Nataliia Cheilytko and Ruprecht von Waldenfels

WORD EMBEDDINGS FOR DETECTING LEXICAL SEMANTIC CHANGE IN UKRAINIAN

Abstract The paper presents a count-based semantic vector space model for Ukrainian, which has been applied for the semantic change detection task. The approach assumes creation of multidimensional vector representations of occurrences for a particular lexeme or a group of related lexemes with further visual and quantitative analysis of the obtained semantic vector space. The multidimensional space has been reduced to 2D for visual data analysis with the Multidimensional Scaling technique. The paper described two case studies to show how the proposed R & D workflow helps revealing potential semantic change events and discuss benefits and limitations of the approach. One case study traces the disappearance of a regional sense, and another identifies the appearance of a new metaphoric sense that is widespread in the Ukrainian media discourse.

Keywords Ukrainian; corpus linguistics; GRAC; word embeddings; vector representation; semantic vector space model; semantic change; semasiological variation; multidimensional scaling; pointwise mutual information; semantic distance

1. Introduction

1.1 Goals and Motivation Semantic Change and Variation in Ukrainian

Ukrainian is a large European language with a high degree of variability and versatile semantic developments. This makes Ukrainian a perfect experimental field for exploring semantic trends in a systematic way, which is of high interest to both diachronic semantics and variationist sociolinguistics. Therefore, the paper focuses on a corpora-based empirical examination of semasiological changes in the Ukrainian lexicon within the XX–XXI centuries.

The project's research questions are: How and with what degree of certainty can we identify semantic change in Ukrainian lexemes given significantly large input textual data? What are the types of such changes? Are there any trends and hidden patterns in how semantic changes are distributed across different regions of the country within various registers and domains?

The current paper looks into two specific extralinguistic circumstances in Ukraine's history that triggered significant instances of semantic change and lexical variation to assess the proposed methodology. The first case is the fusion of two separate Ukrainian standards after WWII due to the unification of the Ukrainian territories within the Soviet Union. Extralinguistic cataclysmic events often significantly influence language development, as language users revisit conceptual representations

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

under such circumstances. The second case is therefore the Russian full-scale invasion of Ukraine that started on February 24th, 2022.

1.2 Corpus Data

Data-wise, the research uses the concordances from the following corpora: the General Regionally Annotated Corpus of Ukrainian (GRAC), which represents Ukrainian starting XIX century and up to date in various domains and regions, including diaspora (Shvedova et al., 2017–2024), and the Ukrainian Trends corpus available within the Sketch Engine platform for the most recent data in the Ukrainian digital media. For the first case study, we used textual data from the three periods: 1920–1939 (approx. 10M tokens), 1940–1969 (approx. 30M tokens), 1970–1990 (approx. 50M tokens). The data included from fiction and periodicals and represented texts published in both the Western and the Eastern region of Ukraine. For the second case study, we considered two sub-corpora with texts from 1960–2019 and 2020–2024 from fiction, periodicals, and media texts. All texts have been processed with the help of the Ukrainian Part-of-Speech tagger and lemmatizer developed by Andrii Rysin[1].

1.3 Semantic Vector Space Representation

Our research is methodologically grounded on the distributional hypothesis that posits that one can distinguish the meanings of a word by their typical occurrences and collocates (Schütze, 1998). For example, in the following two sentences: "*I have a great fan of rock among my friends*" (fan as a person) and "*I have a great fan with several heating options*" (fan as a cooling, heating device), we can see the difference in the senses relying on the contextual surrounding (*rock, friends* vs. *heating options*). In order to identify and explore semantic variation and change from a usage-based perspective and in a bottom-up fashion by performing both quantitative and qualitative comparison of word occurrences, we construct and analyze semantic vector space representations (aka semantic word embeddings) of tokens following the research pipeline proposed by (Hilpert & Correia Saavedra, 2020), (Montes & Geeraerts, 2022) and (Geeraerts et al., 2023). Vector space models treat text as a multidimensional vector space, where a word, a token as a particular utterance of a word, or 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 vectors, apply dimensionality reduction to a vector space, visualize and cluster them.

The calculation of vector representations is often grounded on co-occurrence patterns of a particular linguistic entity (a word, a token, etc.). It considers how frequently other linguistic entities are being used in a particular corpus within a particular surrounding of the target word (Turney & Pantel, 2010, p. 149).

In our research, we build vector representations for tokens and account not only for the immediate surrounding collocates but also for second-order collocates – typical collocates of words found in concordance lines built for a given token of interest. In

[1] https://github.com/brown-uk/nlp_uk

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

such a way, a word occurrence vector includes information not only about collocates observed in its surroundings, but also collocates of the collocates. In other words, if to follow a social network metaphor, one can be distinguished not only by their friends but also by friends of a friend (Hilpert & Correia Saavedra, 2020, p. 394).

1.4 The Semantic Vector Space Pipeline

The text processing and analysis pipeline contains the following steps:

1. **First-order vectors**. For a given corpus, calculate a co-occurrence matrix with first-order type-based vectors for the vocabulary of the 10,000 most frequent words in the corpus, with the exception of the most frequent grammatical words (prepositions, conjunctions, and the likes). Normalize the matrix with the Pointwise Mutual Information (PMI) index, keep only those columns and rows in the matrix that contain at least one PMI no less than 3.

2. **Second-order vectors**. Build a second-order token-level vector representation for each utterance of a lexeme:

3. Get first-order vectors from step 1 for each collocate found in the surrounding of the lexeme under examination;

4. Average those vectors so that each lexeme utterance gets its averaged vector that accounts for second-order collocates found in the corresponding utterance of a particular size.

5. Apply **multidimensional scaling (MDS)** technique to the second-order embeddings representing utterances of a particular lexeme under examination and **visualize semantic vector space** for the occurrences as 2D scatterplots as proposed in (Wheeler, 2005), where each dot stands for a second-order vector of a single utterance. Perform visual analysis of the plots to see trends in lexeme usage – across time, regions, and registers.

6. Apply MDS for the lexeme in question with specific **anchor word(s)** to see how lexeme's second-order embeddings are located in a 2D semantic vector space compared to the embeddings built for the anchor word(s). For example, in section 3 of this paper, for the Ukrainian noun *bavovna* 'cotton', we use the anchor word *l'on* 'lien, flax'.

7. Calculate the **average Euclidean distance** within a group of token embeddings of a lexeme within a certain period, as well as average distances to a similar group of token embeddings but of a different period and/or register so that a significant change of the distance may indicate a change in how the target lexeme is being used.

This pipeline was sufficiently successful in identifying evidence for semantic changes here. However, the approach faces complications and uncertainties in certain cases, which are discussed below.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

2. Semantic Change in the Context of Merging Standard Ukrainian

2.1 Ukrainian as a Multi-Standard Language Before WWII

We consider Ukrainian to be a multi-standard language with significant variation and fast language development due to intense and active social and political changes. As shown in (Shvedova, 2021) and (Lahjouji-Seppälä et al., 2022), in the time before WWII Ukrainian had two standard language varieties, as the Ukrainian-speaking territory was historically separated between two neighbouring countries – between Habsburg, then Poland in the West and the Russian Empire and the Soviet Union in the East. The two standards were merged after WWII when the Ukrainian territories were united within the Soviet Union. As a result, a similar language policy was applied to the whole country. Due to this and extensive ethnic cleaning, the influence of the Polish language contact became much less significant.

In order to perform a systematic bottom-up exploration of how those processes influenced semantic variation and change in Ukrainian, we applied the R & D pipeline briefly described in the previous section.

2.2 Disappearance of a Word Meaning

This case study presents how the adjective *povazhnyi*, meaning 'respectable, honorable' and also 'serious, severe', has almost lost the latter sense during the second part of the XX century due to the fusion of the standard Ukrainian varieties. Figure 1 contains the 2D visualization of the semantic vector space obtained with the multidimensional scaling for regionally annotated occurrences of this adjective (K – Kyiv representing the Eastern part of the country, L – Lviv representing the Western part). The plot in Figure 1 shows the complete overlap between the two regional groups of embeddings built for adjective occurrences in 1970–1990. Further qualitative analysis of the occurrences confirmed that the specific Western sense 'serious, severe' is almost absent for both regions, and thus, such overlap indicates that this sense is disappearing.

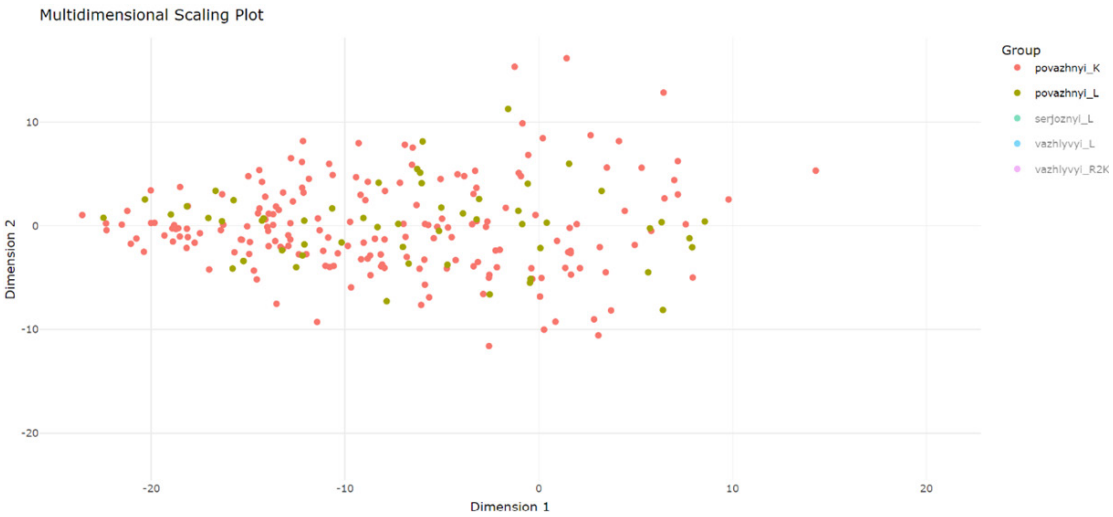

Fig. 1: Embeddings of the *povazhnyi* occurrences in 1970–1990 (pink – Eastern Ukraine, brown – Western)

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Figure 2, on the contrary, demonstrates that before WWII (1940–1969), the Western occurrences have their own area on the plot (the central lower part of the plot) where the Eastern occurrences are not present. Indeed, manual inspection of the occurrences in question shows that this regionally specific area most often involves cases with the 'serious, severe' sense, which is much less usual for the other area on the plot. In this way, the visual representation of embeddings indicates regional differences in usage.

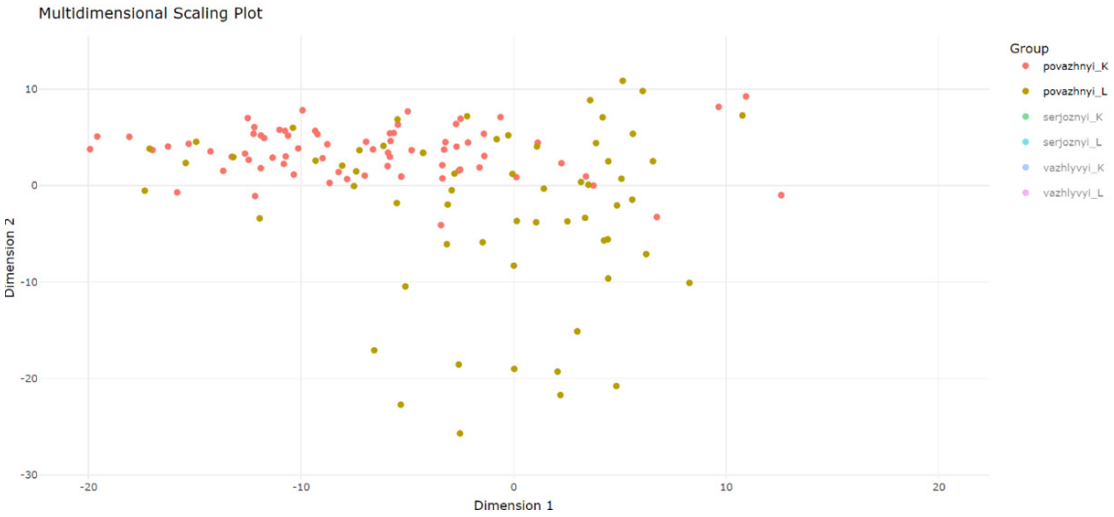

Fig. 2: Embeddings of the *povazhnyi* occurrences in 1940–1969 (pink – Eastern Ukraine, brown – Western)

The same observation is valid also for an earlier period of 1920–1939 (Figure 3): the area on the plot where the data from the two regions do not overlap reflects cases where the adjective is used in the sense 'serious, severe' – most of which are from the Western region.

We conclude that the adjective *povazhnyi* had a regional sense 'serious, severe' in West in the period when Ukrainian was a multi-standard language. We explain this with close and strong linguistic contact with Polish in West, where the cognate the adjective *powazhny* primarily denotes 'serious, severe'. After the war, this sense almost disappeared from use due to the general trend of Ukrainian to become closer to its Eastern standard in this period.

It is important to mention that the proposed case study is just a single example from a systematic lectometric assessment we are performing for standard Ukrainian targeted to show how restructuring of a multi-standard Language results in semantic structure changes of a word or a concept.

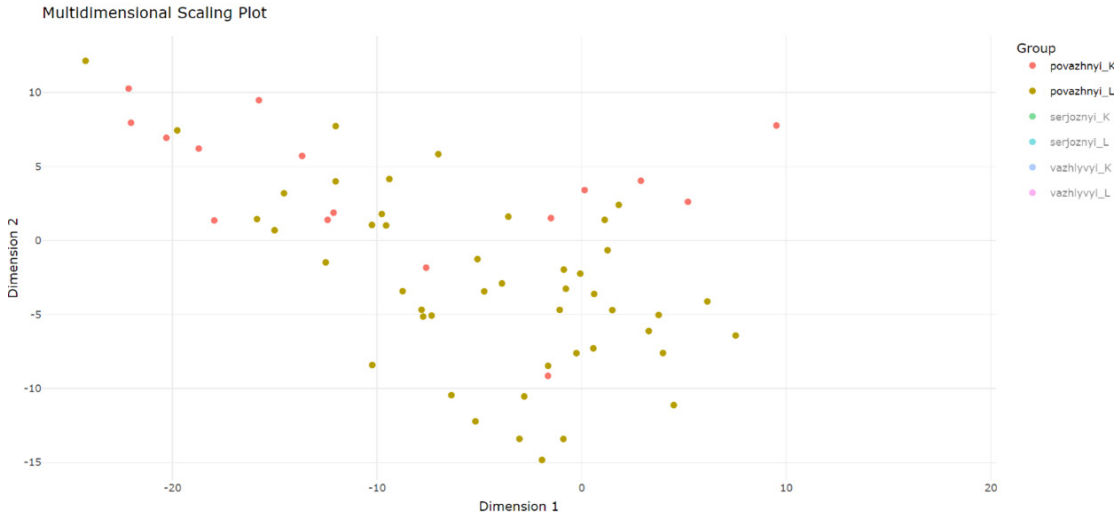

Fig. 3: Embeddings of the *povazhnyi* occurrences in 1920–1939 (pink – Eastern Ukraine, brown – Western)

3. Semantic Change in Cataclysmic Events

3.1 Appearance of a New Metaphoric Sense

With the MDS-based visualisation of second-order collocation vectors of a particular lexeme done for different periods, it becomes possible to identify a significant change in how words are used, and reveal new meanings. Figure 4 demonstrates the vector space for the Ukrainian noun *bavovna* 'cotton flower, cotton material' for two periods: 1960–2019 and 2020–2024 – before and during the ongoing war in Ukraine. The plot shows that most of the occurrences of the same word do not overlap but are located in different areas in the vector space, which is a strong indicator of semantic change. Indeed, this noun has acquired a completely new metaphoric sense 'explosion', since the word is being used as a euphemism referring to the results of a fire attack.

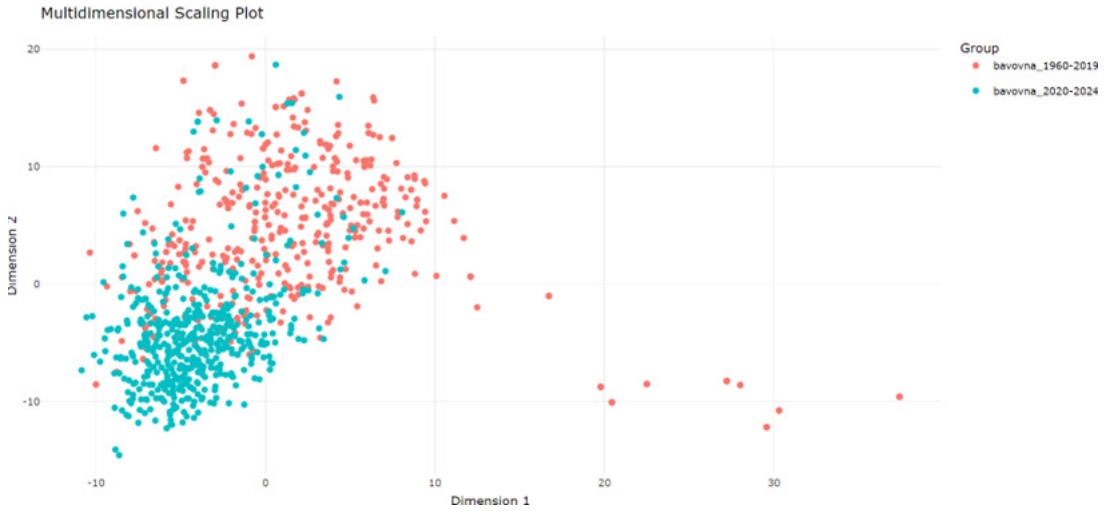

Fig. 4: Embeddings of the *bavovna* occurrences (pink – 1960–2019, blue – 2020–2024)

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

This usage originated from a wordplay with Russian *hlopok*, a homograph for two words meaning 'cotton' (same as Ukrainian *bavovna*) and 'a clap', implying an undisturbing and quiet sound, which Russian media often use to disguise the facts of fire attacks on Russian military and infrastructure objects, so that to avoid panic among their citizens. The Ukrainian media confronting and mocking this fact-hiding strategy of the opponents uses the Ukrainian equivalent of 'cotton', i.e., *bavovna,* to denote events of explosions caused by fire attacks. As a result, the noun has acquired a new metaphoric sense 'explosion' in Ukrainian media texts. In other registers, however, the noun continues to refer to the classical meanings 'a cotton flower' and 'cotton material'.

3.2 Anchor Words to Explore Semantic Change

A further technique to identify and examine evidence of semantic change is to build a vector space plot for a target word with certain anchor words (near-synonyms, antonyms) to see how the target word is being used in comparison to these other words over time. Figure 5 explicitly shows that before the Russian invasion (1960–2019), the near-synonyms *bavovna* 'cotton flower', 'cotton material' and *l'on* with the meanings 'flax flower', 'linen material', and 'flax seeds as a food product' have significantly larger overlapping area, so they were used as near-synonyms. With the invasion, however, the dots on Figure 6 representing the *bavovna* embeddings become more concentrated on the left part of the semantic vector space and almost do not overlap with the *l'on* embeddings. This differentiation from a word with a similar meaning is another indication of the semantic change.

Fig. 5: Embeddings of the *bavovna* (pink) and *l'on* (blue) occurrences in 1960–2019

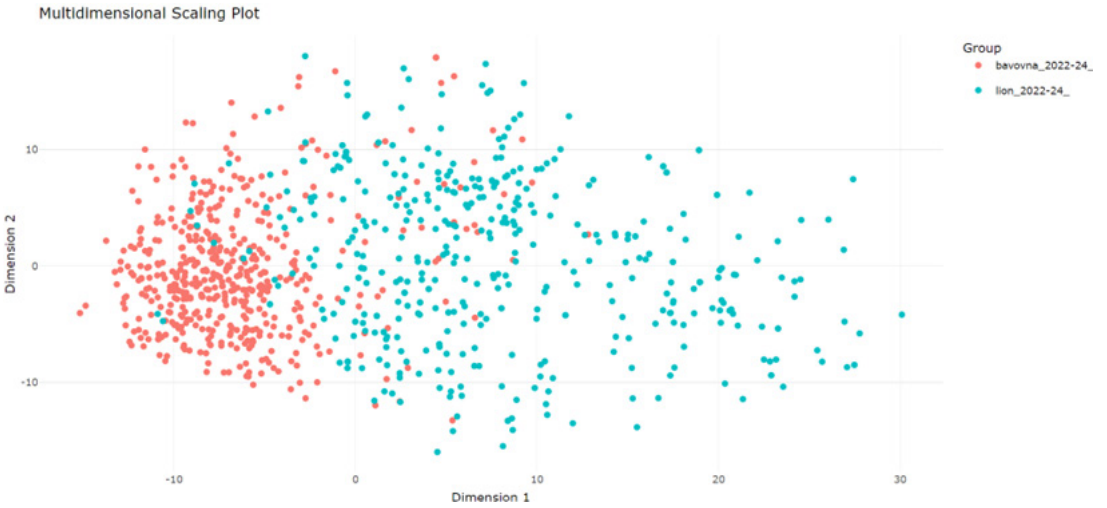

Fig. 6: Embeddings of the *bavovna* (pink) and *l'on* (blue) occurrences in 2022–2024

Similarly, Figures 7–8 demonstrate how *bavovna* relates to the noun *vybuh* 'explosion'. Before the Russian invasion (Figure 7), the embeddings of those two words take mostly distinct areas on the semantic vector space. In contrast, after the invasion, their embeddings significantly overlap (Figure 8), proving that these words start sharing their usage patterns and thus meaning – 'explosion'.

Therefore, the distributional methodology grounded on the second-order vector representations helped to explore how particular lexemes function over time and across different regions and registers and, thus, to reveal potential indications of conceptual reorganization in the Ukrainian lexicon.

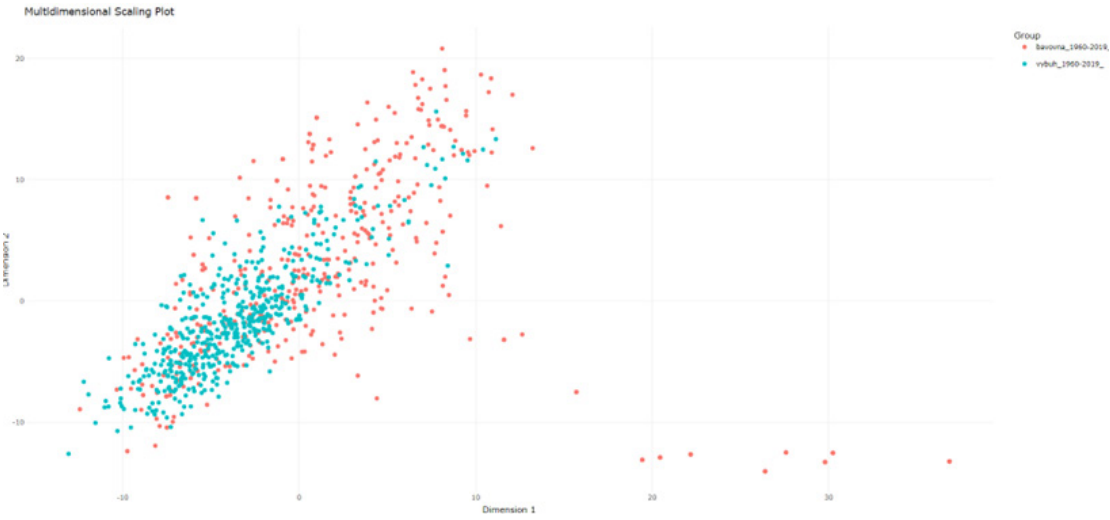

Fig. 7: Embeddings of the *bavovna* (pink) and *vybuh* (blue) occurrences in 1960–2019

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Fig. 8: Embeddings of the *bavovna* (pink) and *vybuh* (blue) occurrences in 2020–2024

4. Moving on: LLM Representations vs. Count-Based Vectors

The current approach is based on the semantic vector space model obtained by counting frequencies of collocates in a contextual window for a target word. Such models are considered to be good from the explanatory perspective. However, they have several limitations, among which is the inability of a single generic count-based vector space model to represent the entire lexicon as used in a multitude of registers, and therefore, a plurality of models would be required for a systematic exploration of language change. Moreover, such models face certain complications with the words-sense disambiguation task, which is an essential step of any semasiological modelling, as collocations are indicative on many other factors besides word senses.

To overcome the limitations in the next phase of our research, we are going to expand the solution by adding vector representations (embeddings) obtained from different transformer models fine-tuned for Ukrainian (Laba et al., 2023). Our preliminary test study showed that ChatGPT-4o is able to distinguish various senses for *bavovna*, including the new metaphoric sense, as well as for a group of other Ukrainian polysemous words.

5. Conclusions

The second-order vector representations, together with the MDS-based visualisation technique, make it possible to identify events of semantic change in Ukrainian across different periods, registers and geographic regions. The paper presented two case studies to demonstrate the research workflow for exploring such changes. The approach is being used for a large-scale bottom-up systematic exploration of lexical variation and change in Ukrainian.

In order to further tune and assess the methodology, the authors are preparing a representative test dataset with known cases of Ukrainian variation and semantic

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

change. However, the preliminary goal is to monitor and identify unknown cases and, therefore, to reveal hidden patterns and trends in Ukrainian language development.

A systematic study of linguistic variation and semantic change in Ukrainian will contribute to the fundamental understanding of universal mechanisms of language development by comparing the findings to similar studies done for other multi-standard languages like German, Dutch, English, French, Spanish, and Portuguese.

Apart from that, such a consistent dynamic large-scale view on semantic evolution may bring additional insights with respect to neighbour language contact influence, in the case of Ukrainian, with Polish and Russian.

Last but not least, the proposed R & D pipeline is not language-specific and can be applied to various languages.

References

Geeraerts, D., Speelman, D., Heylen, R., Montes, M., de Pascale, S., Franco, K., & Lang, M. (2023). *Lexical Variation and Change. A Distributional Semantic Approach.* Oxford Academic.

Hilpert, M., & Correia Saavedra, D. (2020). Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory*, *16*(2), 393–424. https://doi.org/10.1515/cllt-2017-0009

Laba, Y., Mudryi, V., Chaplynskyi, D., Romanyshyn, M., & Dobosevych, O. (2023). Contextual Embeddings for Ukrainian: A Large Language Model Approach to Word Sense Disambiguation. In M. Romanyshyn (Eds.), *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)* (pp. 11–19). Association for Computational Linguistics.

Lahjouji-Seppälä, M. Z., Rabus, A., & von Waldenfels, R. (2022). Ukrainian standard variants in the 20th century: stylometry to the rescue. *Russian Linguistics*, *46*(3), 217–232. https://doi.org/10.1007/s11185-022-09262-9

Montes, M., & Geeraerts, D. (2022). How vector space models disambiguate adjectives: A perilous but valid enterprise. *Yearbook of the German Cognitive Linguistics Association*, *10*(1), 7–32. https://doi.org/10.1515/gcla-2022-0002

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24*(1), 97–124.

Shvedova, M., von Waldenfels, R., Yarygin, S., Rysin, A., Starko, V., & Nikolajenko, T. (2017–2024): *GRAC: General Regionally Annotated Corpus of Ukrainian*. Retrieved July 31, 2024, from https://uacorpus.org/

Shvedova, M. (2021). Lexical Variation in the Language of the Ukrainian Press of the 1920s-1940s and the Development of a New Lexical Norm. *Movoznavstvo*, *1*, 16–35. https://doi.org/10.33190/0027-2833-316-2021-1-002

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Wheeler, S. (2005). Multidimensional scaling for linguistics. In R. Köhler, G. Altmann, & R. G. Piotrowski (Eds.), *Quantitative linguistics. An international handbook* (pp. 548–553). Walter de Gruyter.

Contact information

Nataliia Cheilytko
Friedrich Schiller University Jena
natalia.cheilytko@gmail.com

Ruprecht von Waldenfels
Friedrich Schiller University Jena
ruprecht.waldenfels@uni-jena.de