

КОРПУСНО-СТАТИСТИЧЕСКИЙ АНАЛИЗ ЛЕКСИКИ РУССКОГО ЯЗЫКА С ПРИМЕНЕНИЕМ МЕТОДОВ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

<https://doi.org/10.70728/phil.v05.i02.018>

Болтаева Манзура Шариповна

Бухарский инновационный образовательный и медицинский университет

Аннотация: В статье рассматриваются теоретические и прикладные аспекты корпусно-статистического анализа лексики русского языка на основе методов компьютерной лингвистики. Исследование направлено на выявление количественных закономерностей функционирования лексических единиц в различных типах текстов с использованием инструментов автоматической обработки данных. В качестве эмпирической базы используется Национальный корпус русского языка. В работе описываются методы частотного анализа, морфологической разметки, коллокационного анализа и векторного моделирования семантики. Делается вывод о высокой эффективности интеграции статистических и вычислительных методов в исследовании лексической системы русского языка.

Ключевые слова: корпусная лингвистика, компьютерная лингвистика, лексика русского языка, частотный анализ, коллокации, морфологическая разметка, семантическое моделирование

CORPUS-STATISTICAL ANALYSIS OF RUSSIAN VOCABULARY USING COMPUTER LINGUISTICS METHODS

Manzura Sharipovna Boltaeva

Bukhara Innovative Educational and Medical University

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.

Keywords: corpus linguistics, computational linguistics, Russian vocabulary, frequency analysis, collocations, morphological annotation, semantic modeling

ВВЕДЕНИЕ

Современный этап развития языкознания характеризуется активным внедрением цифровых технологий в гуманитарные исследования. Традиционные методы анализа текста, основанные на интуитивных наблюдениях и качественном описании, постепенно дополняются количественными методами обработки больших массивов данных. Это особенно актуально для исследования лексики русского языка, обладающего сложной морфологической структурой и высокой степенью вариативности.

Корпусная лингвистика как самостоятельное направление сформировалась во второй половине XX века и сегодня является одним из ключевых инструментов изучения языковых явлений. Электронные текстовые корпуса позволяют анализировать реальные примеры употребления слов в различных контекстах, жанрах и исторических периодах. Использование компьютерных алгоритмов обеспечивает объективность и воспроизводимость результатов.

Актуальность исследования обусловлена необходимостью системного анализа лексики русского языка с применением современных вычислительных технологий.

Цель статьи - рассмотреть возможности корпусно-статистического анализа лексики русского языка и продемонстрировать эффективность методов компьютерной лингвистики при исследовании языковых данных.

Задачи исследования:

1. Охарактеризовать теоретические основы корпусной лингвистики.
2. Описать методы компьютерной обработки текста.
3. Проанализировать статистические подходы к изучению лексики.
4. Выявить особенности функционирования лексических единиц в корпусе текстов.

Теоретические основы корпусного анализа

Корпусная лингвистика основывается на использовании репрезентативных текстовых массивов, снабжённых специальной разметкой. Корпус представляет собой структурированную базу данных, включающую тексты различных жанров: художественные произведения, публицистику, научные статьи, разговорную речь и др.

Одним из наиболее авторитетных ресурсов является Национальный корпус русского языка, включающий основной, газетный, поэтический, устный и исторический подкорпуса. Корпус снабжен морфологической, синтаксической и частично семантической разметкой, что позволяет проводить комплексный анализ языковых единиц.

Принцип репрезентативности корпуса предполагает равномерное распределение текстов по жанрам и временным периодам. Это обеспечивает объективность статистических выводов и минимизирует искажения, связанные с доминированием определенного типа текстов.

Методы компьютерной лингвистики в анализе лексики

Морфологический анализ и лемматизация

Русский язык характеризуется развитой флективной системой, что приводит к множеству словоформ одной и той же лексемы. Автоматическая лемматизация позволяет объединять словоформы в единую лексическую единицу (лемму), что необходимо для корректного частотного анализа.

Морфологическая разметка включает определение части речи, рода, числа, падежа и других грамматических характеристик. Это позволяет исследовать грамматические закономерности употребления слов в разных контекстах.

Частотный анализ

Частотность является одним из основных показателей функционирования лексики. Анализ частоты употребления слов позволяет выявить ядро лексической системы языка. Как правило, наиболее частотными являются служебные слова (предлоги, союзы, частицы), тогда как знаменательная лексика демонстрирует более широкую вариативность.

Частотные словари, построенные на основе корпуса, отражают реальную языковую практику и используются в лексикографии и преподавании русского языка.

Например, анализ сочетаний существительных с прилагательными позволяет выявить типичные семантические модели и устойчивые словосочетания, характерные для определенных жанров.

Методы машинного обучения Алгоритмы машинного обучения позволяют:

- выявлять семантические группы слов;
- проводить автоматическую классификацию текстов;
- моделировать семантические поля;
- анализировать динамику лексических изменений.

Корпусный анализ показывает, что частотность и распределение лексики существенно зависят от жанра текста. В научных текстах преобладает терминологическая лексика, в публицистике - общественно-политическая, в художественной литературе - эмоционально-экспрессивная.

Диахронический аспект

Использование исторических подкорпусов позволяет отслеживать изменения частотности слов в разные периоды. Это дает возможность изучать процессы неологизации и архаизации.

Например, слова, связанные с цифровыми технологиями, демонстрируют резкий рост частотности в текстах XXI века.

На основе статистических данных возможно формирование семантических кластеров, отражающих тематические группы лексики. Такой подход позволяет количественно описывать структуру семантических полей.

Обсуждение результатов

Корпусно-статистический анализ демонстрирует объективные закономерности функционирования лексики русского языка. Количественные методы позволяют подтвердить или уточнить выводы, полученные в рамках традиционной лингвистики.

Интеграция статистики и компьютерных технологий расширяет исследовательские возможности и способствует формированию междисциплинарного подхода к изучению языка.

Практическая значимость

Результаты корпусного анализа применяются в:

- создании современных словарей;
- автоматическом переводе;
- разработке систем обработки естественного языка;
- обучении русскому языку как иностранному;
- создании интеллектуальных поисковых систем.

ЗАКЛЮЧЕНИЕ

Корпусно-статистический анализ лексики русского языка представляет собой эффективный инструмент современного языкознания. Применение методов компьютерной лингвистики позволяет проводить масштабные исследования, основанные на реальных языковых данных.

Использование электронных корпусов, таких как Национальный корпус русского языка, обеспечивает высокую степень объективности и воспроизводимости научных результатов.

Таким образом, интеграция корпусных и статистических методов открывает новые перспективы для комплексного изучения лексической системы русского языка в условиях цифровой эпохи.

Использованная литература

1. Захаров, В. П. Корпусная лингвистика : учеб. пособие / В. П. Захаров, С. Ю. Богданова. - Санкт-Петербург : Изд-во СПбГУ, 2011. - 272 с.
2. Кибрик, А. А. Методы количественного анализа текста / А. А. Кибрик. - Москва : Языки славянской культуры, 2003. - 304 с.
3. Плунгян, В. А. Общая морфология: Введение в проблематику / В. А. Плунгян. - Москва : Едиториал УРСС, 2000. - 384 с.

4. Национальный корпус русского языка [Электронный ресурс]. - Режим доступа: <https://www.ruscorpora.ru> (дата обращения: 10.02.2026).
5. Абдуазизов, А. А. Тилшунослик назариясига кириш / А. А. Абдуазизов. - Тошкент : Ўқитувчи, 2010. - 256 б.
6. Махмудов, Н. М. Ўзбек тили лексикологияси / Н. М. Махмудов. - Тошкент : Фан, 2012. - 320 б.
7. Сафаров, Ш. С. Когнитив тилшунослик асослари / Ш. С. Сафаров. - Тошкент : Akademnashr, 2013. - 288 б.
8. Йўлдошев, Б. Ю. Матн лингвистикаси / Б. Ю. Йўлдошев. - Тошкент : Фан ва технология, 2015. - 240 б.
9. Раҳматуллаев, Ш. Р. Ҳозирги ўзбек адабий тили / Ш. Р. Раҳматуллаев. - Тошкент : Университет, 2006. - 384 б.
10. Ҳожиёв, А. Ҳ. Ўзбек тили морфологияси / А. Ҳ. Ҳожиёв. - Тошкент : Фан, 2008. - 296 б.
11. Нурмонов, А. Н. Лингвистик тадқиқот методологияси / А. Н. Нурмонов. - Тошкент : Янги аср авлоди, 2014. - 224 б.
12. Шарапов, Ш. М. Компьютер лингвистикаси асослари / Ш. М. Шарапов. - Тошкент : Тафаккур, 2018. - 208 б.