Improving Zero-Shot-Learning for German Particle Verbs by using Training-Space Restrictions and Local Scaling

Maximilian Köper and **Sabine Schulte im Walde** and **Max Kisselew** and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Pfaffenwaldring 5B, 70569 Stuttgart, Germany
{koepermn,schulte,kisselmx,pado}@ims.uni-stuttgart.de

Abstract

Recent models in distributional semantics consider derivational patterns (e.g., $use \rightarrow use + ful$) as the result of a compositional process, where base term and affix are combined. We exploit such models for German particle verbs (PVs), and focus on the task of learning a mapping function between base verbs and particle verbs. Our models apply particle-verb motivated training-space restrictions relying on nearest neighbors, as well as recent advances from zero-shot-learning. The models improve the mapping between base terms and derived terms for a new PV derivation dataset, and also across existing derivation datasets for German and English.

1 Introduction

Lazaridou et al. (2013) were the first to apply distributional semantic models (DSMs) to the task of deriving the meaning of morphologically complex words from their parts. They relied on high-dimensional vector representations to model the derived term (e.g., *useful*) as a result of a compositional process that combines the meanings of the base term (e.g., *to use*) and the affix (e.g., *ful*). For evaluation, they compared the predicted vector of the complex word with the original, corpus-based vector.

More recently, Kisselew et al. (2015) put the task of modeling derivation into the perspective of zero-shot-learning: instead of using cosine similarities they predicted the derived term by learning a mapping function between the base term and the derived term. Once the predicted vector was computed, a nearest neighbor search was applied to validate if the prediction corresponded to the derived term. In zero-shot-learning the task is to predict novel values, i.e., values that were never seen in training. More formally, zero-shot-learning trains a classifier $f: X \rightarrow Y$ that predicts novel values for Y (Palatucci et al., 2009). It is often applied across vector spaces, such as different domains (Mikolov et al., 2013; Lazaridou et al., 2015).

The experiments by Kisselew et al. (2015) were performed over six derivational patterns for German (cf. Table 1), including particle verbs (PVs) with two different particle prefixes (*an* and *durch*), which were particularly difficult to predict. PVs such as *anfangen* (to start) are compositions of a base verb (BV) such as *fangen* (to catch) and a verb particle such as *an*. Predicting PV meaning is challenging because 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; Kliche, 2011; Springorum, 2011). Furthermore, the particles often trigger meaning shifts when they combine with base verbs (Springorum et al., 2013b), so the resulting PVs represent frequent cases of non-literal meaning.

In this paper, we focus on predicting the meanings of German PV derivations. Our models provide two contributions to the research field of predicting derivations: (i) We suggest a novel idea of restricting the available training data, which has a positive impact on the mapping quality. (ii) We integrate a correction method for popular nearest neighbors into our models, so-called *hubs* (Radovanović et al., 2010), to improve the prediction quality.

*Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*, pages 91–96,
Berlin, Germany, August 11-12, 2016.

POS	Affix	Example	Inst.
adj/adj	un-	sagbar - unsagbar	80
adj/adj	anti-	religiös - antireligiös	80
noun/noun	-in	Bäcker - Bäckerin	80
noun/noun	-chen	Schiff - Schiffchen	80
verb/verb	an-	backen - anbacken	80
verb/verb	durch-	sehen - durchsehen	80

Table 1: German dataset (Kisselew et al., 2015).

POS	Affix	Example	Inst.
verb/verb	auf-	nehmen - aufnehmen	171
verb/verb	ab-	setzen - absetzen	287
verb/verb	mit-	streiken - mitstreiken	216
verb/verb	ein-	laufen - einlaufen	185
verb/verb	zu-	drücken - zudrücken	50
verb/verb	an-	legen - anlegen	221
verb/verb	aus-	malen - ausmalen	280

Table 2: New German PV derivation dataset.

2 Prediction Experiments

As in Kisselew et al. (2015), we treat every derivation type as a specific learning problem: we take a set of word pairs with a particular derivation pattern (e.g., "-in", Bäcker::Bäcker**in**), and divide this set into training and test pairs by performing 10-fold cross-validation. For the test pairs, we predict the vectors of the derived terms (e.g., $\overrightarrow{\textit{Bäckerin}}$). The search space includes all corpus words across parts-of-speech, except for the base term. The performance is measured in terms of recall-out-of-5 (McCarthy and Navigli, 2009), counting how often the correct derived term is found among the five nearest neighbors of the predicted vector.

2.1 Derivation Datasets

We created a new collection of German particle verb derivations[1] relying on the same resource as Kisselew et al. (2015), the semi-automatic derivational lexicon for German *DErivBase* (Zeller et al., 2013). From DErivBase, we induced all pairs of base verbs and particle verbs across seven different particles. Non-existing verbs were manually filtered out. In total, our collection contains 1410 BV–PV combinations across seven particles, cf. Table 2.

In addition, we apply our models to two existing collections for derivational patterns, the German dataset from Kisselew et al. (2015), comprising six derivational patterns with 80 instances each (cf. Table 1), and the English dataset from Lazaridou et al. (2013), comprising 18 derivational patterns (3 prefixes and 15 suffixes) and 7449 instances (cf. Table 3).

POS	Affix	Example	Inst.
verb/adj	-able	believe - believable	227
noun/adj	-al	doctor - doctoral	295
verb/noun	-er	repeat - repeater	874
noun/adj	-ful	use - useful	103
noun/adj	-ic	algorithm - algorithmic	330
verb/noun	-ion	erupt - eruption	687
noun/noun	-ist	drama - dramatist	294
adj/noun	-ity	accessible - accessibility	422
noun/verb	-ize	cannibal - cannibalize	155
noun/adj	-less	word - wordless	172
adj/adv	-ly	diagonal - diagonally	1,897
verb/noun	-ment	equip - equipment	215
adj/noun	-ness	empty - emptiness	652
noun/adj	-ous	religion - religious	207
noun/adj	-y	sport - sporty	454
adj/adj	in-	dispensable - indispensable	151
verb/verb	re-	write - rewrite	136
adj/adj	un-	familiar - unfamiliar	178

Table 3: English dataset (Lazaridou et al., 2013).

2.2 Word Embedding Vectors

We relied on the German and English *COW* web corpora[2] (Schäfer and Bildhauer, 2012) to obtain vector representations. The corpora contain 20 billion words and 9 billion words, respectively. We parsed the corpora using state-of-the-art pipelines integrating the MarMoT tagger and the MATE parser (Müller et al., 2013; Bohnet, 2010), and induced window co-occurrences for all corpus lemma–POS pairs and co-occurring nouns, verbs and adjectives in a 5-lemma window. We then created 400-dimensional word representations using the *hyperwords* toolkit (Levy et al., 2015), with context distribution smoothing of 0.75 and positive point-wise mutual information weighting together with singular value decomposition. The resulting vector space models contain approximately 460 000 lemmas for German and 240 000 lemmas for English.

2.3 Prediction Methods

2.3.1 Baseline

A baseline method that simply guesses the derived term has a chance of approx. $\frac{1}{460000}$ for German and $\frac{1}{240000}$ for English to predict the correct term. We thus apply a more informed baseline, the same as in Kisselew et al. (2015), and

[1] The dataset is available from http://www.ims.uni-stuttgart.de/data/pv-deriv-dataset/.

[2] http://corporafromtheweb.org

predict the derived term at exactly the same position as the base term.

2.3.2 Additive Method (AvgAdd)

AvgAdd is a re-implementation of the best method in Kisselew et al. (2015):[3] For each affix, the method learns a difference vector by computing the dimension-wise differences between the vector representations of base term A and derived term B. The method thus learns a centroid $\vec{c}$ for all relevant training pairs (N) with the same affix:

$$\vec{c} = \frac{1}{N} \sum_{i=0}^{n} (B_i - A_i) \qquad (1)$$

For each PV test instance with this affix, the learned centroid vector is added dimension-wise to the vector representation of the base term to predict a position for the derived term.

2.3.3 Restricting the Training Space (BestAdd)

Avg-Add learns a vector representation based on the full available training data for each derivational pattern. In this paper, we suggest a method *BestAdd*$_k$ that restricts the training items of a given base term to those BV–PV training instances that include the k nearest base verbs (using $k = 1,3,5$) according to their *cosine*. The motivation for our adjusted method relies on the observation that particles are very ambiguous and thus differ in their meanings across particle verbs. For example, the meanings of 'an' include a directed contact as in *sprechen::ansprechen* (to speak/to speak to s.o.) and in *schreiben::anschreiben* (to write/to write to s.o.), and also a start of an action as in *spielen::anspielen* (to play/to start playing) and in *stimmen::anstimmen* (to pitch/to start singing). We assume that base verbs that are distributionally similar also behave in a similar way when combined with a specific particle, and that a more restricted training set that is however specified for BV semantics outperforms a larger training set across wider BV meanings.

2.3.4 3CosMul

We also re-implemented *3CosMul* (Levy and Goldberg, 2014), a method that has been proven successful in solving analogy tasks, such as man (A) is to $king$ (B) as $woman$ (C) is to $queen$ (D). *3CosMul* does not explicitly predict a position in space but selects a target D in space that is close to B and C but not close to A. We applied *3CosMul* by always using the most similar training instance (as for *BestAdd* with $k = 1$).

2.4 Local Scaling

All methods introduced in the previous section perform a nearest neighbor search at the predicted position. We suggest to improve the prediction quality at this stage by mitigating the hubness problem (Dinu et al., 2015). *Hubs* are objects in vector space that are likely to appear disproportionately often among nearest neighbors, without necessarily being semantically related. Hubness has been shown an intrinsic problem of high-dimensional spaces (Tomasev, 2014). In order to reduce hubness, three unsupervised methods to re-scale the high-dimensional distances have been proposed (Schnitzer et al., 2014): local scaling, global scaling, and shared nearest neighbors. We focus on a local scaling (LS) type of hubness-correcting distance measure, namely the non-iterative contextual measure NI (Jégou et al., 2007):

$$NI(x, y) = \frac{d_{xy}}{\sqrt{\mu_x \cdot \mu_y}} \qquad (2)$$

NI relies on the average distance μ of x and y to their k nearest neighbors. It increases the similarity between x and y in cases where we observe low average similarities between x, y and its k nearest neighbors. Intuitively, if a word x is not even close to its nearest neighbors but comparably close to y then we increase the similarity between x and y.

For *3CosMul*, we adapt local scaling by scaling over the neighborhood information for all four parts (A, B, C and D) in the analogy:

$$\textit{3CosMul+LS (D)} = \frac{\textit{3CosMul(D)}}{\sqrt[4]{\mu_A \cdot \mu_B \cdot \mu_C \cdot \mu_D}}$$

3 Results

3.1 *BestAdd* and Local Scaling

Table 4 presents macro-averaged recall-out-of-5 scores, giving equal weight to each derivation regardless of the number of instances. Across the three datasets, the default results (i.e., without local scaling) obtained with our novel method

[3] We also conducted experiments with the least-squares error objective method *LexFun* but the results were clearly inferior to the *AvgAdd* method.

Method	**Particle Verbs (DE)**		**Kisselew (DE)**		**Lazaridou (EN)**	
	Default	+ NI_{15}	Default	$+\text{NI}_{15}$	Default	+ NI_{15}
Baseline	10.79%		16.08%		15.36%	
AvgAdd	11.82%	+1.28%	24.26%	+3.14%	24.19%	+2.95%
BestAdd_1	10.22%	+1.19%	33.91%	+3.97%	27.32%	+1.87%
BestAdd_3	**14.26%**	**+2.24%**	**38.50%**	**+4.17%**	37.06%	+1.40%
BestAdd_5	14.44%	+1.97%	38.07%	+4.61%	**38.49%**	**+2.12%**
3CosMul	10.06%	-0.73%	33.91%	+ 1.04%	27.88%	+0.90%

Table 4: Macro-averaged recall-out-of-5 across methods, with and without local scaling NI_{15}.

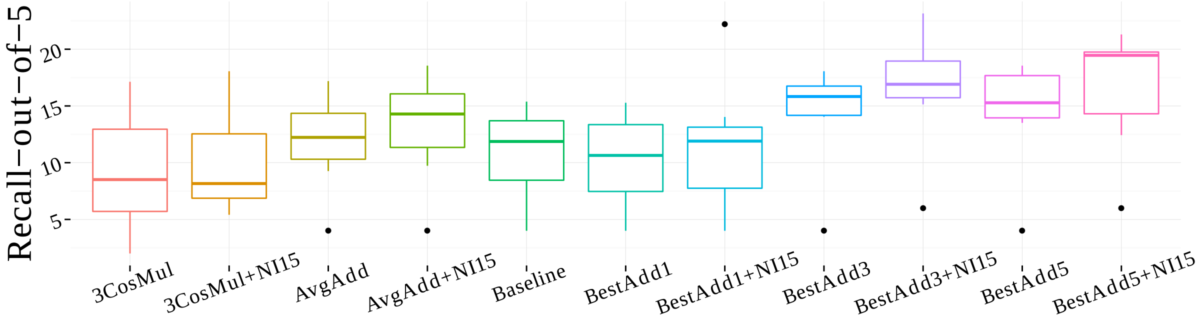


Figure 1: Recall-out-of-5 results across methods, for the German PV derivation dataset.

BestAdd (with $k = \{3,5\}$) are significantly[4] above *AvgAdd* ($p < 0.01$), the previously best method for the existing German and English datasets. *BestAdd* with $k = 1$ and *3CosMul* perform at a similar level than *AvgAdd*, but for our new PV derivation dataset do not even outperform the baseline. Restricting the training process to a small selection of nearest neighbors therefore has a positive impact on the prediction quality.

Furthermore, local scaling relying on $k = 15$ nearest neighbors (NI_{15}) improves the prediction results in all but one cases. These improvements are however not significant.

The results in Table 4 also demonstrate that predicting particle verbs is the most challenging derivation task, as the results are significantly lower than for the other two datasets. Figure 1 once more illustrates the recall-out-of-5 results for our new PV dataset. In the following, we zoom into dataset derivation types.

3.2 Improvement across Derivation Types and Languages

Figures 2 to 4 break down the results from Table 4 across the German and English derivation types.

The blue bars show the *BestAdd*$_3$ results, and the green stacked bars represent the additional gain using local scaling (NI_{15}). The yellow points correspond to baseline performance, and the dotted black lines to the *AvgAdd* results.

We can see that *BestAdd*$_3$ not only outperforms the previously best method *AvgAdd* on average but also for each derivation type. Also, local scaling provides an additional positive impact for all but one particle type in German, *ab-*, and for all but three derivation types in English, *-able, -al, -less.*

At the same time, we can see that the impact of local scaling is different across derivation types. For example, looking into the data we observe that *mit* PVs are often wrongly mapped to nouns, and *BestAdd* and local scaling correct this behavior: The nearest neighbors of the verb *erledigen* (to manage sth.) with *BestAdd*$_3$ are *Botengang* (errand), *Haushaltsarbeit* (domestic work), *Hausmeisterarbeit* (janitor work), and further six compounds with the nominal head *Arbeit* (work). Additional local scaling predicts the correct PV *miterledigen* (to manage sth. in addition) as second nearest neighbor.

[4] Significance relies on χ^2.

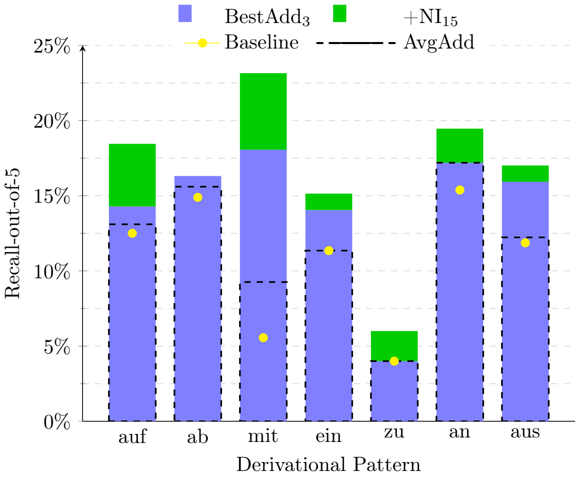


Figure 2: Performance gain across particle types.

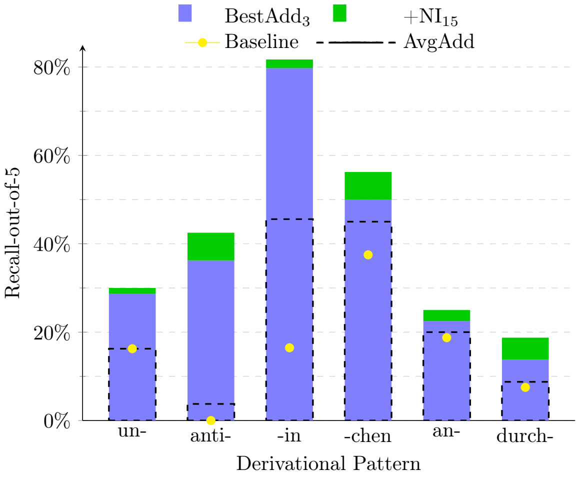


Figure 3: Performance gain for derivation types in Kisselew et al. (2015).

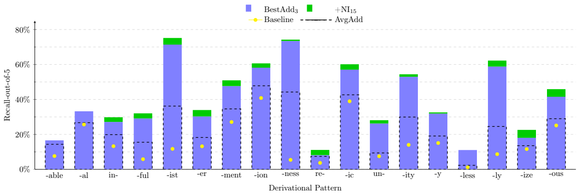
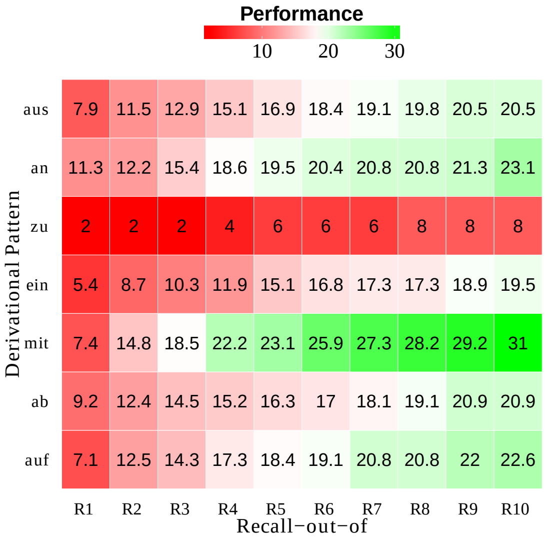


Figure 4: Performance gain for derivation types in Lazaridou et al. (2013).

3.3 Recall-out-of-x across Particle Types

Figure 5 focuses on the particle types, but varies the strength of the evaluation measure. Relying on *BestAdd*$_3$ with local scaling NI$_{15}$, we apply recall-out-of-x with $x \in [1, 10]$. With one exception (*zu*), all particle types achieve a performance of 15-23% for recall-out-of-5, so *zu* had a negative impact on the average score in Table 4. Looking at recall-out-of-10, the performances go up to 20-30%. While PVs with the rather non-ambiguous *mit* are again modeled best, also PVs with strongly ambiguous particles (such as *an* and *auf*) are modeled well.



	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
aus	7.9	11.5	12.9	15.1	16.9	18.4	19.1	19.8	20.5	20.5
an	11.3	12.2	15.4	18.6	19.5	20.4	20.8	20.8	21.3	23.1
zu	2	2	2	4	6	6	6	8	8	8
ein	5.4	8.7	10.3	11.9	15.1	16.8	17.3	17.3	18.9	19.5
mit	7.4	14.8	18.5	22.2	23.1	25.9	27.3	28.2	29.2	31
ab	9.2	12.4	14.5	15.2	16.3	17	18.1	19.1	20.9	20.9
auf	7.1	12.5	14.3	17.3	18.4	19.1	20.8	20.8	22	22.6

Figure 5: Recall-out-of-[1,10] across particles.

4 Conclusion

We suggested two ways to improve the prediction of derived terms for English and German. Both (i) particle-verb motivated training-space restrictions and (ii) local scaling to address hubness in high-dimensional spaces had a positive impact on the prediction quality of derived terms across datasets. Particle-specific explorations demonstrated the difficulty of this derivation, and differences across particle types.

Acknowledgments

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

References

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.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of the International Conference on Learning Representations, Workshop Track*, San Diego, CA, USA.

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.

Hervé Jégou, Hedi Harzallah, and Cordelia Schmid. 2007. A contextual dissimilarity measure for accurate and efficient image search. In *Proceedings of the Conference on Computer Vision & Pattern Recognition*, pages 1–8, Minneapolis, MN, USA.

Max Kisselew, Sebastian Padó, Alexis Palmer, and Jan Šnajder. 2015. Obtaining a better understanding of distributional models of German derivational morphology. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 58–63, London, UK.

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

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1517–1526, Sofia, Bulgaria.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 270–280, Beijing, China.

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 and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the 18th Conference on Computational Natural Language Learning*, pages 171–180, Baltimore, MD, USA.

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.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

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.

Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems 22*, pages 1410–1418.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531.

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.

Dominik Schnitzer, Arthur Flexer, and Nenad Tomasev. 2014. A case for hubness removal in high-dimensional multimedia retrieval. In *Advances in Information Retrieval - 36th European Conference on IR Research*, pages 687–692.

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, Leuven, Belgium.

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.

Nenad Tomasev. 2014. *The Role Of Hubness in High-dimensional Data Analysis*. Ph.D. thesis.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1201–1211, Sofia, Bulgaria.