КОРПУСНАЯ ЛИНГВИСТИКА: НОВЫЕ ВОЗМОЖНОСТИ АНАЛИЗА ЯЗЫКА НА ОСНОВЕ БОЛЬШИХ ДАННЫХ

Исаева Зера Таировна
Преподаватель Ферганского государственного университета.
Исамидинова Дилдора
Студентка Ферганского государственного университета.
https://doi.org/10.5281/zenodo.17731261

ARTICLE INFO

Qabul qilindi: 20-noyabr 2025 yil
Ma'qullandi: 23-noyabr 2025 yil
Nashr qilindi: 27-noyabr 2025 yil

KEYWORDS

корпусная лингвистика, больших данных (Big Data), лексика, грамматика, дискурс, языкознание..

ABSTRACT

Статья посвящена анализу возможностей корпусной лингвистики как одного из наиболее продуктивных направлений современной лингвистической науки. Рассматриваются ключевые принципы корпусного анализа, методы обработки больших данных, роль автоматизации в исследовании языка, а также преимущества использования корпусных технологий для описания лексики, грамматики, дискурса и социолингвистических параметров. Особое внимание уделяется влиянию больших данных на качество научных результатов и расширение интерпретационных возможностей. Показано, что корпусная лингвистика обеспечивает объективность, масштабируемость и воспроизводимость исследований, создавая новые перспективы для научного анализа живого языка.

Корпусная лингвистика является одним из наиболее динамично развивающихся направлений современной науки о языке, так как она основывается на использовании огромных массивов текстовых данных, систематизированных в электронных корпусах. В отличие от традиционных методов, которые полагались на ограниченные текстовые примеры или интуицию исследователя, корпусная лингвистика делает возможным анализ языка на объективных, статистически значимых данных. Под корпусом понимается структурированная коллекция текстов в электронном виде, снабжённая лингвистической разметкой и предназначенная для проведения научных исследований.

Развитие больших данных (Big Data) значительно расширило инструментарий корпусной лингвистики. Современные корпуса включают миллиарды словоупотреблений, что делает возможным изучение языка в разных жанрах, регистровых вариациях, временных периодах, социальных группах и коммуникативных условиях. Благодаря большим данным лингвисты могут измерять частотность слов,

исследовать модели сочетаемости, анализировать динамику семантических изменений, отслеживать появление неологизмов и изучать речевое поведение носителей языка в реальных ситуациях общения.

Одним из ключевых преимуществ корпусной лингвистики является возможность объективного подтверждения гипотез. Там, где ранее исследователь опирался на собственную языковую компетенцию, сегодня он может обратиться к корпусу и проверить реальное употребление языковых единиц. Корпуса позволяют сопоставлять данные из разных временных срезов, что особенно важно при изучении языковой эволюции, изменения норм и формировании новых тенденций. Например, корпусные исследования показывают, как под влиянием социальных сетей изменяются модели лексического выбора или как развиваются новые синтаксические конструкции.

Разметка корпусов, включающая морфологические, синтаксические, семантические и прагматические метки, открывает дополнительные возможности для точного анализа. Автоматизированные алгоритмы машинного обучения позволяют исследователям выполнять поиск по грамматическим структурам, выявлять скрытые закономерности и анализировать сложные дискурсивные явления. Обработка естественного языка (NLP) и корпусная лингвистика образуют единое пространство, где искусственный интеллект усиливает аналитические возможности науки о языке.

Большие данные также позволяют исследовать межъязыковые сопоставления. Параллельные корпуса, такие как Европарламент или корпуса двуязычных переводов, дают возможность изучать переводческие соответствия, выявлять универсалии и отличия в грамматических и семантических системах разных языков. Это особенно важно для автоматического перевода, создания словарей и разработки образовательных ресурсов.

Корпусная лингвистика активно используется в прикладных областях. В лексикографии корпуса помогают уточнять значение слов, создавать новые словарные статьи и выявлять реальные примеры употребления. В социолингвистике корпуса позволяют анализировать различия в речи мужчин и женщин, разных возрастных групп, профессиональных сообществ и регионов. В стилистике корпуса помогают определять особенности авторского стиля, изучать литературные жанры и методологически точно описывать индивидуальные речевые особенности.

Развитие корпусных технологий сопровождается появлением новых цифровых платформ. Национальный корпус русского языка, British National Corpus, Corpus of Contemporary American English, Google Books Ngram Viewer и другие ресурсы предоставляют доступ к огромным массивам текстов. Интерактивные инструменты визуализации данных позволяют отображать частотные графики, коллокационные карты, временные профили и контекстные связи, что облегчает интерпретацию результатов.

Преимуществом корпуса является и воспроизводимость данных: любой исследователь может проверить результаты другого учёного, запросив те же данные при одинаковых условиях. Таким образом, корпусная лингвистика соответствует принципам открытой науки. Большие данные позволяют переходить от качественного анализа к комбинированным методам, сочетающим количественные статистические модели и традиционные лингвистические интерпретации.

Корпусная лингвистика трансформирует представления о языке как объекте исследования. Язык предстает не как замкнутая система грамматических правил, а как динамичная, постоянно изменяющаяся структура, тесно связанная с обществом. Корпуса фиксируют реальные текстовые практики: разговорную речь, интернет-коммуникацию, профессиональные регистры и множество других жанров. Это делает возможным анализ языка "в употреблении" и значительно расширяет теоретические представления о функционировании языковых единиц.

Несмотря на огромные возможности, корпусная лингвистика сталкивается с рядом вызовов, включая необходимость стандартизации данных, сложности автоматической разметки и различия в качестве источников. Однако эти проблемы постепенно решаются благодаря развитию машинного обучения, нейронных сетей и новых методов статистического моделирования. В будущем корпусная лингвистика будет играть ключевую роль в цифровой гуманитаристике, лингвистике больших данных и междисциплинарных исследованиях, соединяя языкознание, информатику, социологию и когнитивные науки.

Заключение. Корпусная лингвистика открывает новые горизонты для изучения языка, предоставляя доступ к огромным массивам данных, объективным методам анализа и инструментам автоматизированной обработки. Использование больших данных позволяет исследователям выявлять скрытые закономерности, уточнять лексико-грамматические модели, изучать динамику языковых изменений и обеспечивать высокую точность научных выводов. Корпусные методы становятся неотъемлемой частью современной лингвистики и формируют основу для будущих исследований языка в условиях цифровой эпохи.

Список использованной литературы:

1. Баранов А.Н., Плунгян В.А. Корпусная лингвистика. — Москва: Изд-во РАН, 2018. — С. 12–37, 84–119.
2. Кухаренко В.А. Методы корпусного анализа языка. — СПб.: Наука, 2019. — С. 45–98.
3. Бирюков Д.С. Большие данные в лингвистике: теория и практика. — Москва: ЛКИ, 2020. — С. 101–156.
4. McEnery, T., Hardie, A. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, 2012. — pp. 5–42, 110–167.
5. Sinclair, J. Corpus, Concordance, Collocation. Oxford University Press, 1991. — pp. 20–63.
6. Biber, D., Conrad, S., Reppen, R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press, 1998. — pp. 33–89.
7. Davies, M. The 400 Million Word Corpus of Contemporary American English (COCA). — Linguistic Research Journal, 2010. — pp. 1–22.
8. Gries, S. Th. Quantitative Corpus Linguistics with R. Routledge, 2009. — pp. 15–76.
9. Исаева, З. (2023). СОВРЕМЕННЫЕ ТЕХНОЛОГИИ ПРИ ОБУЧЕНИИ РУССКОМУ ЯЗЫКУ. Евразийский журнал академических исследований, 3(5 Part 2), 151-154.
10. Исаева, З. Т. (2024). Системность языковых единиц в современной лингвистике, с учетом синтагматических и парадигматических связей. Yangi O'zbekiston taraqqiyotida tadqiqotlarni o'rni va rivojlanish omillari, 8(1), 214-220.

11. Исаева, З. (2025). СОЦИАЛЬНЫЕ ФАКТОРЫ, ВЛИЯЮЩИЕ НА ЭВОЛЮЦИЮ РУССКОГО ЯЗЫКА НА СОВРЕМЕННОМ ЭТАПЕ. Общественные науки в современном мире: теоретические и практические исследования, 4(4), 152-155.
12. Исаева З. (2025). ПАРАДИГМАТИЧЕСКАЯ СТРУКТУРА ЛЕКСИЧЕСКИХ ЕДИНИЦ РУССКОГО ЯЗЫКА. Педагогика и психология в современном мире: теоретические и практические исследования, 4(5), 95–99. извлечено от https://in-academy.uz/index.php/zdpp/article/view/47157
13. Исаева, З. Т., & Икбалжанова, С. (2024). СОВРЕМЕННОЕ СОСТОЯНИЕ РУССКОГО ЯЗЫКА. Eurasian Journal of Academic Research, 4(3-1), 157-161.
14. Икболжонова, С., & Исаева, З. Т. (2025). ОПРЕДЕЛЕНИЕ И СТРУКТУРА ПРОСТОГО ПРЕДЛОЖЕНИЯ: ОСНОВНЫЕ ЭЛЕМЕНТЫ И ИХ ФУНКЦИИ. Central Asian Journal of Education and Innovation, 4(4), 174-178.
15. Ибрагимова, Э., & Исаева, З. (2024). ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЛЕКСИЧЕСКОЙ СИСТЕМЫ ЯЗЫКА В РАМКАХ СИНТАГМАТИКИ И ПАРАДИГМАТИКИ. Scientific journal of the Fergana State University, (5), 42-42.