Advances in Science and Humanities



Advances in Science and Humanities, 2026, 39–41

doi: 10.70728/human.v02.i04.010
Volume 02, Issue 04
ISSN: 3060-5393
Paper

PAPER

CORPUS-STATISTICAL ANALYSIS OF RUSSIAN VOCABULARY USING COMPUTER LINGUISTICS METHODS

Manzura Sharipovna Boltaeva[1,*]

[1]Head of the Department of General Education Bukhara Innovative Educational and Medical University

* sharipovna@gmail.com

Abstract

This article examines the theoretical and applied aspects of corpus-based statistical analysis of Russian vocabulary using computational linguistics methods. The study aims to identify quantitative patterns in the functioning of lexical units in various types of texts using automated data processing tools. The National Corpus of the Russian Language serves as the empirical base. The paper describes methods of frequency analysis, morphological annotation, collocation analysis, and vector modeling of semantics. It concludes that the integration of statistical and computational methods is highly effective in studying the lexical system of the Russian language.

Key words: Corpus linguistics, computational linguistics, Russian vocabulary, frequency analysis, collocations, morphological annotation, semantic modeling.

INTRODUCTION

The current stage of linguistic development is characterized by the active integration of digital technologies into humanities research. Traditional text analysis methods based on intuitive observations and qualitative descriptions are gradually being supplemented by quantitative methods for processing large data sets. This is particularly relevant for studying the vocabulary of the Russian language, which has a complex morphological structure and a high degree of variability.

Corpus linguistics as an independent field emerged in the second half of the 20th century and today is one of the key tools for studying linguistic phenomena. Electronic text corpora allow for the analysis of real examples of word usage in various contexts, genres, and historical periods. The use of computer algorithms ensures the objectivity and

Compiled on: March 2, 2026.
Copyright:©2026 by the authors. Submitted to Advances in Science and Humanities for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) 4.0 license.

reproducibility of results.

The relevance of this study stems from the need for a systematic analysis of Russian vocabulary using modern computing technologies.

The purpose of this article is to examine the possibilities of corpus-based statistical analysis of Russian vocabulary and to demonstrate the effectiveness of computational linguistics methods in studying linguistic data. Research objectives:

1. To characterize the theoretical foundations of corpus linguistics.

2. To describe methods of computer text processing.

3. To analyze statistical approaches to vocabulary study.

4. To identify the functional characteristics of lexical units in a corpus of texts.

Theoretical Foundations of Corpus Analysis

Corpus linguistics is based on the use of representative text arrays provided with special markup. A corpus is a structured database containing texts of various genres: fiction, journalism, scientific articles, colloquial speech, etc.

One of the most authoritative resources is the National Corpus of the Russian Language, which includes the main, newspaper, poetry, oral, and historical subcorpora. The corpus is annotated with morphological, syntactic, and partially semantic data, enabling a comprehensive analysis of linguistic units.

The principle of corpus representativeness presupposes an even distribution of texts across genres and time periods. This ensures the objectivity of statistical conclusions and minimizes distortions associated with the dominance of a particular text type.

Computational Linguistics Methods in Lexicon Analysis

Morphological Analysis and Lemmatization The Russian language is characterized by a developed inflectional system, resulting in multiple word forms for the same lexeme. Automatic lemmatization allows for the combination of word forms into a single lexical unit (lemma), which is necessary for accurate frequency analysis.

Morphological annotation includes the identification of part of speech, gender, number, case, and other grammatical characteristics. This allows us to explore the grammatical patterns of word use in different contexts.

Frequency Analysis

Frequency is one of the main indicators of vocabulary functioning. Analyzing word frequency reveals the core of a language's lexical system. Functional words (prepositions, conjunctions, particles) are typically the most frequent, while notional vocabulary exhibits greater variability.

Frequency dictionaries based on a corpus reflect real-world linguistic practice and are used in lexicography and Russian language teaching.

For example, analyzing noun-adjective combinations reveals typical semantic patterns and set phrases characteristic of certain genres.

Machine Learning Methods Machine learning algorithms allow:

identifying semantic word groups;

automatically classifying texts;

modeling semantic fields;

analyzing the dynamics of lexical change. Corpus analysis shows that the frequency and distribution of vocabulary significantly depend on the text genre. Scientific texts are dominated by terminological vocabulary, journalism by socio-political vocabulary, and fiction by emotionally expressive vocabulary.

Diachronic Aspect

Using historical subcorpora allows us to track changes in word frequency over different periods. This makes it possible to study the processes of neologization and archaization. For example, words related to digital technologies show a sharp increase in frequency in 21st-century texts.

Based on statistical data, it is possible to form semantic clusters reflecting thematic groups of vocabulary. This approach allows for a quantitative description of the structure of semantic fields.

Discussion of Results

Corpus-statistical analysis demonstrates objective patterns in the functioning of Russian vocabulary. Quantitative methods allow us to confirm or refine the conclusions drawn within traditional

linguistics.

The integration of statistics and computer technology expands research capabilities and promotes an interdisciplinary approach to language study. Practical Significance

The results of corpus analysis are used in:

the creation of modern dictionaries;

automatic translation;

the development of natural language processing systems;

teaching Russian as a foreign language;

the creation of intelligent search engines.

CONCLUSION

Corpus-statistical analysis of Russian vocabulary is an effective tool in modern linguistics. The application of computational linguistics methods allows for large-scale research based on real language data.

The use of electronic corpora, such as the National Corpus of the Russian Language, ensures a high degree of objectivity and reproducibility of scientific results. Thus, the integration of corpus and statistical methods opens up new prospects for a comprehensive study of the lexical system of the Russian language in the digital age.

References

1. Zakharov, V. P. Corpus Linguistics: A Manual / V. P. Zakharov, S. Yu. Bogdanova. – Saint Petersburg: SPbGU Publishing House, 2011. – 272 p.
2. Kibrik, A. A. Methods of Quantitative Text Analysis / A. A. Kibrik. – Moscow: Languages of Slavic Culture, 2003. – 304 p.
3. Plungyan, V. A. General Morphology: An Introduction to the Problematic / V. A. Plungyan. – Moscow: Editorial URSS, 2000. – 384 p.
4. The National Corpus of the Russian Language [Electronic resource]. – Access mode: https://www.ruscorpora.ru (date of access: 02/10/2026).
5. Abduazizov, A. A. Tilshunoslik nazariyasiga kirish / A. A. Abduazizov. – Tashkent: Okituvchi, 2010. – 256 b.
6. Makhmudov, N. M. Ozbek tili lexicology / N. M. Makhmudov. – Tashkent: Fan, 2012. – 320 b.
7. Safarov, Sh. S. Cognitive tilshunoslik asoslari / Sh. S. Safarov. – Tashkent: Akademnashr, 2013. – 288 b.
8. Yўldoshev, B. Yu. Matn linguistics / B. Yu. Yўldoshev. – Tashkent: Fan va technology, 2015. – 240 b.
9. Rakhmatullaev, Sh. R. Ҳozirgi ўzbek adabii tili / Sh. R. Rakhmatullaev. – Tashkent: University, 2006. – 384 b.
10. Khozhiev, A. G. Ozbek tili morphology / A. H. Khozhiev. – Tashkent: Fan, 2008. – 296 b.
11. Nurmonov, A. N. Linguistic tadkikot methodology / A. N. Nurmonov. – Tashkent: Yangi asr avlodi, 2014. – 224 b.
12. Sharapov, Sh. M. Computer linguistics asoslari / Sh. M. Sharapov. – Toshkent: Tafakkur, 2018. – 208 b.