

# Development of a text-analytic tool based on Corpus of Russian academic texts

Anna Dmitrieva<sup>1</sup>, Mariia Fedorova<sup>1</sup>, Aleksandr Klimov<sup>1</sup>, Anastasiia Baranchikova<sup>2</sup>  
<sup>1</sup>*Higher School of Economics*  
*black-letter@yandex.ru, maria.fjodorowa@gmail.com, aleksklimow@gmail.com, six3.danika@gmail.com*

## Introduction

This research is a part of the CAT project, which main goal is to develop 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. The tool compares novice texts to standard academic texts along the lists of pre-set criteria and runs a series of “error analysis” tests. The patterns of deviations are identified along lexical, collocational, morphological, and syntactic planes.

## Corpus description

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, 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).

## Creation of the academic style checking tool

The task of the academic style checking can be reduced to the task of checking for grammar and stylistic mistakes. When a mistake is found, it is supposed to be marked as wrong by the tool to let the author know which parts of the text might need correction. The checking for grammar and stylistic mistakes was implemented as follows:

1. Long genitive chains are found in the texts, as they are considered to make the text overly complicated for the reader. The threshold for assuming that a genitive chain is "long" is set by comparison with the gold standard texts from the CAT.
2. Mixtures of synthetic and analytical comparatives (such occurrences as "более лучше"/"more better") are found.
3. Heads of coordinate groups are checked for the possibility to occur in the same context to avoid syllepsis (such phrases as "они покрывались пылью и славой"/"They covered themselves with dust and glory").

For this task, we used word2vec bigram model that was taught on 21000 texts (ca. 42 mln tokens) crawled from Cyberleninka, biggest Russian on-line resource containing millions of scientific papers.

4. All the words are checked for occurrences in the Cyberleninka vocabulary (e.g. obscene words never appear in scientific texts, excluding linguistic examples).
5. The usage of personal pronouns is checked. It was usual to use "we" as a pronoun referenced to the author in the Soviet academic writing tradition; nowadays the tendency is to use "I" and occurrences of using both pronouns in the same text are happened. Such occurrences are marked as mistakes.
6. Since the usage of imperatives and subjunctives in scientific texts is discouraged, verbs are checked for their mood.
7. The lengths of all the sentences are checked for being too long. The maximums of sentence lengths in the corresponding domains of the CAT are used as thresholds.

The style checking algorithm has been tested on the CAT for finding false positives and on some artificial examples for finding false negatives.

## CAT collocation lists and finding miscollocations

For collocational analysis, we have collected lists of word ngrams (from unigrams to sixgrams), including lists obtained from the whole corpus and lists obtained for each domain. In order to determine if a certain cooccurrence is an actual collocation of Academic Russian, we used a few measurements for determining if a probability of a certain word constellation is statistically important - t score, PMI and likelihood ratio - that have been proven to be effective on Russian data (Kopotev et al, 2017:155). CAT collocation lists can be used for different purposes such as data source for the task of finding miscollocations in student texts and suggesting possible replacements.

In this task, all NN, NV, VN and VV collocations of the input text are compared to the CAT collocation list, and if one of the words in the pair is not found as a possible collocate to the other, the word sequence is considered to be a possible miscollocation and a few replacements are suggested. The replacements are chosen from the list by a set of criteria described in (Liu et al., 2009) that includes PMI between the collocates, their semantic similarity (computed using Word2Vec model mentioned above) and the percentage of shared collocates in collocation cluster (Liu et al., 2009:48). For example, a stylistically inappropriate miscollocation “автор думает” gets the following suggestions: “автор выделяет”, “автор предлагает”, “автор выражает”, “автор придерживается”, “автор публикует”, “автор апеллирует”, “автор наталкивается”, “автор посвящает”, “автор сосредоточивает”, “автор стремился”.

## Conclusion

Our future plans include implementing checking for more complicated types of mistakes and evaluating the whole system on the Corpus of Russian Student Texts (Zevakhina, Dzhakupova 2015). Besides, we plan to evaluate the ngrams in CAT collocation lists with the help of human assessors and use this data later for developing a more advanced style checking algorithm allowing for capturing miscollocations and suggesting stylistically more appropriate options from the lists.

## References

1. Kopotev M., Kormacheva D., Pivovarova L. Evaluation of collocation extraction methods for the Russian language //Quantitative Approaches to the Russian Language. – Routledge, 2017. – C. 137-157.
2. Liu A. L. E., Wible D., Tsao N. L. Automated suggestions for miscollocations //Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications. – Association for Computational Linguistics, 2009. – C. 47-50.
3. Mediankin N., Droганova 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.
4. Zevakhina N., Dzhakupova S. Corpus of Russian student texts: design and prospects, in: Труды 21-й Международной конференции по компьютерной лингвистике "Диалог".