Authors:

Anastasiia Baranchikova[1], Anna Dmitrieva[1], Mariia Fedorova[1], Aleksandr Klimov[1], Svetlana Toldova[1], Natalia Zevakhina[1], Olesya Kisselev[2], Mikhail Kopotev[3]

[1] Higher School of Economics, Moscow, Russia
[2] Pennsylvania State University
[3] University of Helsinki, Finland

Title: CAT&kittens: a corpus-based text-analytic tool for Russian academic writing

Corpus linguistics has contributed significantly to the study of academic discourse in the past two decades, with studies ranging from descriptions of specific grammatical features (Swales, 1990; Hyland, 1994) to general investigations of linguistic patterns, syntactic or lexical (Biber et al. 2004; Durrant & Mathews-Aydinli, 2011; Gray & Biber, 2013), to the development of specific academic vocabulary lists and academic phrase lists (Simpson-Vlach & Ellis, 2010; Ackerman & Chen, 2013). Similar studies for the Russian academic genre, however, have been lacking. The project described in this proposal intends to fulfill this gap.

The paper describes the development of a representative Russian Corpus of Academic Texts (CAT) outfitted with a built-in data processing tool, which allows for evaluation of texts written by novice writers of Academic Russian, both native and non-native, along a set of criteria in relation to the CAT corpus. Consequently, the goal of this paper is twofold: a) to describe the Corpus, and b) to discuss the criteria, upon which a novice text can be evaluated against the Corpus.

The project is currently being developed by a team of researchers from the Higher School of Economics (HSE) in Moscow, the University of Helsinki, and the Pennsylvania State University. The development of the CAT corpus follows established corpus development procedures (e.g., BAWE). It was collected by extracting recently published texts sourced from textbooks, academic journals, and collecting high-quality master's theses from available sources. All texts entered in CAT are divided into six disciplinary fields: social studies and history, political science and international relations, law, general and applied linguistics, economics, psychology and education science. Every discipline sub-corpus consists of about 300 to 400 thousand tokens, amounting to appr. 2 million tokens in the corpus in general. CAT is supplied with metalinguistic information, as well as morphological and syntactic annotation, carried out with the help of the annotation software RU Syntax (Mediankin et al. 2016). Further corpus improvement is also planned.

Since the main goal of the project is to create a tool that compares novice texts to standard academic texts along the lists of pre-set criteria, the tool will run a series of "error analysis" test. The patterns of deviations are identified along lexical, collocational, morphological, and syntactic planes. Their full list is still under discussion, therefore, we present a preliminary set,

1. **The general observation** of an analyzed novice text includes text readability test, average sentence length, and TTR — all as compared to the CAT.
2. **Lexical analysis** includes identifying recurring tokens/lemmas in the student texts and comparing their frequencies to the frequency lists based on the CAT corpus. This analysis, based on low-frequency items and hapax legomena, identifies overuse/underuse of specific vocabulary, highlights terminology that are unattested in the discipline, and suggests alternatives.
3. **Collocational analysis**. Based on n-gram frequencies, a specific type of errors, namely, non-standard word choice selection, will also be identified, and more standard collocational alternatives will be provided. This part consists of two steps: first we extract domain-specific collocations using standard measures (LL, (p)MI, t-score, etc.). Second, we determine non-standard collocations in a student text and suggest an alternative, based on more regular collocations and on distributionally close alternatives calculated with reference to the word2vec model trained on the semantically similar data.
4. **Grammar check**. Having morphological and syntactic annotations both in the CAT and in a student text under examination, checking morphological and syntactic errors is a two-step task. Unlike available spell-checkers, our tool is focused on detecting deviations that feature in academic writing— specifically those written by non-native speakers, e.g. genitive chains and ProDrop.

The results of these multidimensional analyses are provided in two ways: the general information about the whole text and highlighted fragments supplied with recommendations for correction. Although the robustness of the proposed analysis and the implementation of the tool require extensive testing, our project and lessons learnt from its development have implications for methodology of corpus linguistics already at this stage. Being a well-developed, deeply annotated representative corpus of Russian academic texts for the fields of Humanities and Social Studies, the CAT provides language researchers studying academic genres with an indispensible data set. Furthermore, the tool will, upon completion, be a useful to Russian teachers and students, who are seeking to improve their writing skills in this specific register.

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 Feb, 15, 2018.

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.

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes*, 13, 239-56

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

Mediankin N., & Droganova K. (2016). Building NLP Pipeline for Russian with a Handful of Linguistic Knowledge. In: Proceedings of the Workshop on Computational Linguistics and Language Science, Copyright © CEUR-WS, Aachen, Germany, ISSN 1613-0073, pp. 48-56.

Simpson-Vlach, R., and Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, *31*(4), 487-512.

Swales, J.M. (1990). Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press