Patterns in Russian political discourse through the prefix *pro-* using contextual word embeddings

Thomas Samuelsson
(Stockholm University)

This study investigates patterns in the use of the prefix *pro-* 'pro-' in Russian online media using contextual language models. In the project, a corpus of Russian online media texts has been built from more than 60 Russian influential media resources. The texts belong to political topics and are published from 2012 until 2020. Automatic text preprocessing performed on the data include tokenization, lemmatization and morphological analysis. The annotation contains linguistic information such as lemma and part of speech as well as publication date. The corpus consists of more than 500 million tokens.

In the Russian language, the use of nominal prefixes has seen an accelerating trend in the last few centuries. The most prominent of these prefixes are of Greek and Latin origin, such as *anti-* 'anti-', *de-* 'de-', *post-* 'post-' etc. While many of these word formation elements have gained attention by linguists, the prefix *pro-* 'pro-' has mostly gone under the radar. The prefix was productive in Soviet newspaper discourse. Since the dissolution of the Communist monopoly of political power, it spread in the area of politics (Ryazanova-Clarke and Wade 1999). For example, one can find post-Soviet prefixed words like *prorossijskij* 'pro-Russian', *proputinskij* 'pro-Putin' and *protrampovskij* 'pro-Trump' that reflect the current zeitgeist.

The corpus data is analyzed using contextual word embeddings obtained by utilizing a Russian Bert model (Devlin et al. 2019). In contextualized word embeddings, each word is represented by a semantic vector that depends on the context of the word. Since the representations of the words depend on their contexts, the vectors can be used to study contextual patterns of the prefixed words. The dimension reduction technique UMAP (McInnes et al. 2018) is used to visualize the context embeddings. The UMAP projection is explored to understand the dataset and to reveal relevant patterns. The visualization displays clear clusters related to the meaning of the words but also spatially separate clusters for some ambiguous prefixed words of the same type. On the larger scale, the different clusters tend to structure according to contextual similarity in the reporting. The study demonstrates the opportunities on how language models can contribute to the research on prefixal morphology in political language.

Discourse studies, language model, prefix, Russian, visualization

References

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina (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, doi: 10.48550/arXiv.1810.04805, url: https://arxiv.org/abs/1810.04805.

McInnes, Leland, Healy, John and Melville, James (2018), UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, doi: 10.48550/arXiv.1802.03426, url: https://arXiv:1802.03426.

Ryazanova-Clarke, Larissa and Wade, Terence (1999), *The Russian language today*. London: Routledge.