV–PV analogies with semantically highly similar PVs.

In total, we collected 794 analogy questions.\(^3\) The BV–PV pairs were distributed over four lists according to the four particle types, and annotated by five German native speakers with a background in linguistics. To avoid a sense-specific bias, we provided no contextual information and therefore conducted the annotation on the type level. The annotators were asked to classify the analogies into four categories to distinguish between meaning shifts in no/one/both BV–PV pairs:

1. **COMP**: no BV–PV pair has a meaning shift, i.e., both PVs are compositional regarding their BVs, and therefore all four verbs are (near-)synonyms; example: (ab)feilen:(ab)schleifen ‘to grind (off)’

2. **ASYM COMP**: only one of the BV–PV pairs undergoes a meaning shift; in this case, the annotators also indicated that pair; example: (auf)wühlen:(auf)graben lit. ‘to churn::dig (up)’, where aufwühlen includes a strong emotion component

3. **SHIFT DIFF**: both BV–PV pairs show a meaning shift, but the BVs are not semantically similar; example: (aus)baden:(aus)bügeln ‘to pay for an error’ with baden ‘to take a bath’ and bügeln ‘to iron’

4. **SHIFT REG**: both BV–PV pairs undergo a meaning shift, and the BVs are semantically similar; example: (an)graben:(an)baggern ‘to hit on so.’ with both graben and baggern ‘to dig’

For practical reasons, we merged the left/right asymmetric cases ASYM COMP such that the annotated meaning shift was always on the left-hand side (by swapping the asymmetric-right pairs), since these cases represent instances of the same phenomenon, i.e., where just one of the pairs underwent a meaning shift.

Despite a distinction into four categories per instance, we obtained a moderate Fleiss’ \( \kappa \) agreement of 0.43 as the mean across the four particles:

![Figure 2: Number of majority class instances for four meaning shift categories by particle type.](image)

We transformed the annotations to actual class assignments by removing all instances from the dataset without a category majority, i.e., we only included BV–PV analogy pairs where at least 3 out of 5 annotators agreed on the shift category. We assigned the majority decision as class label. The final collection still contains 685 analogy pairs.

The distribution across the four particles and the four categories is illustrated in Figure 2. While meaning shifts have been observed across all four particle types, the analogical case SHIFT REG mentioned in previous corpus explorations represents the smallest class overall (8.5%). For the particle an, the cases with two meaning shifts (SHIFT DIFF+SHIFT REG) are especially rare (16.2%).

A manual inspection revealed that etymology and semantic change often led to opaque PVs annotated as SHIFT DIFF; an example is abkupfern ‘to plagiarise’. The origin of this meaning is based on the 18th century engravers who etched replicas of text and images into copper (Kupfer) plates.
Table 1: Example of BV–PV analogies across the four meaning shift categories.

| Comp          | AsymComp         | ShiftDiff      | ShiftReg       |
|---------------|------------------|----------------|----------------|
| abfeilen::abschleifen | abbauoen::abmontieren | abschreiben::abkupfern | abstottern::abratfern |
| abkuppeln::abhangen  | abchecken::abprüfen     | abschweifen::ablehren      | abrauschen::abzischen   |
| aneignen::anlernen   | anfeuern::anbrennen    | ankreiden::anlasten       | anheizen::anfeuern      |
| anbrüllen::anschreien| anhängen::anheften     | anfechten::angreifen      | anwerfen::anschmeifen   |
| auftupfen::auftröpfen| aufdrehen::aufzupulen  | aufreiben::aufspüren      | aufwirbeln::aufführen   |
| aufschnuppern::aufessen| aufnotzen::aufstylen  | aufkreuzen::auftauchen    | aufbrumen::aufandonnern |
| aufritzen::aufschlimmen| aufwählen::aufgraben  | auferlegen::aufbrumen     | aufkeimen::aufblühen    |
| ausrupfen::ausjäten  | ausposaumen::ausplaudern| ausfeilen::ausbrüten      | ausweinen::ausheulen    |
| ausschauen::ausätmen | ausaugen::auspumpen   | ausstecken::ausbrensen    | auskochen::ausbrüten    |

3 Representations of BV–PV Analogies

The parallelogram in Figure 1 illustrates the (dis-)similarities between BVs and PVs that come into play when distinguishing between the four types of (non-)shifts in our dataset: Comp requires all four sides in the parallelogram to provide strong similarities; ShiftReg requires the BV$_i$–BV$_j$ and the PV$_i$–PV$_j$ sides to provide strong similarities, and both BV$_i$–PV$_j$ sides to provide strong dissimilarities, etc. An obvious option to address the classification of the BV–PV analogies is thus by relying on standard cosine scores, when representing the verbs in a distributional semantic model (DSM). The following paragraphs describe such a basic cosine-similarity model that we used as a baseline, as well as alternative features which we added as potentially salient regarding our task.

3.1 Basic Distributional Similarity Model

We created a basic DSM to represent all BVs and all PVs by using a corpus-derived 300-dimensional vector representation. As corpus resource we relied on DECOW14AX, a German web corpus containing 12 billion tokens (Schäfer and Bildhauer, 2012; Schäfer, 2015). The verb vectors were obtained by looking at all context words within a symmetrical window of size 3. We applied positive pointwise mutual information (PPMI) feature weighting together with singular value decomposition (SVD). Measuring the cosine similarities between the BVs and PVs as suggested by Figure 1 then represents our basic distributional similarity model containing four cosine values.

Figure 3 looks into cosine values across combinations of meaning shift categories. Figure 3 (a) shows box plots for BV-PV pairs in the two compositional categories vs. the meaning-shifted categories. It illustrates that BV-PV combinations with a meaning shift indeed have lower cosine values between BVs and PVs than BV-PV combinations without meaning shifts. The similarity between BVs is expected to be higher for the regularly shifted cases, where the base verbs have something in common, in contrast to the irregular cases. This is also confirmed, cf. Figure 3 (b).

3.2 Generalisation Models

Classes and clusters are powerful techniques to generalise over unseen or infrequent events. We therefore extended the basic similarity model by adding class label features for the four involved verbs. We compared three different classifications. (1) We used the 15 verb classes from GermaNet (Hamp and Feldweg, 1997; Kunze, 2000). For particle verbs not covered by GermaNet, we used the existing verbs as a seed set and applied a nearest-prototype (centroid) classifier to all other BVs and PVs, with a centroid for each of the 15 classes. Thus we were able to assign class labels to all verbs in our dataset. (2) For three out of our four particle types (ab, an, auf), we found existing manual semantic classifications with 9, 8 and 11 classes, respectively (Lechler and Roßdeutscher, 2009; Kliche, 2011; Springorum, 2011). To obtain class labels for all verbs, we applied the same
nearest-centroid technique as for the GermaNet classes. (3) We compared the two resource-based methods with an unsupervised k-Means clustering based on the verbs’ vector representations. Unlike the other methods, k-Means learns the centroids without manually defined seed assignments. We set the number of clusters to \( k = 10 \), as this granularity was similar to the manual classifications.

3.3 Affect Models

A BV–PV meaning shift often involves a change in emotion and/or sentiment. For example, while the BV *servieren* ‘to serve’ is perceived as rather neutral or slightly positive, the PV *abservieren* ‘to dump sb.’ has a clearly negative meaning and correlates with the emotion sadness. On the other hand, the BV *motzen* ‘to grumble’ is associated with a negative sentiment and the emotion anger, while its PV *aufmotzen* ‘to shine up, soup up’ indicates a positive change.

In a slightly different vein, non-literal word usage often correlates with the degree of abstractness of the word’s contexts (Turney et al., 2011; Tsvetkov et al., 2014; Köper and Schulte im Walde, 2016). For example, the PV *abschminken* with the BV *schminken* ‘to put on make-up’ has a literal, very concrete meaning (‘to remove make-up’) and also a shifted, very abstract non-literal meaning (‘to forget about something’).

We enriched the basic similarity model by integrating affective information from human-created lexicons. Since affective datasets are typically small-scale and mostly exist for English, we applied a cross-lingual approach (Smith et al., 2017) to learn a linear transformation that aligns monolingual vectors from two languages in a single vector space. We took off-the-shelf word representations\(^4\) for German and English that live in the same semantic space, learned a regression model based on the English data, and applied it to the German data by relying on findings from Köper and Schulte im Walde (2017), who showed that a feed-forward neural network obtained a high correlation with human-annotated ratings.

The procedure was applied to a range of affective norm datasets in isolation: The NRC Hashtag Emotion Lexicon (Mohammad and Kiritchenko, 2015) contains emotional ratings for 17k words; we used anger, disgust, fear, joy, and sadness.

Warriner et al. (2013) collected 14k ratings for valence and arousal. For concreteness, we relied on the 40k ratings from Brysbaert et al. (2014). Finally, we used the 10k ratings for happiness from Dodds et al. (2011). In total, we obtained nine affective values for 2.2 million words.\(^5\)

We added the affective features to our basic similarity model by first looking up the 9-dimensional affect vectors for all four verbs involved in an analogy, and then calculating for each of the four similarities in the analogy parallelogram (Figure 1) the element-wise differences between the nine affective dimensions of the respective two verbs, resulting in \( 4 \times 9 = 36 \) extra vector dimensions.

In addition to looking at the verbs’ affective values we also looked at the affect of the respective context words: For each verb we created a second 9-dimensional vector with average affective values across the 500 most associated context words, according to PPMI. With respect to the four verbs in the analogy, this resulted in another \( 4 \times 9 = 36 \) extra vector dimensions.

We further added affect information restricted to the common context words of the involved verbs (red and blue intersections in Figure 4): For each intersection of the two BVs and the two PVs as well as the two BV–PV combinations, we learned another 9-dimensional emotional centroid, now only based on the shared context words, and provided the element-wise differences between the two centroids as a feature.

4 Experiments on BV–PV Analogies

Two classification scenarios were implemented: a four-class distinction between our four shift categories (4-Classes), and a binary distinction between cases where both BV–PV pairs include a

\[ \text{PV}_1 \cap \text{PV}_2 - \text{BV}_1 \cap \text{BV}_2, \]
\[ \text{PV}_1 \cap \text{BV}_1 - \text{PV}_2 \cap \text{BV}_2. \]

Figure 4: Venn diagrams with intersections.

---

\(^4\)https://github.com/Babylonpartners/fastText_multilingual

\(^5\)These ratings are also available at www.ims.uni-stuttgart.de/data/pv-meaning-shift.
meaning shift vs. BV–PV pairs including cases of compositionality (Shift-vs-Comp).

We applied a supervised classification setting based on support vector machines (SVMs) with an RBF kernel (Chang and Lin, 2011), using 10-fold cross-validation. Next to the similarity, generalisation and affect features, we provided the particle type as a feature in all settings. Table 2 reports the results across feature sets. As evaluation metric we report accuracy and a macro-average (equally-weighted) f-score ($F_1$) over all classes.

| Feature Set                      | 4 Classes | Shift-vs-Comp |
|----------------------------------|-----------|---------------|
| Majority baseline                | 31.24 .12 | 60.29 .38     |
| Basic Sim                        | 40.73 .32 | 65.10 .60     |
| Sim+GermaNet                     | 43.36 .34 | 67.15 .59     |
| Sim+ManClass                     | 45.55 .36 | 69.05 .62     |
| Sim+k-Means                      | 52.99 .71 | 70.51 .66     |
| Affect (full)                    | 57.08 .44 | 76.49 .74     |
| Affect only verbs                | 47.73 .37 | 69.05 .65     |
| Affect only context              | 58.39 .45 | 78.54 .77     |
| Combination                      | 56.20 .44 | 77.08 .75     |

Table 2: Results for 4- and 2-class distinctions, reporting accuracy and macro-$F_1$.

All models perform significantly better than the majority baseline. In addition, the full and the context-only affective models perform significantly better than the similarity models with and without generalisation, even though the unsupervised k-Means clustering improves the basic similarity model significantly (Sim+k-Means). Finally, the context-only affective model outperforms the verb-only affective model. Interestingly, a combination of all features (Combination) does not perform better than the context-only affective model in isolation.

A leave-one-out classification using the best classifier Affect only context as starting point revealed that most performance (accuracy) is lost when removing the emotion fear (-2.77), followed by the emotion joy (-1.46) and arousal (-0.88). In contrast, features related to disgust showed no impact on the overall performance.

Figure 5 illustrates that we can spot changes in affect and emotion even on the verb level: For three BV–PV verb pairs with particle $ab$, it plots the nine affective and emotion ratings for both verbs, after rescaling to an interval of $[0, 10]$. In the compositional case (a) the PV is highly similar to the BV in all dimensions, creating roughly the same shape as the BV. In the shift cases (b) and (c), the PVs are less concrete and evoke less happiness and joy than the BVs, while they evoke more fear, anger and sadness in comparison to their BVs.

Figure 5: Changes in affect and emotion for one compositional and two shifted BV–PV pairs. The affect/emotion values are based on the top associated context words according to PPMI.

5 Conclusion

This paper presented a computational model of meaning shifts for German particle verbs. Relying on a novel dataset, we found that shifts were observed across all our four particle types, but the analogical case mentioned in previous corpus explorations only represented the smallest class overall (8.5%). SVM models successfully distinguished between shift categories, with verb classes boosting standard cosine similarity performance, and affective context features representing the most salient indicators.

6Significance relies on $\chi^2$ with $p < 0.001$. 

154
Acknowledgments

The research was supported by the DFG Collaborative Research Centre SFB 732.

References

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand generally known English Word Lemmas. Behavior Research Methods 64:904–911.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2:1–27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Peter Sheridan Dodds, Kameron D. Harris, Isabel M. Merrill, and Pamela S. Morgan. 1997. Figuring out figure out: Metaphor and the Semantics of English Verb-Particle Constructions. Cognitive Linguistics 8(4):327–357.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In Piotr Bańkowski, 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.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, GA, USA, pages 746–751.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using Hashtags to Capture Fine Emotion Categories from Tweets. Computational Intelligence 31(2):301–326.

Peter D. Turney. 2006. Similarity of Semantic Relations. Computational Linguistics 32(3):379–416.
Peter D. Turney. 2012. Domain and Functions: A Dual-Space Model of Semantic Relations and Compositions. *Journal of Artificial Intelligence Research* 44:533–585.

Peter D. 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*. Edinburgh, UK, pages 680–690.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods* 45(4):1191–1207.