ЦИФРОВЫЕ КОРПУСЫ, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ЛИНГВИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В СОВРЕМЕННЫХ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Бахромжонова Шахло Бахтиеровна
1 Факультет английского языка и литературы, 3 курс
Shahlobahramjanova38@gmail.com
ORCID: 0009-0005-1564-7258
Научный руководитель: Алимханова Нигорахон Амилевна
Узбекский государственный университет мировых языков

Аннотация. В статье рассматриваются актуальные вопросы современных лингвистических исследований, связанные с цифровыми корпусами, искусственным интеллектом и лингвистическим моделированием. Анализируется их роль в обработке и интерпретации языковых данных, а также в выявлении закономерностей функционирования языка. Подчёркивается значение данных технологий для развития прикладной лингвистики и совершенствования методов обучения иностранным языкам.

Ключевые слова: цифровые корпусы, искусственный интеллект, лингвистическое моделирование, корпусная лингвистика, обработка естественного языка.

Annotation. This article examines key issues in modern linguistic research related to digital corpora, artificial intelligence, and linguistic modeling. It analyzes their role in processing and interpreting language data, as well as in identifying patterns of language use. The study highlights their importance for the development of applied linguistics and the improvement of foreign language teaching methods.

Keywords: digital corpora, artificial intelligence, linguistic modeling, corpus linguistics, natural language processing.

Annotatsiya. Mazkur maqolada zamonaviy lingvistik tadqiqotlarda raqamli korpuslar, sun’iy intellekt va lingvistik modellashtirish bilan bog‘liq masalalar ko‘rib chiqiladi. Ularning til ma’lumotlarini qayta ishlash va tahlil qilishdagi roli hamda til qonuniyatlarini aniqlashdagi ahamiyati tahlil etiladi. Shuningdek, ushbu texnologiyalarning amaliy lingvistika rivoji va xorijiy tillarni o‘qitish samaradorligini oshirishdagi o‘rni yoritiladi.

Kalit so‘zlar: raqamli korpuslar, sun’iy intellekt, lingvistik modellashtirish, korpus lingvistikasi, tabiiy tilni qayta ishlash.

Современный этап развития лингвистической науки в основном характеризуется активной интеграцией цифровых технологий, которые существенно расширяют традиционные методы анализа языка. Особое место в этом процессе занимают цифровые корпусы, которые позволяют работать с большими массивами аутентичных текстов и выявлять языковые закономерности на основе эмпирических данных [1-6]. Параллельно развивается искусственный интеллект(ИИ), который, благодаря методам обработки естественного языка, обеспечивает автоматизацию анализа и интерпретации различной текстовой информации [5]. Взаимодействие этих направлений формирует основу для лингвистического моделирования, направленного на формализацию и компьютерное представление языковых процессов[2]. Актуальность исследования обусловлена тем, что в условиях цифровизации науки и образования традиционные

DOI: 10.5281/zenodo.20377566

методы лингвистического анализа уже не в полной мере соответствуют требованиям обработки больших объемов языковых данных. Возникает необходимость в применении более точных, автоматизированных и масштабируемых инструментов, способных обеспечить высокую скорость и объективность анализа языковых явлений. В этом контексте цифровые корпусы и технологии ИИ становятся ключевыми ресурсами для развития современной лингвистики и образовательных практик. Проблема исследования заключается в недостаточной систематизации знаний о комплексном использовании цифровых корпусов, ИИ и лингвистического моделирования в единой исследовательской парадигме, а также в отсутствии чётких методических подходов к их интеграции в лингвистический анализ и обучение языкам. Научная новизна исследования заключается в рассмотрении цифровых корпусов, ИИ и лингвистического моделирования как взаимосвязанной системы инструментов, обеспечивающих многоуровневый анализ языка. Таким образом, в работе акцентируется внимание на их совместном потенциале для повышения точности лингвистических исследований и эффективности образовательного процесса, а также на возможностях их интеграции в прикладные задачи современной лингвистики и преподавания иностранных языков.

Методы исследования. Данное исследование выполнено в рамках теоретико-аналитического подхода, основанного на обзоре и систематизации научной литературы в области корпусной лингвистики, ИИ и лингвистического моделирования. Такой тип исследования позволяет не только описать существующие научные направления, но и выявить основные тенденции развития цифровых методов анализа языка в условиях современной цифровизации науки и образования. Методологическую основу работы составляют современные исследования в области цифровой лингвистики, обработки естественного языка и корпусных технологий. Особое внимание уделяется междисциплинарному характеру исследования, поскольку анализ языка в цифровую эпоху требует интеграции лингвистики, информатики и когнитивных наук [3][5]. В качестве эмпирической базы исследования использовались международные и национальные научные ресурсы, а также корпусные платформы, позволяющие анализировать реальные языковые данные и проверять лингвистические гипотезы на больших текстовых массивах[4]. В исследовании использовались корпусные ресурсы, которые являются ключевыми инструментами современной корпусной лингвистики и позволяют проводить анализ языка на основе реальных текстовых данных. Их использование обеспечивает эмпирическую достоверность и объективность результатов лингвистического анализа. Кроме того, позволяют осуществлять комплексное исследование языковых единиц, включая анализ частотности, контекстов употребления и сочетаемости слов. Благодаря этому становится возможным выявление устойчивых языковых моделей и закономерностей функционирования языка в реальной коммуникации, что значительно повышает научную обоснованность выводов. В работе применялись следующие методы исследования, обеспечивающие комплексный характер анализа и достоверность полученных результатов: анализ научной литературы — для изучения теоретических основ корпусной лингвистики, ИИ и языкового моделирования, а также выявления современных научных подходов в данной области; сравнительный анализ — для сопоставления традиционных, корпусных и цифровых подходов к исследованию языка, а также выявления их преимуществ и ограничений; описательный метод — для детального анализа функций цифровых корпусов и интеллектуальных систем обработки языка, включая их

возможности в лингвистическом моделировании; обобщение и систематизация — для выявления общих закономерностей развития цифровой лингвистики и формирования целостного представления о взаимодействии рассматриваемых технологий. Использование данных методов позволяет обеспечить многоуровневый анализ исследуемой проблемы и повысить научную обоснованность полученных выводов [1].

Результаты исследования. Проведённый теоретико-аналитический обзор научных источников и корпусных ресурсов показал, что цифровые корпусы, технологии ИИ и лингвистическое моделирование образуют взаимосвязанную и функционально интегрированную систему современного языкового анализа. Полученные результаты позволяют выделить несколько ключевых направлений их влияния на развитие лингвистики. Анализ корпусных данных подтвердил, что цифровые корпусы являются одним из наиболее надёжных инструментов изучения реального функционирования языка [6]. Их основное преимущество заключается в том, что они основаны на больших массивах аутентичных текстов, отражающих живое употребление языка в различных коммуникативных ситуациях. Таким образом, цифровые корпусы обеспечивают переход от интуитивного описания языка к эмпирически обоснованному и статистически подтверждённому анализу языковых явлений, что значительно повышает объективность лингвистических исследований [6]. Возможности ИИ в обработке языка на современном этапе развития науки демонстрируют значительный прогресс и широкий спектр применения. Результаты анализа показывают, что технологии ИИ играют ключевую роль в автоматизации обработки естественного языка и существенно расширяют функциональные возможности лингвистического анализа [5]. К примеру, современные языковые модели способны выполнять разнообразные задачи, обеспечивая более глубокое и точное взаимодействие с текстовой информацией. В частности, одним из наиболее распространённых направлений является машинный перевод, который позволяет автоматически преобразовывать тексты с одного языка на другой с учётом контекста и семантических особенностей высказывания. Не менее важной является функция генерации текста, при которой интеллектуальные модели создают связные, логически структурированные и содержательно полноценные тексты на основе заданных пользователем запросов. Кроме того, технологии ИИ активно применяются для анализа тональности, что даёт возможность определять эмоциональную окраску текста и выявлять отношение автора к описываемым событиям или объектам. Важное значение имеет классификация текстов, предполагающая автоматическое распределение текстовой информации по различным тематическим и функциональным категориям. Наряду с этим, системы ИИ используются для обработки грамматики и синтаксиса, что находит отражение в разработке современных инструментов проверки и редактирования текстов. Таким образом, полученные результаты свидетельствуют о том, что ИИ не только способствует автоматизации процессов языкового анализа, но и открывает новые перспективы для взаимодействия человека с текстом. Это проявляется, в частности, в создании интеллектуальных образовательных и коммуникативных систем, которые обеспечивают более эффективное усвоение информации и расширяют возможности языкового общения. Одним из наиболее значимых результатов исследования стало выявление тесной взаимосвязи между цифровыми корпусами и технологиями ИИ. Установлено, что эти направления не существуют отдельно, а формируют единый цикл обработки языковых данных. С одной стороны, цифровые корпусы выполняют функцию основного источника

обучающих данных для языковых моделей [7]. Именно на основе корпусных данных формируются алгоритмы машинного обучения, позволяющие моделям понимать структуру языка, его грамматику и семантику. С другой стороны, ИИ используется для улучшения корпусного анализа, так как позволяет автоматически: выявлять скрытые закономерности в текстах; классифицировать большие объёмы данных; проводить семантическое моделирование; ускорять обработку языковых корпусов. Таким образом, наблюдается двусторонняя взаимосвязь: корпусы обеспечивают данные, а ИИ обеспечивает их интеллектуальную обработку, что формирует основу современной вычислительной лингвистики. В целом, результаты исследования подтверждают, что цифровые корпусы обеспечивают эмпирическую базу для изучения языка, ИИ обеспечивает автоматизацию и интеллектуализацию анализа, а их интеграция формирует новый уровень лингвистического моделирования. Данный подход позволяет значительно повысить точность, масштабируемость и эффективность лингвистических исследований, а также открывает новые возможности для развития цифрового образования.

Обсуждение. Полученные результаты подтверждают высокую эффективность интеграции цифровых корпусов, ИИ и методов лингвистического моделирования в современных лингвистических исследованиях. Однако наряду с очевидными преимуществами данных технологий необходимо учитывать ряд ограничений, проблем и этических аспектов, влияющих на их практическое применение. Несмотря на высокий уровень развития языковых моделей, их использование сопровождается рядом существенных ограничений. Одной из ключевых проблем являются так называемые “галлюцинации моделей”, когда система генерирует правдоподобную, но фактически некорректную информацию. Это снижает надёжность автоматического анализа текстов и требует дополнительной проверки результатов человеком. Другой важной проблемой является смещение данных, возникающее из-за неравномерности обучающих корпусов. Если данные содержат культурные, социальные или языковые перекосы, модель воспроизводит их в своих ответах, что может приводить к искажению результатов анализа. Также существенным ограничением является зависимость качества работы моделей от обучающего корпуса [1] [4]. Чем более репрезентативны и разнообразны данные, тем точнее результаты обработки языка, и наоборот. Несмотря на высокую эффективность, корпусная лингвистика также имеет ряд ограничений. Во-первых, многие корпуса содержат неполные или ограниченные данные, что не всегда отражает всю полноту языковой реальности [1]. Во-вторых, существует проблема устаревания корпусов, так как язык постоянно развивается, появляются новые слова, выражения и коммуникативные практики, которые не всегда своевременно фиксируются в существующих базах данных [3]. В-третьих, в некоторых корпусах наблюдается недостаточное представление разговорной речи, особенно спонтанной коммуникации, что ограничивает анализ живого языка в реальных ситуациях общения [2]. С развитием ИИ и цифровых корпусов возникает ряд этических проблем. Одной из них является вопрос авторских прав на тексты, используемые для обучения моделей. Не всегда ясно, каким образом и с согласия ли авторов используются текстовые данные. Другой проблемой является использование данных пользователей, особенно в онлайн-системах обработки языка, где информация может применяться для дальнейшего обучения моделей. Кроме того, важным аспектом остаётся прозрачность алгоритмов ИИ, так как многие модели функционируют как

“чёрный ящик”, и пользователю сложно понять, каким образом принимаются те или иные решения [5].

Перспективы развития. Несмотря на существующие ограничения, перспективы развития цифровой лингвистики являются крайне широкими. Одним из ключевых направлений является развитие будущих технологий, направленных на повышение точности и адаптивности языковых моделей. Большое значение имеет развитие адаптивных языковых моделей, способных подстраиваться под индивидуальные особенности пользователя, его уровень языка и контекст общения. Перспективным направлением также является развитие мультимодальных моделей, которые объединяют текст, звук и изображение, обеспечивая более полное понимание языкового и культурного контекста. Кроме того, важной тенденцией является интеграция культурного и контекстуального анализа, что позволит моделям лучше учитывать социокультурные особенности языка, а не только его формальную структуру. В будущем ожидается также значительное улучшение точности анализа языка, что будет достигнуто за счёт более качественных корпусов, развитие нейросетей и усиления взаимодействия между лингвистикой и ИИ.

В заключение, необходимо упомянуть, что проведённое исследование показало, что цифровые корпусы, ИИ и лингвистическое моделирование представляют собой взаимосвязанную систему современных методов анализа языка, которая существенно трансформирует как теоретическую, так и прикладную лингвистику. Использование корпусных ресурсов обеспечивает эмпирическую основу для изучения языковых явлений, позволяя анализировать реальные модели употребления языка, его частотные характеристики, коллокации и грамматические структуры. Технологии ИИ значительно расширяют возможности обработки естественного языка, обеспечивая автоматизацию перевода, генерации текста, классификации и семантического анализа. Взаимодействие ИИ с корпусными данными способствует повышению точности и эффективности лингвистических исследований, а также развитию интеллектуальных систем обработки языка. В то же время выявлены определённые ограничения, связанные с качеством обучающих данных, наличием смещений в моделях, устареванием корпусов и этическими вопросами использования текстовых ресурсов. Эти факторы требуют критического подхода к применению цифровых технологий и их методически обоснованной интеграции в научную и образовательную практику. Таким образом, можно сделать вывод, что цифровые технологии не заменяют традиционную лингвистику, а дополняют её, формируя новую исследовательскую парадигму, основанную на синергии эмпирических данных, вычислительных методов и языкового моделирования. Перспективы дальнейших исследований связаны с развитием адаптивных и мультимодальных моделей, а также с углублением интеграции культурного и контекстуального анализа в системы искусственного интеллекта.

Список литературы

1. Чилингарян К. А. (2020). Корпусная лингвистика: теория и методология. Вестник РУДН. Серия: Теория языка. Семиотика. Семантика.
2. Потапова, Р. К. (2012). Дискурсивная составляющая современной корпусной лингвистики. Вестник МГЛУ.
3. Зубов, А. В. (2006). Корпусная лингвистика: возможности и перспективы. В Русский язык: система и функционирование. Минск.
4. Национальный корпус русского языка. http://www.ruscorpora.ru

5. Обработка естественного языка. (б. д.). В Википедия. https://ru.wikipedia.org/wiki/Обработка_естественного_языка
6. Корпус текстов. (б. д.). В Википедия. https://ru.wikipedia.org/wiki/Корпус_текстов
7. СинТагРус. (б. д.). В Википедия. https://ru.wikipedia.org/wiki/СинТагРус