Abstract for SlaviCorp 2020

Designing and Building a Corpus of Russian On-line Media

The talk presents the development of a corpus that represents Russian political journalism in online media during 2012–2020, a time period marked by major changes. First, the content of the corpus and the tools used to build the corpus will be outlined. Secondly, the topic is how the corpus can serve as research material for studies on Russian political language, with the focus on diachrony in the Russian media discourse.

There are other Russian newspapers corpora such as the newspaper subcorpus of the Russian National Corpus (RNC) and the Moscow State University corpus of Russian newspapers of the end of the 20th century (Vinogradova et al. 2001), but none of them cover the investigated time frame.

The corpus is composed of on-line written texts, first published between the years 2012 and 2020, from the most influential news outlets on the Russian internet. The publications are selected according to a citation index, compiled by the leading Russian media monitoring company Medialogiya (2020). The index measures the most cited internet resources, including general news outlets, tabloid-styled magazines, niche media, as well as a variety of local, regional and national outlets. The substantial share of the texts belongs to news articles, but text types as opinion columns and featured stories are also sampled. The corpus is automatically tokenized, morphological annotated and lemmatized, using a UDPipe pretrained model (Straka and Straková 2017). It is also annotated with metadata as URL, publication date and title.

The corpus build is within the methodology of Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS), a subdiscipline of Corpus-Assisted Discourse Studies (CADS) (Partington et al. 2013). The aim of the corpus is to study changes of word-formation patterns in political discourse in Russian on-line media.

References

Medialogiya. 2020. Available at: https://www.mlg.ru/ratings/media/federal/ (accessed April 13, 2020).

Partington A. S., Duguid A. and Taylor C. 2013. *Patterns and Meanings in Discourse. Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Russian National Corpus (RNC). 2020. Available at: http://www.ruscorpora.ru/ (accessed April 17, 2020).

Straka, M. and Straková, J. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Vinogradova, V. B., Kukushkina, O. V., Polikarpov, A. A. and Savchuk, S. O. 2001. Kompjuternyj korpus tekstov russkich gazet konca 20-go veka. [The Computer Corpus of Russian Newspapers of the End of the 20th Century]. Available at: http://www.philol.msu.ru/~lex/corpus/ (accessed April 17, 2020).