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

Anastasiia Baranchikova, Anna Dmitrieva, Mariia Fedorova, Aleksandr Klimov, Svetlana Toldova, Natalia Zevakhina, Olesya Kisselev, and Mikhail Kopotev

Higher School of Economics, University of Pennsylvania, University of Helsinki

six3.danika@gmail.com, black-letter@yandex.ru, maria.fjodorowa@gmail.com, aleksklimow@gmail.com, toldova@gmail.com, natalia.zevakhina@gmail.com, olesyakisselev@gmail.com, mihail.kopotev@helsinki.fi

26 September 2018

Overview

Introduction

Corpus linguistics has contributed significantly to the study of academic discourse in the past decades, with studies ranging from descriptions of specific grammatical features (Swales, 1990; Hyland, 1994) to general investigations of syntactic patterns or lexical classes (Biber et al. 2004; Durrant & Mathews-Aydinli, 2011; Gray & Biber, 2013), and to the development of specific academic vocabulary or phrase lists (Simpson-Vlach & Ellis, 2010; Ackerman & Chen, 2013). However, corpus-based studies of Russian academic texts have been lacking mainly due to lack of corpus resources. The CAT&kittens project intends to fulfill this gap.

CAT&kittens: Goals

- To create a corpus of academic texts (=CAT)
- To develop a set of tools for evaluating student papers (='kittens') against CAT

Corpus description

CAT corpus

- is collected by extracting recently published texts sourced from academic journals
- is divided into six thematic subcorpora: social studies, political science and international relations, law, **general and applied linguistics**, economics, psychology, and education science.
- each sub-corpus consists of about 300 to 400 thousand tokens, which gives appr. 2 million tokens in sum. (And more to come!)
- is annotated both morphologically and syntactically based on the RU-Syntax tagger (Mediankin et al. 2016).
- is supplied with meta-linguistic information

Cyberleninka corpus: 21000 scientific articles (ca. 42 mln tokens) crawled from Cyberleninka

CAT&kittens: tools

CAT&kittens is not only a corpus, but a set of tools to evaluate a student text against standard academic texts. The features that we aim to evaluate include:

- Grammatical, e.g. genitive chains,
- Lexical, e.g. non-standard lexemes
- Collocational: collocations non-attested in CAT
- Readability: relative text complexity

NB! Recommendations given by the system are not binding obligations! They are scaled from bright-line rules ("Please, change this according to the style") to soft recommendations ("This might be slightly inaccurate. Consider that instead")

Tools: grammar

Genitive chains

Genitive chains (>2 nested Genitive phrases) make the text overly complicated. A genitive chain is considered too long, if it exceeds an average attested in CAT.

Example: “какого-либо лингвистического механизма контактного влияния одного языка”/"of some linguistic mechanism of contact influence of one language"

Comparatives

Non-standard co-occurrences of synthetic and analytical comparatives, e.g. *"более лучше"* ("more better").

Tools: grammar

Syllepsis

Heads of coordinate groups are checked for coherence to avoid syllepsis (e.g. "Он взял шляпу и отпуск"/"He took his hat and his leave"). The word2Vec bigram model that was taught on 21000 texts (ca. 42 mln tokens) crawled from Cyberleninka was used.

Mood

Since the usage of imperatives and subjunctives in scientific texts is unrecommended, mood of verbs are checked and non-indicative forms are highlighted for further consideration.

Tools: vocabulary

Personal pronouns

In Soviet academic writings, using of so-called "academic *we*"was obligatory. Nowadays the tendency goes towards using "I". Occurring of both pronouns in the same text happens quite often. Such occurrences are marked by the system as inaccuracy. **Example**: "...насколько **мне** известно..."/"...as far as **I** know..."and "предлагаемыми **нами** в данной работе"/"that **we** suggest in this paper"(both phrases are from the same article)

Vocabulary

Many words, e.g. obscene or colloquial ones, never appear in scientific texts, excluding linguistic examples. To check this, all words in a student text are looked up in the vocabulary, compiled on the basis of Cyberleninka corpus. If a word is not found, the system suggests the closest synonyms based on vector similarity (word2vec model).

Tools: Collocations

For collocation analysis, we have collected lists of n-grams (from uni- to six-grams): obtained from the whole corpus and from each domain. Standard methods are used, which includes: t-score, PMI and log-likelihood ratio, which have been proven to be effective on Russian data (Kopotev et al, 2017:155).

Tools: Collocations

CAT collocation lists are used for different purposes:

- determining the most specific collocations in a given domain
- determining those collocations in student texts, which are not attested in CAT, and suggesting possible replacements (by Word2vec)
- Spin-off: verifying theoretical observations on academic language, made in manuals of style.
- Spin-off: creating a dictionary of academic collocations

Tools: Collocations

- **Collocation list**: “автор выделяет”("author marks out"), “автор предлагает” ("author suggests"), “автор выражает” ("author expresses"), “автор придерживается”("author holds (the opinion)"), “автор публикует” ("author publishes"), “автор апеллирует”("author appeals"), “автор наталкивается”("author comes across"), “автор посвящает”("author dedicates"), “автор сосредоточивает”("author focuses"), “автор стремился”("author endeavoured").
- **Non-academic collocation from a student text**: "автор думает"("author thinks")

Tools: collocations

Non-attested collocations

All verbal and nominal collocations in a student text are compared to the domain-specific collocation list based on CAT. If a collocation is not attested, replacements are suggested. The replacements are chosen based on criteria described in (Liu et al., 2009). This includes PMI value, their vector similarity (Word2Vec model mentioned above) and the percentage of shared collocates in collocation cluster (Liu et al., 2009:48).

Text readability (beta)

Text readability is measured using various readability metrics (Flesch-Kincaid index, SMOG etc.); the obtained readability values are compared to those of the texts in reference corpus. TTR measures the lexical diversity of a text. If a TTR value is below an average for a specific domain, a warning is thrown.

Evaluation

- False positives: testing on CAT
- Both false positives and false negatives: testing on CoRST (Corpus of Russian Student Texts)
  `http://web-corpora.net/learner_corpus/`

References

Ackerman, K., & Chen, Y-H.

Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach.

Journal of English for Academic Purposes,2013, 12(4), 235-247.

Biber, D., Conrad., & Cortes, V.

If you look at . . . : Lexical bundles in University teaching and textbooks.

Applied Linguistics, 2004, 25(3), 371-405.

Durrant, P., & Mathews-Aydinli, J.

A function-first approach to identifying formulaic language in academic writing.

English for Specific Purposes, 2011, 30(1), 58-72.

Gray, B., & Biber, D.

Lexical frames in academic prose and conversation.

International Journal of Corpus Linguistics, 2013, 18(1), 109-136.

Hyland, K.

Hedging in academic writing and EAP textbooks.

English for Specific Purposes, 1994, 13, 239-56

References

Kopotev, M., O. Kisselev, and M. Polinsky

Collocations and near-native competence: Lexical strategies of heritage speakers of Russian

Balancing bilingualism: linguistic implications of input limitations. London: Oxford University Press. 2018. (In press).

Kopotev, M., Kormacheva, D., Pivovarova, L.

Evaluation of collocation extraction methods for the Russian language'.

Quantitative approaches to the Russian language. Routledge, 2018. pp. 137-157.

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.

Mediankin N., & Droganova K. (2016).

Building NLP Pipeline for Russian with a Handful of Linguistic Knowledge.

Proceedings of the Workshop on Computational Linguistics and Language Science, Copyright © CEUR-WS, Aachen, Germany, ISSN 1613-0073, pp. 48-56

References

Simpson-Vlach, R., and Ellis, N. C.

An academic formulas list: New methods in phraseology research.

Applied linguistics, 2010, 31(4), 487-512.

Swales, J.M.

Genre analysis: English in academic and research settings.

Cambridge: Cambridge University Press, 1990.

Zevakhina N., Dzhakupova S.

Corpus of Russian student texts: design and prospects.

Труды 21-й Международной конференции по компьютерной лингвистике "Диалог"

The End