Predictability of Distributional Semantics in Derivational Word Formation

Sebastian Padó* **Aurélie Herbelot+** **Max Kisselew*** **Jan Šnajder†**
* Institute for Natural Language Processing, University of Stuttgart
{sebastian.pado,max.kisselew}@ims.uni-stuttgart.de
+ Center for Mind/Brain Sciences, University of Trento
{aurelie.herbelot}@unitn.it
† Faculty of Electrical Engineering and Computing, University of Zagreb
jan.snajder@fer.hr

Abstract

Compositional distributional semantic models (CDSMs) have successfully been applied to the task of predicting the meaning of a range of linguistic constructions. Their performance on semi-compositional word formation process of (morphological) derivation, however, has been extremely variable, with no large-scale empirical investigation to date. This paper fills that gap, performing an analysis of CDSM predictions on a large dataset (over 30,000 German derivationally related word pairs). We use linear regression models to analyze CDSM performance and obtain insights into the linguistic factors that influence how predictable the distributional context of a derived word is going to be. We identify various such factors, notably part of speech, argument structure, and semantic regularity.

1 Introduction

Compositional models of distributional semantics, or CDSMs (Mitchell and Lapata, 2010; Erk and Padó, 2008; Baroni et al., 2014; Coecke et al., 2010), have established themselves as a standard tool in computational semantics. Building on traditional distributional semantic models for individual words (Turney and Pantel, 2010), they are generally applied to *compositionally compute phrase meaning* by defining combination operations on the meaning of the phrase's constituents. CDSMs have also been co-opted by the deep learning community for tasks including sentiment analysis (Socher et al., 2013) and machine translation (Hermann and Blunsom, 2014). A more recent development is the use of CDSMs to model meaning-related phenomena above and below syntactic structure; here, the term "composition" is used more generally to apply to processes of meaning combination from multiple linguistic units, e.g., above and below syntactic structure. Above the sentence level, such models attempt to predict the unfolding of discourse (Kiros et al., 2015). Below the word level, CDSMs have been applied to model word formation processes like compounding (*church* + *tower* $\rightarrow$ *church tower*) and (morphological) derivation (Lazaridou et al., 2013) (*favor* + *able* $\rightarrow$ *favorable*). More concretely, given a distributional representation of a basis and a derivation *pattern* (typically an affix), the task of the CDSM is to predict a distributional representation of the derived word, without being provided with any additional information. Interest in the use of CDSMs in this context comes from the observation that derived words are often less frequent than their bases (Hay, 2003), and in the extreme case even completely novel; consequently, distributional evidence is often unreliable and sometimes unavailable. This is confirmed by Luong et al. (2013) who compare the performance of different types of word embeddings on a word similarity task and achieve poorer performance on data sets containing rarer and more complex words. Due to the Zipfian distribution there are many more rare than frequent word types in a corpus, which increases the need for methods being able to model derived words.

In this paper, we ask to what extent the application of CDSMs to model derivation is a success story. The record is unclear on this point: Lazaridou et al. (2013), after applying a range of CDSMs to an English derivation dataset, report success, while Kisselew et al. (2015) found very mixed results on German derivation and generally high variance across words and derivation patterns. The analyses in both studies

POS + ID	Pattern	Sample word pair	English translation
A → N 16	*+ität*	produktiv → Produktivität	productive → productivity
A → V 04	*(umlaut)*	kurz → kürzen	short → to shorten
N → A 26	*-ung +end*	Einigung → einigend	agreement → agreeing
N → V 07	*be+ +ig*	Ende → beendigen	end → to end
V → N 09	*(null)*	aufatmen → Aufatmen	to breathe → sigh of relief
V → V 14	*auf+*	holen → aufholen	to fetch → to catch up

Table 1: Examples of derivation patterns from DErivBase

were also limited in scope (cf. Section 2.2). Furthermore, from a linguistic point of view, it is not at all obvious that it is reasonable to model derivation as a fully compositional process, as CDSMs do. Indeed, the classic linguistic definition of derivation distinguishes it from inflection by appealing to its *semantic irregularity*: the meaning changes that accompany derivation are not supposed to be completely predictable (Plank, 1981; Laca, 2001; Plag, 2003; Dressler, 2005).

More specifically, our goal is to gain a more precise understanding of the linguistic factors that govern the success or failure of CDSMs to predict distributional vectors for derived words. To this end, we conduct a broad-coverage analysis of the performance of CDSMs on more than 30,000 German derivationally related word pairs instantiating 74 derivation patterns. As a first step, we build *CDSM prediction models* for each of these patterns. The second step is a linear regression analysis with linguistic properties of patterns and word pairs as predictors and the models' performances as dependent variables. We formulate and test a number of hypotheses about the linguistic properties and establish that, notably, derivations that create new argument structure are generally hard to predict – although the difficulty is mediated by the regularity of the semantic shift involved. Subsequently, we exploit the regression results to combine several state-of-the-art CDSMs into an ensemble. Unfortunately, we do not see improvements over the individual models, which we trace back to a lack of complementarity among the CDSMs.

2 Background: Modeling Morphological Derivation

2.1 Derivational Lexicons

Morphological derivation is a word formation process that produces new words and which, at the word surface-level, can be described by means of an orthographic *pattern* applied to basis words. Table 1 shows that in the simplest case (row 1) this means attaching an affix (*+ität*). The other rows show that the pattern can be more complex, involving stem alternation (row 2; note that the infinitive suffix *+en* is inflectional), deletion of previous affixes (row 3), circumfixation (row 4), or no overt changes, i.e., conversion (row 5).[1] Derivation can take place both within parts of speech (row 6) and across parts of speech.

Derivation is a very productive process in many languages, notably Slavic languages. Thus, natural language processing (NLP) for these languages can profit from knowledge about derivational relationships (Green et al., 2004; Szpektor and Dagan, 2008; Padó et al., 2013). Nevertheless, derivation is a relatively understudied phenomenon in NLP, and few lexicons contain derivational information. For English, there are two main resources. CatVar (Habash and Dorr, 2003) is a database that groups 100K words of all parts of speech into 60K *derivational families*, i.e., derivationally related sets of words. The other is CELEX (Baayen et al., 1996), a multi-level lexical database for English, German, and Dutch, which covers about 50K English words and contains derivational information in its morphological annotation. For German, DErivBase (Zeller et al., 2013) is a resource focused on derivation that groups 280K lemmas into 17K derivational families. As opposed to CatVar and CELEX, it also provides explicit information about the applicable derivation pattern at the level of word pairs. The examples in Table 1 are from DErivBase.

[1] We write patterns as sequences of orthographic operations, using '+' for addition and '-' for deletion, and place the operator before or after the affix to distinguish prefixation and suffixation.

2.2 Modeling the Semantics of Derivation with CDSMs

Lazaridou et al. (2013) were the first to predict distributional vectors for derived words using CDSMs and experimented with a range of established CDSMs. In their paper, all models are supervised, i.e., some word pairs for each pattern are used as training instances, and others serve for evaluation. Also, all models assume that the base word (input) $\mathbf{b}$ and derived word (output) $\mathbf{d}$ are represented as vectors in some underlying distributional space. The *simple additive model* predicts the derived word from the base word as $\mathbf{d} = \mathbf{b} + \mathbf{p}$ where $\mathbf{p}$ is a vector representing the semantic shift accompanying the derivation pattern. The *simple multiplicative model*, $\mathbf{d} = \mathbf{b} \odot \mathbf{p}$ is very similar, but uses component-wise multiplication ($\odot$) instead of addition to combine the base and pattern vectors (Mitchell and Lapata, 2010). The third model, the *weighted additive model*, enables a simple reweighting of the contributions of basis and pattern ($\mathbf{d} = \alpha\mathbf{b} + \beta\mathbf{p}$). Finally, the *lexical function model* (Baroni and Zamparelli, 2010) represents the pattern as a matrix P that is multiplied with the basis vector: $\mathbf{d} = \mathrm{P}\mathbf{b}$, essentially modelling derivation as linear mapping. This model is considerably more powerful than the others, however its number of parameters is quadratic in the number of dimensions of the underlying space, whereas the additive and multiplicative models only use a linear number of parameters.

For their empirical evaluation, Lazaridou et al. (2013) considered a dataset of 18 English patterns defined as simple affixes – 4 within-POS (such as *un-*) and 14 across-POS (such as *+ment*) – and found that the lexical function model is among the top performers, followed by the weighted additive and multiplicative models, all substantially better than baseline. From our perspective, their evaluation has a number of limitations, though: they only included "high-quality" vectors (using human judgments of nearest neighbors to determine quality), thereby focusing on a relatively well-behaved subset of the vocabulary and potentially missing out on highly polysemous words. Furthermore, they evaluated mainly by computing mean cosine similarities between predicted and corpus-observed ("gold") vectors for the derived words – this is not highly informative, as the closeness of the prediction to the actual vector is also dependent on the density of the target's neighborhood.[2] A follow-up study on German (Kisselew et al., 2015) attempted to address these limitations by including all word pairs without prefiltering, and introducing a new evaluation metric that measured how often the predicted vector $\mathbf{d}$ was among the five nearest neighbors of the corpus-observed ("gold") vector $\mathbf{d}$. Kisselew et al.'s evaluation obtained fairly different results: the lexical function model performed worse than the simple additive model, and both CDSMs often had problems outperforming the baseline. This study had its own limitations, though, since it considered only 6 derivation patterns, all within-POS. Thus, it remains unclear to what extent the differences between the two studies are due to (a) the language difference, (b) prefiltering word pairs, or (c) the choice of derivation patterns under consideration.

3 Analyzing Models of Morphological Derivation

Given these conflicting results, we propose to empirically investigate the factors that influence how well CDSMs can predict the semantics of derived words. Note that when we talk about 'predictability' of a derived form, we refer to the ability to model an otherwise already established term, for which a distributional analysis can be performed. That is, we investigate to which extent a one-off compositional procedure can capture the meaning of a word, as in a situation where a speaker encounters an existing term for the first time. Further, we assume that the individual items in the observed data will naturally have different frequencies (from rare to highly frequent) and that this will affect both the learning process and the certainty we can have about the meaning of a test vector. We believe this is a realistic setup in terms of modelling first encounters with established derivations, and we therefore make no attempt to control for the frequency of individual word vectors, either at training or test time.

3.1 Overall Workflow

We follow a two-step workflow depicted on the left-hand side of Figure 1. The workflow involves *prediction models* (cf. Section 3.2), i.e., CDSMs that predict a vector for the derived word given the vector for the base word, as well as *analysis models* (cf. Section 3.3), i.e., linear regression models that predict

[2] They also performed a manual judgment study, but only as an additional experiment on "low-quality" word pairs.

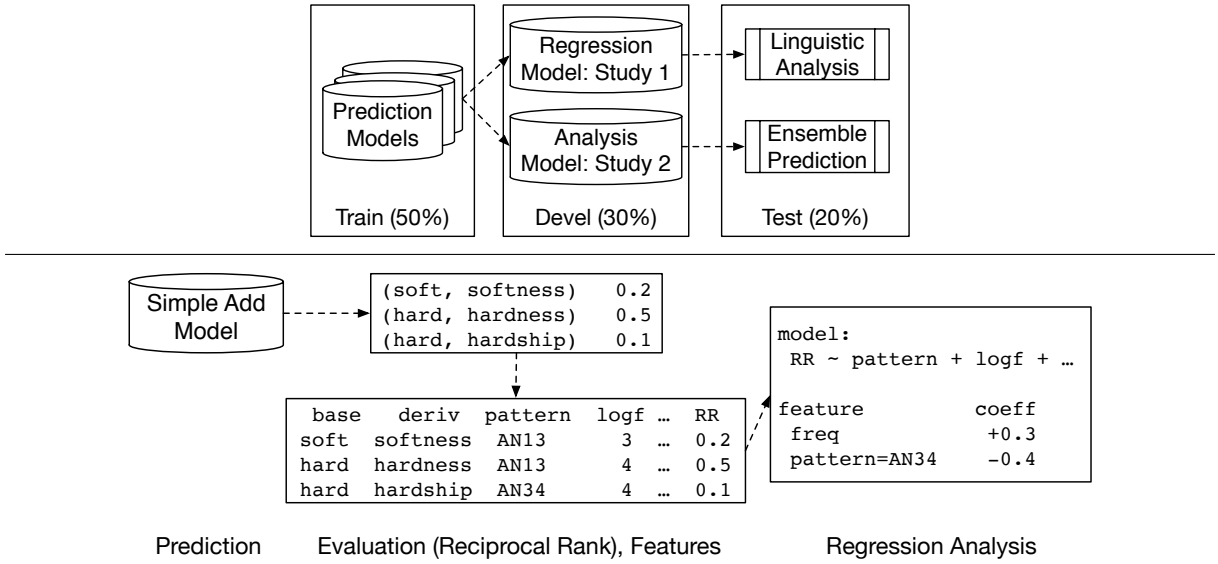

Figure 1: Top: Overall workflow. Below: Toy example

the performance of the CDSMs based on a rich set of linguistic predictors. We build two analysis models, one for linguistic analysis (Experiment 1, Section 4) and one for NLP (Experiment 2, Section 5).

The workflow uses a large set of derivationally related word pairs split into training, development, and test sets (50%, 30%, and 20%, respectively). The splits are stratified by derivation pattern, i.e., each pattern occurs in approximately these ratios in each split. This is a reasonable strategy, assuming that our set of patterns is fairly complete (Zeller et al., 2013) and we can disregard the problem of unseen patterns.

The three sets play different roles. The training set is used to train prediction models. The development set is then used to measure the performance of prediction models on unseen data. These performance numbers are those that the regression model is then trained to predict. Finally, the analysis model itself is evaluated on the test set, that is, on another previously unseen dataset. In this manner, we guarantee that all results we obtain are measured on unseen data and generalize well to novel instances.

We note that the task of the prediction models (constructing the vector for the derived word) incorporates our general assumption that we do not have any information about the derived word. While this is not a reasonable assumption from an NLP point of view, where we would know at least the frequency of the derived word, and may also have a (typically less reliable) distributional vector for it, this "no-knowledge" setup represents, in our opinion, the cleanest setup for an analysis of linguistic factors.

3.2 Prediction (CDSM) Models

Derivationally Related Word Pairs. We draw our derivationally related word pairs from DErivBase[3] (Zeller et al., 2013). As stated above, each word pair is labeled with a derivation pattern, representing the orthographic transformation of the basis word. Since our predictive models are trained for each pattern separately, we ensure that each model will have enough training instances by discarding all patterns with less than 80 word pairs. Out of the 158 patterns in DErivBase, we retain 74 patterns, of which 49 are cross-POS patterns. The 74 patterns cover 30,757 word pairs. Patterns have a median of 194.5 word pairs (min. 83, max. 3028).

Corpus. We derive the frequency counts and the distributional vectors for our analysis from the German web corpus SdeWaC (Faaß and Eckart, 2013), POS-tagged and lemmatized using TreeTagger (Schmid, 1994). Following Kisselew et al. (2015), to mitigate sparsity, for out-of-vocabulary words we back off to the lemmas produced by MATE Tools (Bohnet, 2010), which have higher recall but lower precision than TreeTagger. We also use the MATE dependency analysis to reconstruct lemmas for separated prefix verbs.

[3] http://goo.gl/tiRJy0

Prediction Models. To obtain the vector representations on which we can train our prediction models, we use CBOW, a state-of-the-art predictive distributional semantics space which been shown particularly effective for modelling word similarity and relational knowledge (Mikolov et al., 2013).[4] (Considering the type of semantic space as a parameter is outside the scope of our study.)

As both target and context elements, we use all 280K unique POS-disambiguated lemmas (nouns, adjectives, and verbs) from DErivBase. We use a within-sentence context window of size ± 2 to either side of the target word, 300 context dimensions, negative sampling set to 15, and no hierarchical softmax. On these vector representations, we train four prediction models (cf. Section 2): the simple additive model, the simple multiplicative model, the weighted additive model, and the lexical function model. Each model is trained on each of the 74 patterns separately by minimizing the expected square loss between the predicted and the observed derived vector.[5] For additive models, the shift vector $\mathbf{p}$ is computed as the average of the shift vectors across all word pairs from a single pattern, while the weighted additive model additionally optimizes α and β in a subsequent step. Since the lexical function model is more prone to overfitting due to its many parameters, we train it using ridge regression, employing generalized cross-validation to tune the regularization parameter on the training set. As a fifth, baseline model, we use the identity mapping, which simply predicts the basis vector as the vector of the derived word. Our implementation is based on the DISSECT toolkit (Dinu et al., 2013).

Evaluation. The performance of the CDSMs is measured by how well the predicted vector aligns with the corpus-observed vector for the derived word. More concretely, we quantify the performance on each word pair by *Reciprocal rank (RR)*, that is, 1 divided by the position of the predicted vector in the similarity-ranked list of the observed vector's neighbors. Besides being a well-established evaluation measure in Information Retrieval, RR is also more sensitive than the "Recall out of n" measure used previously (Kisselew et al., 2015), which measures the 0–1 loss and also requires fixing a threshold n. RR also has the advantage of being easily interpretable: a mean reciprocal rank (MRR) of 0.33, e.g., indicates that the correct predicted vector is on average the third-nearest neighbor of the observed vector. The neighbor list for each derived word is POS-specific, that is, it consists of all words in the space that match its part of speech.

3.3 Analysis (Linear Regression) Models

The task of our analysis models is to predict the performance of the CDSM models (measured as reciprocal rank, cf. Section 3.2) at the word pair level, i.e., individual pairs of base and derived words. The goal is to assess which factors have a substantial influence on the prediction of the semantics of derived words. To this end, we use linear regression, which is a well-established analysis method in linguistics and psycholinguistics (Baayen, 2008). Linear regression predicts a dependent variable v as a linear combination of weighted predictors p_i, i.e., $v = \alpha_1 p_1 + \cdots + \alpha_n p_n$. A coefficient α_i can be interpreted as the change in v resulting from a change in the predictor p_i. We use the R statistical environment.

The right-hand side of Figure 1 shows a toy example for a single prediction model (simple additive). We first run the prediction model, then evaluate its reciprocal ranks at the word pair level, then compute features (such as the pattern and the logarithmized frequency of the base). Finally, we perform a regression analysis. It yields the information that higher frequency has a positive impact on performance, while the pattern AN34 has a negative impact.

Our complete set of predictors comprises three classes:

- **Base word level predictors** describe properties of the base word. They include `base_productivity`, the number of derived words known for the base, `base_polysemy`, the number of WordNet synsets, and `base_freq`, its lemma frequency in the SDeWaC corpus.[6] Predictor `base_typicality` is the cosine similarity between the base and the centroid of all bases for the present pattern, as a measure of how semantically typical the base is for the pattern;

[4] `https://code.google.com/p/word2vec/`

[5] For the simple additive and multiplicative models, there are analytical solutions.

[6] All numeric variables (predictors and dependent variable) are z-scaled; frequency variables are logarithmized.

	Baseline	Simple Add	Weighted Add	Mult	LexFun
Mean Reciprocal Rank	0.271	0.309	**0.316**	0.272	0.150
# Predictions used by Oracle (Experiment 2)	2139	954	1613	532	913
# Predictions used by Regression-based Ensemble (Experiment 2)	51	2306	3528	190	76

Table 2: Results for individual prediction models on test set

- **Prediction level predictors** describe properties of the vector that the CDSM outputs. Following work on assessing the plausibility of compositionally constructed vectors by Vecchi et al. (2011), we compute the length of the vector (`deriv_norm`) and the similarity of the vector to its nearest neighbors (`deriv_density`), and the similarity between base vector and derived vector (`base_deriv_sim`);
- **Pattern level predictors**. We represent the identity of the pattern, which is admissible since we can assume that the DErivBase patterns cover the (vast) majority of German derivation patterns (Clark, 1973). Unfortunately, this excludes a range of other predictors, such as the parts of speech of the base and derived words, due to their perfect collinearity with the pattern predictor.

The rest of the paper is concerned with performing a regression analysis based on these features. We perform two separate analyses to satisfy two fairly different motivations. The first one is linguistic, namely to understand which *properties of the base and the pattern* make the prediction easy or difficult. This analysis concentrates on one single CDSM, namely the best individual one: if it included multiple CDSMs, the regression model would spend part of its power on predicting the behavior of the (arguably irrelevant) worse CDSMs. Further, this regression model should include only pattern-level and base-level features, since prediction-level features are arguably not linguistic properties. For this approach, see Section 4.

The second motivation comes from NLP and concerns the possibility to define an *ensemble* over several CDSMs that works better than individual CDSMs by employing a regression model to select the best CDSM at the word pair level. This analysis must by definition include the output of multiple prediction models. Furthermore, it should also include the features at the prediction level since they may help distinguish reasonable from unreasonable predictions. We will pursue this approach in Section 5.

4 Experiment 1: Linguistic Analysis

As outlined above, the first task in the linguistic analysis is to select the best individual prediction model for evaluation. The test set evaluation results are shown in the first row of Table 2. As the numbers show, the weighted additive model is the best model, and our analysis will focus on it. We estimate the following linear regression model to predict reciprocal rank (RR) on the development set:

```
RR ~ pattern + base_productivity + base_typicality + base_polysemy + base_freq
```

Applied to the test set, the model achieves a highly significant fit with the data (F=58.32, $p<10^{-12}$, R^2=0.324). Performance is highly variable across patterns and words pairs: results for word pairs span almost the full range of reciprocal ranks between 0 and 1, and the pattern level results range between 0.03 (pattern VV01, *zucken → zuckeln / twitch → saunter*), i.e., predictions are no good, and 0.69 (pattern AN10, *präsent → Präsenz / present → presence*), i.e., most predictions are ranked first or second. Predicted values are not correlated with the residuals ($r < 10^{-6}$). Our further discussion of this regression model is structured along a set of hypotheses we made regarding the influence of particular factors, or more specifically how they translate into distributional behavior.

Training Data and Polysemy. We start by considering the "usual suspects" in data-driven computational linguistics regarding performance, which leads us to three hypotheses. First, *low-frequency bases are hard* due to the limited reliability of the distributional evidence. Second, *atypical bases are hard*, that is,

Predictor	Estimate	LMG score
pattern	(see Table 4)	87.2%
base_productivity	-0.13***	7.6%
base_freq	0.21***	4.1%
base_polysemy	-0.03**	0.8%
base_typicality	0.04***	0.2%

Table 3: Experiment 1: Coefficients, significances, and effect sizes for the predictors

derivations of instances unlike those seen in the training data are difficult to predict. Third, derivation models must account for the selection of individual word senses in derivation: e.g., the verbal base *absetzen* variously means *depose (an official), drop (a load), deduct (an amount)*, but the derived adjective *absetzbar* is only used in the meaning of *deductable*. Since typical distributional models, including ours, do not disambiguate bases by sense, *highly polysemous bases are hard*.

Consider now Table 3, which lists coefficients, significance, and effect sizes for these predictors. Recall that we predict reciprocal rank (RR), that is, positive coefficients indicate better whereas negative coefficients indicate worse performance.[7] The data bears out our hypotheses fairly well: we find positive effects of frequency and of typicality, and negative effects of base productivity and polysemy. The relative importances of these effects is however only weakly indicated by the sizes of the coefficients. Thus, the column LMG provides normalized Lindeman-Merenda-Gold (LMG) scores (Lindeman et al., 1980), a measure of effect size (Grömping, 2012), applied, e.g., by Marelli et al. (2015) in a similar context. These scores indicate what percentage of the variance explained by the model is due to the individual predictor groups. As we see, most variance is accounted for by the `pattern` predictor. `Productivity` and `frequency` account for respectable amounts of variance, while `polysemy` and `typicality` contribute surprisingly little. This finding needs however to be interpreted taking into account that the `pattern` predictor is categorical, and as such "soaks up" all properties at the level of individual patterns, including polysemy. As a matter of fact, the correlation between RR and polysemy at the level of individual word pairs is only weak (ρ=-0.03), while MRR and average polysemy are strongly correlated at the level of derivational patterns (ρ=-0.30).

Since the bulk of the variance is accounted for by the `pattern` predictor, we now turn to formulating hypotheses about derivation patterns.

Within-POS Derivations. We first start out by considering within-POS derivations. While cross-POS derivations are at least partially motivated by the need to change the base's syntactic category, within-POS derivations primarily reflect proper semantic processes, such as polarity and gradation prefixes (*un+*, *über+* for adjectives) or prefix verbs (*hören* $\rightarrow$ *aufhören / hear* $\rightarrow$ *stop*), which are particularly prominent in German. Such affixes are known to be highly polysemous and hard to characterize both linguistically and computationally (Lechler and Roßdeutscher, 2009; Lehrer, 2009). Thus, we expect that *within-POS derivation is hard to model*.

Table 4 lists all levels of the factor `pattern` that are statistically significant from the grand mean (using contrast coding in the regression model), adopting a threshold of α=0.01. The columns correspond to the parts of speech of the base word, and the rows to the parts of speech of the derived word. Recall that negative coefficients indicate worse performance than average, and positive coefficients better-than-average performance.

The table strongly confirms our hypothesis: all five significant adjective $\rightarrow$ adjective and all seven verb $\rightarrow$ verb derivation patterns come with large negative coefficients.

Argument Structure. For cross-POS derivation, we hypothesize that *argument structure* (Grimshaw, 1990) is a major factor, connected to the largest difference to existing applications of CDSMs for phrase meaning: while in phrasal composition the resulting phrase usually shows the same semantic behavior as

[7] We use standard notation for significance (*: $p < 0.05$, **:$p < 0.01$, ***:$p < 0.001$).

Derived POS	Base POS: A			Base POS: N			Base POS: V		
A	AA03	*anti+*	−0.39 ***	NA02	*+isch*	0.52 ***	VA02	*+end*	0.48 ***
	AA07	*ab+*	−0.41 ***	NA05	*-e/en +ig*	−0.19 **	VA03	*+ig*	−0.26 **
	AA13	*nach+*	−0.41 ***	NA25	*-ung +t*	0.67 ***	VA11	*+lich*	−0.36 ***
	AA15	*über+*	−0.43 ***	NA26	*-ung +end*	0.40 **	VA12	*+end*	0.85 ***
	AA17	*vor+*	−0.42 ***	NA27	*+lich*	0.46 ***	VA13	*+t*	1.02 ***
				NA29	*+los*	−0.42 ***			
				NA31	*ge+*	−0.33 ***			
N	AN01	*+e*	−0.39 ***				VN03	*+er*	−0.24 ***
	AN02	*+heit*	0.45 ***				VN07	*+ung*	1.14 ***
	AN03	*+keit*	0.78 ***				VN09	*+en*	0.45 ***
	AN04	*+igkeit*	0.64 ***						
	AN10	*-t +z*	0.75 ***						
	AN11	*+ie*	0.95 ***						
	AN12	*-isch -ik*	0.71 ***						
	AN16	*+ität*	0.64 ***						
	AN17	*(null)*	−0.38 ***						
V	AV01	*+isieren*	0.69 ***	NV09	*(null)*	0.36 ***	VV01	*-en +eln*	−0.45 **
	AV04	*(null)*	0.28 **	NV15	*an+*	−0.31 ***	VV05	*ver+*	−0.26 ***
				NV17	*aus+*	−0.35 ***	VV12	*(stem)*	−0.36 **
				NV20	*ein+*	−0.36 ***	VV13	*an+*	−0.26 ***
				NV22	*ab+*	−0.34 ***	VV22	*ein+*	−0.21 **
							VV27	*vor+*	−0.41 ***
							VV30	*um+*	−0.28 **

Table 4: Experiment 1: Derivation patterns with significant regression model coefficients (α=0.01), cross-classified by base and derived part of speech (*null*: morphologically null derivation; *stem*: anticausative stem change, as in *legen* → *liegen, put* → *lie*)

its head component (an adjective-noun phrase behaves largely like a noun), this is not always the case in derivation. For example, the agentive nominalization pattern *-er* (*laufen* → *Läufer / run* → *runner*) incorporates the agent noun of the verb, which therefore drops out of the context of the derived word. We hypothesize that *argument structure changes are difficult to learn* for the CDSMs we consider.

Looking at Table 4, we see a mixed picture, with easy and difficult patterns. Adjective → noun derivations, which predominantly generate abstract nouns without argument structure (like AN02, *taub* → *Taubheit / deaf* → *deafness*), are overwhelmingly easy to generate. We hypothesize that the deletion of the adjective's argument is not problematic to learn. For verb → noun patterns, the default event nominalization suffix *+ung* (*umleiten* → *Umleitung / redirect* → *redirection*) and stem nominalizations (*fahren* → *Fahren / drive* → *driving*), both of which preserve argument structure, are easy to model. So are the verb → adjective patterns that form present participles (*+end*) and past participles (*+t*). In contrast, the agentive/instrumental nominalization pattern *+er* (*fahren* → *Fahrer / drive* → *driver*), where argument structure changes, is associated with a loss in performance.

We noted that those verb → adjective patterns that form property adjectives (*beachten* → *beachtlich / notice* → *noticeable*) are more challenging to model. This made us aware that difficulties associated with argument structure are mediated by an another important factor, namely *semantic regularity*. The difficulty of such patterns is related to how uniform the semantic shift is among the instances of the pattern, and how well it can be picked up by distributional analysis. As an example of semantically regular shifts, consider the significant adjective → verb patterns (AV01, AV04) which can be paraphrased as "to cause to have property X" (*anonym* → *anonymisieren / anonymous* → *anonymize*). Since there is a direct mapping from the modified head noun of the adjective onto the direct object of the verb, the distributional mapping is relatively easy, even though the shift even involves the creation of new argument structure. In contrast, some verb → adjective patterns like VA11 (*+lich*) involve the introduction of modality, a complex semantic change whose distributional consequences are hard to grasp and which is similar in nature to within-POS derivations (see above).

At the far end of the difficulty scale, we find bad performance for the noun → verb derivations, because these patterns face challenges on both the argument structure and regularity fronts: they generate verbs

from nouns that are only loosely semantically related (Clark and Clark, 1979). An example is NV22 with instances like *Zweig → abzweigen / (tree) branch → branch off*. The only easy noun → verb pattern, NV09, comes with a particularly regular semantic shift, paraphrasable as "to use X as instrument" (*Hammer → hämmern / (a) hammer → (to) hammer*).

Argument structure being a complex phenomenon, we would require additional work to exactly identify which factors play a role in derivational processes, and how those factors interact with distributional models. For instance, certain types of argument deletion/addition can result in shifting lexical items to other sentence constituents (e.g., *X developed Y over ten years* vs. *Y underwent ten years of development*). This kind of effect can, at least in principle, be captured using variable window sizes in a CDSM. Whilst we leave such questions for further research, the present results seem to support the idea that argument structure is a worthwhile aspect to investigate.

5 Experiment 2: Ensemble Prediction

In our second study, we investigate the use of linear regression models to construct an *ensemble* of CDSMs for derivation prediction. Ensemble learning is well established in NLP to capture phenomena that exceed the power of individual models (Dietterich, 2000). In our case, we want to select one vector from among the predictions of multiple CDSMs. We consider two strategies to perform this selection: The *oracle model* compares all prediction models, and simply picks the one with the highest RR. The oracle thus establishes an upper bound that assesses the theoretical benefit of model combination. It achieves an MRR of 0.362 – a modest, but substantial improvement of four and a half points over the best individual model (weighted additive, MRR=0.316, cf. Table 2).

The second strategy is the *regression model* which predicts the CDSMs' expected performances at the word pair level with a linear regression model trained on the development set (cf. Figure 1). As discussed in Section 3.3, our regression model for this purpose differs in two respects from the first study: it includes features for the prediction, and it is trained on the evaluations of all five CDSMs. The provenance of each evaluation result is coded in a new predictor, `cdsm`, with the values `baseline, simple_add, weighted_add, mult, lexfun`. We introduce interactions between `cdsm` and all base-level features to enable the regression model to learn, e.g., that some CDSMs can deal better with low-frequency bases. We estimate the following model on the development set:

```
RR ~ deriv_density + base_deriv_sim + deriv_norm + pattern +
     (base_productivity + base_typicality + base_freq + base_polysemy) * cdsm
```

On the test set, the model achieves a highly significant fit with the data (F=193.5, p$< 10^{-12}$, R^2=0.305), that is, it achieves a similar model fit to the first study.

Unfortunately, the use of this regression model to define an ensemble does not work particularly well: the ensemble yields an MRR of just 0.321, only half a point above the best-performing individual model, weighted additive, with an MRR of 0.316. This is a negative result: our regression models do not directly translate into better predictions for derived word vectors. To understand the reasons for this failure, we perform two analyses. The first one compares how many predictions of each CDSM the oracle and the ensemble selected, as shown in the lower part of Table 2. The oracle selects substantially from all models, while the regression-based ensemble chooses strictly in proportion to the CDSMs' overall performance: The best model (weighted additive) is selected for over 60% of all cases while the lexical function model is almost ignored. This indicates that the regression model is overly dependent on the `cdsm` predictor, while the base-level and pattern-level predictors are not powerful enough to reverse the bias towards higher-MRR models.

Our second analysis follows Surdeanu and Manning (2010), who found that the complementarity between participating models in an ensemble is more important than the exact combination method. To test the amount of complementarity, we computed rank correlations (Spearman's ρ) between the CDSMs' predictions at the word pair level. The results in Table 5 show that the baseline, additive, and multiplicative models are highly correlated (all pairwise ρs larger than 0.84). Only the lexical function model behaves substantially differently (pairwise ρ less than 0.34). This would make it a good candidate

	Baseline	Simple add	Weighted add	Multiplicative
Simple add	0.923			
Weighted add	0.840	0.930		
Multiplicative	0.967	0.929	0.853	
Lexical function	0.281	0.310	0.333	0.304

Table 5: Experiment 2: Correlations among CDSMs at the word level (Spearman's ρ)

for complementary predictions (as its selection by the oracle also witnesses) – however, its overall bad performance (MRR=0.150) drastically reduces its chance to be picked by the ensemble.

6 Conclusions

In this paper, we presented the first analysis of CDSMs on derivational phenomena that is both detailed and broad-coverage. Our main premise was that the linguistic features of individual lexical items, as well as the nature of the derivation pattern, would affect the extent to which the derived form could be predicted. This led us to establish relationships between linguistic properties and distributional behavior of words, a central topic in distributional semantics that seems to have received very little attention.

To quantify these relationships, we built a linear regression model with CDSM performance as dependent variable and linguistic features as predictors. An effect size analysis showed that the base term's productivity and frequency influence difficulty, but that the derivation pattern has a much larger effect. By analyzing patterns, we found that the three main factors for bad performance were: modifications of argument structure, semantic irregularity, and within-POS derivations.

Regarding the apparent contradictions among previous studies, our analysis can resolve them to some degree. We can attribute the overall bad CDSM results of Kisselew et al. (2015) to an unfortunate choice of hard within-POS derivations. At the same time, we replicate their particularly disappointing results for the lexical function, which contrasts with Lazaridou et al.'s reported performance for that model. To test whether these differences are due to Lazaridou et al.'s prefiltering, we re-evaluated all CDSMs on a "high-quality" subset of our data by throwing away the quartile with the lowest base word frequency (corresponding to a threshold of 420). The results for all models improve by 1–2%, but the lexical function model remains at 15% below the baseline. Obvious remaining differences are the language and the type of the distributional model used. However, these factors were outside the scope of the current study, so we leave them for future work.

We also built an ensemble model over the different CDSMs but did not substantially outperform the best single CDSM. We draw two conclusions from this failure: (a), despite the array of available CDSMs, it makes sense to continue developing new CDSMs to increase complementarity; and (b), the limiting factor in difficult prediction is the idiosyncratic behaviour of base words that our current distributional features capture only imperfectly.

To encourage further research, we make available our dataset with derivationally related word pairs and CDSM performance predictors.[8]

Acknowledgements

We thank the anonymous reviewers for their helpful comments. The first and third authors are supported by Deutsche Forschungsgemeinschaft (SFB 732, project B9). The second author is supported by the ERC Starting Grant COMPOSES (283554). The fourth author has been supported by the Croatian Science Foundation under the project UIP-2014-09-7312.

[8] Available at: `http://www.ims.uni-stuttgart.de/data/derivsem`

References

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. *The CELEX lexical database. Release 2. LDC96L14*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Harald Baayen. 2008. *Analyzing Linguistic Data*. Cambridge University Press.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Cambridge, MA, USA.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in Space: A Program for Compositional Distributional Semantics. *LiLT (Linguistic Issues in Language Technology)*, 9:5–110.

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of COLING*, pages 89–97, Beijing, China.

Eve V. Clark and Herbert H. Clark. 1979. When Nouns Surface as Verbs. *Language*, 55:767–811.

Herbert H Clark. 1973. The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal of verbal learning and verbal behavior*, 12(4):335–359.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical Foundations for a Compositional Distributional Model of Meaning. *Linguistic Analysis*, 36:345–386.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. DISSECT - DIStributional SEmantics Composition Toolkit. In *Proceedings of ACL*, pages 31–36, Sofia, Bulgaria.

Wolfgang Dressler. 2005. Word-Formation in Natural Morphology. In Pavol Štekauer and Rochelle Lieber, editors, *Handbook of Word-Formation*, pages 267–284. Springer.

Katrin Erk and Sebastian Padó. 2008. A Structured Vector Space Model for Word Meaning in Context. In *Proceedings of EMNLP*, pages 897–906, Honolulu, HI, USA.

Gertrud Faaß and Kerstin 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.

Rebecca Green, Bonnie J Dorr, and Philip Resnik. 2004. Inducing Frame Semantic Verb Classes from WordNet and LDOCE. In *Proceedings of ACL*, pages 375–382, Barcelona, Spain.

Jane Grimshaw. 1990. *Argument Structure*. MIT Press, Cambridge.

Ulrike Grömping. 2012. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician*, 61(2):139–147.

Nizar Habash and Bonnie Dorr. 2003. A Categorial Variation Database for English. In *Proceedings of NAACL*, pages 17–23, Edmonton, Canada.

Jennifer Hay. 2003. *Causes and Consequences of Word Structure*. Routledge.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In *Proceedings of ACL*, pages 58–68, Baltimore, Maryland, USA.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Proceedings of NIPS*, pages 3294–3302, Montreal, Canada.

Max Kisselew, Sebastian Padó, Alexis Palmer, and Jan Šnajder. 2015. Obtaining a Better Understanding of Distributional Models of German Derivational Morphology. In *Proceedings of IWCS*, pages 58–63, London, UK.

Brenda Laca. 2001. Derivation. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, editors, *Language Typology and Language Universals: An International Handbook*, volume 1, pages 1214–1227. Walter de Gruyter, Berlin.

Angeliki Lazaridou, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2013. Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics. In *Proceedings of ACL*, pages 1517–1526, Sofia, Bulgaria.

Andrea Lechler and Antje Roßdeutscher. 2009. German Particle Verbs with "auf". Reconstructing their Composition in a DRT-based Framework. *Linguistische Berichte*, 220:439–478.

Adrienne Lehrer. 2009. Prefixes in English Word Formation. *Folia Linguistica*, 21(1–2):133–138.

Richard H. Lindeman, Peter F. Merenda, and Ruth Z. Gold. 1980. *Introduction to Bivariate and Multivariate Analysis*. Scott Foresman, Glenview, IL, USA.

Thang Luong, Richard Socher, and Christopher D. Manning. 2013. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of CoNLL*, pages 104–113, Sofia, Bulgaria.

Marco Marelli, Simona Amenta, and Davide Crepaldi. 2015. Semantic transparency in free stems: The effect of orthography-semantics consistency on word recognition. *The Quarterly Journal of Experimental Psychology*, 68(8).

Tomas Mikolov, Wen-Tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of HLT-NAACL*, pages 746–751, Atlanta, GA, USA.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Sebastian Padó, Jan Šnajder, and Britta Zeller. 2013. Derivational Smoothing for Syntactic Distributional Semantics. In *Proceedings of ACL*, pages 731–735, Sofia, Bulgaria.

Ingo Plag. 2003. *Word-Formation in English*. Cambridge University Press, Cambridge.

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

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of NeMLaP*, Manchester, UK.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of EMNLP*, pages 1631–1642, Melbourne, Australia.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble Models for Dependency Parsing: Cheap and Good? In *Proceedings of NAACL*, pages 649–652, Los Angeles, California.

Idan Szpektor and Ido Dagan. 2008. Learning Entailment Rules for Unary Templates. In *Proceedings of COLING*, pages 849–856, Manchester, UK.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (Linear) Maps of the Impossible: Capturing Semantic Anomalies in Distributional Space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9, Portland, Oregon, USA.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of ACL*, pages 1201–1211, Sofia, Bulgaria.