Exploring Word Sense Distribution in Ukrainian with a Semantic Vector Space Model

Nataliia Cheilytko and **Ruprecht von Waldenfels**
Friedrich Schiller University Jena
natalia.cheilytko@gmail.com, ruprecht.waldenfels@uni-jena.de

Abstract

The paper discusses a Semantic Vector Space Model targeted at revealing how Ukrainian word senses vary and relate to each other. One of the benefits of the proposed semantic model is that it considers second-order collocates of the words and, thus, has more potential to distinguish word senses observed in a unique concordance line, compared to the models that take into account only immediate collocates. Combined with the Multidimensional Scaling technique, this model allows for a lexicographer to explore the Ukrainian word senses distribution on a large scale. The paper describes the first research results and the following steps of the initiative.

1 Introduction

Word Vector Space Model (VSM) is a distributional semantic technique initially developed in statistical natural language processing and is a principal tool in Computational Linguistics (Turney and Pantel, 2010). Such models treat text as multidimensional vector space, where a word, a combination of words, or a sentence are represented as a vector so that it is possible to apply various vector algebra operations: calculate distance between them, apply dimensionality reduction to the vector space, and cluster.

Among VSMs, a group of models targets semantic items with more precise attention. According to Hilpert and Saavedra (2020), a semantic vector is a statistically processed frequency list of all collocates of a particular word in a given corpus, which expresses the idea that one can distinguish the meaning of a word by its context. For example, compare: *"I have a great fan of rock among my friends"* (fan as a person) to *"I have a great fan with several heating options"* (fan as a device).

The representatives of the Leuven variationist school have been advocates for Semantic Vector Space Models as statistical state-of-art for lexicographers to identify semantic patterns in big unstructured corpora, helping linguists to avoid unfeasible manual data exploration (Heylen et al., 2015).

2 Approach

There are many techniques to build and examine a Semantic VSM. The proposed approach is based on a specific kind of Semantic Vector Space Model that, in addition to considering immediate collocates of words, also accounts for the second-order collocates – typical collocates of words found in concordance lines built for a given word of interest. In such a way, a word is represented not only by friends but also by friends of a friend if to follow a social network metaphor.

Heylen et al. (2012) and Montes and Geeraerts (2022) used the second-order collocate vectors to examine semantic variations of the pluricentric Dutch. Hilpert and Saavedra (2020) applied this technique to English to investigate boundaries among various senses of a polysemic word and between different lexemes. The authors showed that Semantic VSM and visual analytics could provide a solid support for lexicological analysis of polysemy in large corpora. In our research, we reproduce and expand this methodology to explore Ukrainian word sense distribution to identify how words belonging to a particular synonymous set are related and contrasted.

The experiment performed contains the following steps:

1) build Semantic Vector Space Model for the given Ukrainian corpora, including a vocabulary of the most frequent words and their co-occurrence matrix;

2) from the corpora, extract concordance lines, and calculate second-order collocate vectors for them;

3) apply dimensionality reduction with the multidimensional scaling technique, as Wheeler (2005) proposed, to the vectors and visualize them on a scatterplot for particular words from the validation

Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 73–78

set;
4) perform the pairwise calculation of cosine similarity to the vectors in question.

3 Data

Given that we are interested in identifying regional variation of Ukrainian over time, we performed our initial set of experiments on the corpus with texts published in the Kyiv region within 1940–1969. The texts in the corpus are fiction and periodic publications. The corpus size is relatively small for the word-sense exploration tasks (30 mln tokens total) but sufficient for the proof-of-concept. The source of the textual data investigated is the General Annotated Corpus of the Ukrainian[1] (Shvedova, 2020).

4 Implementation

For the experiment, a set of Python data science and natural language processing libraries have been used (scikit-learn, pandas, numpy, matplotlib, beautiful soap, NLTK).
The morphological tagger and lemmatizer of the Ukrainian[2] preprocessed the input texts. Specific high-frequency words were removed from the texts. Then a vocabulary model was created based on the word frequency across the corpus. The chosen size of the vocabulary for the experiment is 3,000 words so that the most frequent 3,000 lemmas (with the exclusion of highly-frequent grammatical words) comprised the vocabulary.
As the next step, for each pair of words from the vocabulary, we calculated the amount of time they co-occur in the same 4-token-window context in each corpus. As a result, we obtained a sparse co-occurrence matrix of collocates for the vocabulary elements and then normalized it with the PMI (Pointwise Mutual Information) index. To make the co-occurrence matrix less sparse, we kept only those columns and rows that contain at least one value with a PMI > 1.0.
In the next step, we extracted concordance lines for particular words of interest (with 5+5 and 10+10 words windows), shortened the lines to include only the words from the vocabulary, and calculated second-order collocate vectors for the lines. We consider only concordance lines with five or more vocabulary words for further processing.
The second-order collocate vector calculation is the following: for each word in a shortened concordance line, get its vector representation in the co-occurrence matrix initially built (*i. e.* a corresponding column in the co-occurrence matrix). Then average those word vectors for a particular concordance line. Therefore, each word of interest obtained concordance lines with the corresponding second-order collocate vectors[3].
For the sake of assessing the quality of the model, the multidimensional scaling technique and calculation of cosine distance were applied to the averaged second-order collocate vectors.

5 Model Assessment

Despite a relatively small input corpus, the proposed model turned out sensitive enough to distinguish different words and word senses and see commonalities among them.
The PMI co-occurrence matrix with first-order collocates already gives some understanding of word senses. Let us consider the top 10 collocates for *dvygun* ('an engine'; 'a driving force') with their PMI values (*dvygun* ('an engine') 4.53, *potužnyj* ('powerful') 3.64, *atomnyj* ('nuclear') 3.45, *polit* ('a process of flying') 3.38, *vičnyj* ('eternal') 3.27, *švydkist'* ('speed') 3.15, *raketa* ('a rocket') 3.10, *zamovknuty* ('to become silent') 3.05, *korabel'* ('a ship') 2.97.
And for its synonym *motor* ('a motor', 'an engine'): *gurkit* ('roar') 3.88, *motor* ('a motor') 3.33, *gudity* ('to buzz') 3.16, *kabina* ('a cabin') 3.16, *zavesty* ('to start an engine') 2.996, *litak* ('an airplane') 2.991, *potužnyj* ('powerful') 2.78, *traktor* ('a tractor') 2.77, *avtomašyna* ('a car') 2.73, *avtomobil'* ('an automobile') 2.64.
From these collocates, we can already see the difference between the two synonyms. In the Kyiv corpus 1940-1969, *motor* often denotes common-life vehicles (autos, boats, tractors). In contrast, *dvygun* is associated with "serious" topics like space travel, nuclear power, and metaphorically a driving force and a cause of activities and events.
Trying to distinguish word senses only by immediate collocates may not capture sense similarity if words in the context do not overlap. That is why the second-order collocate approach gives better sensitivity to similar word senses even if direct collocates in the concordance lines differ. To evaluate

[1] http://uacorpus.org/Kyiv/en
[2] https://github.com/brown-uk/nlp_uk
[3] The source code for building the model and data samples are stored at https://github.com/NataliaChey/unlp_2nd_order_vectors

the model's ability to capture word sense commonalities and differences, we have compared the cosine semantic similarity measure calculated for the second-order collocate vectors and an alternative VSM – the 200k Ukrainian floret vectors available via the spaCy framework[4]. For example, for *dvygun* in the two concordance lines:

1) *"Nočamy v tumani gorily svitliačky, i vytt'a zviriv inodi zaglušalo šum dvyguna v tabori."* (*"At night, fireflies burned in the fog, and the howling of the animals sometimes drowned out the noise of the engine in the camp."*)

2) *„Kolia prysluhavsia. Ni, dvyguny dyryzablia gudut' tak samo monotonno j nevpynno."* (*"Kolia listened. No, the airship's engines were buzzing just as monotonously and incessantly."*)

the semantic similarity measure by the second-order collocates is 0.9976, and by the Ukrainian floret vectors – 0.7508.

In this example, the different context words in both lines express the same idea that an engine creates noise, represented with different context words, which the model was able to capture with the help of information about the second-order vectors.

Currently, the vocabulary of the model accounts for 3,000 lemmas. For the first concordance line, the following vocabulary words contributed to the calculation: *zvir* ('an animal'), *inodi* ('sometimes'), *šum* ('noise'), *tabir* ('a camp'), *soldat* ('a soldier'). And for the second line, those collocates are *Kolia* (a person name), *prysluhatysia* ('to listen up'), *gudity* ('to buzz').

Let us consider another example with the pair of synonyms *dvygun* and motor in the following contexts:

1) *"Zarevly dvyguny, dribnyi drіž projšov po mašyni."* (*"The engines roared to life, and a small shudder went through the car."*)

2) *"Do jogo čujnogo vucha doletilo poforkuvann'a motora, srkyp galm. Prybuv komendant taboru – Bil'ava Bestija."* (*"A light engine whirring and brakes screeching reached his sensitive ears. The camp commandant, the Blonde Beast, had arrived."*)

The semantic similarity measure by the second-order collocates is 0.9799, and by the Ukrainian floret vectors – 0.6416.

The vocabulary elements for the former concordance line are *smuga* ('a lane'), *myt'* ('a momemnt'), *zupynytys'a* ('to stop'), *dribnyj* ('small'), *projty* ('to pass through'), *mašyna* ('a machine'). And for the latter: *vuho* ('an ear'), *legkyi* ('light'), *prybuty* ('to arrive'), *komendant* ('commander'), *tabir* ('camp').

The first results are promising. However, the initial version of the model has limitations due to a relatively small vocabulary size. It works well with the concordance lines with at least five collocates from the vocabulary. Therefore, we will significantly increase the model's vocabulary to 20–50k lemmas on the project's next iteration to make it comprehend a wider range of words in more versatile contexts.

It is also important to mention that we calculated the semantic similarity for vectors with already reduced dimensionality and only for the target words instead of the entire vocabulary. Thus, the provided comparison to the alternative language model is made solely to show that the model can find commonalities in non-overlapping concordance lines.

In addition to the cosine similarity metric, the multidimensional scaling technique made it possible to explore the model outputs visually. Figures 1–5 demonstrate the scatterplots for several synonymous word pairs. Multidimensional scaling was applied separately to the second-order collocate vectors of a particular group of words or a single word per each test case.

The dots on the plot are the reduced vectors. For a word or a pair of words, one can investigate close and remote dots to validate whether they denote similar or distinct occurrences. Such vector-space-based visual representations of word concordances bring additional insights for lexicographers targeted at exploring polysemy, various semantic relations, and semantic variation in language.

Figure 1 contains 97 vectors built from the 5+5

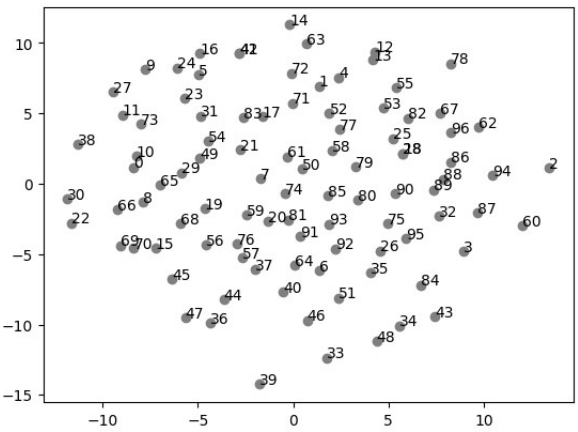

Figure 1: The Second-order Collocate Vectors: *dvygun*

[4] https://spacy.io/models/uk#uk_core_news_lg

concordance lines for *dvygun*. This word has two senses: 'engine as a mechanical device', and an abstract metaphoric sense, 'something that pushes, causes things to happen'. In the corpus, the first sense dominates. However, several vectors on the plot represent the second sense. Let us consider vector d on the right of the plot with the following collocates: *vplyv* ('influence'), *rozym* ('intelligence'), *istoria* ('history'), *rid* ('lineage'), *svitogl'ad* ('a world view').

If to compare vector 2 to vector 10 having the vocabulary collocates: *dokaz* ('a proof'), *Kolia* (a person name), *prysluhatysia* ('to listen up'), *gudity* ('to buzz'), the cosine similarity for these vectors is negative -0.7647, which indicates different senses.

Figure 2 shows 183 vectors built from the 5+5 con-

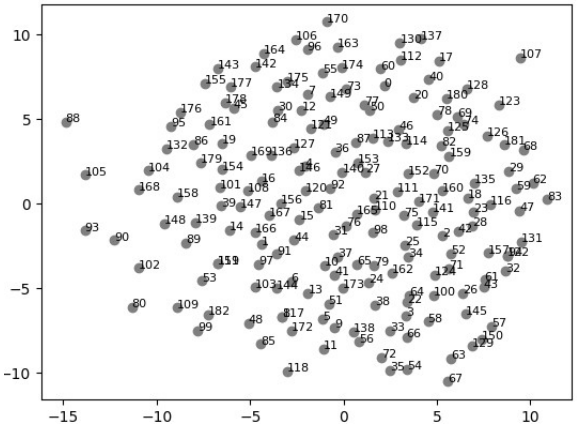

Figure 2: The Second-order Collocate Vectors: *motor*

cordance lines for *motor* ('engine'). The following two vectors demonstrate again that the model is able to capture sense similarity despite non-overlapping collocates. Consider vector 69 with the vocabulary collocates: *prosto* ('easily'), *znyknuty* ('to dissapear'), *prolunaty* ('to resound'), *signal* ('a signal'), *zakryčaty* ('to scream') and vector 74 with the vocabulary collocates: *mašyna* ('a machine'), *movčaty* ('to be silent'), *rušyty* ('to move'), *pravoruč* ('to the right'). For these two vectors, the semantic similarity by the second-order collocates is 0.99, and by the Ukrainian floret vectors is 1.0.

The model has captured the concordance lines with another sense of motor, usually occurring as an exclamation during a movie production: *"Motor!"* as *"Action!"*. The vectors 86, 176–179 are located quite close to each other on the plot. The corresponding concordance lines contain a word denoting exclamation and attention (*kruknuty, kryčaty* ('to cry out'), *vyguknyty* ('to exclaim') and *uvaga* ('attention')), as well as nouns denoting movie production – *režyser* ('director') and the likes.

Vector 107 (with the vocabulary collocates *vidro* ('a bucket'), *krutyty* ('to spin'), *kriz'* ('through'), *dirka* ('a hole')) stays apart on the left for a reason. The context is atypical – a humorous story from a humoristic magazine *Perets, 1961* on how to construct an engine from a bucket: *"Vkladajete v take vidro vyprany bilyznu, motor krutyt' vidro, voda kriz' dirky vylitaje, bilyzna sohne na očah!.."* (*"Put the laundry in the bucket, the motor spins the bucket, the water flies out through the holes, and the laundry dries before your eyes!"*)

Figure 3 shows that *motor* and *dvygun* have over-

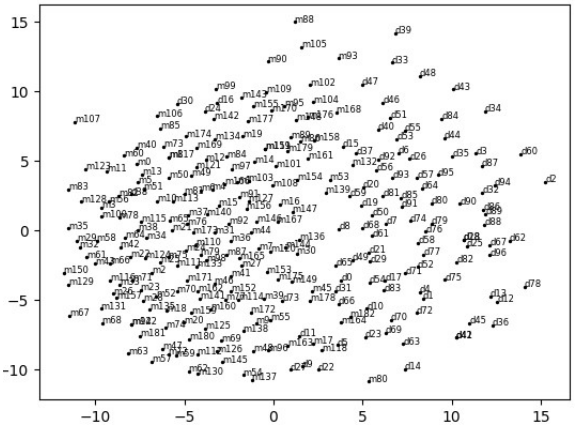

Figure 3: The Second-order Collocate Vectors: *motor* (m) and *dvygun* (d)

lapping areas but also quite distinct ones, which proves the above observation that these two synonyms have separate areas of usage: *motor* for common everyday-life vehicles and devices, whereas *dvygun* is for large industrial machinery.

Figure 4 shows how the items of another synset

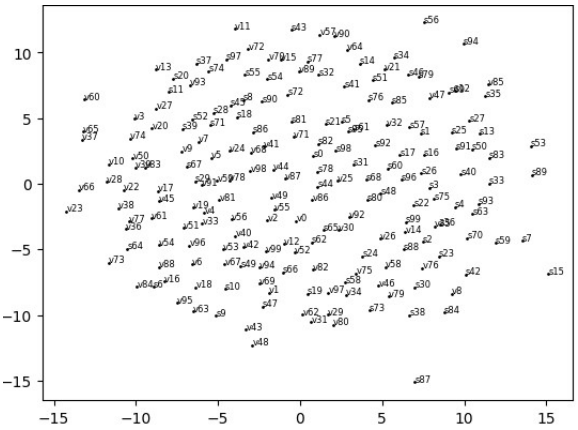

Figure 4: The Second-order Collocate Vectors: *volia* (v) and *svoboda* (s)

volia ('freedom', 'a will') and *svoboda* ('freedom')

are distributed. This time, we calculated the second-order collocate vectors for 10+10 concordance lines.

The top PMI collocates for *svoboda* are *svoboda* ('freedom') 3.58, *borec'* ('a fighter') 3.43, *demokratia* ('democracy') 3.28, *carstvo* ('a kingdom') 2.52, *myr* ('peace') 2.46, *demokratyčnyi* ('democratic') 2.45, *poniatt'a* ('a concept') 2.4, *ideal* ('an ideal') 2.37, *borot'ba* ('a fight') 2.34.

And for *volia*, the top collocates are *pamjataty* ('to remember') 4.16, *odynyca* 2.42 ('a unit'), *volia* ('freedom') 2.37, *nevolia* ('captivity') 2.31, *nacia* ('a nation') 2.26, *zusyll'a* ('an effort') 2.24, *vlada* ('power', 'authorities') 1.99, *borec'* ('a fighter') 1.94, *rozum* ('intelligence') 1.94, *bažann'a* ('a desire') 1.88.

Since both words are relatively frequent ((*volia*5k, *svoboda* 1.6k occurrences in the corpus), we plotted 100 random concordance lines per synonym.

On the plot, the vectors for both words overlap significantly, which indicates that *volia* tends to denote the concept of freedom more often than the idea of will in the texts published in Kyiv in 1940-1969.

Figure 5 provides an example with two semantically unrelated words (*svoboda* and *dvygun*) – to demonstrate how the model distinguishes them. The plot contains two completely separate clusters.

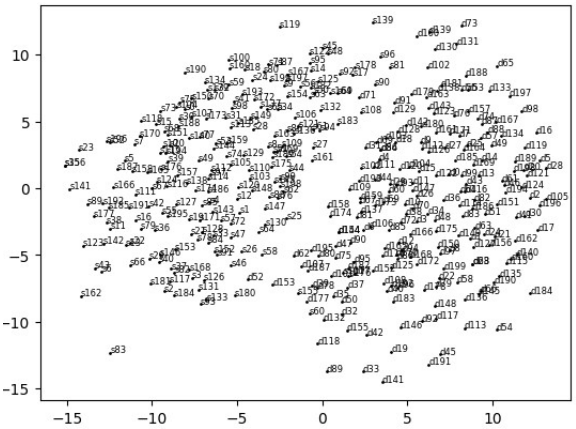

Figure 5: The Second-order Collocate Vectors: *dvygun* (d) and *svoboda* (s)

The provided observations for the test cases make us believe in the potential of the Semantic Vector Space Model with second-order collocate vectors for various semantic explorations of Ukrainian, including but not limited to word-sense disambiguation problems, regional variation investigation, and diachronic semantics.

Combined with profound lexicological analysis, such formal semantic representation applied to large-scale corpora would make it possible to reveal hidden trends and model dynamics of language over time and across different regions.

6 Further Work

The initial set of experiments with the proposed semantic model has opened several directions for the subsequent research and development to enhance and extend this approach. The ambition is to reveal Ukrainian word variation over regions, time, and registers following the prior work of von Waldenfels (2014).

Therefore, we have to deal with the problem of a semantic model being generic enough to represent the Ukrainian language as a whole and simultaneously being sensitive to regional and time-wise peculiarities. The open question that requires further exploration is building time-and-region-specific models vs. a single semantic model.

In addition, certain steps of the current data processing pipeline and analytic modeling require enrichment. We aim to continue experimenting with pipeline configuration decisions, vectorizer algorithms, and dimensionality reduction algorithms, utilizing clustering techniques, and various visualization approaches, including building semantic graphs.

Moreover, to properly represent a wide range of word senses, the model must be trained on significantly larger corpora (ideally, billions of tokens) and consider a vocabulary of greater size (at least 20,000 words).

Another challenge is to make the semantic model able to deal with high-frequency words, like prepositions, since their semantic variation is of high research interest for many lexicographers. Traditionally, such words are excluded from a vector space model as stop words, but we would like to treat them as another valuable target of semantic modeling.

Last but not least, there is a need to tackle several language standards of Ukrainian in specific periods of its history, which requires both additional data normalization and model sensitivity to different standards.

7 Limitations

During this initial phase of the research, we needed more digital textual data, especially for the period before WWII, and a poor representation of various

regions of Ukraine. Therefore, we had to limit our exploration to the texts published in 1940–1969 for the most represented region in the General Annotated Corpus of Ukrainian.

Apart from that, we had to simplify some of the data processing steps to avoid using extensive GPU resources, which, however, is unavoidable in the further stages of the project.

8 Ethics Statement

The broader value of the research is grounded on exploring and showing the versatility and growth of the Ukrainian language with the help of advanced NLP techniques combined with solid linguistic analysis.

References

Kris Heylen, Dirk Speelman, and Dirk Geeraerts. 2012. Looking at word meaning. an interactive visualization of semantic vector spaces for dutch synsets. *Proceedings of the EACL-2012 joint workshop of LINGVIS UNCLH: Visualization of Language Patterns and Uncovering Language History from Multilingual Resources*, pages 16–24.

Kris Heylen, Thomas Wielfaert, Dirk Speelman, and Dirk Geeraerts. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, 157:153–172.

Martin Hilpert and David Correia Saavedra. 2020. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory 2020*, 16(2):393–424.

Mariana Montes and Dirk Geeraerts. 2022. How vector space models disambiguate adjectives: A perilous but valid enterprise. *Yearbook of the German Cognitive Linguistics Association*, 10(1):7–32.

Maria Shvedova. 2020. The general regionally annotated corpus of ukrainian (grac, uacorpus.org): Architecture and functionality. *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020)*, I:489–506.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Ruprecht von Waldenfels. 2014. Explorations into variation across slavic: Taking a bottom-up approach. *Aggregating Dialectology, Typology, and Register Analysis*, pages 290–323.

Eric S. Wheeler. 2005. Multidimensional scaling for linguistics. *Quantitative linguistics. An international handbook*, pages 548–553.