Khokhlova Maria V., PhD, Associate Professor, Department of Mathematical Linguistics St. Petersburg State University

Topic: «The Study of Collocability and Frequency of Lexical Units in Russian Corpora »

1. Khokhlova M. Similarity between the Association Measures: a Case Study of Noun Phrases. In Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018. Edited by Horák A., Rychlý P., Rambousek A. Brno: Tribun EU, 2018. P. 21–27. (**Scopus**)

https://nlp.fi.muni.cz/raslan/2018/paper05-Khokhlova.pdf

2. Khokhlova M. Building a Gold Standard for a Russian Collocations Database. In Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana, 2018. P. 863–869. (**Scopus**)

https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2958-1-10-20180820.pdf

3. Klyshinsky E., Khokhlova M. In Search of Lost Collocations: Combining Measures to Reach the Top Range. In Internet and Modern Society: Proceedings of the International Conference IMS-2017 (St. Petersburg; Russian Federation, 21-24 June 2017). Eds Radomir V. Bolgov, Nikolai V. Borisov, Leonid V. Smorgunov, Irina I. Tolstikova, Victor P. Zakharov. ACM International Conference Proceeding Series. N.Y.: ACM Press, 2017. P. 160–163. DOI: 10.1145/3143699.3143731 (**Scopus**)

https://dl.acm.org/citation.cfm?id=3143731&dl=ACM&coll=DL

4. Khokhlova M. Comparison of High-Frequency Nouns from the Perspective of Large Corpora. In Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016. Edited by Horák A., Rychlý P., Rambousek A. Brno: Tribun EU, 2016. P. 9-17. (**Scopus**)

https://nlp.fi.muni.cz/raslan/2016/paper05-Khokhlova.pdf

5. Khokhlova M., Rosso P., Patti V. Distinguishing between Irony and Sarcasm in Social Media Texts: Linguistic Observations. In Proceedings of the International FRUCT Conference on Intelligence, Social Media and Web, ISMW FRUCT 2016. P. 17-22. (**Scopus**)

https://ieeexplore.ieee.org/document/7584765

Abstract

The research made by Dr Maria Khokhlova deals with to an important problem, i.e. the study of frequency and collocational properties of lexical units based on Russian corpora. In the last decade, linguists have addressed to corpora, which allows them to take a fresh look at phenomena that are described in the literature, including collocability. Despite the fact that there are dictionaries representing collocability, there is a lack of resources that would analyze this phenomenon on corpus data more profoundly.

The purpose of the work is to fully describe the frequency and collocational properties of lexical units on corpus data with the use of interdisciplinary methods, which include distribution analysis and statistical methods.

The phenomenon of collocability itself has a twofold nature, which can be explained, on the one hand, by the linguistic component and, on the other hand, by probabilistic (combinatorial) features. For a statistical evaluation of the strength between the words (based on the frequency of joint occurrence of words that are part of the described linguistic unit), the author applied association measures that count more than 80 metrics. In the papers Maria Khokhlova presents an overview of association measures that have been rarely mentioned in the scientific literature before (Jaccard measure, MS, z-score, log-log, true mutual information, etc.). The author testes several measures on Russian data and the results of the experiments were reflected in the papers (*paper number 1 in the list*). Among the found bigrams, the following groups of phrases were identified as showing different degrees of stability and reproducibility: proper names ("Eberswalde city"), free phrases ("Japanese publisher"), terms ("control chamber"), stable phrases ("undergo deformation") and phraseological units (“short leash”). It is also important to examine how reliable the automatically extracted data is. In this regard, the author began creating a gold standard of lexical collocability for the Russian language, which should be used to evaluate the results (*paper number 2 in the list*). Maria Khokhlova is developing a database that will include phrases from lexicographic sources, combined with information that was obtained during experiments. Thus, there will be a combination of the classical approach, which is understood as a description of collocations found in dictionaries and grammar guides, and modern technologies, including corpus and statistical approaches, as well as big data processing. A comparison was made with the data given in the explanatory and specialized dictionaries of the Russian language, and the metrics were used for the evaluation (MI, t-score, log likelihood coefficient, MI3, MS, Dice coefficient, etc.). It was shown that the best results were achieved by the logDice measure, which highlights both frequency recurrent phrases and combinations described in dictionaries.

Maria Khokhlova did not only study statistical methods, but also described more complex models that were applied to already obtained results. One of the methods of machine learning was implemented in the task of automatic extraction of combinations. It consists in the automatic selection of coefficients for the logistic regression function. The authors made an attempt to select coefficients for the optimal search of collocations found in dictionaries (*paper number 3 in the list*). The resulting quality metric was based on ranks for various statistical measures.

The linguists use large collections of texts called “text corpora” for describing the frequency characteristics of nouns and identifying data on collocability. These for Russian include ruWaC, ruTenTen and Aranea Russicum Maximum corpora. The corpora with volumes exceeding 100 million words have appeared relatively recently. Their appearance is associated with an increase in technical capabilities and a gradual change from “manual” building to a more automatic one. At the moment, we can talk about two types of corpora. The first type includes cases that are created by linguists according to a previously described technology, which can be called a “classical” or a “traditional” one. For such cases, texts are selected, marked and then loaded into a corpus. Corpora of the second type are automatically created based on texts crawled from the Internet. The earliest of these was ruWaC (Russian Web as Corpus). The Aranea Russicum Maximum and ruTenTen corpora are also examples of the second type, although they were built on a different principle.

Maria Khokhlova described the mentioned Russian corpora, analyzed the frequency distributions of high- and low-frequency words on their basis, made a comparison with the data of the frequency dictionary of the Russian language (*paper number 4 in the list*). During the

experiments, the author came to the conclusion that the results for high-frequency nouns obtained on 1 billion sample from the ruTenTen corpus and its full version do not differ, which is not true in the case of low-frequency nouns. In general, in comparison with other text collections, e.g. the ruWaC corpus, showed similar results with the frequency dictionary of the Russian language, built on the basis of the Russian National Corpus.

The studied statistical methods can be applied to various corpus data, including the analysis of social media texts that have been actively studied recently. *The paper number 5* is focused on the definition of statistical parameters, as well as the identification of set phrases, by which it would be possible to automatically identify ironic and sarcastic texts.

Maria Khokhlova has published 16 works indexed by the Web of Science or Scopus databases, 14 of them were published over the past 5 years. 5 works studying the collocability and frequency properties of lexical units in Russian corpora were selected.

Currently, there is a gradual transition from simple tools that allows linguists to work with corpus data to more complex systems that provide wider opportunities for users. The toolkit of statistical measures described in the works of Maria Khokhlova opens up new perspectives in processing corpus data, being a kind of filter for them.

The results of this work can be used in courses on lexicology, morphology and syntax of modern Russian, in compiling dictionaries and grammars, as well as in teaching Russian. They can also be used in machine learning and for automatic language processing, for example, in automatic clustering of phrases or for disambiguation.