Обзорная статья
УДК 81’33

Сферы применения корпусных менеджеров в российском и зарубежном научном пространстве

С. А. Романова
Московский государственный лингвистический университет, Москва, Россия
s.a.romanova@linguanet.ru

Аннотация. Статья посвящена обзору сфер применения корпусных менеджеров в российских и зарубежных научных публикациях за последние 10 лет. Установлено, что корпусные менеджеры и программное обеспечение на основе искусственного интеллекта применяются для анализа изменений, происходящих в языке в течение времени (десемантизации, дизамбигуации, транстерминологизации), тематических исследований конкретных типов текстов, составления терминологических словарей и баз данных, обучения иностранному языку.

Ключевые слова: научный дискурс, корпусный менеджер, лингвистический корпус, обучение иностранному языку, дизамбигуация, транстерминологизация

Для цитирования: Романова С. А. Сферы применения корпусных менеджеров в российском и зарубежном научном пространстве // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2025. Вып. 6 (900). С. 105–112.

Review article

Application Areas of Corpus Managers in the Russian and Foreign Scientific Space

Svetlana A. Romanova
Moscow State Linguistic University, Moscow, Russia
s.a.romanova@linguanet.ru

Abstract. The article is devoted to the review of the areas of application of corpus managers in Russian and foreign scientific publications over the past 10 years. It has been established that corpus managers and artificial intelligence-based software are used to analyze changes occurring in a language over time (desemantization, disambiguation, transterminologization), case studies of specific types of texts, compilation of terminological dictionaries and databases, and teaching a foreign language.

Keywords: scientific discourse, corpus manager, linguistic corpus, teaching a foreign language, disambiguation, transterminologization

For citation: Romanova, S. A. (2025). Application areas of corpus managers in the Russian and foreign scientific space. Vestnik of Moscow State Linguistic University. Humanities, 6(900), 105–112. (In Russ.)

ВВЕДЕНИЕ

Сегодня лингвисты, занимающиеся темой корпусной лингвистики, или разрабатывают локально для своих исследовательских задач закрытые корпусные менеджеры и корпусы, или используют корпусные менеджеры (AntConc, Sketch Engine, WordSmith и др.) и корпусы (например, НКРЯ, ГИКРЯ, COCA, FLEURON, CQP и др.), находящиеся в открытом доступе в Интернете. И в том, и в другом случае корпусные менеджеры применяются для лингвистического анализа оперируемых ими корпусов текстов различного жанра и объема, изучения изменений языка в реальной жизни, оценки качества перевода и для обучения студентов иностранным языкам [Gorozhanov, Guseynova, 2020; Палийчук, 2022; Горожанов, Гусейнова, Степанова, 2022; Титова, Игнатова, 2024]. Отдельного интереса заслуживают исторические корпусы [Баранов, 2023].

При этом корпусные менеджеры могут включать в себя технологии искусственного интеллекта, расширяя технические возможности по сбору и частотному анализу корпусов текстов больших объемов, визуализации текстовых данных и постоянному их обновлению [Клочихин, 2024].

Целью исследования является обзор сфер применения корпусных менеджеров в гуманитарных исследованиях на примере современных российских и зарубежных научных публикаций.

В соответствии с поставленной целью сформулированы следующие задачи:

1) отобрать и проанализировать русскоязычные и англоязычные научные публикации о применении корпусных менеджеров в гуманитарных исследованиях (корпусной лингвистике);

2) определить особенности использования корпусных менеджеров в гуманитарных науках.

МАТЕРИАЛ И МЕТОДЫ ИССЛЕДОВАНИЯ

Материалом исследования послужили русскоязычные и англоязычные научные публикации о применении корпусных менеджеров в корпусной лингвистике, изданные за последние 10 лет. Критериями выбора научных публикаций являются следующие: а) опубликованы в открытом доступе в российских рецензируемых научных изданиях; б) опубликованы в открытом доступе в англоязычных рецензируемых научных изданиях по лингвистике (рубрика по классификатору научных направлений OECD «Languages and literature», а также по классификатору ASJC в базе данных Scopus «Arts and Humanities. 1203. Language and Linguistics», «Social Sciences. 3310. Linguistics and Language», «Computer Science. 1702. Artificial Intelligence»).

Отобраны статьи на русском языке по ключевым словосочетаниям «корпусный менеджер», «язык для специальных целей» и англоязычные статьи из базы данных Scopus по ключевым словосочетаниям «artificial intelligence in language learning», «corpus linguistics».

В работе использовались общенаучные методы описания и обобщения, анализ источников.

ХОД ИССЛЕДОВАНИЯ И ЕГО РЕЗУЛЬТАТЫ

Рассмотрим сферы применения корпусных менеджеров в русскоязычных научных публикациях.

Корпусное исследование Т. Р. Беляевой дисциплинарных подкорпусов академической части Корпуса современного американского английского (*далее* COCA) с использованием корпусного менеджера «Word and Phrase» Академического подкорпуса COCA выявило, что трех-, четырех- и многокомпонентные образования, употребляемые в терминологии прикладной математики и статистики, взаимодействуют с терминологией гуманитарных наук и общенаучной лексикой, способствуя появлению новых многокомпонентных номинативных терминов. При этом коэффициент лексической плотности новых терминов в гуманитарных науках отличается. По результатам исследования лексически более плотными являются тексты на английском языке точных, естественных, психологических и педагогических наук, менее плотными – тексты по истории, филологии, культурологии и философии [Беляева, 2021].

В 2022 году Т. Р. Беляева с помощью корпусного менеджера Sketch Engine создала репрезентативный корпус научных текстов BIOMED медико-биологической тематики (молекулярной и клеточной биологии, генетики, эндокринологии и др.). Корпус содержал более пяти млн словоупотреблений из 872 научных статей и предназначался для проведения сравнительного количественного анализа наиболее частотных академических прилагательных на английском языке в общенаучной и медицинской лексике, а также выявления у прилагательных корреляции высокой или низкой частотности с тремя разновидностями научного дискурса – социологическим, историческим и гуманитарным [Беляева, 2022]. Как показало исследование, плотность лексических единиц общенаучного характера (на примере изученных Т. Р. Беляевой глаголов-коллокатов прилагательного *significant*) возрастает при переходе от гуманитарного к естественно-научному дискурсу.

Е. И. Шпит и ее соавторы исследовали студенческий научный корпус с использованием инструментов компьютерной лингвистики Coh-Metrix

и Gramulator, чтобы продемонстрировать отличие студенческих научных работ российских студентов-инженеров, написанных на английском языке, от работ состоявшихся международных ученых, экспертов в своей области.

Исследователи применили метод машинного определения дифференциалов (Machine Differential Diagnosis) в совокупности с методом сравнительного корпусного анализа (Contrastive Corpus Analysis) для получения инструмента – «дифференцирующих наборов n-грамм», т. е. сочетаний слов, свойственных только одному из анализируемых корпусов текстов, в данном случае написанных студентами или состоявшимися учеными (экспертами) на английском языке. Этот инструмент позволил обнаружить множество расхождений в выборе российскими студентами лексико-грамматических и риторических средств на английском языке.

В итоге студенческие научные работы отличились от экспертных научных работ плотностью именных фраз, разбросом в длине предложений, низким лексическим разнообразием, отсутствием употребления предикативных придаточных предложений и средств хеджирования, что привело исследователей к предположению об использовании российскими студентами автоматического переводчика на английский язык для предварительно составленных текстов научных статей на русском языке. Таким образом, Е. И. Шпит и ее соавторы выявили проблему на стыке лингвистики и педагогики, для решения которой необходимо продолжать обучать российских студентов устному и письменному английскому академическому языку, чтобы их научные публикации и выступления были понятны международному научному сообществу [Шпит, Куровский, 2022].

В процессе взаимодействия преподаватели и студенты для улучшения письменной и устной научной коммуникации на английском языке могут использовать различное программное обеспечение (чат-ботов, тематические онлайн-платформы и социальные сети) [Беспалова, Тастемирова, Волкова, 2024]. Это программное обеспечение может также помочь в исследовании корпусов текстов [Uchida, 2024].

Большой интерес для исследователей, например, представляет проверка эффективности обработки запросов и точности работы модулей корпусного менеджера при вводе материалов на языке изолирующего типа – китайском.

В эксперименте по обработке в закрытом корпусном менеджере частеречной разметки комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» текстов китайских электронных СМИ с учетом типологических свойств китайского языка был установлен уровень точности работы модуля – 7 % [Горожанов, 2024]. Наиболее частотными оказались запросы по идентификации нарицательного имени существительного и глагола, менее частотными – собственное имя существительное, частица, числительное, имя прилагательное. Погрешность работы модуля «китайский язык» составила 7 %, что побудило продолжить исследование с помощью специального ручного запроса и формирования списка «стоп-слов» для уменьшения погрешности. Специальный ручной запрос позволяет лингвистам находить последовательности токенов в корпусе по заданным ими параметрам, например, многокомпонентные имена собственные, которые впоследствии включаются в банк данных на основе частотного списка [Горожанов, Красикова, 2024].

Корпусный менеджер применяется и для создания текстовых заданий во время обучения иностранным языкам. Текстовое задание создается в формате XML на основе загруженных в модуль корпусов текстов на выбранном для обучения языке – английском, французском, немецком и пр. Функция генерации тестового задания прописывается в коде корпусного менеджера на языке программирования Python и обязательно содержит временную метку для защиты XML-файла от «затирания» [Горожанов, Гусейнова, 2024; Горожанов, Степанова, 2024].

Корпусные менеджеры используются при изучении языка в специальных целях. Во-первых, для составления из подготовленного корпуса электронного терминологического переводного словаря, т. е. списков слов, словосочетаний и определений для конкретной предметной области, помогающих студентам в освоении лексики и терминологии изучаемой дисциплины на иностранном языке [Васильева, Салимов, 2023]. Во-вторых, для создания из корпуса электронной терминологической базы лексики по конкретной научной дисциплине на иностранном языке [Шмелева, 2021]. При составлении терминологического словаря и базы лингвисты должны учитывать транстерминологизацию и экстралингвистические факторы, обусловливающие перенос термина из одной дисциплины в другую [Мусаева, Сложеникина, 2024].

В корпусной лингвистике также применяется NLP-инструментарий (например, библиотека обработки естественного языка spaCy и др.), который позволяет автоматически распознавать человеческую речь, анализировать, рубрицировать и обобщать текст и осуществлять машинный перевод. Несмотря на постоянные программные обновления и модификации инструментов по обработке естественного языка, ученые отмечают

наличие общей для всех систем на базе искусственного интеллекта проблемы – дизамбигуации – лексической многозначности слов [Гаджиев, Хмелёв, 2019; Зарипова, Лукашевич, 2023]. Слова в разных языках могут иметь как одно, так и несколько значений и употребляться в переносном значении. К тому же, омонимия, заимствованная узкоспециализированная терминология из точных и естественных наук, многозначность и неоднозначность сокращений и аббревиатур, нехватка аннотированных по значениям текстовых данных затрудняют устранение лексической многозначности слов в гуманитарных науках [Большина, 2022; Awotunde, 2025].

В случае непонимания искусственным интеллектом значения слова и контекста, в котором оно употребляется, возможны серьезные ошибки в машинном переводе [Коврижкин, 2022]. Для решения проблемы дизамбигуации (многозначности слов) необходимо выявление гипо-гиперонимических отношений в лексике [Alexeyevsky, Temchenko, 2016], разработка специализированных аннотированных наборов текстовых данных и создание стандартизированных контрольных показателей [Большина, 2022; Awotunde, 2025].

Перейдем к рассмотрению англоязычных научных публикаций о применении в гуманитарных науках корпусных менеджеров и об анализе корпусов.

Наибольшую сложность для зарубежных исследователей представляет работа по созданию и анализу корпусов текстов разных эпох, например на ранненовоанглийском [Säily et al., 2024]; текстов на урду (язык Пакистана и Индии) [Sadia et al., 2024]; арабском языке [Hassanein, Moustafa, 2024]; лингала (язык Демократической Республики Конго) [Sene-Mongaba, 2015] и на редких языках, например, микенском (древнегреческом языке) [Aurora, 2015] с применением как корпусных менеджеров (Sketch Engine), так и различного программного обеспечения на базе искусственного интеллекта (ChatGPT 3.5, mBART-large и др.). В некоторых случаях (например, при изучении антонимов) исследователи обращаются непосредственно к носителям языка, поскольку только они могут дать достоверную информацию о значении того или иного слова, включенного в корпус.

В исследовании потенциала большой языковой модели ChatGPT 3.5 по составлению списка наиболее частотных слов в сравнении с возможностями корпусного менеджера COCA по аналогичной исследовательской задаче было выявлено, что ChatGPT 3.5 демонстрирует на 75 % совпадение результатов с COCA [Uchida, 2024]. Тем не менее делается вывод о таких существенных ограничениях большой языковой модели ChatGPT для исследования, как невозможность точно определить жанровую принадлежность слов, отсутствие воспроизводимости результатов и риск возникновения галлюцинаций (неподтвержденного текста) [Curry, Baker, Brookes, 2024].

При работе с корпусом исследователь точно знает, из какой предметной области взят текст, чего нельзя сказать о результатах обработки текстового запроса пользователя, выдаваемых большими языковыми моделями ChatGPT, GenAI и др. [Crosthwaite, Baisa, 2023]. В больших языковых моделях количество токенов в обучающих данных исчисляется миллиардами и триллионами, тогда как в закрытых корпусах – тысячами и миллионами. При этом сочетание корпусного подхода («жесткости» корпусных данных) и выдаваемых результатов больших языковых моделей («дифференциации» языка для пользователей с разным уровнем владения иностранным языком и знаний по запросу / промту «перепиши этот текст») при изучении языка дает, по мнению Crosthwaite и Baisa, больше сведений о его изменении во времени и контексте и сокращает время на подготовку тестовых заданий для обучающихся.

Из последних достижений в области корпусной лингвистики можно отметить создание Сиднейским центром информатики вместе с Сиднейской лабораторией корпусов текстов семантического разметчика (Semantic Tagger. v1.0., 2022)[1] для извлечения семантических тегов на уровне токенов из помеченного текста, лемм и тегов частей речи и многословных выражений.

ЗАКЛЮЧЕНИЕ

Обзор сфер применения корпусных менеджеров в гуманитарных исследованиях на примере современных российских и зарубежных научных публикаций показал, что распространение корпусных менеджеров на основе искусственного интеллекта и NLP-инструментария облегчает работу исследователям-лингвистам при условии, что это программное обеспечение протестировано на наличие возможных ограничений производительности для хранения и обработки больших данных и защищено от DDoS-атак. При этом использование открыто размещенных в интернете корпусных менеджеров и корпусов, а также программного обеспечения на основе искусственного интеллекта во время обучения иностранным языкам (ChatGPT, чат-ботов и др.) должно гарантировать пользователям защиту их персональных данных.

[1]URL: https://github.com/Australian-Text-Analytics-Platform/semantic-tagger (дата обращения: 13.03.2025).

СПИСОК ИСТОЧНИКОВ

1. Gorozhanov A. I., Guseynova I. A. Korpusanalyse der Konstituenten Grammatischer Kategorien im Literarischen Text mit Berücksichtigung der Linguoregionalen Komponente // Журнал Сибирского федерального университета. Серия: Гуманитарные науки. 2020. Vol. 13. № 12. P. 2035–2048. DOI 10.17516/1997-1370-0702. EDN KFHPJI.
2. Палийчук Д. А. Корпусные технологии в лингвистических исследованиях // Гуманитарные исследования. История и филология. 2022. № 6. С. 72–79. DOI: 10.24412/2713-0231-2022-6-72-79. EDN VLMGJT.
3. Горожанов А. И., Гусейнова И. А., Степанова Д. В. Стандартизированная процедура получения статистических параметров текста (на материале цикла рассказов Дж. Лондона «Смок Белью. Смок и Малыш») // Вестник Минского государственного лингвистического университета. Серия 1: Филология. 2022. № 4 (119). С. 7–13. EDN PXAVUX.
4. Титова С. В., Игнатова С. Д. Технология применения мультимодальных лингвистических корпусов для развития умений иноязычной интеракции // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2024. Т. 29. № 6. С. 1539–1549. DOI 10.20310/1810-0201-2024-29-6-1539-1549. EDN CBJDQP.
5. Баранов В. А. Кирилло-Мефодиевская и Восточноболгарская лексика в рукописях X–XV вв. (корпусное исследование) // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2023. Т. 22. № 6. С. 5–20. DOI 10.15688/jvolsu2.2023.6.1. EDN CRRTSY.
6. Клочихин В. В. Корпусные технологии искусственного интеллекта в обучении сочетаемости слов и исследовательской работе // Иностранные языки в школе. 2024. № 3. С. 39–46. EDN JFYLHF.
7. Беляева Т. Р. Атрибутивное существительное во множественном числе в многокомпонентных синтагмах английской научной речи // Ученые записки национального общества прикладной лингвистики. 2021. № 1 (33). С. 57–72. EDN NZQDSU.
8. Беляева Т. Р. Корпусный подход к изучению общенаучной лексики английского языка: имя прилагательное // Казанская наука. 2022. № 3. С. 91–97. EDN WCSSBE.
9. Шпит Е. И., Куровский В. Н. Англоязычное научное письмо: затруднения начинающих русскоязычных авторов // Вестник Красноярского государственного педагогического университета имени В. П. Астафьева. 2022. № 3 (61). С. 193–219. DOI 10.25146/1995-0861-2022-61-3-363. EDN EXQKUF.
10. Беспалова Ю. Е., Тастемирова З. К., Волкова М. В. Корпусный анализ перефразирования в научном дискурсе: закономерности, стратегии и последствия для улучшения письма и общения в академических текстах // Гуманитарные исследования. 2024. № 1 (89). С. 23–29. EDN MRAPDZ.
11. Uchida S. Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. Applied Corpus Linguistics. 2024. Vol. 4. Iss. 1. P. 100089. DOI 10.1016/j.acorp.2024.100089.
12. Горожанов А. И. Архитектура сбалансированного лингвистического корпуса, полученного автоматическим путем (опыт Московского государственного лингвистического университета) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 11 (892). С. 24–30. EDN BCSCXQ.
13. Горожанов А. И., Красикова Е. А. Получение значимых данных из неподготовленного текста путем его автоматической обработки авторскими лингвистическими инструментами (на материале электронных китайских СМИ) // Вопросы прикладной лингвистики. 2024. № 54. С. 115–138. DOI 10.25076/vpl.54.05. EDN GGUQXL.
14. Горожанов А. И., Гусейнова И. А. Применение элементов технологии управляемого искусственного интеллекта для наполнения онлайн-курса иностранного языка // Язык в эпоху цифровых трансформаций и развития искусственного интеллекта: сб. науч. ст. Международной научной конф., Минск, 23–24 октября. 2024 года. Минск: Минский государственный лингвистический университет, 2024. С. 26–32. EDN RQRSJA.
15. Горожанов А. И., Степанова Д. В. Лингводидактический потенциал сбалансированного корпуса текстов (на материале английского языка) // Актуальные проблемы лингвистики и лингводидактики иностранного языка делового и профессионального общения: сб. материалов XI Международной науч.-практ. конф. Москва, 17–18 апреля 2024 года. М.: РУДН, 2024. С. 343–345. EDN TNKAPX.
16. Васильева Т. В., Салимов Ф. Н. Предпосылки создания цифрового переводного терминологического словаря тезаурусного типа для иностранных пользователей вузов инженерного профиля // Русский язык за рубежом. 2023. № 1 (296). С. 38–45. DOI 10.37632/PI.2023.296.1.006. EDN OPTNFC.
17. Шмелева О. Н. Краткий обзор переводных словарей противопожарной лексики (на материале немецкого и русского языков) // Культура и безопасность. 2021. № 4. С. 64–70. DOI 10.25257/KB.2021.4.64-70. EDN IPPZAH.
18. Мусаева А. С., Сложеникина Ю. В. Когнитивная обработка термина в процессе транстерминологизации (на примере терминов искусственного интеллекта) // Вестник Череповецкого государственного университета. 2024. № 6 (123). С. 103–113. DOI 10.23859/1994-0637-2024-6-123-9. EDN VBWFAM.

19. Гаджиев А. А., Хмелев А. К. Алгоритм Леска и система Babelfy для дизамбигуации // Вопросы прикладной лингвистики. 2019. № 4 (36). С. 54–77. DOI 10.25076/vpl.36.03. EDN PVWXOG.
20. Зарипова Д. А., Лукашевич Н. В. Подходы к автоматическому разрешению многозначности на основе неравномерности распределения значений слов в корпусе // Вестник Московского университета. Серия 9: Филология. 2023. № 6. С. 40–51. DOI 10.55959/MSU0130-0075-9-2023-47-06-4. EDN ALEVII.
21. Большина А. С. Методы автоматического формирования семантически размеченных корпусов // Вестник Московского университета. Серия 9: Филология. 2022. № 2. С. 173–183. EDN QHGJLK.
22. Awotunde J. B. Chapter 28 – Word sense disambiguation in biomedical applications // Mining Biomedical Text, Images and Visual Features for Information Retrieval. Editor(s): Sujata Dash, Subhendu Kumar Pani, Wellington Pinheiro Dos Santos, Jake Y. Chen. Academic Press. 2025. P. 587–605. DOI 10.1016/B978-0-443-15452-2.00028-5.
23. Коврижкин А. А. Семантические проблемы машинного перевода // Роль и место лингвокультурной адаптации художественного текста в теории и практике перевода. Переводческие стратегии и тактики: материалы Всероссийской науч.-практ. конф. с международным участием, Москва, 23 декабря 2021 года. М.: МГОУ, 2022. С. 295–302. EDN EZPPCR.
24. Alexeyevsky D. A., Temchenko A. V. WSD in monolingual dictionaries for Russian Word Net: Proceedings of the 8th Global Word Net Conference, GWC 2016: 8, Bucharest, 27–30 January 2016. Bucharest, 2016. P. 10–15. EDN WSTVXB.
25. Säily T. et al. Changing styles of letter-writing? Evidence from 400 years of early English letters in a POS-tagged corpus / Säily T., Vartiainen T., Siirtola H., Nevalainen T. // Unlocking the History of English: Pragmatics, prescriptivism and text types. Ed. by Luisella Caon, Moragh S. Gordon and Thijs Porck. Amsterdam: John Benjamins Publishing Company, 2024. P. 154–179. DOI 10.1075/cilt.364.07sai.
26. Sadia B. et al. Meeting the challenge: A benchmark corpus for automated Urdu meeting summarization / Sadia B., Adeeba F., Shams S., Javed K. // Information Processing & Management. 2024. Vol. 61. Iss. 4. P. 103734. DOI 10.1016/j.ipm.2024.103734.
27. Hassanein H. S. A., Moustafa B. S. M. Sequential order of antonym pairs in Modern Standard Arabic: A corpus-based analysis // Lingua. 2024. Vol. 306. P. 103742. DOI 10.1016/j.lingua.2024.103742.
28. Sene-Mongaba B. The Making of Lingala Corpus: An Under-resourced Language and the Internet // Procedia – Social and Behavioral Sciences. 2015. Vol. 198. P. 442–450. DOI 10.1016/j.sbspro.2015.07.464.
29. Aurora F. DĀMOS (Database of Mycenaean at Oslo). Annotating a Fragmentarily Attested Language // Procedia – Social and Behavioral Sciences. 2015. Vol. 198. P. 21–31. DOI 10.1016/j.sbspro.2015.07.415.
30. Curry N., Baker P., Brookes G. Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT // Applied Corpus Linguistics. 2024. Vol. 4. Iss. 1. P. 100082. DOI 10.1016/j.acorp.2023.100082.
31. Crosthwaite P., Baisa V. Generative AI and the end of corpus-assisted data-driven learning? Not so fast! // Applied Corpus Linguistics. 2023. Vol. 3. Iss. 3. P. 100066. DOI 10.1016/j.acorp.2023.100066.

REFERENCES

1. Gorozhanov, A. I., Guseynova, I. A. (2020). Korpusanalyse der Konstituenten Grammatischer Kategorien im Literarischen Text mit Berücksichtigung der Linguoregionalen Komponente. J. Sib. Fed. Univ. Humanit. Soc. Sci., 13(12), 2035–2048. DOI 10.17516/1997-1370-0702.
2. Paliychuk, D. A. (2022). Corpus technologies in linguistic research. Humanitarian Research. History and Philology, 6, 72–79. 10.24412/2713-0231-2022-6-72-79. (In Russ.)
3. Gorozhanov, A. I., Guseinova, I. A., Stepanova, D. V. (2022). Standardized procedure for obtaining statistical parameters of a text (on the material of the stories by J. London “Smoke Bellew. Smoke and Shorty”). Minsk State Linguistic University Bulletin. Series 1. Philology, 4(119), 7–13. (In Russ.)
4. Titova, S. V., Ignatova, S. D. (2024). Technology of using multimodal linguistic buildings for the development of foreign language interaction skills. Bulletin of Tambov University. Series: Humanities, 29(6), 1539–1549. DOI 10.20310/1810-0201-2024-29-6-1539-1549. (In Russ.)
5. Baranov, V. A. (2023). Cyril-Methodian and Eastern Bulgarian Words in the Manuscripts of the 10th-15th Centuries (Text Corpus Study). Science Journal of Volgograd State University. Linguistics, 22(6), 5–20. DOI 10.15688/jvolsu2.2023.6.1. (In Russ.)
6. Klochikhin, V. V. (2024). Corpus technologies of artificial intelligence in teaching word compatibility and research work. Foreign Languages at School, 3, 39–46. (In Russ.)

7. Belyaeva, T. R. (2021). Attributive noun in the plural in multicomponent syntagmas of English scientific speech. Scientific Notes of the National Society of Applied Linguistics, 1(33), 57–72. (In Russ.)
8. Belyaeva, T. R. (2022). A corpus approach to the study of the general scientific vocabulary of the English language: An adjective. Kazan Science, 3, 91–97. (In Russ.)
9. Shpit, E. I., Kurovsky, V. N. (2022). English-language scientific writing: The difficulties of novice Russian-speaking authors. Bulletin of the Krasnoyarsk State Pedagogical University Named After V. P. Astafiev, 3(61), 193–219. DOI 10.25146/1995-0861-2022-61-3-363. (In Russ.)
10. Bespalova, Y. E., Tastemirova, Z. K., Volkova, M. V. (2024). Corpus analysis of paraphrasing in scientific discourse: Patterns, strategies, and consequences for improving writing and communication in academic texts. Humanitarian Studies, 1(89), 23–29. (In Russ.)
11. Uchida, S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. Applied Corpus Linguistics, 4(1), 100089. DOI 10.1016/j.acorp.2024.100089.
12. Gorozhanov, A. I. (2024). Architecture of a balanced linguistic corpus built automatically (experience of Moscow State Linguistic University). Vestnik of Moscow State Linguistic University. Humanities, 11(892), 24–30. EDN BCSCXQ. (In Russ.)
13. Gorozhanov, A. I., Krasikova, E. A. (2024). Obtaining meaningful data from an unprepared text by automatically processing it with author's linguistic tools (based on the material of electronic Chinese media). Questions of Applied Linguistics, 54, 115–138. DOI 10.25076/vpl.54.05. (In Russ.)
14. Gorozhanov, A. I., Guseynova, I. A. (2024). Application of elements of controlled artificial intelligence technology to complete an online foreign language course. Language in the era of digital transformations and the development of artificial intelligence (pp. 26–32): Proceedings of the International scientific conference. Minsk. (In Russ.)
15. Gorozhanov, A. I., Stepanova, D. V. (2024). Lingvodidactic potential of a balanced corpus of texts (based on the English language).Current issues of linguistics and linguodidactics of foreign language for business and professional communication (pp. 343–345): Proceedings of theXI International scientific and practical conference. Moscow. (In Russ.)
16. Vasilyeva, T. V., Salimov, F. N. (2023). Prerequisites for creating a digital translated thesaurus-type terminological dictionary for foreign users of engineering universities. Russian Language Abroad, 1(296), 38–45. DOI 10.37632/PI.2023.296.1.006. (In Russ.)
17. Shmeleva, O. N. (2021). A brief overview of translated dictionaries of fire vocabulary (based on the material of German and Russian languages). Culture and Safety, 4, 64–70. DOI 10.25257/KB.2021.4.64-70. (In Russ.)
18. Musayeva, A. S., Skladenikina, Yu. V. (2024). Cognitive processing of a term in the process of transterminologization (using the example of artificial intelligence terms). Bulletin of Cherepovets State University, 6(123), 103–113. DOI 10.23859/1994-0637-2024-6-123-9.(In Russ.)
19. Gadzhiev, A. A., Khmelev, A. K. (2019). The Leska algorithm and the Babelfy system for disambiguation. Questions of Applied Linguistics, 4(36), 54–77. DOI 10.25076/vpl.36.03. (In Russ.)
20. Zaripova, D. A., Lukashevich, N. V. (2023). Approaches to automatic resolution of ambiguity based on the uneven distribution of word meanings in the corpus. Bulletin of the Moscow University. Episode 9: Philology, 6, 40–51. DOI 10.55959/MSU0130-0075-9-2023-47-06-4. (In Russ.)
21. Bolshina, A. S. (2022). Methods of automatic formation of semantically marked buildings. Moscow University Bulletin. Series 9: Philology, 2, 173–183. (In Russ.)
22. Awotunde, J. B. (2025). Word sense disambiguation in biomedical applications. In Dash, S., Pani, S. K., Dos Santos, W. P., Chen, J. Y. (Eds.), Mining Biomedical Text, Images and Visual Features for Information Retrieval (pp. 587–605). Academic Press. DOI 10.1016/B978-0-443-15452-2.00028-5.
23. Kovrizhkin, A. A. (2022). Semanticheskie problemy mashinnogo perevoda = Semantic problems of machine translation. In The Role and Place of Linguistic and Cultural Adaptation of Literary Text in the Theory and Practice of Translation. Translation Strategies and Tactics (pp. 295–302): Proceedings of the All-Russian Scientific and Practical Conference with International Participation, Moscow, December 23, 2021. Moscow: Moscow State Regional University. (In Russ.)
24. Alexeyevsky, D. A., Temchenko, A. V. (2016). WSD in monolingual dictionaries for Russian WordNet. In Proceedings of the 8th Global WordNet Conference (pp. 10–15), GWC 2016, Bucharest, January 27–30. Bucharest.
25. Säily, T. et al. (2024). Changing styles of letter-writing? Evidence from 400 years of early English letters in a POS-tagged corpus. In Caon, L., Gordon, M. S., Porck, T. (Eds.), Unlocking the History of English: Pragmatics, Prescriptivism and Text Types (pp. 154–179). Amsterdam: John Benjamins Publishing Company. DOI 10.1075/cilt.364.07sai.
26. Sadia, B., Adeeba, F., Shams, S., Javed, K. (2024). Meeting the challenge: A benchmark corpus for automated Urdu meeting summarization. Information Processing & Management, 61(4), 103734. DOI 10.1016/j.ipm.2024.103734.

27. Hassanein, H. S. A., Moustafa, B. S. M. (2024). Sequential order of antonym pairs in Modern Standard Arabic: A corpus-based analysis. Lingua, 306, 103742. DOI 10.1016/j.lingua.2024.103742.
28. Sene-Mongaba, B. (2015). The making of Lingala corpus: An under-resourced language and the Internet. Procedia – Social and Behavioral Sciences, 198, 442–450. DOI 10.1016/j.sbspro.2015.07.464.
29. Aurora, F. (2015). DĀMOS (Database of Mycenaean at Oslo). Annotating a fragmentarily attested language. Procedia – Social and Behavioral Sciences, 198, 21–31. DOI 10.1016/j.sbspro.2015.07.415.
30. Curry, N., Baker, P., Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1), 100082. DOI 10.1016/j.acorp.2023.100082.
31. Crosthwaite, P., Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! Applied Corpus Linguistics, 3(3), 100066. DOI 10.1016/j.acorp.2023.100066.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Романова Светлана Андреевна
специалист отдела научного менеджмента и наукометрии
Московского государственного лингвистического университета

INFORMATION ABOUT THE AUTHOR

Romanova Svetlana Andreevna
Specialist of the Department of Scientific Management and Scientometrics
Moscow State Linguistic University

Статья поступила в редакцию	12.04.2025	The article was submitted
одобрена после рецензирования	04.05.2025	approved after reviewing
принята к публикации	15.05.2025	accepted for publication