Modeling Derivational Morphology in Ukrainian

Mariia Melymuka, Gabriella Lapesa, Max Kisselew, Sebastian Padó
University of Stuttgart, Institute for Natural Language Processing
{melymuma,lapesaga,kisselmx,pado}@ims.uni-stuttgart.de

Abstract

We report on a study applying compositional distributional semantic models (CDSMs) to a set of Ukrainian derivational patterns. Ukrainian is an interesting language as it is morphologically rich, and low-resource. Our study aims at resolving inconsistent results from previous studies which employed CDSMs for derivation; we provide evidence for a cross-lingual advantage of CBOW over NMF representations, as well as a simple additive over a lexical function model. In addition, we present two case studies in which we test the capabilities of CDSMs to deal with pattern-level ambiguity and apply the same CDSMs to inflectional patterns.

1 Introduction

The potential and limitations of compositional distributional semantic models (CDSMs, Mitchell and Lapata (2010)) in modeling the semantic shifts associated with derivational processes have been explored in a number of evaluation studies on English (Lazaridou et al., 2013) and German (Kisselew et al., 2015; Köper et al., 2016; Padó et al., 2016). By considering high-resource languages, these studies can take advantage of large standard corpora, well-established pre-processing tools, and derivational lexicons to select experimental items, such as CELEX (Baayen et al., 1996). In this paper, we target Ukrainian, a language that is both morphologically rich and computationally low-resource. To the best of our knowledge, this is the first study employing CDSMs for modeling (aspects of) the derivational morphology of a Slavic language, for which most studies followed either knowledge-based (Pala and Hlaváčková, 2007; Šnajder, 2014) or classification approaches (Piasecki et al., 2012). We believe that derivational morphology is a suitable starting point to gather evaluation data for CDSMs in new languages, given that the semantic relations between pairs are encoded straightforwardly at the surface level.

Following the previous studies listed above, we operationalize derivation as a compositional operation: the vector for a base word (e.g., $\overrightarrow{work}$) is the input for the CDSM which implements a function modeling the derivational meaning shift (e.g., f_{er}) whose output is its prediction for the vector of the derived word ($f_{\text{er}}(\overrightarrow{work}) = \overrightarrow{worker}$). We consider two types of CDSMs that performed well in previous evaluations. The first one is the simple additive model (*SimpleAdd*), which represents the shift induced by a derivational pattern as the difference vector (e.g., $\overrightarrow{er}$) between the vector of the derived word and the vector of the base. It is added to new bases yielding a predicted vector for derived words ($\overrightarrow{love} + \overrightarrow{er} = \overrightarrow{lover}$). The second CDSM we test is the more expressive lexical function model (*LexFun*, Baroni and Zamparelli (2010)). It represents the shift as a matrix (i.e., M_{er}) to be multiplied with new bases to predict derived words ($M_{\text{er}} \cdot \overrightarrow{love} = \overrightarrow{lover}$). This enables the model to account for interactions between dimensions in the shift. The downside is that *LexFun* uses $O(d^2)$ parameters instead of *SimpleAdd's* $O(d)$, where d is the dimensionality of the space. Thus, *LexFun* requires considerably more training data.

The literature on CDSMs for derivation shows a contradictory picture: for English, Lazaridou et al. (2013) showed that *LexFun* outperforms additive (and multiplicative) models; for German, Kisselew et al. (2015) and Padó et al. (2016) found *LexFun* to be almost consistently outperformed by additive models. Kisselew et al. could not distinguish between the influence of the language or the experimental setup in explaining the differences in outcome to Lazaridou et al. (2013). Our goal is to resolve this uncertainty, by bringing a third language to the table and mediating between the experimental setups of the two previous

studies. Our pattern inventory was defined to keep our study parallel to Kisselew et al. (2015). We however consider both CBOW spaces (as considered by Kisselew et al. (2015)) and NMF spaces (as considered by Lazaridou et al. (2013)). Our results show a surprising cross-lingual (Ukrainian/German) consistency in the behavior of the aligned patterns. As a further contribution, we present two case studies which test the capability of our CDSMs to handle ambiguous patterns (section 4.1) and inflectional patterns (4.2).

2 Experimental setup

Experimental items. As stated above, we selected our derivation patterns to be as parallel as possible to Kisselew et al. (2015) both at the qualitative level (trying to find patterns with similar meaning) and at the quantitative level (we employed the same sampling methodology). This resulted in the selection of patterns displayed in table 1.

Pos	Pattern	Example
N→N	NEGATIVE: не- (*ne-*)	*druh*, friend → *ne-druh*, enemy
	FEMALE (for occupations): -к- (*-k-*)	*likar*, doctor (male) → *likar-k-a*, doctor (female)
	DIMINUTIVE (for female nouns): -к- (*-k-*)	*mashyna*, car → *mashyn-k-a*, small car
V→V	PERFECTIVE (prefix): про- (pro-)	*testuvaty*, to test → *pro-testuavaty*, to complete testing
	PERFECTIVE (suffix): -ну- (-nu-)	*znykaty*, to disappear → *znyk-nu-ty*, to be vanished
A→A	NEGATIVE: не- (*ne-*)	*solodkyi*, sweet → *ne-solodkyi*, not sweet
	PRIVATIVE: без- (*bez-*)	*shumnyi*, noisy → *bez-shumnyi*, noiseless

Table 1: Selection of experimental items

For each pattern, we extracted pairs of base/derived words in which each word occurs at least 80 times in the corpus (see next paragraph) and randomly selected 70 pairs matching our frequency threshold. A native speaker manually checked the correctness of the base/derived pairs.

Corpus. Our corpus is a concatenation of four Ukrainian corpora: three raw corpora (`web-2012`, `news-2011`, and `wikipedia-2013`) available through the University of Leipzig[1], described in Goldhahn et al. (2012), and the corpus described in Babych and Sharoff (2016).[2] The corpus was part-of-speech tagged and lemmatized using LanguageTool[3]. Our concatenated corpus contains approximately 131 million tokens. Albeit relatively small by modern standards, this corpus size should enable the construction of reliable distributional representations (for comparison, the British National Corpus contains 100 million tokens).

Distributional representations. The first distributional representation we consider is CBOW as implemented by word2vec (Mikolov et al., 2013) and used in previous studies (Kisselew et al., 2015; Padó et al., 2016). The second one is Non-Negative Matrix Factorization or NMF (Lee and Seung, 2000) applied to standard count-based vectors, as used by Lazaridou et al. (2013). All DSMs evaluated in this paper were built adopting a 5-words symmetric window. We extracted vectors for POS-disambiguated lemmas of open-class words (nouns, verbs, adjectives and adverbs) occurring at least 20 times in our corpus (target vocabulary: 77.707 words).[4]

To the best of our knowledge, this is the first study to directly compare CBOW and NMF as representations for CDSMs. We believe such comparison to be of particular interest because of the different nature of the spaces: dense and negative for CBOW, sparse and positive for NMF. Moreover, the fact that our target language is under-resourced and under-explored made the "best practices" adopted in previous

[1] `wortschatz.uni-leipzig.de/en/download/`

[2] We thank Serge Sharoff for providing the corpus.

[3] `github.com/languagetool-org/languagetool/tree/master/languagetool-standalone`

[4] Hyperparameters are set as follows: CBOW negative sampling is 15, no hierarchical softmax. NMF is based on a count space built with the 10k most frequent open-class words as contexts, scored with PPMI.

studies potentially less portable. Thus, we decided not to adopt the standard 300 dimensions of Padó et al. (2016) or 350 of Lazaridou et al. (2013), but to experiment with distributional representations of different size. For both CBOW and NMF, we build models with between 200 to 600 dimensions (step size 100).

Experimental Setup and Evaluation. We test the ten different DSM configurations (DSM class $\in$ {CBOW, NMF} $\times$ dimension $\in$ {200, 300, 400, 500, 600}) as the input for the two CDSMs (*LexFun* and *SimpleAdd*).[5] The CDSMs are tested in the supervised task of predicting the vector for a derived word given a base vector and a set of base/derived pairs as a training set for the compositional function. We model each derivational pattern separately and perform ten-fold cross-validation on each pattern.

The quality of the predicted vector (the performance of each DSM+CDSM combination) is quantified in terms of two measures which have already been employed in the reference literature. The first measure is the average cosine similarity (CosSim) between predicted vectors for derived words and corresponding gold vectors, as employed by Lazaridou et al. (2013). Cosine has however a crucial limitation: The interpretation of a cosine score is highly dependent on the density or sparsity of the region in question; consequently, it also does not support a comparison of the results from the NMF and CBOW models. The second measure is a rank-based measure that directly quantifies the goodness of a predicted vector in terms of the position of the gold vector in the predictions' nearest neighbor list. Concretely, we compute *Mean Reciprocal Rank (MRR)*, a measure from information retrieval for evaluating ranking task. In our case, it is the average inverse of the positions of the gold vectors in the nearest-neighbor lists of the predicted vectors, as adopted by Padó et al. (2016). For example, a value around 0.5 indicates that the predicted and gold vectors are on average second-nearest neighbours, with higher values indicating better and lower values worse performance.[6] Our baseline model uses the base word's vector as the vector for the derived word. This reflects the assumption, shared by previous studies, that the base is often semantically similar to the derived word.

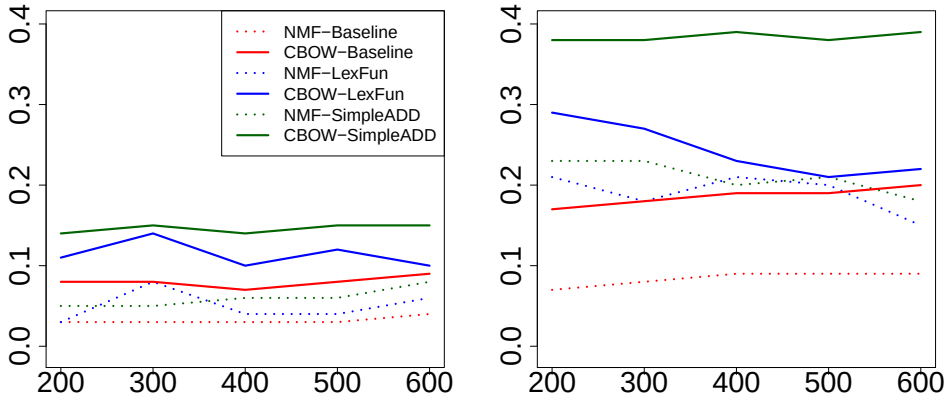
3 Results

This section focuses on the MRR metric (Cosine results are provided in Table 2). Figure 1 shows MRR for all patterns, using dimensionality as the x axis, color for models (*LexFun*: blue, *SimpleAdd*: green, Baseline: red), and line style for the distributional representation (CBOW: solid, NMF: dotted).

We first discuss the baseline. Patterns with high baselines are those where the derived words exhibit a relatively minor semantic shift from the base. We find the highest baselines for the verbal perfectivizer suffix *-nu-*, the female occupational *-k-*, and the adjective negative *ne-*. These observations match linguistic intuitions. Among the two verbal affixes for perfectivization (the action denoted by the base verb is completed), *-nu-* is the one whose contribution is more systematic; the contribution of *pro-* is more nuanced and hence likely to affect the lexical semantics of the base in a more marked way: besides denoting that the action came to an end, it can, for example, contribute a protractedness feature (from *krychaty*, "to yell" to *pro-krychaty* "to stop yelling continuously, after having done it for a while") or a directional one (from *bihty*, "to run" to *pro-bihty* "to run by, to pass"). In our experiments, *pro-* moves the vectors of the derived word much further away than *-nu-* (interestingly, the two perfectivizers are almost in a complementary distribution). Comparably, for German, Kisselew et al. (2015) report a very low baseline for the verbal prefix *durch-* ("through"), which is not semantically aligned to Ukrainian *pro-* but at least comparable to it in terms of the nature and variety of the semantic contribution (compare *braten (fry)* $\rightarrow$ *durchbraten (cook through), blättern (turn page)* $\rightarrow$ *durchblättern (skim)*). In the adjectival domain, negation (*ne-*) also shows a high baseline, the high similarity of antonyms being a known problem in distributional semantics. Once again, this result is in line with the findings of Kisselew et al. (2015), who found the highest baseline for the adjectival negation (*un-*). In this context, the low baseline exhibited by nominal negation (*ne-*) seems to suggest that a higher degree of lexicalisation is at work for nouns

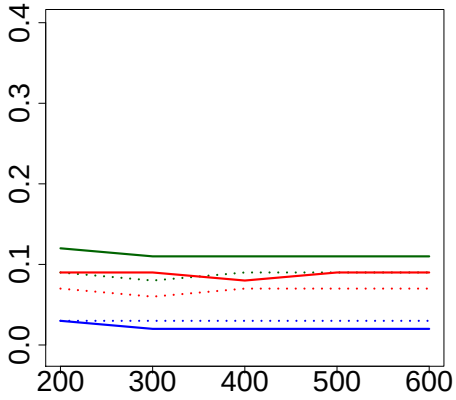
[5] Compositional models were trained using the DISSECT toolkit (Dinu et al., 2013). For comparability with Kisselew et al. (2015) and Padó et al. (2016), vectors are normalized to unit length.

[6] When calculating MRR, we restrict the neighbor lists to the words with the same part-of-speech of the derived word.

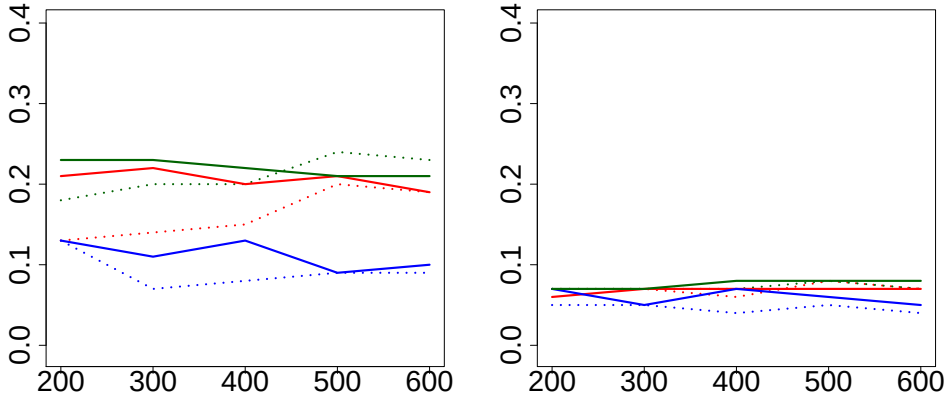


(a) ne-

(b) -k- (female)

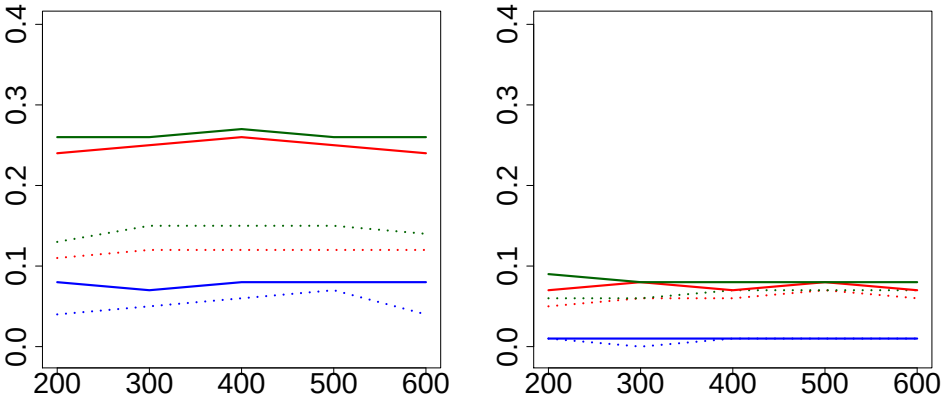


(c) -k- (diminutive)



(d) -nu-

(e) pro-



(f) ne-

(g) bez-

Figure 1: Result plots for (a)-(c): noun-noun; (d), (e): verb-verb; (f), (g): adjective-adjective. MRR performance (y-axis) by DSM dimensionality (x-axis).

("sweet"/"not sweet" vs. "friend"/"enemy"). Lower similarities for the pairs in *bez-* may be ascribed to the fact that this is, indeed, a "two-step" derivation, as *bez-* accesses the nominal component of a denominal adjective, in a derivational chain for which at times the base adjective is not even attested:[7] "noise" → "noisy" → "without noise". Within the nominal patterns, the female occupation *-k-* produces derived vectors that are closer to their bases than the other patterns. A comparable trend was found by Kisselew et al. (2015): the female pattern (*-in*) had a stronger baseline than the diminutive (*-chen*), and the second strongest baseline after the adjectival negation (*un-*).

We now proceed to the performance of the actual non-baseline CDSMs. This performance quantifies the extent to which the predicted derived vector is a good approximation of the corpus-observed derived vector and can be interpreted as the extent to which the targeted derivational shift is *predictable* (Padó et al., 2016). In additional to (lack of) predictability, other contributing factors are:

(a) lack of training data (absent here, since the data is balanced);

(b) semantic ambiguity of patterns and base words (e.g., like Ukrainian *pro-*), which forces the CDSM to learn multiple transformations at the same time (we return to this point in Section 4.1);

(c) limitations of the CDSM model (particularly true for the additive model, compare the discussion of adjective classes in Baroni et al. (2014)).

The main observations are as follows:

(a) *SimpleAdd* almost always outperforms *LexFun* independently on the underlying DSM representation (CBOW or NMF);

(b) CBOW performs better than NMF (except for *-nu-*);

(c) the best CDSMs outperform the baseline for the noun-noun patterns, for the adjectival patterns, and for the verbal perfectivizer *-nu-*;

(d) performance is relatively constant across dimensionalities, at least for the best model, *SimpleAdd* on CBOW, while *LexFun* appears to be more affected.

These results are strikingly similar to the findings of Kisselew et al. (2015) for German, to the extent that they were considered in that study (i.e., (a) and (c)). At the pattern level, the Ukrainian and German female patterns both show the strongest improvement over the baseline. In Ukrainian, the female pattern is followed by the nominal negative *ne-* (no comparable affix was tested for German) and by the diminutive *-k-* (*-chen*, in German).[8] In contrast, the verbal perfectivizer *pro-*, like the German *durch-*, and the adjectival privative *bez-* (comparably to the German noun-adjective pattern *-los*) appear to be very difficult to learn: the nuanced lexical semantics of *pro-* and the derivational chain at work in *bez-* are likely reasons.

Table 2 summarizes, in its left-hand side, the performance of the best models (CBOW, 400 dimensions, the most robust choice across patterns), in terms of MRR and Cosine (in brackets). It confirms that the SimpleAdd model shows the best performance for all patterns. MRR and Cosine correlate well but not perfectly (compare *nu-*, *ne-*). The right-hand side aligns Ukrainian patterns with German patterns from Kisselew et al. (2015) where available. Since Kisselew et al. used a different evaluation metric, we only report the relative ranking of the patterns' performances (1: best pattern, 4: worst pattern).[9] The table confirms the good correspondence: the two best-performing and worst-performing patterns align perfectly.

[7] This is obviously not the case for our experimental pairs, for which both base and derived are attested.

[8] Note that *-k-* only takes female nouns as bases, while *-chen* is unrestricted.

[9] *-los* was not tested in that study, but was included in Padó et al. (2016), where it was found that *SimpleAdd* outperformed *LexFun* on this pattern.

	Ukrainian			German			
Pattern	Baseline	LexFun	SimpleAdd	Affix	Best	CDSM Rank	Baseline Rank
V-PERF (*pro-*)	.07 (.35)	.07 (.33)	**.08 (.40)**	*durch-*	SimpleAdd	4	4
V-PERF (*-nu-*)	.20 (.52)	.13 (.47)	**.22 (.58)**	–	–	–	–
A-NEG (*ne-*)	.26 (.45)	.08 (.32)	**.27 (.48)**	*un-*	Baseline	2	1
A-PRIV (*bez-*)	.07 (.29)	.01 (.19)	**.08 (.33)**	*-los*	SimpleAdd	–	–
N-FEM (*-k-*)	.19 (.54)	.23 (.59)	**.39 (.65)**	*-in*	SimpleAdd	1	2
N-DIM (*-k-*)	.08 (.39)	.02 (.09)	**.11 (.47)**	*-chen*	SimpleAdd	3	3
N-NE (*ne-*)	.07 (.38)	.10 (.44)	**.14 (.47)**	–	–	–	–

Table 2: CDSM performance per pattern (CBOW, 400 dim.) for Ukrainian (left side: MRR and Cosine (in brackets)) and corresponding German results (Kisselew et al. (2015), right side: rank of pattern in terms of performance)

4 Case studies

In this section, we present two follow-up studies which extend our investigation of CDSMs for Ukrainian morphology in two directions. In section 4.1, we rely on the same set of items of our main experiments, but test the CDSMs in a more difficult task: that of handling ambiguous derivational shifts – in other words, the task of learning, potentially, *two shifts at once*. In section 4.2, we extend the scope of our investigation, bringing two inflectional patterns into the picture: the comparative and the superlative adjectival affixes.

4.1 Ambiguous derivation patterns

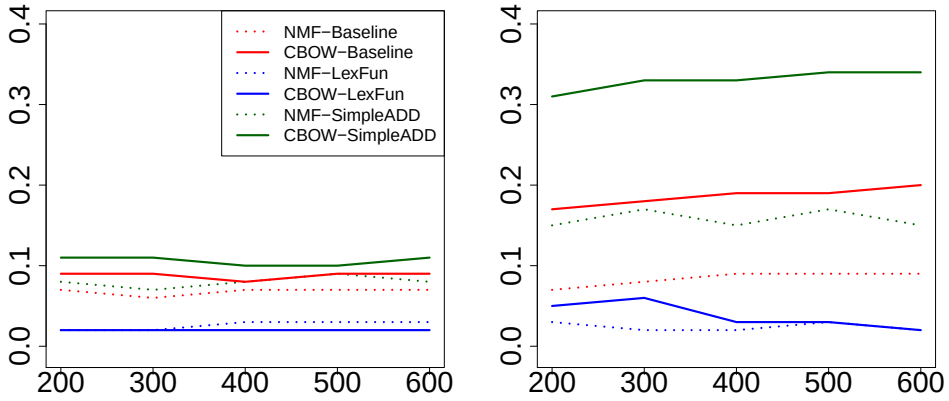
In this case study we test the capability of the CDSMs to deal with derivation patterns that express different semantic transformations, like English *-ment* which can have eventive (*enjoyment*) as well as non-eventive (*pavement*) readings (Plag, 2003).

Two patterns in our selection are suitable for this analysis: (a), *-k-* female and *-k-* diminutive (which we see as homonymy: they look identical on the surface, but have very different meanings); and (b), the adjectival and nominal *ne-* (which we see as polysemy: the negation shifts for the two parts of speech are presumably closely related). For each of the two cases, our study *combines* the datasets for the two patterns, and performs cross-validation as before; we however tease apart the results for the test items of the two original patterns. Results are shown in Figure 2 (compare to the original results in Figure 1).

For the *-k-* patterns, combined training negatively affects performance, but not dramatically so. Evidently, the DSMs manage to represent the different meanings, presumably because the two patterns are clearly distinct in terms of their semantics and of the set of bases on which they apply. This is true even for the very limited SimpleAdd model.

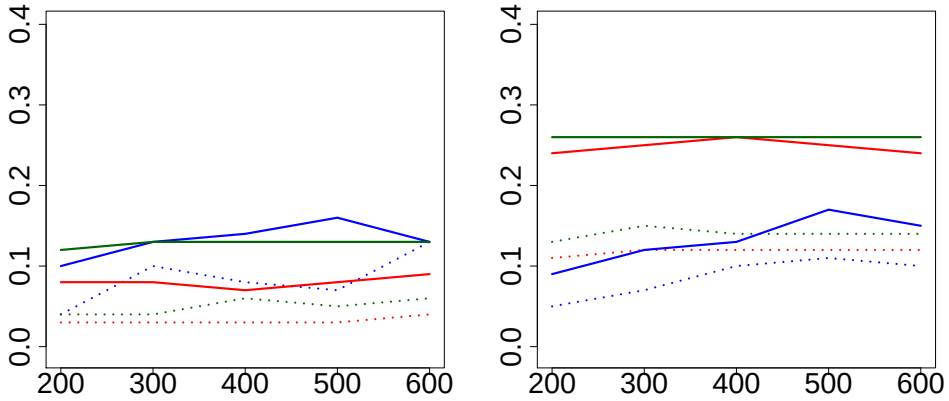
On the *ne-* patterns, we even found an improvement from combined training: performance improved for nouns, remaining constant for adjectival data; even LexFun became competitive (at higher dimensionalities); we interpret the latter result in light of the higher computational power of LexFun with respect to SimpleAdd, which naturally leads to the need of larger training data: in this set of experiments, we use twice as much training data as in the main experiments presented in the previous section. In sum, LexFun requires more training (even if slightly noisier), and when more training is available, makes better use of larger dimensionality. Note that, albeit the double amount of training, nothing could be done for the female/diminutive ambiguity: unsurprising, as, by design, in that case there was no semantic shift to learn.

Overall, our results indicate that combining patterns for modeling, maybe even beyond ambiguity, is a promising strategy not only to improve generalization and efficiency but, in the ideal case, even prediction quality.



(a) -k-: Training on diminutive+female occupation, test on diminutive

(b) -k-: Training on diminutive+female occupation, test on female occupation



(c) ne-: Training on noun+adjective, test on noun

(d) ne-: Training on noun+adjective, test on adjective

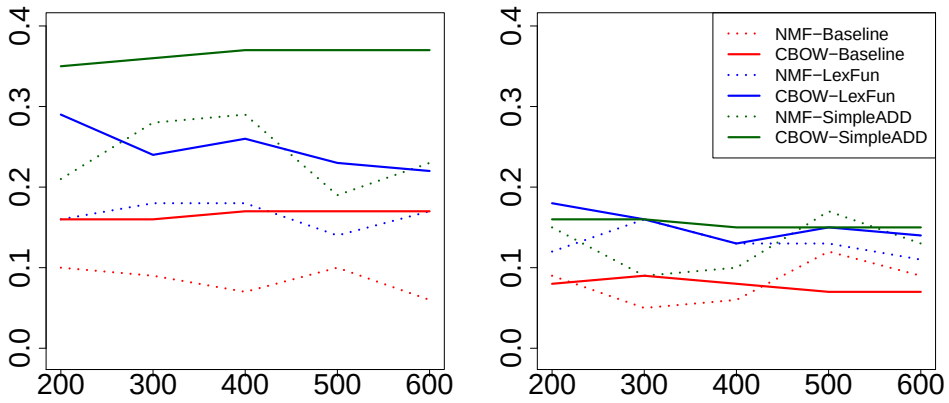
Figure 2: Case study: Combined learning of ambiguous patterns. MRR performance (y-axis) by DSM dimensionality (x-axis).

4.2 Modeling inflection with CDSMs

In addition to the set of patterns discussed above we also explore two inflectional affixes: the comparative -(і)ш- (-(i)sh-) and the superlative най- (nai-).

From a purely linguistic perspective, inflection is known to give rise to more regular shifts with respect to derivation, and we want to test this hypothesis with our CDSM methodology. From a distributional modeling perspective, Ukrainian degree morphology is interesting for our methodology because the superlative builds on the comparative as a complex base: our experiments provide additional insight into the impact of complex bases on the performance of CDMSs. From the perspective of the contribution of distributional modeling to linguistic theory, a potential future application of our quantitative methodology is that of getting more insight into the directionality of the two word-formation processes (which would have been encoded in linear order, had we been dealing with two prefixes).

Figure 3 shows the performance of our models: the comparative has a higher baseline than the superlative, and it is more predictable. For both inflectional patterns, CDSMs manage to improve on the baseline. Overall, these preliminary results can be taken as a hint of a stronger regularity (to be thoroughly tested in future work) of comparative with respect to superlative: this can be possibly interpreted at a



(a) -ish- (b) nai-

Figure 3: Case study: Inflectional patterns

purely semantic level (superlatives are more prone to undergo lexicalisation, which affects both baseline and predictability) or pragmatically (superlatives are used in more marked contexts, and more marked in a less systematic way).

5 Conclusion

In this paper, we have tested the learning capabilities of CDSMs on a morphologically rich and low-resource language: Ukrainian. Our study, albeit a pilot in terms of the restricted set of selected patterns, helped us getting a better understanding of the parameters regulating the performance of CDSMs, produced results which match linguistic intuitions, opened avenues for future work (e.g., exploring the aspectual system more thoroughly and taking systematic advantage of ambiguity to group patterns), and uncovered striking parallelism with respect to previous findings for German. We interpret the difference between our results and those by Lazaridou et al. (2013) for English (*LexFun* competitive with *SimpleAdd*) as a by-product of the different experimental setup: they adopt much larger training sets (often above 100 pairs per pattern), which probably allows the more expressive *LexFun* to take full advantage of the richer data.

Acknowledgments

The authors gratefully acknowledge funding from the DFG (SFB 732, project B9).

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.

Babych, B. and S. Sharoff (2016). Ukrainian part-of-speech tagger for hybrid MT: Rapid induction of morphological disambiguation resources from a closely related language. In *Proceedings of HyTra*.

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

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, pp. 1183–1193.

Dinu, G., N. T. Pham, and M. Baroni (2013). DISSECT - DIStributional SEmantics Composition Toolkit. In *Proceedings of ACL*, Sofia, Bulgaria, pp. 31–36.

Goldhahn, D., T. Eckart, and U. Quasthoff (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of LREC*.

Kisselew, M., S. Padó, A. Palmer, and J. Šnajder (2015). Obtaining a Better Understanding of Distributional Models of German Derivational Morphology. In *Proceedings of IWCS*, London, UK, pp. 58–63.

Köper, M., S. Schulte im Walde, M. Kisselew, and S. Padó (2016). Improving Zero-Shot-Learning for German Particle Verbs by using Training-Space Restrictions and Local Scaling. In *Proceedings of STARSEM*, Berlin, Germany, pp. 91–96.

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, pp. 1517–1526.

Lee, D. D. and H. S. Seung (2000). Algorithms for Non-negative Matrix Factorization. In *Proceedings of NIPS*, pp. 556–562.

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.

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

Padó, S., A. Herbelot, M. Kisselew, and J. Šnajder (2016). Predictability of Distributional Semantics in Derivational Word Formation. In *Proceedings of COLING*, Osaka, Japan, pp. 1285–1296.

Pala, K. and D. Hlaváčková (2007). Derivational Relations in Czech WordNet. In *Proceedings of the ACL Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, Prague, Czech Republic, pp. 75–81.

Piasecki, M., R. Ramocki, and M. Maziarz (2012). Recognition of Polish Derivational Relations Based on Supervised Learning Scheme. In *Proceedings of LREC*, Istanbul, Turkey, pp. 916–922.

Plag, I. (2003). *Word-Formation in English*. Cambridge: Cambridge University Press.

Šnajder, J. (2014). Derivbase.HR: A High-Coverage Derivational Morphology Resource for Croatian. In *Proceedings of LREC*, Reykjavík, pp. 3371–3377.