A corpus-based text-analytic tool for novice writers of Academic Russian

Mikhail Kopotev[1], Olesya Kisselev[2], Mariia Fedorova[3], Alexandr Klimov[3],
Anna Dmitrieva[3], Anastasiia Baranchikova[3]
[1] University of Helsinki, [2] University of Texas at San Antonio, [3]
Higher School of Economics in Moscow
mihail.kopotev@helsinki.fi, olesya.kisselev@utsa.edu

The study of English academic discourse has benefited greatly from the application of corpus-based tools and analysis devoted to it in the past few decades (Ackerman & Chen, 2013; Biber et al. 2004; Durrant & Mathews-Aydinli, 2011; Gray & Biber, 2013). Similar studies of academic registers of languages other than English have been lagging behind. The project, titled CAT&kittens, described in this paper intends to address this gap, as well as to contribute to a general exploration of (semi)automated tools available to the learning of academic genres. The service is intended to help a user with two tasks: to highlight fragments, which differs from a reference corpus, and to offer, when possible, a substitute that better serves in a given context.

The central part of the project involves the development of the comprehensive representative Russian Corpus of Academic Texts (CAT). Following well-established corpus development procedures (e.g., BAWE). Texts in the CAT corpus are sourced from six general disciplinary fields: social studies, political science and international relations, law, linguistics, economics, psychology and education science. The discipline sub-corpora consist of about 370 to 480 thousands tokens, amounting to approximately 2 mln. tokens in the corpus in general. Texts entered in CAT are supplied with metalinguistic, morphological and syntactic annotations, carried out with the help of the Universal Dependencies pipeline (Straka et al. 2017).

CAT is outfitted with built-in data processing tools, which allows for evaluation of texts written by novice writers of Academic Russian, both native and non-native:

- *General statistics* of an analyzed novice text: a readability test, average length of words, sentences, and paragraphs, and TTR.
- *Lexical analysis* highlights terminology that are unattested in the discipline domain, and suggests alternatives.
- *Collocational analysis*. Based on n-gram frequencies, all non-attested word choice selections in the novice texts will be identified; attested collocational alternatives extracted from CAT will be provided.
- *Grammar check*. Unlike available spell-checkers, the tool is focused on detecting deviations that feature in academic writing, e.g. genitive chains, mixtures of synthetic and analytical comparatives etc.

Tools like the one described above are routinely evaluated in terms of recall and precision, when both measures are taken equally important. We believe, however, that for many CALL tools, precision is more instrumental. While a complete automatic correction of every error is an impossible task, focusing on a precise improvement based on a shortlist is likely to be realistic (if challenging) and pedagogically useful.

References

Ackerman, K., & Chen, Y-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12(4), 235–247.

BAWE (The British Academic Written English), available at:
http://www2.warwick.ac.uk/fac/soc/al/research/collections/bawe/how_to_cite_bawe. Last retrieved April, 15, 2019.

Biber, D., Conrad., & Cortes, V. (2004). If you look at …: Lexical bundles in University teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.

Durrant, P., & Mathews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30(1), 58–72.

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1), 109–136.

Straka, M., and Straková J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.