DOI: 10.18137/RNU.V925X.25.03.P.064 УДК 81-115

Фигура Елена Владимировна
ассистент кафедры теоретической и прикладной лингвистики, Байкальский государственный университет, город Иркутск. ORCID: 0009-0007-4008-7998, SPIN-код: 9268-3969, AuthorID: 1215416
Электронный адрес: elena.figura.00@mail.ru
Elena V. Figura
Assistant at the Department of Theoretical and Applied Linguistics, Baikal State University, Irkutsk. ORCID: 0009-0007-4008-7998, SPIN-code: 9268-3969, AuthorID: 1215416
E-mail address: elena.figura.00@mail.ru

ГЕНЕРАЦИЯ ТЕКСТОВ КАК ОБЪЕКТ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

Аннотация. Статья посвящена обзору исследований в области генерации текстов искусственным интеллектом. Актуальность темы обусловлена широким применением нейросетей в медиа, образовании и других сферах, что требует систематического изучения создаваемых текстов. *Цель* работы – охарактеризовать основные направления исследований сгенерированных текстов, включая технологические, этические и собственно лингвистические направления. Используемые *методы* включают описательный, сопоставительный и контекстуальный анализ. В *результате* изучения существующих исследований выявлены наиболее значимые аспекты изучения: обучение языковых моделей, обработка естественного языка, правовое регулирование в области института авторского права и исключение культурных искажений. Особое внимание уделяется лингвистическим аспектам, а также возможностям практического применения сгенерированных текстов. *Научная новизна* заключается в комплексном подходе к исследованию феномена генерации текстов, сочетающем лингвистический, технологический и этический аспекты. Практическая значимость работы заключается в определении направлений дальнейших исследований, для лингвистического мониторинга одной из способностей нейросетей в виде генерации текста.

Ключевые слова: лингвистика, искусственный интеллект, генерация текста, языковая модель, нейросеть.

Для цитирования: *Фигура Е.В.* Генерация текстов как объект лингвистических исследований // Вестник Российского нового университета. Серия: Человек в современном мире. 2025. № 3. С. 64–70. DOI: 10.18137/RNU.V925X.25.03.P.064

TEXT GENERATION AS AN OBJECT OF LINGUISTIC RESEARCH

Abstract. The article is an overview of research in the field of text generation by artificial intelligence. The relevance of the topic is due to the wide application of neural networks in media, education and other spheres, which requires a systematic study of generated texts. The *aim* of the paper is to characterise the main directions of research of generated texts, including technological, ethical and linguistic directions. The *methods* used include descriptive, comparative and contextual analyses. As a result of the study of existing research, the most significant aspects are identified: training of language models, natural language processing, legal regulation of authorship of texts generated by artificial intelligence and elimination of cultural distortions. Special attention is paid to linguistic aspects, as well as the possibilities of practical application of generated texts. The *scientific novelty* consists in a comprehensive approach to the study of the phenomenon of text generation, combining linguistic, technological and ethical perspectives. The *practical significance* of the work is to identify directions for further research, for linguistic monitoring of one of the abilities of neural networks in the form of text generation.

Keywords: linguistics, artificial intelligence, text generation, language model, neural network.
For citation: Figura E.V. (2025) Text generation as an object of linguistic research. *Vestnik of Russian New University. Series: Man in the Modern World.* No. 3. Pp. 64–70. DOI: 10.18137/RNU.V925X.25.03.P.064 (In Russian).

За последние годы развитие искусственного интеллекта (далее – ИИ) вызвало большой интерес у исследователей, в частности, одной из важнейших областей исследования стала генерация текстов. Эта технология интересует специалистов уже давно. Еще в 1950-х годах Алан Тьюринг предложил эмпирический тест, состоящий из трех вопросов, с целью определить, умеет ли компьютер мыслить. Один из запросов заключался в том, чтобы сгенерировать стихотворение [1]. Сейчас ChatGPT легко сможет решить эту задачу и, возможно, справится лучше среднестатистического человека.

Так, в 2025 году итальянская газета *Il Foglio* опубликовала номер, который был полностью сгенерирован ИИ[1]. По словам главного редактора, цель такого эксперимента заключалась в том, чтобы показать изменения, происходящие в журналистике по причине развития искусственного интеллекта. Среди других изданий, использующих искусственный интеллект для подготовки статей, можно выделить агентство *Associated Press*[2], которое использует нейросеть для генерации идей или заголовков по теме написанной статьи[3]. Таким образом, генерация текстов при помощи искусственного интеллекта имеет очень много перспектив, следовательно, можно обусловить важность изучения данной области ИИ.

Изучение искусственного интеллекта тесно связано с наукой о языке. В настоящее время исследования охватывают технологические, собственно лингвистические и даже этические вопросы. Рассмотрим их далее.

Технологические основы генерации текста в основном включают рассмотрение вопросов обучения языковых моделей (LLM) и обработки естественного языка (NLP) с целью сделать тексты, созданные ИИ, качественными и похожими на те, которые мог бы написать человек. Современные технологии генерации текстов основаны на архитектуре Transformer, глубокой нейросетевой языковой модели, которая, в отличие от предыдущих моделей, использует методы кодирования и декодирования. При кодировании решается поставленная задача, при этом используется механизм, позволяющий анализировать связь между словами во всем предложении; при декодировании на основе проведенного анализа формируется сгенерированный текст [2]. Исследования в области обработки естественного языка стремятся улучшить понимание семантики предложений, связи между словами, имеющими один и тот же объект (coreference resolution), а также

[1] *Giuffrida A.* Italian newspaper says it has published world's first AI-generated edition // The Guardian. 2025. March 8. URL: https://www.theguardian.com/technology/2025/mar/18/italian-newspaper-says-it-has-published-worlds-first-ai-generated-edition (дата обращения: 28.04.2025).
[2] *David E.* Associated Press outlines AI guidelines for journalists // The Verge. 2023. August 16. URL: https://www.theverge.com/2023/8/16/23834586/associated-press-ai-guidelines-journalists-openai (дата обращения: 28.04.2025).
[3] Associated Press разработало руководство для журналистов по применению ИИ // ТАСС. 2023. 17 августа. URL: https://tass.ru/obschestvo/18528009 (дата обращения: 28.04.2025).

Вестник Российского нового университета
Серия: Человек в современном мире. 2025. № 3

решить проблему полисемии [3; 4]. Другой вектор исследований направлен на создание предобученных моделей языка (pretraining). Подразумевается, что такие языковые модели уже обладают определенными навыками и знаниями, в том числе о семантике слов и синтаксисе, и в дальнейшем их можно настроить для решения более узких задач (transfer learning) [5].

Этический аспект исследования генерации текстов основывается на определении авторского права текста, написанного нейросетью. На сегодняшний день в российской науке и правоприменительной практике единое мнение о том, кому принадлежит написанный нейросетью текст, отсутствует. Среди существующих теорий можно выделить следующие: право на произведение принадлежит самой нейросети, право на произведение принадлежит пользователю нейросети, автором сгенерированного произведения является владелец нейросети, сгенерированный текст является общественным достоянием [6]. Проблемной областью генерации текстов также являются культурные искажения. Например, результаты исследования 2024 года, посвященного анализу генерируемых текстов на тему политического статуса Тайваня, показали, что нейросеть чаще предлагала тексты с маркерами, выражающими суверенитет острова Тайвань, чем его принадлежность к Китайской Народной Республике [7]. Следовательно, учитывая сложную политическую ситуацию в этом регионе, с точки зрения Китая, это является искажением информации и может спровоцировать цифровой конфликт. Данный пример свидетельствует о необходимости пересмотра данных, используемых для обучения нейросетей.

В рамках данной статьи наиболее важным является собственно лингвистический аспект исследований в области генерации текста. В настоящее время ученые пытаются выявить сходства и различия текстов, сгенерированных ИИ, и текстов, написанных человеком. В подобных исследованиях делается вывод, что нейросети не обладают той же креативностью, что реальный человек, а также не способны сознательно использовать языковые средства в текстах определенного жанра [8; 9]. Ученые считают, что сгенерированные тексты довольно просты по своей структуре и использованным лексическим средствам, в них отсутствуют сложные синтаксические конструкции, метафоры, эпитеты и другие средства художественной выразительности, присутствует шаблонность и клишированность формулировок, повторяются отдельные части текста [10; 11]. Исследования с целью выявить типовые различия естественных и сгенерированных текстов показывают, что распределение ключевых слов в сгенерированных текстах неравномерно и что в них не раскрывается проблематика в полной мере, а лишь содержится общая информация [12; 13].

Существуют также узконаправленные исследования, например, проводится анализ понимания человеком и искусственным интеллектом контекстуальных значений лексических единиц; ставится цель сгенерировать тексты юмористического характера и изучить использованные в текстах лексические приемы [14; 15]. Несмотря на активное обучение языковых моделей выразительным средствам, на момент проведения исследований нейросети не способны в полной мере понимать и воссоздавать коннотативные оттенки и грамотно использовать их в процессе генерации текстов. Этот вопрос относится к областям литературоведения, лингвостилистики и смежным с ними областями риторики и жанроведения.

Большое внимание также уделяется сферам применения сгенерированных текстов. Генерация текстов в основном применяется в области медиа и в сфере образования. Использование инструментов искусственного интеллекта для написания текстов имеет достаточное количество преимуществ, среди которых экономия времени, помощь в генерации идей, структурирование материала, но в то же время формирует необходимость создания дополнительных требований к оригинальности текстов и их более тщательному анализу с целью исключить так называемые «галлюцинации нейросетей», использование ложной информации, упоминание несуществующих ученых, цитат и экспериментов [16; 17]. Данное обстоятельство следует отнести к факторам, в значительной степени влияющим на качество текстового продукта.

Несмотря на результаты исследований, которые показывают, что возможности ИИ намного ниже, чем таковых у реального автора, нельзя отрицать, что нейросети развиваются весьма стремительно и, возможно, в обозримом будущем они смогут если не превзойти, то точно приблизиться к человеческим способностям. В качестве иллюстрации этого тезиса приведем следующий пример.

В 2015 году языковую модель Seq2seq, которая применялась в основном для машинного перевода, решили использовать для синтеза текстов. Поскольку модель была полностью обучена на произведениях У. Шекспира, в качестве текстового материала был сгенерирован белый стих (стих, не имеющий рифмы, но обладающий размером):

Capulet:
No, good sir,
And make a pen and meeding down, trouble me
Here provide it: your breast of any other use
Wearing behind the old looks of a man begin:
But, for your city, as the cases of the art
I have held to go to you [1, с. 83].

Данный фрагмент не соотносится с историей, описанной в «Ромео и Джульетте», в нем отсутствует связность между элементами, и больше это напоминает простой набор слов.

Теперь сравним стих, приведенный выше, со сгенерированным современным GPT-4, который является примером нейросети с трансформерной архитектурой («Т» в GPT обозначает «transformer»):

Capulet:
No, good sir,
Your words, though sweet, have not found favour here.
In matters of my house, I hold the sway,
Decisions made are mine, and mine to bear.
Your counsel, though well meant, I shan't adhere [1, с. 84].

Стих, сгенерированный GPT-4, имеет структуру, схожую с оригинальным произведением У. Шекспира, и, несмотря на то, что подлинность текста установить несложно (это НЕ трагедия У. Шекспира), данный фрагмент имеет свойственный драматургу ритм и рифму, а семантика фрагмента отражает характер сеньора Капулетти [1]. Очевидны значительные различия в текстах, сгенерированных десять лет назад, и тех, которые можно создать с помощью ресурсов ИИ сейчас.

Таким образом, способность нейросетей генерировать тексты представляет собой динамичный процесс, который требует пристального внимания со стороны языковедов. Безусловно, с каждым годом открывается все больше областей для применения данного инструмента ИИ. Учитывая сказанное, отметим, что лингвистика обладает значительным научным потенциалом, который позволяет как фиксировать возникающие проблемы в плоскости оппозиции «язык программирования vs естественный язык», так и адаптировать имеющиеся наработки к возникающим по-

Вестник Российского нового университета
Серия: Человек в современном мире. 2025. № 3

требностям. Характеризуя процесс генерации текстов как объект лингвистических исследований, выскажем мнение о необходимости разработки внутридисциплинарного мониторинга, который бы учитывал телеологию изучения этого объекта, а также практические инструменты в области генерации текстов.

Те области, которые мы эскизно очертили в настоящей статье, можно отнести к таким дисциплинам, как стилистика, текстология, риторика, теория жанров, литературоведение. Каждая из этих дисциплин обладает сложившимся инвентарем методов. Например, аналитические компоненты «точка зрения», «фокальный персонаж» с одной стороны откроют доступ к тому, кто ведет повествование, с другой – расширят репертуар текстового продукта ИИ. Стилистика и, в частности, метафорология дадут доступ к пониманию интерпретации ИИ метафоричных текстов. Так, можно обучить языковую модель схемам концептуальной интеграции для прогнозирования стратегий интерпретации и развертывания метафоры. Работу можно построить, используя дискурсивный анализ, актуально провести сравнительно-сопоставительный анализ используемых в текстах ИИ лексико-семантических и грамматических средств, фреймов, маркеров, а также изучить соответствие параметрам определенных видов дискурса.

Литература

1. *Summerfield C.* These Strange New Minds: How AI Learned to Talk and What It Means. New York : Viking, 2025. 384 p. ISBN 0593831713.
2. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., et al.* Attention is All You Need // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. P. 6000–6010. DOI: https://doi.org/10.48550/arXiv.1706.03762
3. *Боярский К.К.* Введение в компьютерную лингвистику : учеб. пособие. СПб. : НИУ ИТМО, 2013. 72 с. URL: https://books.ifmo.ru/file/pdf/1470.pdf (дата обращения: 30.04.2025).
4. Использование промежуточных языков представления для упрощения процесса перевода естественного языка в запросы к базе данных / М.А. Сельмах, В.Г. Миснянкин, А.Ю. Кунац, А.В. Костина // Наука настоящего и будущего. 2017. Т. 1. С. 114–116. EDN ZDQVHR
5. *Raffel C., Shazeer N., Roberts A., et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research. 2020. DOI: https://doi.org/10.48550/arXiv.1910.10683
6. *Драгунова С.А.* Проблемы правового регулирования авторских прав на произведения, созданные нейросетью // Вестник Воронежского государственного университета. Серия: Право. 2024. № 1 (56). С. 104–110. DOI: 10.17308/law/1995-5502/2024/1/104-110. EDN BXUKGZ.
7. *Петухов А.Ю., Каминченко Д.И.* Анализ текстов о спорном статусе Тайваня, созданных языковой моделью ChatGPT // Вестник Российского университета дружбы народов. Серия: Литературоведение. Журналистика. 2024. Т. 29. № 3. С. 593–611. DOI :10.22363/2312-9220-2024-29-3-593-611. EDN GICACA.
8. *Henrickson L.* Computer-generated fiction in a literary lineage: Breaking the hermeneutic contract // Logos. 2018. Vol. 29. № 2/3. Pp. 54–63. DOI: 10.1163/18784712-02902007

9. *Bringsjord S., Ferrucci D.* Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine. Hillsdale, NJ : Lawrence Erlbaum, 2000. 262 p. ISBN 0805819878.
10. *Jurafsky D., Martin J.H.* Speech and Language Processing. 3rd edition. Prentice Hall, 2023. 636 p.
11. *Калдыбекова Н.Б.* Искусственный интеллект в лингвистике: влияние ChatGPT и нейросетей на письменную речь и стилистику // Endless Light in Science. 2025. № 28. С. 328–331.
12. *Cohen A., Mantegna R., Havlin S.* Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts // Fractals. 2011. Vol. 5. No. 1. Pp. 1–19. DOI: https://doi.org/10.1142/S0218348X97000103
13. *Тельпов Р.Е., Ларцина С.В.* Типовые различия естественных и сгенерированных нейронной сетью текстов в квантитативном аспекте // Научный диалог. 2023. Т. 12, № 7. С. 47–65. DOI: 10.24224/2227-1295-2023-12-7-47-65. EDN ZZRRKN.
14. *Лихачёв Э.В.* Проблема конструирования юмористических высказываний генеративным искусственным интеллектом (на материале немецкого языка) // Ученые записки Крымского федерального университета имени В.И. Вернадского. Филологические науки. 2023. Т. 9. № 2. С. 92–100. EDN LONKBD
15. *Пустоведова В.А., Быкова Н.О., Тупикова С.Е.* Между строк: понимание контекстуальных окказиональных значений человеком и искусственным интеллектом // Universum: филология и искусствоведение. 2025. № 1 (127). С. 45–48. DOI: 10.32743/UniPhil.2025.127.1.19079. EDN EPADMH.
16. *Горина Е.В., Уфимцева С.М.* Особенности использования текстов нейросетей в медиа и образовании // Russian Linguistic Bulletin. 2024. № 1 (49). DOI: 10.18454/RULB.2024.49.27. EDN VEPNLU.
17. *Черкасова М.Н., Тактарова А.В.* Искусственно сгенерированный академический текст (лингвопрагматический аспект) // Филологические науки. Вопросы теории и практики. 2024. Т. 17. № 7. С. 2551–2557. DOI: 10.30853/phil20240363. EDN YYWRDN.

References

1. Summerfield C. (2025) *These Strange New Minds: How AI Learned to Talk and What It Means.* New York : Viking. 384 p. ISBN 0593831713.
2. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., et al. (2017) Attention is All You Need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems.* Pp. 6000–6010. DOI: https://doi.org/10.48550/arXiv.1706.03762
3. Boyarskiy K.K. (2013) *Vvedenie v komp'yuternuyu lingvistiku* [Introduction to Computational Linguistics]: A Textbook. St. Petersburg : ITMO University Publ. 72 p. URL: https://books.ifmo.ru/file/pdf/1470.pdf (accessed: 30.04.2025). (In Russian).
4. Selmakh M.A., Misnyankin V.G., Kunats A.Yu., Kostina A.V. (2017) Using Intermediate Representation Languages to Simplify the Process of Translating Natural Language into Database Queries. *Nauka nastoyashchego i budushchego* [Science of the Present and Future]. Vol. 1. Pp. 114–116. (In Russian).
5. Raffel C., Shazeer N., Roberts A., et al. (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research.* DOI: https://doi.org/10.48550/arXiv.1910.10683
6. Dragunova S.A. (2024) Problems of Legal Regulation of Copyright for Works Created by Neural Networks. *Proceedings of Voronezh State University. Series: Law.* No. 1 (56). Pp. 104–110. DOI: 10.17308/law/1995-5502/2024/1/104-110 (In Russian).

Вестник Российского нового университета
Серия: Человек в современном мире. 2025. № 3

7. Petukhov A.Yu., Kaminschenko D.I. (2024) Analysis of texts about the disputed status of Taiwan created by the ChatGPT language model. *RUDN Journal of Studies in Literature and Journalism.* Vol. 29. No. 3. Pp. 593–611. DOI :10.22363/2312-9220-2024-29-3-593-611 (In Russian).
8. Henrickson L. (2018) Computer-generated fiction in a literary lineage: breaking the hermeneutic contract. *Logos.* Vol. 29, No. 2/3. Pp. 54–63. DOI: 10.1163/18784712-02902007
9. Bringsjord S., Ferrucci D. (2000) *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine.* Hillsdale, NJ : Lawrence Erlbaum. 262 p. ISBN 0805819878.
10. Jurafsky D., Martin J.H. (2023) *Speech and Language Processing*. 3rd edition. Prentice Hall. 636 p.
11. Kaldybekova N.B. (2025) Artificial intelligence in linguistics: the impact of ChatGPT and neural networks on written speech and stylistics. *Endless Light in Science*. No. 28. Pp. 328–331. (In Russian).
12. Cohen A., Mantegna R., Havlin S. (2011) Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts. *Fractals*. Vol. 5. No. 1. Pp. 1–19. DOI: https://doi.org/10.1142/S0218348X97000103
13. Telpov R.E., Lartsina S.V. (2023) Typological differences between natural and neural network-generated texts in the quantitative aspect. *Nauchnyy dialog* [Scientific Dialogue]. Vol. 12. No. 7. Pp. 47–65. DOI: 10.24224/2227-1295-2023-12-7-47-65 (In Russian).
14. Likhachyov E.V. (2023) The problem of constructing humorous utterances by generative artificial intelligence (in German language). *Scientific Notes of V.I. Vernadsky Crimean Federal University. Philological sciences*. Vol. 9. No. 2. Pp. 92–100. (In Russian).
15. Pustovedova V.A., Bykova N.O., Tupikova S.E. (2025) Between the lines: Understanding contextual occasional meanings by people and artificial intelligence. Universum: Philology and Art History. No. 1 (127). Pp. 45–48. (In Russian).
16. Gorina E.V., Ufimtseva S.M. (2024) Specifics of using neural network texts in media and education. *Russian Linguistic Bulletin*. No. 1 (49). DOI: 10.18454/RULB.2024.49.27 (In Russian).
17. Cherkasova M.N., Taktarova A.V. (2024) Artificially generated academic text (linguopragmatic aspect). *Philology. Theory & Practice.* Vol. 17. No. 7. Pp. 2551–2557. DOI: 10.30853/phil20240363 (In Russian).

Поступила в редакцию: 20.06.2025
Поступила после рецензирования: 02.07.2025
Принята к публикации: 14.07.2025

Received: 20.06.2025
Revised: 02.07.2025
Accepted: 14.07.2025