Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО

Кафедра теории, истории языка и прикладной лингвистики

Лингвистические особенности учебно-научных текстов, сгенерированных ИИ

АВТОРЕФЕРАТ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ БАКАЛАВРА

студентки 4 курса 441 группы
направления 45.03.03 «Фундаментальная и прикладная лингвистика»
Института филологии и журналистики
Рединой Екатерины Алексеевны

Научный руководитель
к.ф.н., доцент _____ Е.В. Старостина

Зав. кафедрой
д.ф.н., профессор _____ О.Ю. Крючкова

Саратов 2025

Актуальность исследования обусловлена стремительным развитием генеративных языковых моделей (ChatGPT, Gemini, DeepSeek и др.) и их активным проникновением в академическую сферу. Возникает острая необходимость в выявлении лингвистических маркеров, позволяющих отличать ИИ-тексты от человеческих, особенно в жанре учебно-научного стиля, для обеспечения академической добросовестности, оценки качества ИИ-генерации и совершенствования самих алгоритмов.

Объект исследования: учебно-научные тексты, сгенерированные искусственным интеллектом.

Предмет исследования: лингвистические особенности (морфологические, синтаксические, лексические, структурные) текстов, созданных ИИ, в сопоставлении с текстами человека; методы их идентификации.

Гипотеза исследования: Несмотря на высокий уровень имитации, тексты, сгенерированные ИИ в учебно-научном стиле, обладают устойчивыми лингвистическими особенностями (маркерами), отличающими их от текстов, написанных человеком, и выявление этих маркеров необходимо для разработки эффективных методов детекции и улучшения генеративных моделей.

Цель работы: Выявить и систематизировать лингвистические различия между учебно-научными текстами, написанными людьми и сгенерированными ИИ, и оценить их значимость для идентификации авторства.

Задачи исследования:

1. Провести теоретический анализ современных исследований в области компьютерной лингвистики, генерации текста ИИ и методов атрибуции авторства.
2. Выявить и систематизировать лингвистические характеристики текстов, сгенерированных ИИ, описанные в научной литературе.
3. Сформировать корпус текстов для анализа: 25 учебно-научных текстов, написанных человеком (из учебных пособий), и 25 аналогичных по тематике и объему текстов, сгенерированных различными ИИ-моделями (ChatGPT-3.5, Gemini, DeepSeek).

4. Провести сопоставительный лингвистический анализ корпуса с использованием:

   количественных методов: автоматизированный анализ метрик читаемости (Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, Dale Chall Readability Score и др.) с помощью платформы Open Brain AI.

   лексического анализа: оценка лексического разнообразия (Type Token Ratio) и частотности слов с помощью сервиса Textometr.ru.

   качественного анализа: выявление структурных, стилистических и содержательных паттернов.
5. Провести лингвистический эксперимент (онлайн-опрос 55 респондентов) по идентификации происхождения текстов (человек/ИИ) с последующим анализом аргументации респондентов для выявления субъективно воспринимаемых маркеров.
6. Сравнить результаты теоретического анализа, автоматизированной обработки текстов и эксперимента с респондентами.
7. Систематизировать выявленные лингвистические особенности ИИ-текстов и оценить их значимость как маркеров для идентификации.
8. Сформулировать выводы о возможностях и ограничениях ИИ в имитации учебно-научного стиля и предложить направления для дальнейших исследований и совершенствования генеративных моделей.

Материал исследования: корпус из 50 текстов учебно-научного стиля (25 человеческих, 25 ИИ-генераций) на естественнонаучные и технические темы; результаты автоматизированного анализа метрик Open Brain AI и Textometr.ru; данные онлайн-опроса 55 респондентов.

Методы исследования: отбор и формирование корпуса текстов, генерация текстов ИИ по заданным параметрам, количественный анализ (статистическая обработка метрик читаемости, лексического разнообразия, частотности), качественный анализ (описание структурных, стилистических, содержательных особенностей), сопоставительный анализ; проведение онлайн-опроса с

последующей интерпретацией результатов и комментариев респондентов; использование платформ Open Brain AI, Textometr.ru, онлайн-сервисов для анализа результатов опроса.

Структура работы. Работа состоит из введения, двух глав, заключения, списка использованной литературы и приложений.

Основное содержание работы. Первая глава содержит теоретическую информацию по теме исследования и состоит из 4 разделов.

Раздел 1.1 «Влияние информационных технологий на развитие современной лингвистики» состоит из 5 параграфов. Параграф 1.1.1. «Возможности использования искусственного интеллекта в лингвистических исследованиях и обработке естественного языка» содержит информацию о том, что современные информационные технологии и искусственный интеллект коренным образом изменили лингвистические исследования, открыв новые возможности для анализа языка. В корпусной лингвистике ИИ позволяет обрабатывать огромные массивы текстовых данных, выявляя статистические закономерности и тенденции развития языка. Технологии обработки естественного языка (NLP) значительно улучшили машинный перевод, научившись учитывать контекст, идиомы и сленг. Современные системы могут не только переводить, но и генерировать тексты, анализировать их стилистические особенности и эмоциональную окраску. В лексикографии ИИ используется для создания "живых" электронных словарей, которые автоматически обновляются по мере изменения языка. Особую ценность представляет возможность количественного анализа языковых явлений - с помощью ИИ исследователи могут обрабатывать такие объемы данных и выявлять такие закономерности, которые были недоступны при традиционных методах анализа.

В параграфе 1.1.2. «Компьютерная лингвистика: история развития, перспективы и направления исследований» рассматривается становление и современное состояние этой междисциплинарной области. Компьютерная лингвистика, особенно направление NLP (обработка естественного языка), занимается созданием алгоритмов для взаимодействия компьютеров с

человеческим языком, решая практические задачи в отличие от теоретической лингвистики.

В параграфе 1.1.3 «Роль нейронных сетей и ИИ в генерации текстов» рассматривается стремительное развитие технологий искусственного интеллекта и нейронных сетей в области генерации текстового контента. Современные системы, такие как ChatGPT, демонстрируют впечатляющие возможности по созданию связных и логически структурированных текстов, включая научные работы. Однако массовое распространение ИИ-генераторов текста породило серьезные вопросы контроля качества и аутентичности контента.

В параграфе 1.1.4 «Персонализация и практическое применение компьютерной лингвистики в различных сферах» рассматривается персонализация и практическое применение компьютерной лингвистики в современных технологиях. Ключевым аспектом является адаптация языковых моделей под индивидуальные потребности пользователей, что особенно важно в условиях информационной перегрузки. В образовательной сфере это проявляется в системах персонализированного обучения, где контент динамически подстраивается под уровень и особенности ученика.

Параграф 1.1.5 «Технологические, теоретические и этические вызовы в компьютерной лингвистике» посвящен вызовам компьютерной лингвистики. Среди технологических проблем выделяются: многозначность языковых единиц, необходимость больших объемов, обучающих данных, высокие требования к вычислительным ресурсам, динамическая природа языка

Авторы подчеркивают необходимость баланса между технологическим прогрессом и этическими нормами, что особенно актуально в контексте растущего влияния ИИ на коммуникационные процессы. Компьютерная лингвистика, несмотря на существующие ограничения, продолжает трансформировать способы нашего взаимодействия с информацией и друг с другом.

Раздел 2 состоит из четырех параграфов. Параграф 1.2.1 «Развитие методов обработки естественного языка: от статистических моделей к нейросетевым технологиям» посвящен эволюции подходов к анализу текстов. Современные методы детекции ИИ-генерации сочетают статистический анализ языковых характеристик (частотность слов, длина предложений) с алгоритмами машинного обучения и семантическим анализом. Исторически развитие прошло три этапа: в 1970-х доминировали rule-based системы (SYSTRAN), в 1990-х произошел переход к статистическим методам (скрытые марковские модели), а современный этап характеризуется доминированием нейросетевых технологий (GPT, трансформеры). Особое внимание уделяется мультимодальным моделям типа CLIP, способным анализировать текст в сочетании с визуальными и аудиоданными, что открывает новые возможности для контент-анализа.

Параграф 1.2.2 «Методы идентификации авторства текстов» рассматривает концепцию «авторского инварианта» как ключевого инструмента атрибуции текстов. В фокусе анализа три базовых характеристики: массовость (неконтролируемые автором параметры), устойчивость (стабильность параметров у одного автора) и различающая способность (вариативность между разными авторами). Особую актуальность приобретает проблема различения человеческих и ИИ-текстов, для решения которой предлагается четырехэтапная методология: 1) выявление уникальных параметров ИИ-генерации, 2) формирование специфического «авторского инварианта», 3) оценка его эффективности, 4) анализ гибридных текстов. Подчеркивается важность объема текста для достоверного анализа.

Параграф 1.2.3 «Экспериментальный анализ текстов, подвергшихся синонимизации» представляет результаты сравнительного исследования естественных и искусственно преобразованных текстов. Эксперимент с использованием программ SyMonym и Article Clone Easy выявил значительные различия в степени уникальности (22,03% против 68,71%) и текстовых параметрах: увеличение средней длины слов при сокращении количества

предложений и служебных слов. Анализ показал, что наиболее чувствительными к синонимизации оказались параметры частоты служебных слов и количества коротких слов, тогда как лексическое разнообразие изменялось минимально. Результаты демонстрируют принципиальные различия между человеческими и машинно-обработанными текстами.

Параграф 1.2.4 «Автоматическое извлечение лингвистических характеристик» описывает инновационные инструменты лингвистического анализа на базе ИИ, такие как платформа Open Brain AI. Эти решения автоматизируют традиционно трудоемкие процессы анализа по шести ключевым направлениям: фонология (анализ слоговой структуры), морфология (распределение частей речи), синтаксис (структура предложений), лексика (индекс разнообразия), семантика (смысловые единицы) и читаемость (грамматическая сложность). Особое значение такие системы приобретают в клинической лингвистике и образовании, где позволяют проводить объективную оценку речевых особенностей и выявлять отклонения в развитии. Подчеркивается потенциал автоматизированного анализа для масштабных лингвистических исследований.

Раздел 3 «Особенности текстов, сгенерированных ИИ» раскрывает ключевые характеристики искусственно созданных текстов. Современные языковые модели типа ChatGPT, основанные на архитектуре GPT-3.5 с 175 миллионами параметров, демонстрируют впечатляющую способность имитировать человеческую речь. Однако сравнительные исследования выявляют существенные различия: ИИ-тексты отличаются структурной униформностью, шаблонными вступлениями/заключениями, частым повторением ключевых слов из заголовка и поверхностным раскрытием темы. Эксперименты с академическими эссе показали, что человеческие тексты обладают большей вариативностью и глубиной, тогда как ИИ-генерации следуют жестким шаблонам (например, конструкции "Название. Развернутое определение"). Важной проблемой остается отсутствие у ИИ способности использовать контекст для разрешения двусмысленностей и предпочтения более коротких

слов для менее информативного контента, что характерно для человеческой речи.

В разделе 4 кратко подводится итог теоретической главы.

Во второй главе «Сопоставительное и экспериментальное изучение учебно-научных текстов, написанных человеком и сгенерированных ИИ» приводится описание эксперимента и анализ его результатов. Она состоит из 4 разделов.

Параграф 2.1.1 «Методика проведения исследования» описывает первый этап исследования, в котором для сопоставительного изучения текстов, написанных профессиональными преподавателями из учебного пособия по русскому языку для студентов технических специальностей, и текстов, сгенерированных нейросетями Chat-3.5, ChatGPT и DeepSeek с использованием той же тематики и стиля, было использовано по 25 текстовых образцов. Далее тексты были добавлены в онлайн-приложение Open Brain AI для извлечения показателей читаемости, морфологических, синтаксических и лексических характеристик, при этом для анализа были выбраны только параметры, касающиеся количества лингвистических компонентов и значительного присутствия в текстах. Параграф «Многоаспектный анализ текстовых характеристик: лингвистические метрики в сравнении человеческих и ИИ-генераций» объясняет, что из-за комплексности традиционных показателей читаемости, таких как индекс удобочитаемости Флеша и индекс туманности Ганнинга, которые учитывают как морфологические, так и синтаксические компоненты, детальное рассмотрение отдельных лексических и синтаксических характеристик может оказаться избыточным, поэтому исследование сосредотачивается на непосредственном сравнении текстов с использованием существующих метрик читаемости.

В параграфе 2.1.2, посвященном описанию отличительных признаков текстов, сгенерированных искусственным интеллектом, отмечается, что такие тексты характеризуются структурной униформностью (шаблонное начало и конец, предсказуемая композиция), повторяющимися словосочетаниями и отсутствием синтаксического варьирования на лексическом уровне, поверхностным

анализом и слабой связью между фрагментами на содержательном уровне, более частым использованием слов из заголовка и низким лексическим разнообразием, а также отсутствием индивидуального стиля и эмоциональной окраски, что делает их шаблонными и облегчает выявление ИИ-генераций.

В параграфе 2.1.3 представлен сравнительный анализ 25 текстов, созданных человеком, и 25 текстов, сгенерированных ИИ, с использованием различных метрик читаемости, таких как Flesch Reading Ease, Flesch-Kincaid Grade Level, и других. Тексты ИИ длиннее и имеют более сложную структуру предложений. В целом, тексты ИИ близки к текстам человека.

Раздел 2 «Экспериментальное исследование идентификации ИИ-текстов: анализ лингвистических маркеров». В параграфе 2.2.1, посвященном методике организации эксперимента и анализу демографических характеристик респондентов, описывается онлайн-опрос, в котором 55 участникам предлагалось определить происхождение текстовых материалов (человек/ИИ), аргументируя свои решения в комментариях, при этом опрос проводился анонимно. Основную группу респондентов составили лица в возрасте 18-25 лет (80% выборки), с преобладанием мужских ответов (60%), а также отмечается минимальное представительство старших возрастных групп.

В параграфе 2.2.2, посвященном общему анализу результатов опроса, сообщается, что в первой части, где было представлено два текста ИИ, доля правильных ответов составила 9,1%. Респонденты отмечали шаблонность изложения, формальные клишированные структуры, грамматические и синтаксические аномалии, злоупотребление вводными конструкциями, использование узкоспециальных терминов без пояснений, дисбаланс между обобщенностью и детализацией, а также отсутствие элементов человеческой речи. Во второй части, с двумя человеческими текстами, доля правильных ответов выросла до 46,3%. В третьей части, где был представлен один текст ИИ, доля верных ответов составила 47,3%. Рассматривается, как респонденты выявляли сгенерированный текст, выделяя особенности композиционной организации (стандартизированные заголовки, шаблонные заключения типа

"Таким образом", жесткая структурированность, дисбаланс в детализации), стилистико-языковые характеристики (отсутствие местоимений первого лица, комбинация формализации с внезапными стилистическими сбоями, неестественные синтаксические конструкции, канцелярские обороты) и содержательные аспекты (механическое соединение информации, противоречие между целевой аудиторией и сложностью материала). Подчеркивается, что грань между машинной и человеческой генерацией размывается по мере совершенствования языковых моделей, требуя разработки более тонких критериев идентификации.

В параграфе 2.2.3, посвященному анализу ответов респондентов о текстах ИИ, отмечается, что респонденты определяли тексты ИИ по следующим признакам: избыточная структурированность (одинаковая длина абзацев, шаблонные заголовки, финальные абзацы с "Таким образом", маркированные списки), стилистические особенности (чрезмерная формальность, обилие вводных конструкций, отсутствие индивидуального голоса), лексико-синтаксические особенности (неестественные фразы, шаблоны с "таких как", избыточная точность грамматики), а также информационное наполнение ("слишком заумные" описания, отсутствие примеров и контекста, шаблонные саммари).

В параграфе 2.2.4, посвященном анализу ответов респондентов о текстах человека, отмечается, что респонденты определяли тексты, написанные человеком, опираясь на стиль изложения (использование личных местоимений, обращение к читателю, риторические вопросы), речевые обороты (метафоры, непринужденные фразы, "шероховатости"), структуру подачи материала (менее формализованная, творческие примеры), эмоциональную окраску и юмор, а также интуитивное восприятие.

В параграфе 2.2.5, подводящем итоговый анализ результатов эксперимента, отмечается, что респонденты отличали ИИ-тексты по шаблонной структуре, формальному стилю, языковым аномалиям и информационной перегруженности, а человеческие тексты – по живому стилю, творческой

неоднородности и интуитивной естественности. При этом констатируется, что граница между ИИ и человеческими текстами размыта.

В разделе 2.3, посвященном сравнению характеристик ИИ-текстов, описанных в научной литературе и выявленных в ходе эксперимента, отмечается, что целью исследования было определение степени соответствия между теоретическими моделями из академических источников и практическими наблюдениями пользователей. Был проведен компаративный анализ признаков текстов ИИ, выявленных в ходе теоретического обзора научной литературы и эмпирического эксперимента с респондентами. Анализ ответов респондентов выявил, что основные маркеры ИИ-текстов включают жесткую структурированность, безличный стиль, языковые особенности (кальки с английского, шаблоны), перегруженность терминами, а человеческие тексты отличаются использованием личных местоимений, наличием "шероховатостей" языка и образных выражений, менее формальной структурой, а также интуитивным восприятием. Подчеркивается, что граница между ИИ и человеческими текстами становится менее очевидной.

В последнем разделе подводятся итоги анализа материала, обобщаются полученные данные. Так, нами были выявлены устойчивые лингвистические маркеры, позволяющие отличать ИИ-тексты (жесткая структура, формальный стиль, неестественные синтаксические конструкции, перегруженность терминами, отсутствие эмоций) от человеческих текстов (живая вариативность, эмоциональные оценки, метафоры, естественные "шероховатости"). Отмечено, что ИИ-модели успешно имитируют научный стиль, но сохраняют алгоритмические паттерны, и граница между искусственными и человеческими текстами размывается. Результаты имеют практическое значение для совершенствования генеративных моделей и проверки академической добросовестности.

Наконец, в **заключении** подводятся итоги исследования. Исследование выявило различия между текстами, созданными человеком, и сгенерированными искусственным интеллектом. Анализ показал, что ИИ-

тексты имеют маркеры, которые помогают их идентифицировать. Эксперименты подтвердили структурную шаблонность и клишированные формулировки ИИ-текстов, а также их отсутствие синтаксического разнообразия и глубины анализа темы.

Квантитативные методы доказали, что, несмотря на формальную правильность, ИИ-тексты уступают человеческим по естественности. Опрос респондентов выявил признаки, указывающие на искусственность ИИ-текстов, такие как избыточная формальность и отсутствие эмоций.

Перспективы дальнейших исследований связаны с разработкой новых методов для более точной идентификации ИИ-генераций. Практическая значимость работы заключается в её применении для обеспечения академической добросовестности и актуализации технологий обработки языка.

Таким образом, исследование вносит вклад в понимание лингвистических особенностей ИИ-текстов и предлагает инструментарий для их анализа, что актуально как в теории, так и на практике.