Obtaining a Better Understanding of Distributional Models of German Derivational Morphology

Max Kisselew* Sebastian Padó* Alexis Palmer* Jan Šnajder†
*Institut für maschinelle Sprachverarbeitung, Stuttgart University, Germany
†Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
{kisselmx,pado,palmeras}@ims.uni-stuttgart.de jan.snajder@fer.hr

Abstract

Predicting the (distributional) meaning of derivationally related words (*read* / *read+er*) from one another has recently been recognized as an instance of distributional compositional meaning construction. However, the properties of this task are not yet well understood. In this paper, we present an analysis of two such composition models on a set of German derivation patterns (e.g., *-in*, *durch-*). We begin by introducing a rank-based evaluation metric, which reveals the task to be challenging due to specific properties of German (compounding, capitalization). We also find that performance varies greatly between patterns and even among base-derived term pairs of the same pattern. A regression analysis shows that semantic coherence of the base and derived terms within a pattern, as well as coherence of the semantic shifts from base to derived terms, all significantly impact prediction quality.

1 Introduction

Derivation is a major morphological process of word formation (e.g., *read* → *read+er*), which is typically associated with a fairly specific *semantic shift* (*+er*: agentivization). It may therefore be surprising that the semantics of derivation is a relatively understudied phenomenon in distributional semantics. Recently, Lazaridou et al. (2013) proposed to consider the semantics of a derived term like *read+er* as the result of a compositional process that combines the meanings of the base term *read* and the affix *+er*. This puts derivation into the purview of *compositional distributional semantic models* (CDSMs). CDSMs are normally used to compute the meaning of phrases and sentences by combining distributional representations of the individual words. A first generation of CDSMs represented all words as vectors and modeled composition as vector combination (Mitchell and Lapata, 2010). A second generation represents the meaning of predicates as higher-order algebraic objects such as matrices and tensors (Baroni and Zamparelli, 2010; Coecke et al., 2010), which are combined using various composition operations.

Lazaridou et al. predict vectors for derived terms and evaluate their approach on a set of English derivation patterns. Building on and extending their analysis, we turn to German derivation patterns and offer both qualitative and quantitative analyses of two composition models on a state-of-the-art vector space, with the aim of better understanding where these models work well and where they fail. Our contributions are as follows. First, we perform all analyses in parallel for six derivation patterns (two each for nouns, adjectives, and verbs). This provides new insights, as we can cross-reference results from individual analyses. Secondly, we evaluate using a rank-based metric, allowing for better assessment of the practical utility of these models. Thirdly, we construct a regression model that is able to explain performance differences among patterns and word pairs in terms of differences in *semantic coherence*.

2 Modeling Derivation as Meaning Composition

Morphological derivation is a major morphological process of word formation that combines a *base term* with functional morphemes, typically a single affix, into a *derived term*, and may additionally involve

Proceedings of the 11th International Conference on Computational Semantics, pages 58–63,
London, UK, April 15-17 2015.

stem changes. In contrast to inflection, derivation produces new lexical items, and it is distinct from composition, which combines two bases. Derivation comprises a large number of distinct *patterns*. Some cross part-of-speech boundaries (nominalization, verbalization, adjectivization), but many do not (gender indicators like *actor / actress* or (de-)intensifiers like *red / reddish*). In many languages, such as German or the Slavic languages, derivational morphology is extremely productive (Štekauer and Lieber, 2005).

Particularly relevant from a semantic perspective is that the meanings of the base and derived terms are often, but not always, closely related to each other. Consequently, derivational knowledge can be used to improve semantic processing (Luong et al., 2013; Padó et al., 2013). However, relatively few databases of derivational relations exist. CELEX (Baayen et al., 1996) contains derivational information for several languages, but was largely hand-written. A recent large-coverage resource for German, DErivBase (Zeller et al., 2013), covers 280k lemmas and was created from a rule-based framework that is fairly portable across languages. It is unique in that each base-derived lemma pair is labeled with a sequence of derivation patterns from a set of 267 patterns, enabling easy access to instances of specific patterns (cf. Section 3).

Compositional models for derivation. Base and derived terms are closely related in meaning. In addition, this relation is coherent to a substantial extent, due to the phenomenon of productivity. In English, for example, the suffix *-er* generally indicates an agentive nominalization (*sleep / sleeper*) and *un-* is a negation prefix (*well / unwell*). Though Mikolov et al. (2013) address some inflectional patterns, Lazaridou et al. (2013) were the first to use this observation to motivate modeling derivation with CDSMs. Conceptually, the meaning of the base term (represented as a distributional vector) is combined with some distributional representation of the affix to obtain a vector representing the meaning of the derived term. In their experiments, they found that the two best-motivated and best-performing composition models were the *full additive model* (Zanzotto et al., 2010) and the *lexical function model* (Baroni and Zamparelli, 2010). Botha and Blunsom (2014) use a related approach to model morphology for language modeling.

The additive model (ADD) (Mitchell and Lapata, 2010) generally represents a derivation pattern p as a vector computed as the shift from base term vector $\boldsymbol{b}$ to the derived term vector $\boldsymbol{d}$, i.e., $\boldsymbol{b} + \boldsymbol{p} \approx \boldsymbol{d}$. Given a set of base-derived term pairs (b, d) for p, the best $\hat{\boldsymbol{p}}$ is computed as the average of the vector difference, $\hat{\boldsymbol{p}} = \frac{1}{N}\sum_i(\boldsymbol{d_i} - \boldsymbol{b_i})$.[1] The lexical function model (LEXFUN) represents the pattern as a matrix $\boldsymbol{P}$ that encodes the linear transformation that maps base onto derived terms: $\boldsymbol{Pb} \approx \boldsymbol{d}$. The best matrix $\hat{\boldsymbol{P}}$ is typically computed via least-squares regression between the predicted vectors $\hat{\boldsymbol{d}}_i$ and the actual vectors $\boldsymbol{d}_i$.

3 Experimental Setup

Distributional model. We build a vector space from the SdeWaC corpus (Faaß and Eckart, 2013), part-of-speech tagged and lemmatized using TreeTagger (Schmid, 1994). To alleviate sparsity arising from TreeTagger's lexicon-driven lemmatization, we back off for unrecognized words to the MATE Tools (Bohnet, 2010), which have higher recall but lower precision than TreeTagger. We also reconstruct lemmas for separated prefix verbs based on the MATE dependency analysis. Finally, we get a word list with 289,946 types (content words only). From the corpus, we extract lemmatized sentences and train a state-of-the art predictive model, namely CBOW (Mikolov et al., 2013). This model builds distributed word vectors by learning to predict the current word based on a context. We use lemma-POS pairs as both target and context elements, 300 dimensions, negative sampling set to 15, and no hierarchical softmax.

Selected patterns and word pairs. We investigate six derivation patterns in German and the word pairs associated with them in DErivBase (see Table 1). We consider only patterns where base and derived terms have the same POS, and we prefer patterns encoding straightforward semantic shifts. Such patterns tend to encode meaning shifts without corresponding argument structure changes; thus they are represented appropriately in composition models based on purely lexical vector spaces. Per pattern, we randomly select 80 word pairs for which both base and derived lemmas appear at least 20 times in SdeWaC.[2]

[1] Lazaridou et al. (2013) use a slightly different formulation of the additive model. We experimented with both theirs and the standard version of the additive model. Since we obtained best results with the latter, we use the standard version.

[2] We replace a small number of erroneous pairs (e.g., *Log* → *Login* for NN02) found by manual inspection.

ID	Pattern	Sample word pair	English translation	BL	ADD	LEXFUN
AA02	*un-*	*sagbar → unsagbar*	*sayable → unspeakable*	42.5% (.46)	41.25% (.49)	18.75% (.31)
AA03	*anti-*	*religiös → antireligiös*	*religious → antireligious*	7.5% (.51)	37.5% (.58)	47.5% (.58)
NN02	*-in*	*Bäcker → Bäckerin*	*baker → female baker*	35.0% (.56)	66.25% (.65)	26.25% (.51)
NN57	*-chen*	*Schiff → Schiffchen*	*ship → small ship*	20.0% (.55)	28.75% (.57)	15.0% (.49)
VV13	*an-*	*backen → anbacken*	*to bake → to stick, burn*	18.75% (.43)	18.75% (.43)	5% (.27)
VV31	*durch-*	*sehen → durchsehen*	*to see → to peruse*	3.75% (.40)	7.5% (.40)	1.25% (.27)
Mean				21.25% (.49)	33.33% (.52)	18.96% (.41)

Table 1: Derivation patterns, representative examples (and translations), and prediction performance in terms of R_{oof} percentages and mean similarity between derived and gold vectors, 10-fold cross-validation.

Experimental design and baseline. We experiment with the two composition models described in Section 2 (ADD and LEXFUN) as implemented in the DISSECT toolkit (Dinu et al., 2013). As baseline (BL), again following Lazaridou et al. (2013), we predict the base term of each word pair as the derived term. With six derivation patterns, our investigation thus includes 18 experiments. In each experiment, we perform 10-fold cross-validation on the 80 word pairs for each pattern.

All these models predict some point in vector space for the derived term, and we compare against the gold standard position of the derived term with cosine similarity. Like Lazaridou et al. (2013), we consider this average similarity directly, but believe that it is not informative enough since it does not indicate concretely how many correct derivations are found. Therefore, we adopt as our primary evaluation metric the R_{oof} (*Recall out of five*) metric proposed by McCarthy and Navigli (2009) for lexical substitution. It counts how often the correct derived term is found among the five nearest neighbors of the prediction (selected from all words of the same POS). R_{oof} is motivated by rank-based evaluation metrics from IR (such as Precision at n), but our setup differs in that there can be at most one true positive in each list.

4 Results and Discussion

Global observations. Table 1 shows R_{oof} performance and mean similarities, pattern-by-pattern, of the two composition models (ADD and LEXFUN) and the baseline. Measured by R_{oof} score, ADD strongly outperforms BL for four patterns; for the other two, it achieves (nearly-)equivalent performance. LEXFUN, on the other hand, beats BL for one pattern (AA03) and in all other cases is much worse. ADD outperforms LEXFUN for all but one pattern. A comparison of R_{oof} and mean similarity indicates that similarity alone is not a good indicator of how reliably a model will include the actual derived vector in the nearest neighbors of its prediction. This validates our call for a more NLP-oriented evaluation.

The mean similarities are sufficient to make some comparisons across languages, though. Lazaridou et al. (2013) find that both additive and lexical function models yield higher mean similarities than the baseline. For our German data, this is true only for ADD. This shows that the semantic shifts underlying derivation patterns are, to some extent, expressible as vector addition in the CBOW space, while it is more difficult to capture them as a lexical function. The overall worse performance is, in our view, related to some specific characteristics of German. First, due to the general capitalization of nouns, named entities are not orthographically recognizable. Consequently, for *Strauß* (*bouquet*), BL and ADD return terms related to the composers Richard (e.g., *Alpensinfonie*) or Johann (e.g., *Walzerkönig* (waltz king)) Strauss. Secondly, nominal compounds introduce higher sparsity and more confounders. For example, for the derived term *Apfelbäumchen* (*~apple treelet*), LEXFUN's closest returned neighbor is the noun *Bäumchen*, which is a case of *combined* derivation and composition, yet is counted as incorrect. In English, compounds such as *apple tree(let)* are considered neither as base nor as potential derived terms.

Semantic coherence appears to be an important determinant of prediction quality. The best-performing pattern for ADD is NN02, the gender affix *-in* (turning masculine into feminine nouns), which applies to fairly coherent classes of people (nationalities, roles, and professions). We see a similar effect for the

Norweger* → *Norweger+in (*male* → *female Norwegian*)		NN02	***pluralistisch* → *anti+pluralistisch*** (*pluralistic* → *antipluralistic*)		AA03
BL	ADD	LEXFUN	BL	ADD	LEXFUN
1. Norweger	*Norweger*	*Schwed+in*	*1. pluralistisch*	*pluralistisch*	*anti+demokratisch*
2. Däne	*Schwed+in*	*Australier+in*	*2. plural*	*plural*	*anti+liberal*
3. Schwede	***Norweger+in***	***Norweger+in***	*3. demokratisch*	*demokratisch*	*anti+modernistisch*
4. Isländer	*Däne*	*Dän+in*	*4. säkular*	*anti+totalitär*	***anti+pluralistisch***
5. Solberg	*Dän+in*	*Landsfrau*	*5. freiheitlich*	*säkular*	*anti+modern*

Table 2: Five nearest neighbors to the predicted vector for the derived term. Correct derived term appears in bold; + marks instances of the relevant derivational affix.

best-performing pattern for LEXFUN, AA03 (the adjectival affix *anti-*). While the base term meanings of this pattern vary quite a bit, the meanings of the derived terms are distributionally coherent, with many shared typical context words (*demonstration*, *protest*, etc.). In contrast, the more difficult diminutive affix *-chen* (NN57) can be applied to nearly any noun, leading to less coherent sets of base and derived terms.

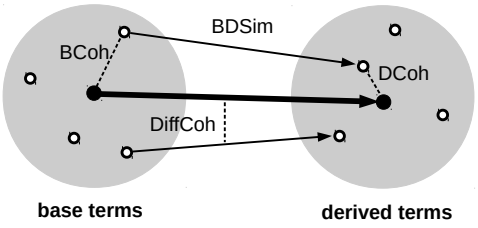
Both models have the most difficulty with verb-verb patterns. Analysis of word pairs for patterns VV13 and VV31, both of which are *prefix verbs*, indicates that the semantic shifts expressed by verb prefixation are often highly ambiguous (Lechler and Roßdeutscher, 2009) and therefore difficult to capture.

Nearest-neighbor analysis. We next examined the five nearest neighbors (5-NN) produced by the different models, as shown in Table 2. Interestingly, ADD and LEXFUN appear to capture different aspects of the semantic composition. The neighbors of LEXFUN are not random, despite its bad R_{oof} performance, but largely composed of morphologically complex words combining semantically related bases with the derivational affix in question. ADD generally finds terms that are semantically related to the base, both morphologically simple and complex. Analysis of the 5-NNs for each model confirms these impressions: (a), ADD always returns the base term as the nearest neighbor. That is, in an evaluation considering only the top neighbor, BL and ADD would have identical performance; (b), 54% of the LEXFUN 5-NNs are other derived terms of the same derivation pattern. This is quite high in comparison to BL (8%) and ADD (14%); (c), both ADD (48%) and BL (46%) are much more likely than LEXFUN (11%) to return neighbors that include the base term or its stem, as *plural (pluralistic)* above.

Regression analysis of ADD. Finally, we perform a quantitative analysis of the performance of ADD, our best model. Based on the discussion above, our hypothesis is that the additive model works best for patterns that are *coherent*. We operationalize this by defining four *coherence features* at the word pair level (cf. Figure 1): (a) Coherence of base terms (*BCoh*) – cosine similarity between a base term and the centroid of all base terms of the pattern; (b) Coherence of derived terms (*DCoh*) – cosine similarity between a derived term and the centroid of all derived terms of the pattern; (c) Similarity between base and derived term (*BDSim*) – cosine similarity between the two vectors of the base and derived terms; (d) Coherence of difference vectors (*DiffCoh*) – cosine similarity between the difference vector for a base-derived term pair (its semantic shift) and the centroid of all difference vectors of the pattern.

For our analysis, we use a mixed effects logistic regression (MELR, Jaeger (2008)). Logistic regression is used in linguistics to investigate quantitative relationships (Bresnan et al., 2007). It predicts the probability of a binary response variable y depending on a set of predictors $\mathbf{x}$ as $P(y = 1) = \sigma(\mathbf{bx})$ where σ is the sigmoid function. Its coefficients are interpretable as *log odds*: given a positive coefficient of size b_i, every one-unit increase of x_i increases the odds of $y = 1$ by a factor of e^{b_i}; correspondingly, negative values increase the odds of $y = 0$. MELR is a generalization that distinguishes traditional *fixed effects* from a novel category of predictors, so-called *random effects*, $\mathbf{x}'$, so that $P(y = 1) = \sigma(\mathbf{bx} + \mathbf{cx'})$. The coefficients $\mathbf{c}$ of random effects are drawn from a normal distribution with zero mean, which is appropriate for many predictors (Clark, 1973) and makes the model generalizable to unseen values.

In our MELR model, each word pair is a datapoint. We use 0/1 (success in the R_{oof} evaluation) as y, the coherence features as fixed effects, and the pattern identity as random effect. The resulting coefficients are shown in Figure 1. The negative intercept results from the overall predominance of failures. The next



Feature name	Coefficient	p-value
Intercept	-15.0	<0.0001
BCoh	-4.6	<0.01
DCoh	$+2.0$	n.s.
BDSim	$+6.7$	<0.0001
DiffCoh	$+26.8$	<0.0001

Figure 1: Illustration of coherence in vector space (left) and regression coefficients (right)

two features are somewhat surprising: We find a negative coefficient for BCoh, indicating semantically more coherent base terms are correlated with more difficult derivation patterns. Indeed, the most difficult pattern for the model (VV31) has the highest average base term coherence (0.40), and the simplest pattern (NN02) the lowest (0.29). DCoh, the coherence among derived terms, does have a positive coefficient, but it is too small to reach significance. We tend to see these two results as artefacts of our small sample.

The remaining two features (BDSim, and DiffCoh) have strong positive coefficients, indicating that patterns where (a) base terms and derived terms are similar within pairs, or (b) the difference vectors all point into the same direction, are easier to model. The last feature is particularly strong – not surprising, given that ADD uses the centroid of the difference vectors to make predictions. Finally, the random effect coefficients of the patterns are small (between -0.7 and $+0.7$), indicating the model's robustness.

We also evaluate the regression model by predicting the R_{oof} percentages from Table 1, simply counting the number of successes for each pattern. The mean difference to the actual numbers is 2.5%. The highest individual difference is 3.75 (32.5 vs. 28.75 for NN57), the lowest 0% (for NN02). This means that the regression model does quite a good job at predicting top-5 accuracy. We take this as evidence that the features' coefficients capture a relevant and substantial aspect of the phenomenon.

5 Conclusion

In this paper, we have analyzed compositional distributional models for predicting the meaning of derived terms from their base terms with a focus on in-depth analysis. We found that this prediction task is challenging, at least for the derivation patterns we considered. This may not be surprising, given the relatively subtle semantic differences introduced by some patterns (e.g., the gender of the term, or the polarity), which may be hard to recover distributionally. In that sense, our choice of (putatively easy) within-POS derivations may actually have worked against us: in cross-POS derivations, base and derived terms should have more clearly distinguished distributional profiles. At any rate, it seems that additional modeling efforts are necessary to produce more robust models of derivational morphology.

We believe that two results of our analyses are particularly noteworthy. The first is the correlation between the coherence of the derivation patterns and the performance of the additive composition model. While the existence of such correlations may seem obvious given the way the additive model works, we hope, on account of their strength, that we can predict the difficulty of modeling novel derivations, as well as link this approach to theoretical work on (ir-)regularity (Plank, 1981). The second result is the complementarity of the additive and lexical function models, which capture the base meaning and the affix meaning well, respectively. This suggests combining the two models as an interesting avenue.

References

Baayen, H. R., R. Piepenbrock, and L. Gulikers (1996). *The CELEX lexical database. Release 2. LDC96L14*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baroni, M. and R. Zamparelli (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, Cambridge, MA, USA.

Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING*, Beijing, China.

Botha, J. A. and P. Blunsom (2014). Compositional morphology for word representations and language modelling. In *Proceedings of ICML*, Beijing, China.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (2007). Predicting the dative alternation. In *Cognitive Foundations of Interpretation*. Royal Netherlands Academy of Science.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior 12*(4).

Coecke, B., M. Sadrzadeh, and S. Clark (2010). Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis 36*.

Dinu, G., N. T. Pham, and M. Baroni (2013). DISSECT – DIStributional SEmantics Composition Toolkit. In *Proceedings of ACL*, Sofia, Bulgaria.

Faaß, G. and K. Eckart (2013). SdeWaC – a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web*, Lecture Notes in Computer Science. Springer.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs and toward logit mixed models. *Journal of Memory and Language 59*(4).

Lazaridou, A., M. Marelli, R. Zamparelli, and M. Baroni (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of ACL*, Sofia, Bulgaria.

Lechler, A. and A. Roßdeutscher (2009). German particle verbs with *auf-*. Reconstructing their composition in a DRT-based framework. *Linguistische Berichte 220*.

Luong, M.-T., R. Socher, and C. D. Manning (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, Sofia, Bulgaria.

McCarthy, D. and R. Navigli (2009). The English lexical substitution task. *Language Resources and Evaluation 43*(2).

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ, USA.

Mikolov, T., W.-T. Yih, and G. Zweig (2013). Linguistic regularities in continuous space word representations. In *Proceedings of HLT-NAACL*, Atlanta, GA.

Mitchell, J. and M. Lapata (2010). Composition in distributional models of semantics. *Cognitive Science 34*(8).

Padó, S., J. Šnajder, and B. Zeller (2013). Derivational smoothing for syntactic distributional semantics. In *Proceedings of ACL*, Sofia, Bulgaria.

Plank, F. (1981). *Morphologische (Ir-)Regularitäten. Aspekte der Wortstrukturtheorie*. Tübingen: Narr.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NeMLaP*, Manchester, UK.

Štekauer, P. and R. Lieber (Eds.) (2005). *Handbook of Word-Formation*, Volume 64 of *Studies in Natural Language and Linguistic Theory*. Springer.

Zanzotto, F. M., I. Korkontzelos, F. Fallucchi, and S. Manandhar (2010). Estimating linear models for compositional distributional semantics. In *Proceedings of COLING*, Beijing, China.

Zeller, B., J. Šnajder, and S. Padó (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of ACL*, Sofia, Bulgaria.