Microsyntactic Unit Detection using Word Embedding Models: Experiments on Slavic Languages

Iuliia Zaitova, Irina Stenger, Tania Avgustinova
Department of Language Science and Technology, Saarland University, Germany
izaitova@lsv.uni-saarland.de
ira.stenger@mx.uni-saarland.de
avgustinova@coli.uni-saarland.de

Abstract

Microsyntactic units have been defined as language-specific transitional entities between lexicon and grammar, whose idiomatic properties are closely tied to syntax. These units are typically described based on individual constructions, making it difficult to understand them comprehensively as a class. This study proposes a novel approach to detect microsyntactic units using Word Embedding Models (WEMs) trained on six Slavic languages, namely Belarusian, Bulgarian, Czech, Polish, Russian, and Ukrainian, and evaluates how well these models capture the nuances of syntactic non-compositionality.

To evaluate the models, we develop a cross-lingual inventory of microsyntactic units using the lists of microsyntantic units available at the Russian National Corpus. Our results demonstrate the effectiveness of WEMs in capturing microsyntactic units across all six Slavic languages under analysis. Additionally, we find that WEMs tailored for syntax-based tasks consistently outperform other WEMs at the task. Our findings contribute to the theory of microsyntax by providing insights into the detection of microsyntactic units and their cross-linguistic properties.

1 Introduction

Microsyntactic units, which include syntactic idioms and non-standard syntactic constructions, have been defined as language-specific transitional entities between the lexicon and the grammar, idiomatic properties of which are closely tied to syntax (Iomdin, 2017). These units include all the syntactic units that have very specific and even syntactic properties and do not fit into the standard syntax (Iomdin, 2015). Recent research efforts have resulted in the development of several linguistic resources for microsyntactic analysis, such as a microsyntactic dictionary of Russian, a microsyntactically annotated corpus of Russian texts, and a typology of relevant phenomena (Marakasova and Iomdin, 2016; Iomdin, 2016, 2017; Avgustinova and Iomdin, 2019).

Given the vast number and diverse nature of microsyntactic phenomena, it is not surprising that they are often described on the basis of individual constructions or small classes of syntactic phrases. In order to gain a more comprehensive and systematic understanding of these phenomena, it is crucial to attempt an analysis of microsyntactic phenomena at scale, rather than in isolation. In this study, we add to the line of research on microsyntax by adapting quantitative and computational methods used in idiom recognition for identification of microsyntactic units in large corpora of texts and across different languages. We apply different types of Word Embedding Models (WEMs) to the task of microsyntactic unit detection, and test their performance on five functional categories of microsyntactic unit (prepositions, adverbials and predicatives, parenthetical expressions, conjunctions, and particles) in six Slavic languages (Belarusian, Bulgarian, Czech, Polish, Russian, Ukrainian).

Concretely, the contributions of this paper are as follows:

1. We demonstrate that the methods used for idiom recognition can be applied for microsyntactic unit recognition.

2. We find that embedding models adapted for syntactic tasks outperform other WEMs at the task of microsyntactic unit detection.

3. We show that the behavior of embedding models across different types of microsyntactic units has similarities across all six Slavic languages under analysis and is readily generalizable.

Proceedings of Recent Advances in Natural Language Processing, pages 1265–1273
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_134

Our study not only contributes to the theory of microsyntax but also has practical applications in Natural Language Processing, Machine Translation, and other areas of Computational Linguistics where effective handling of non-standard syntactic structures is required.

After presenting the relevant background in Section 2, we introduce the used methods, data and models in Section 3. The obtained results are discussed in Section 4, and finally, the conclusions are drawn in Section 5. The code used for our experiments is available at github.com/IuliiaZaitova/Microsyntactic-Unit-Detection-using-Word-Embedding-Models-Slavic-Languages.

2 Background

2.1 Cross-lingual Comparison of Microsyntactic Units

The cross-linguistic comparability of microsyntactic phenomena has been demonstrated for both closely related and distant languages.

Apresjan (2014) conducts a corpus study to assess the translatability of Russian syntactic idioms, which are a sub-type of microsyntactic units, into English. The study concludes that syntactic idioms are language-specific, but acknowledges the borderline situations in which a syntactic idiom in a first language and its correlate in a second language have partially different properties, implying that it is still possible to compare microsyntactic phenomena cross-linguistically, albeit indirectly.

The study by Avgustinova and Iomdin (2019) provides further evidence for the cross-linguistic comparability of microsyntactic units. The authors investigate the typology of microsyntactic units in four Slavic languages – Bulgarian, Czech, Polish, and Russian – and find that many of the peculiarities of microsyntactic units in one language can be partially reproduced in cognate languages. They propose an approach that uses an existing database of microsyntactic units in Russian available at the Russian Natonal Corpus (rus, 2003–2023) as the pivot source and present a method for parallel examination of microsyntactic units, which could be utilized to create multilingual resources for dealing with non-standard syntactic phenomena.

Even though direct cross-linguistic comparison of microsyntactic units may not always be possible, the use of partial correlates for comparative analysis can provide valuable insights into the nature of microsyntactic phenomena across different languages.

2.2 Word Embedding Models

While current research lacks a specific focus on computational at-scale analysis of microsyntactic units, previous studies suggest that the non-compositionality of idioms and microsyntactic units are closely intertwined. As such, Apresjan (2014) claims that possibly all or the majority of idioms also possess certain compositional properties either on a syntax level or a semantic level or both. We assume that research on semantic compositionality, and in particular, the computational methods and techniques utilized in idiomatic unit recognition, could provide valuable insights for addressing the problem of syntactic idiomaticity.

Despite recent advancements in transformer-based architectures, WEMs remain a popular choice in tackling non-compositionality detection tasks (Salehi et al., 2015; Cordeiro and Candito, 2019; Nandakumar et al., 2019; Hashempour and Villavicencio, 2020). WEMs use context information and represent the meaning of lexical units as vectors based on the idea that words occurring in similar contexts tend to have a similar meaning. At present, research does not agree on a definitive metric to measure the modeling capabilities of WEMs as applied to the non-compositionality detection task. Consequently, different studies have also produced different results when comparing the performance of different WEMs.

Among the WEMs available, research on idiom detection highlights the effectiveness of the Word2Vec CBOW model (Mikolov et al., 2013). As such, in their large-scale evaluation of 816 WEMs Cordeiro et al. (2016) show that Word2Vec CBOW-based architectures produce the best results in detection of semantic non-compositionality in nominal compounds. Additionally, Nandakumar et al. (2019), in their study on how well seven different embedding methods capture the nuances of non-compositional data, also find that the Word2Vec model (the default configuration of Word2Vec is CBOW) performs the best. Moreover, they show that recently-proposed contextualized word embeddings (CWEs) such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) are not adept at handling non-compositionality.

In defense of CWEs, Hashempour and Villavicencio (2020) find that the Context2Vec model prep – prepositions, adv & pred – adverbial and predicative, parenth – parenthetical, conj – conjunctions, part – particles.

Type	**BE**	**UK**	**BG**	**CS**	**PL**	**RU**
Prep	ў канцы	у кінці	в края на	na konec	w końcu	в конце
Eng. trans.	*at the end of*	*at the end of*	*at the end of*	*at the end of*	*at the end of*	*at the end of*
Adv & Pred	не раз	не раз	не веднъж	ne jednou	niejednokrotnie	не раз
Eng. trans.	*not once*	*not once*	*not once*	*not once*	*not once*	*not once*
Parenth	такім чынам	таким чином	по такъв начин	tímto způsobem	w taki oto sposób	таким образом
Eng. trans.	*in this way*	*in this way*	*in this way*	*in this way*	*in this way*	*in this way*
Conj	хіба толькі	хіба що	освен да	snad jen	chyba że	разве что
Eng. trans.	*except (only) that*	*except (only) that*	*except (only) that*	*except (only) that*	*except (only) that*	*except (only) that*
Part	усе ж	все же	все пак	asi spíš	więc jednak	все же
Eng. trans.	*nonetheless*	*nonetheless*	*nonetheless*	*nonetheless*	*nonetheless*	*nonetheless*

We use ISO 639-1 codes for the languages: Belarusian – be, Ukrainian – uk, Bulgarian – bg, Czech – cs, Polish – pl, Russian – ru.

Table 1: Microsyntactic units in six Slavic languages.

(Melamud et al., 2016) outperforms the Word2Vec and BERT models due to its ability to place potentially idiomatic expressions into distinct regions of the embedding space (idiomatic/literal) depending on the particular sense of the expression in context.

3 Methodology

3.1 Slavic Languages

We focus on six Slavic languages that belong to the three main sub-groups of the Slavic language family: Belarusian, Ukrainian, and Russian (East Slavic); Bulgarian (South Slavic); and Polish and Czech (West Slavic). This language selection was made to ensure the inclusion of diverse typological variations across the Slavic languages. Each of the chosen languages has publicly-available large-scale corpora, as well as parallel multilingual data, providing a rich resource for our analysis. By including languages from different sub-groups, we aim to capture a broad range of syntactic and semantic phenomena within the Slavic language family. This allows us to conduct a comprehensive analysis of microsyntactic units from a typological perspective.

3.2 Inventory of Microsyntactic Units

To develop a cross-lingual inventory of microsyntactic units, we adopted the methods proposed by Avgustinova and Iomdin (2019) and utilized the Russian National Corpus (RNC) and its parallel sub-corpora (rus, 2003–2023) as the primary linguistic resource. The microsyntactic dictionary[1] provided by the RNC, which includes prepositions, adverbials and predicatives, parenthetical expressions, conjunctions, and particles, served as our pivot database for the development of a multilingual comparative resource of microsyntactic phenomena. Although on the website the dictionary is called 'corpus dictionary of multi-word lexical units', for the purpose of this work we use the name 'microsyntactic dictionary' to emphasize the syntactic idiomaticity of given expressions. For each Russian expression, the database also provides its frequency score in the RNC sub-corpora and the syntactic function that the expression has.

We sorted the available expressions by their frequency scores and selected the 50 most frequent microsyntactic units from each syntactic category except for particles, for which only 27 distinct expressions are available. This yielded a total of 227 microsyntactic units in Russian for further analysis. For each expression, we used the search function of the RNC to extract translational correlates together with two parallel bilingual context sentences from the parallel sub-corpora. We acknowledge that direct correlates of microsyntactic units in different languages are not always available. Thus, we opt for using partial correspondence whenever required, which we believe, despite its limitations, allows us to compare microsyntactic units at scale. In a similar way, we used the search function of the Czech National Corpus (Machálek, 2020). We obtained six parallel sets of 227 microsyntactic units with parallel bilingual context sentences for each unit in all of the six Slavic languages under analysis. The bilingual sentences can be used for future research on microsyntactic units in context. It is important to mention that in contrast to Avgustinova and Iomdin (2019), we had to choose only one equivalent for each of the microsyntactic units in Russian to enable quantitative and computational analysis. Each of the translated expressions and sentences was proofread and, when required, corrected by professional linguists who are also native speakers of the target language.

Our multilingual database of microsyntactic phe-

[1] https://ruscorpora.ru/page/obgrams/

nomena enables us to compare these phenomena across different languages and can be later used for further research on microsyntactic units and syntactic idiomaticity. To the best of our knowledge, this is the first database of its kind that allows for quantitative and computational analysis of microsyntactic units across different languages. Further examples of the obtained data for each type of microsyntactic unit are provided in Table 1, which showcases the microsyntactic units in Russian along with their corresponding translations to other languages under analysis. Our custom dataset is fully open-sourced and is available at huggingface.co/datasets/izaitova/slavic_fixed_expressions.

3.3 Inventory of Syntactically Compositional Counterparts

For each target microsyntactic unit, we have drawn compositional (non-idiomatic) constructions from the training data as counterparts using random sampling. For the purpose of normalization, we ensured that they have the same number of constituent tokens and share at least one word with the counterpart microsyntactic unit. For instance, for the microsyntactic unit *ne jednou* in Czech, the compositional counterpart should be two words in length and contain either the word *ne* or *jednou*. To refine the selection of the non-microsyntactic counterparts, we manually removed any non-compositional units from the initially sampled list and conducted further random sampling until we obtained the full set of compositional counterparts.

3.4 Training Data

The training data for our experiments is sourced from the Leipzig Corpora Collection (LCC) (Goldhahn et al., 2012), which is a publicly available corpus containing text data generated from newspapers and web resources in 293 languages. For each language under analysis, we utilized 500,000 sentences sourced from language-specific news corpora of LCC.

3.5 Word2Vec CBOW

We chose to use the CBOW architecture of the Word2Vec model due to its demonstrated effectiveness in semantic non-compositionality detection, as highlighted in previous research (Section 2.2). Word2Vec CBOW predicts the center word given a representation of the surrounding words, whereas its counterpart Word2Vec Skip-gram predicts contextual words given the representation of the center word[2]. To train the Word2Vec CBOW model, we use Gensim's implementation of the CBOW algorithm (Řehůřek and Sojka, 2010). We ignore all words that occur less than five times in the training corpus, and use a window size of five.

3.6 Context2Vec

Following Hashempour and Villavicencio (2020), we decided to use Context2Vec (Melamud et al., 2016) due to its ability to capture variable-length sentential contexts using a bidirectional LSTM recurrent neural network. We use an optimized implementation of Context2Vec by Aoki (2018) with the original parameters of the model. It is important to note that Hashempour and Villavicencio (2020) show a superior performance of this model when applied to different senses of a token. Although in our experiments we use a single embedding for each token for better comparison with other models, we anticipate that Context2Vec's improved representation of context will contribute to the detection of syntactic compositionality.

3.7 Structured Skip-gram Word2Vec and Word2Vec CWindow

To enhance the quality of word embeddings for syntax-based tasks, we included Structured Skip-gram Word2Vec[3] and Word2Vec CWindow models (Ling et al., 2015) in our methodology. These modified versions of the Word2Vec Skip-gram and Word2Vec CWindow algorithms take into account the relative positions of context words and have been shown to improve parsing accuracy for part-of-speech tagging and dependency parsing tasks. We anticipate these models will offer valuable insights into the detection of syntactic non-compositionality due to their enhanced understanding of token relationships. For both models, we ignore all words that occur less than five times in the training corpus, and use a window size of five.

3.8 Graph-based Syntactic Word Embeddings with Node2Vec

Incorporating a graph-based approach, we utilize the Node2Vec algorithm (Grover and Leskovec,

[2] Due to existing evidence for better performance of CBOW as compared to Skip-gram in compositionality detection, we use only the CBOW configuration

[3] The Structured Skip-gram model is different from Word2Vec Skip-gram

2016) which learns syntactic embeddings based on information derived from dependency parse trees. Previous research by Al-Ghezi and Kurimo (2020) has demonstrated competitive performance of Node2Vec embeddings in part-of-speech tagging tasks compared to other WEMs. By employing dependency parse trees generated by DiaParser (Zhang and Attardi, 2020), we aim to explore the dependencies between tokens in a sentence and leverage Node2Vec's ability to preserve network neighborhoods of nodes for syntactic non-compositionality detection. To train the Node2Vec models, we use PecanPy (Liu and Krishnan, 2021), an accelerated implementation of the Node2Vec algorithm, with default parameters.

3.9 Experimental Setup

For each of the six Slavic languages under analysis, we construct word embeddings using five models: Word2Vec CBOW, Context2Vec, Structured Skip-gram Word2Vec, Word2Vec CWindow, and Node2Vec. Additionally, we generate these word embeddings for two different dataset sizes, one consisting of 100,000 sentences and another of 500,000 sentences.

To pre-process the datasets, we 1) lower-cased the texts; 2) removed punctuation and non-alphanumerical tokens; 3) randomly selected from 5 to 100 sentences containing occurrences of each of the target expressions, including both microsyntactic and compositional phrases; 4) supplemented the data with additional sentences from the corpus up to either 100,000 or 500,000 sentences, depending on the type of experiment being conducted; 5) following Cordeiro et al. (2016), retokenized all target expressions as a single token with a separator (underscore) between the phrase constituents (e.g. *so far* → *so_far*) to represent target expressions as one unit both in training and testing.

3.10 Non-compositionality Prediction

To predict the non-compositionality of an expression, we use cosine similarity between the expression vector representation $v(w1w2)$ and the sum of the vector representations of the component words $v(w1 + w2)$. This method has been extensively used in previous research on non-compositionality prediction (Mitchell and Lapata, 2010; Salehi et al., 2015; Cordeiro et al., 2016; Loukachevitch and Gerasimova, 2017; Nandakumar et al., 2018, 2019), formally:

$$cos(v(w1w2), v(w1 + w2))$$

where for $v(w1 + w2)$ we use the normalized sum

$$v(w1 + w2) = \frac{v(w1)}{||v(w1)||} + \frac{v(w2)}{||v(w2)||}$$

Intuitively, an expression appearing in different contexts from its components is likely to be non-compositional. In this framework, a phrase is compositional if its representation is close to the sum of its component representations (cosine similarity is close to 1), and it is idiomatic otherwise.

In order to compare the results and analyze the variations in performance, all expressions are arranged in ascending order based on their similarity scores. The aim is to examine whether the compositional phrases would have higher similarity values compared to non-compositional phrases. To evaluate the ordering quality, the measure of mean average precision (MAP) is employed – this way, MAP = 1 would correspond to all microsyntactic units ordered lower than compositional expressions, and MAP = 0 would mean that all microsyntactic units are ordered higher than compositional ones.

4 Results and Discussion

The experimental findings for the five models trained on 100,000 and 500,000 sentences are summarized in Tables 2 and 3, respectively. Table 2 shows the MAP scores on 100,000 sentences and Table 3 shows the MAP scores on 500,000 sentences. The best scores per language are presented in bold.

On the datasets of 100,000 sentences (Table 2), Node2Vec achieves the highest score for four out of six languages, while Word2Vec CWindow performs best on Belarusian and Russian for a dataset size of 100,000 sentences. On a larger dataset size of 500,000 sentences (Table 3), the models' performance generally improves, but with less uniform results, which suggests that some of the studied models might require more data to make meaningful generalizations. Overall, the results show that syntax-adapted models (except for Word2Vec Structured Skip-gram) tend to perform better in identifying microsyntactic units, which aligns with our expectations related to the nature of these units. Surprisingly, Node2Vec, which is based on dependency-parsed graphs, does not consistently

	Word2Vec CBOW	Word2Vec CWindow	Word2Vec Structured Skip-gram	Context2Vec	Node2Vec
Czech	0.608*** (0.607–0.61)	0.643*** (0.645–0.648)	0.524 (0.52–0.523)	0.559*** (0.556–0.559)	**0.678***** (0.685–0.688)
Polish	0.594*** (0.596–0.599)	0.595*** (0.596–0.599)	0.604* (0.609–0.612)	0.507*** (0.504–0.507)	**0.626***** (0.627–0.63)
Bulgarian	0.652*** (0.652–0.655)	0.674*** (0.675–0.677)	0.559** (0.557–0.56)	0.542** (0.543–0.546)	**0.709***** (0.707–0.71)
Ukrainian	0.564*** (0.572–0.575)	0.617*** (0.616–0.619)	0.537* (0.53–0.533)	0.573*** (0.566–0.575)	**0.718***** (0.716–0.718)
Belarusian	0.568*** (0.565–0.568)	**0.674***** (0.672–0.675)	0.546* (0.54–0.543)	0.533** (0.532–0.54)	0.639*** (0.636–0.639)
Russian	0.656*** (0.654–0.657)	**0.705***** (0.707–0.709)	0.643*** (0.64–0.643)	0.564*** (0.561–0.564)	0.551** (0.552–0.555)

*95% Bootstrapping Confidence Intervals in parentheses; *p<0.1; **p<0.05; ***p<0.01*

Table 2: MAP results on 100,000 sentences.

	Word2Vec CBOW	Word2Vec CWindow	Word2Vec Structured Skip-gram	Context2Vec	Node2Vec
Czech	0.63*** (0.628–0.631)	0.652*** (0.65–0.653)	0.617*** (0.614–0.619)	0.546** (0.542–0.548)	**0.678***** (0.676–0.679)
Polish	0.665*** (0.668–0.671)	0.634*** (0.637–0.64)	0.577*** (0.581, 0.584)	0.612*** (0.596–0.599)	**0.683***** (0.675–0.686)
Bulgarian	0.67*** (0.664–0.667)	**0.718***** (0.715–0.718)	0.674*** (0.677–0.68)	0.537* (0.535–0.538)	0.66*** (0.655–0.658)
Ukrainian	0.665*** (0.658–0.666)	**0.705***** (0.706–0.708)	0.652*** (0.651–0.654)	0.595*** (0.592–0.595)	0.66*** (0.665–0.668)
Belarusian	0.621*** (0.619–0.622)	**0.7***** (0.696–0.702)	0.639*** (0.64–0.643)	0.537* (0.531–0.538)	0.533 (0.524–0.538)
Russian	0.67*** (0.671–0.674)	0.718*** (0.717–0.72)	**0.744***** (0.743–0.746)	0.586*** (0.583–0.586)	0.66*** (0.657–0.66)

*95% Bootstrapping Confidence Intervals in parentheses; *p<0.1; **p<0.05; ***p<0.01*

Table 3: MAP results on 500,000 sentences.

outperform other syntax-based WEMs that only account for word order. For most languages, it also produces similar or worse results when trained on larger sets of sentences. The Context2Vec model performs poorly on all languages even compared with Word2Vec CBOW, indicating that variable-length sentential context generated by a bidirectional LSTM recurrent neural network is not beneficial for syntactic non-compositionality detection. As for Word2Vec Structured Skip-gram, we know that similarly to the Word2Vec Skip-gram architecture (Section 3.5), it predicts the context tokens given the center token. In the case of syntactic compositionality prediction, where the relationships between words within a phrase are crucial, it could be more advantageous to predict the center token and capture information from its sentential context, which helps in understanding the sentence's structure.

To better interpret why Word2Vec Structured Skip-gram, despite its generally low performance, significantly outperforms other models in the case of the Russian language (500,000 sentences), it is helpful to compare the results by category of microsyntactic units. Figures 1 and 2 depict violin plots of cosine similarity scores by category for the two best performing models trained on the 500,000 sentence dataset in Russian. A clear difference is observable between the distributions of microsyntactic units and compositional units on both plots. Moreover, we can see that the distribution of cosine similarity scores for microsyntactic units is wider for the Structured Skip-gram model, while there is an opposite tendency in the CWindow plots, where compositional units seem to have a wider range of scores. The wider distribution of cosine similarity scores, which influences the quality of ordering, could be one of the factors that contributed to the observed outlier in the MAP score.

From the violin plots, we can also see that some unit types show a higher difference from compositional units. One explanation for that is that some types, such as adverbial and predicative constructions, additionally possess a lower degree of semantic non-compositionality, to which our models are sensitive.

Figure 3 represents the average MAP scores for models trained on 500,000 sentences, grouped by category and averaged across languages. This figure further supports the observation of varying model performance across different linguistic categories. Certain categories (adverbial and predicative, particles) consistently exhibit higher scores across all models, indicating that their non-compositionality is easier for the models to predict. Similarly, prepositions consistently yield lower scores.

4.1 Cross-Lingual Comparison

Cross-lingual comparison of microsyntactic unit recognition is essential for assessing the behavior and scalability of the non-compositionality detection techniques. To get a better representation of the results on microsyntactic unit recognition across languages, we generated heatmaps of MAP scores by category produced by Word2Vec CWindow and Node2Vec models trained on 500,000 sentences (Figure 4). The heatmaps show the performance of each unit type for each language, with darker colors indicating better performance.

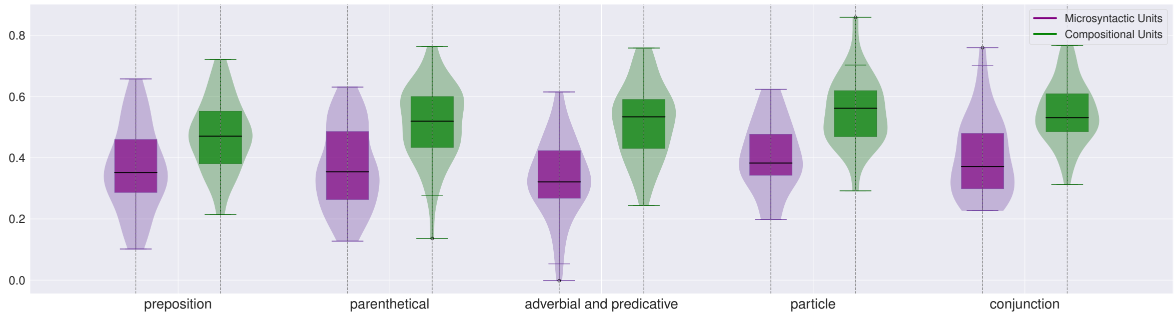


Figure 1: Cosine similarity by type of unit – Word2Vec CWindow trained on 500,000 sentences in Russian.

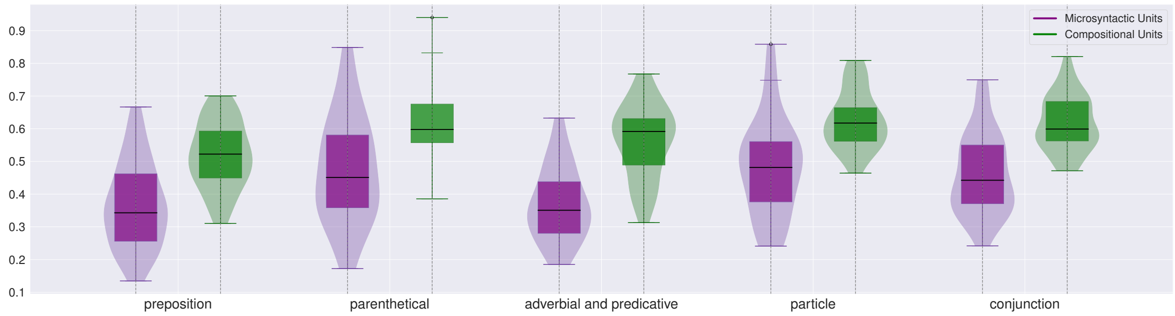
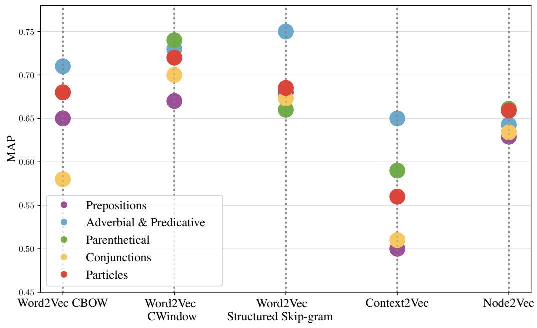


Figure 2: Cosine similarity by type of unit – Word2Vec Structured Skip-gram trained on 500,000 sentences in Russian.



Results are averaged across languages

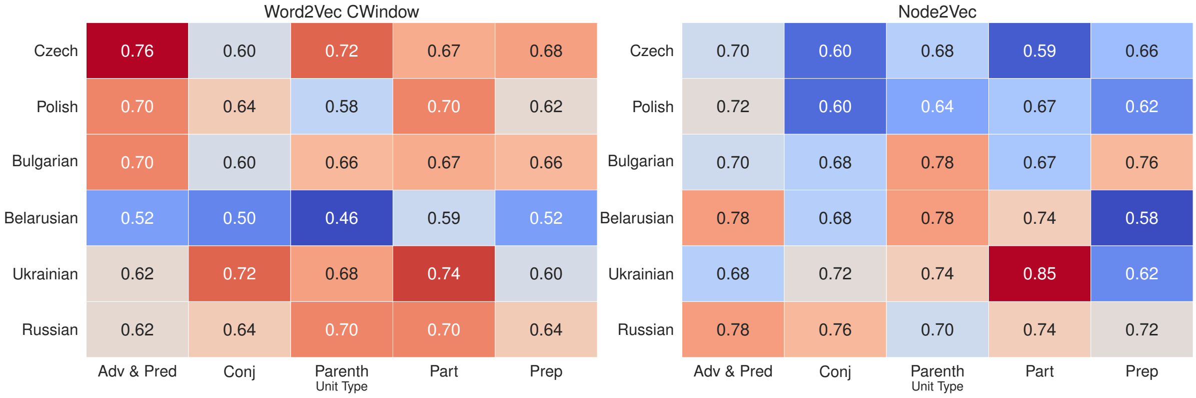
Figure 3: MAP by category for 500,000 sentences.

Across the heatmaps, we observe similarities in performance scores among different languages. For instance, adverbial and predicative constructions, as well as particles, exhibit higher MAP scores compared to other categories. These patterns suggest the presence of shared structural and/or semantic features in types of microsyntactic constructions across different languages.

5 Conclusion and Future Work

In this paper, we presented a novel approach for using WEMs for microsyntactic unit recognition in six Slavic languages. We have built a multilingual comparative database of microsyntactic units in six Slavic languages, each with six sets of parallel bilingual context sentences. Our comparative evaluation of Word2Vec CBOW, Word2Vec CWindow, Word2Vec Structured Skip-gram, Context2Vec and Node2Vec models suggests that WEMs can be effective for non-compositionality prediction, and that WEMs adapted to syntax-based tasks outperform other types of WEMs. The analysis of results shows that there are some differences in the performance of microsyntactic unit recognition across types of these units. In this vein, we have observed that different languages tend to produce similar results across different types of microsyntactic units.

In our future work, we are interested in improving the results for microsyntactic unit recognition. This includes investigating the use of additional features or data sources to improve model performance, as well as exploring different modeling architectures, such as large language models. Additionally, the inconsistent results of microsyntactic unit recognition when split by category also highlight the importance of evaluating models on different types of syntactic non-compositionality. Finally, we plan to explore the use of our database in practical applications, such as improving machine translation systems and using our models as



prep – prepositions, adv & pred – adverbials and predicatives, parenth – parentheticals, conj – conjunctions, part – particles.

Figure 4: Heatmaps of MAP scores by language for Word2Vec models trained on 500,000 sentences.

predictors for intercomprehension experiments.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive and insightful feedback on the paper.

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102 and by Saarland University (UdS-Internationalisierungsfonds, Transform4Europe).

References

2003–2023. Russian National Corpus. http://ruscorpora.ru. Accessed 25.07.2023.

Ragheb Al-Ghezi and Mikko Kurimo. 2020. Graph-based syntactic word embeddings. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 72–78, Barcelona, Spain (Online). Association for Computational Linguistics.

Tatsuya Aoki. 2018. PyTorch implementation of context2vec from Melamud et al., CoNLL 2016. https://github.com/tatsuokun/context2vec.

Valentina Apresjan. 2014. Syntactic idioms across languages: corpus evidence from Russian and English. *Russ Linguist*, 52(2):319–358.

Tania Avgustinova and Leonid Iomdin. 2019. *Towards a Typology of Microsyntactic Constructions*, volume 11755 of *Lecture Notes in Computer Science*, pages 15–30. Springer, Cham.

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1986–1997, Berlin, Germany. Association for Computational Linguistics.

Silvio Ricardo Cordeiro and Marie Candito. 2019. Syntax-based identification of light-verb constructions. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 97–104, Turku, Finland. Linköping University Electronic Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable Feature Learning for Networks.

Reyhaneh Hashempour and Aline Villavicencio. 2020. Leveraging contextual embeddings and idiom principle for detecting idiomaticity in potentially idiomatic expressions. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 72–80, Online. Association for Computational Linguistics.

Leonid Iomdin. 2015. Microsyntactic constructions formed by the russian word *raz*. *SLAVIA časopis pro slovanskou filologii*, 84(3).

Leonid Iomdin. 2016. Microsyntactic Phenomena as a Computational Linguistics Issue. In *Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces (GramLex)*, pages 8–17, Osaka, Japan. The COLING 2016 Organizing Committee.

Leonid Iomdin. 2017. Microsyntactic Annotation of Corpora and its Use in Computational Linguistics Tasks. *Journal of Linguistics/Jazykovedný casopis*, 68.

Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2015. Two/too simple adaptations of Word2Vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, Denver, Colorado. Association for Computational Linguistics.

Renming Liu and Arjun Krishnan. 2021. PecanPy: a fast, efficient, and parallelized Python implementation of node2vec. *Bioinformatics*.

Natalia Loukachevitch and Anastasia Gerasimova. 2017. Human associations help to detect conventionalized multiword expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 459–466, Varna, Bulgaria. INCOMA Ltd.

Tomáš Machálek. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Anna Marakasova and Leonid Iomdin. 2016. Mikrosintaksičeskaja razmetka v korpuse russkix tekstov SynTagRus [microsyntactic tagging in the SynTagRus corpus of Russian texts.]. In *Informacionnye texnologii i sistemy 2016 (ITiS'2016). Sbornik trudov 40-oj meždisciplinarnoj školy–konferencii IPPI RAN*, pages 445–449, Repino, Saint Petersburg, Russia.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. 2019. How well do embedding models capture non-compositionality? a view from multiword expressions. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 27–34, Minneapolis, USA. Association for Computational Linguistics.

Navnita Nandakumar, Bahar Salehi, and Timothy Baldwin. 2018. A comparative study of embedding models in predicting the compositionality of multiword expressions. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 71–76, Dunedin, New Zealand.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Yu Zhang and Giuseppe Attardi. 2020. Direct Attentive Dependency Parser. `https://github.com/Unipisa/diaparser`.