OPEN ACCESS
[Научный диалог = Nauchnyi dialog = Nauchnyy dialog, 14(7), 2025]
[ISSN 2225-756X, eISSN 2227-1295]
Check for updates

Информация для цитирования:

Шадрина О. В. Корпусный анализ репрезентации терминологии искусственного интеллекта в русском языке с использованием инструмента AntConc (на материале альманаха «Искусственный интеллект») / О. В. Шадрина, О. В. Маруневич // Научный диалог. — 2025. — Т. 14. — № 7. — С. 133—160. — DOI: 10.24224/2227-1295-2025-14-7-133-160.

Shadrina, O. V., Marunevich, O. V. (2025). Corpus Analysis of Artificial Intelligence Terminology in Russian: Insights from Almanac "Artificial Intelligence" Using AntConc. *Nauchnyi dialog, 14* (7): 133-160. DOI: 10.24224/2227-1295-2025-14-7-133-160. (In Russ.).

Clarivate Web of Science · Scopus Q1 · DOAJ · ERIHPLUS · eLIBRARY.RU РИНЦ

Перечень рецензируемых изданий ВАК при Минобрнауки РФ

Корпусный анализ репрезентации терминологии искусственного интеллекта в русском языке с использованием инструмента AntConc (на материале альманаха «Искусственный интеллект»)

Шадрина Олеся Владимировна
orcid.org/0000-0003-1980-3754
старший преподаватель,
департамента иностранных языков,
корреспондирующий автор
shadrina.ov@mipt.ru

Маруневич Оксана Викторовна
orcid.org/0000-0002-4480-6642
кандидат филологических наук, доцент
marunevich.ov@mipt.ru

Московский физико-технический институт (национальный исследовательский университет)
(Долгопрудный, Россия)

Corpus Analysis of Artificial Intelligence Terminology in Russian: Insights from Almanac "Artificial Intelligence" Using AntConc

Olesya V. Shadrina
orcid.org/0000-0003-1980-3754
senior lecturer,
Department of Foreign Languages,
corresponding author
shadrina.ov@mipt.ru

Oksana V. Marunevich
orcid.org/0000-0002-4480-6642
PhD in Philology, Associate Professor
marunevich.ov@mipt.ru

Moscow Institute of Physics and Technology
(National Research University)
(Dolgoprudny, Russia)

ОРИГИНАЛЬНЫЕ СТАТЬИ

Аннотация:

Исследование выполнено на стыке корпусной лингвистики и терминоведения. Отмечается, что корпусная лингвистика прошла значительный путь от ранних форм текстовых коллекций до создания крупных национальных и специализированных корпусов в XXI веке. Акцентируется внимание на важности современных технологий, таких как машинное обучение и обработка естественного языка, которые открывают новые возможности для анализа больших массивов данных. Статья освещает методологические аспекты исследования терминологических единиц в области искусственного интеллекта (ИИ) на основе современных аналитических сборников. Цель исследования заключается в выявлении моделей образования составных обозначения, орфографических и стилистических норм использования терминов ИИ в русском языке. Для достижения этой цели использованы методы частотного анализа и контент-анализа с применением сервиса AntConc, что позволило выделить 100 ядерных терминов, а также коллокации, конструируемые на основе таких терминов. Результаты исследования показывают, что терминология ИИ в русском языке активно развивается. Констатируется преобладание англицизмов и гибридных форм. Обсуждаются стилистические особенности текстов, отражающие технический контекст и целевую аудиторию. В заключение подчеркивается необходимость установления норм употребления терминов ИИ в связи с их интеграцией в русский язык.

Ключевые слова:
платформа AntConc; терминосистема искусственного интеллекта; корпусная лингвистика; коллокация; конкорданс; языковой корпус.

ORIGINAL ARTICLES

Abstract:

This study is situated at the intersection of corpus linguistics and terminology studies. It highlights the significant evolution of corpus linguistics, from early text collections to the establishment of large national and specialized corpora in the 21st century. The importance of contemporary technologies, such as machine learning and natural language processing, is emphasized for their role in opening new avenues for analyzing large data sets. The article addresses the methodological aspects of researching terminological units within the field of artificial intelligence (AI) based on modern analytical compilations. The aim of the research is to identify patterns in the formation of compound designations, as well as the orthographic and stylistic norms governing the use of AI terms in the Russian language. To achieve this goal, frequency analysis and content analysis methods were employed using AntConc, resulting in the identification of 100 core terms, along with collocations constructed from these terms. The findings indicate that AI terminology in Russian is actively evolving, with a predominance of Anglicisms and hybrid forms. The stylistic features of texts reflecting the technical context and target audience are discussed. In conclusion, the necessity for establishing norms for the use of AI terms in light of their integration into the Russian language is underscored.

Key words:
AntConc platform; artificial intelligence terminology system; corpus linguistics; collocation; concordance; language corpus.

OPEN ACCESS
[Научный диалог = Nauchnyi dialog = Nauchnyy dialog, 14(7), 2025]
[ISSN 2225-756X, eISSN 2227-1295]

УДК 811.161.1’373.46
DOI: 10.24224/2227-1295-2025-14-7-133-160

Научная специальность ВАК
5.9.5. Русский язык. Языки народов России
5.9.8. Теоретическая, прикладная и
сравнительно-сопоставительная лингвистика

Корпусный анализ репрезентации терминологии искусственного интеллекта в русском языке с использованием инструмента AntConc (на материале альманаха «Искусственный интеллект»)

© Шадрина О. В., Маруневич О. В., 2025

1. Введение = Introduction

Одним из наиболее перспективных направлений современной науки о языке является корпусная лингвистика, занимающаяся изучением языка на основе анализа больших массивов текстовых данных, собранных в корпусах. При этом, по мнению А. П. Кононенко и Л. А. Недосека, ученые-лингвисты все больше обращаются к инструментарию в виде корпусных технологий для решения поставленных задач [Кононенко и др., 2023]. В настоящее время корпусные методы и технологии широко используются как отечественными, так и зарубежными учеными для лингвистического анализа текстов и дискурсивных практик, преподавания и изучения языка, а также в культурологических исследованиях. Многообразие реализованных подходов и достижений в области корпусной лингвистики наглядно иллюстрирует широкий потенциал и значительные возможности, предоставляемые корпусными методами [Corpus …, 2020].

Традиционно под ***корпусами*** понимаются структурированные коллекции текстов, которые могут быть аннотированы на различных уровнях (морфологическом, синтаксическом, семантическом) [O’Keeffe et al., 2010]. Авторы первого языкового корпуса текстов («Брауновский корпус», 1963 год) трактовали понятие «корпус» как совокупность текстов, считающихся релевантными для данного языка, диалекта или иного подмножества естественного языка, предназначенных для лингвистического анализа [Francis et al., 1964]. Согласно Дж. Синклеру, корпус представляет собой коллекцию текстов естественного языка, выбранных для исследования разнообразия языка [Sinclair, 1991]. В свою очередь Н. В. Козлова подчеркивает, что корпус включает в себя оцифрованные как письменные, так и устные высказывания, хранящиеся на специальных платформах и доступ-

ные в электронном виде [Козлова, 2013, с. 79]. Т. А. Архангельский [Архангельский, 2019, с. 528] и Л. Селиван [Selivan, 2023] отмечают, что языковой корпус — это собрание реальных образцов использования определенного языка в виде текстов, специально предназначенное для изучения данного языка. Это связано с тем, что в процессе преподавания и изучения как родного, так и иностранного языка активно используются словари активной лексики, учебники с содержащимися в них грамматическими правилами и упражнениями, тесты и другие обучающие ресурсы [Boulton et al., 2016].

Наиболее развернутая дефиниция термина *языковой корпус*, на наш взгляд, представлена в работе В. П. Захарова: «Под названием лингвистический, или языковой, корпус текстов понимается большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач» [Захаров, 2005, с. 3]. Автор также указывает, что понятие «корпус» включает в себя систему управления текстовыми и лингвистическими данными (так называемый корпус-менеджер), главной задачей которой является поиск данных в корпусе для последующего получения статистической информации и предоставления результатов пользователю в удобном формате [Там же, с. 3].

Следует отметить, что как научное направление корпусная лингвистика прошла долгий путь становления и развития. Ее теоретические основы были заложены еще в начале XX века, например в работе Ф. Боаса о языке американских индейцев (1911—1922) [Boas, 2013]. Несмотря на то, что он не занимался корпусной лингвистикой в современном понимании этого термина, поскольку эта область сформировалась значительно позже, он записывал тексты, мифы, песни и повседневную речь коренных народов Америки, создавая обширные коллекции языковых материалов. Эти коллекции можно рассматривать как ранние формы лингвистических корпусов, хотя они и не были оцифрованы или структурированы так, как это делается в современной корпусной лингвистике. Также следует отметить «Структурную лингвистику» З. Харриса (1951), в которой автор подчеркивал важность анализа реальных языковых данных, а не абстрактных теоретических построений. Кроме того, дистрибутивный анализ, активно применявшийся З. Харрисом, предвосхитил подходы корпусной лингвистики, где важную роль играет анализ контекстов и сочетаемости слов (например, коллокаций) [Harris, 1960].

Статус отдельной дисциплины корпусная лингвистика получила в 1950—60-х годах, когда были предприняты первые опыты создания корпусов. Однако сам термин *корпусная лингвистика* прочно вошел в научный обиход после публикации Дж. Аартсом и В. Мейхсом своей работы

«Corpus Linguistics In Recent Developments in the Use of Computer Corpora» [Aarts et al., 1984].

Одним из первых крупных корпусов является Brown Corpus, созданный Г. Кучерой и У. Н. Фрэнсисом в американском университете Брауна в 1963 году. Объем данного корпуса составил около 1 млн словоупотреблений (500 фрагментов объемом по 2000 слов), извлеченных из текстов разнообразных жанров за 1961 год, включая литературные произведения, публицистику, деловую переписку и даже тесты религиозного характера. В частности, в корпусе было представлено 80 текстов научной тематики, 30 текстов правительственных документов, 75 отрывков из художественной литературы в стихах и прозе и т. д. При этом, как пишет А. Стефанович, корпус строился на основе статистических данных, но это было бы невозможно без опоры на интуицию ученых. Список категорий был составлен на конференции, состоявшейся в Университете Брауна в феврале 1963 года. Участники конференции независимо друг от друга высказали свои мнения относительно количества образцов, которые должны быть в каждой категории. Впоследствии эти цифры были усреднены для получения оптимального набора образцов текстов [Stefanowitsch, 2020, p. 30—32].

Выделяется несколько причин создания Brown Corpus. Во-первых, реализация системного анализа разножанровых текстов, написанных на американском варианте английского языка. Во-вторых, обеспечение исследователей достаточным количеством материала для последующего сопоставления представленных данных. Наконец, привлечение интереса к новой отрасли знания [Assunção et al., 2019].

Это был настоящий прорыв в прикладной лингвистике, вызвавший многочисленные споры и дискуссии. Известно, что Н. Хомский неоднократно подвергал корпусную лингвистику резкой критике, полагая, что собственная интуиция — это единственное, чем может руководствоваться лингвист при изучении различных языковых явлений. Он также подчеркивал, что синтаксис должен быть основным объектом лингвистических исследований [Chomsky, 1969]. Р. Лису создание корпусов представлялось бессмысленной тратой времени и правительственных денег. Он безапелляционно заявлял, что любой носитель английского языка за 10 минут способен представить больше примеров на то или иное явление английской грамматики, чем ученые смогут найти в миллионах отрывков [Biber et al.,1991, p. 206]. Д. Аберкромби считал корпусные исследования языка «псевдометодом» (pseudoprocedure), не имеющим ничего общего с классическими методами прикладной лингвистики: «If “procedure” is taken to mean “way or method of conducting an investigation”, then a “pseudoprocedure” is something which is put forward as a way of conducting an investiga-

tion, but which in fact is an impossible, or at best a completely impracticable, way)» [Abercrombie, 1965] / «Если “метод” понимается как «способ проведения исследования”, то “псевдометод” — это нечто, что декларируется как способ проведения исследования, но на самом деле является невозможным или, в лучшем случае, совершенно неосуществимым» (здесь и далее перевод наш. — *О. Ш., О. М.*).

Кроме того, некоторые ученые писали о практических проблемах, с которыми часто сталкивались составители первых корпусов, например, о медленной обработке данных и ее высокой стоимости при высокой степени ошибочности [Dash et al., 2018].

Несмотря на то, что первый языковой корпус был создан лишь в середине XX века, нам кажется целесообразным выделить и доэлектронный период развития корпусной лингвистики. Начало создания корпуса текстов датируется XIII веком. В этот период появляется «Concordantiae morales sacrae scripturae» («Нравственная конкорданция Священного Писания») (1220-е годы) — первый предметный конкорданс, составленный одним из Учителей Церкви Антонием Падуанским к «Вульгате», латинскому переводу Библии, сделанному в конце IV — начале V веков Иеронимом Стридонским. Указанный конкорданс содержал тексты с параллельными цитатами, однако только по нравственным вопросам. Немногим позже (около 1230 года) первым кардиналом доминиканского монастыря святого Иакова в Париже Гуго де Сен-Шером был составлен «Concordantiae Sancti Jacobi», содержащий короткие цитаты отрывков, где было найдено то или иное слово. Существенным недостатком данного конкорданса является то, что цитируемые слова сопровождались лишь указаниями книг и глав Библии, откуда они были взяты, без приведения самих параллельных текстов [Bataillon et al., 2004]. Впоследствии данный конкорданс неоднократно совершенствовался. В частности, в 1250—1252 годах монахи-доминиканцы Иоанн из Дарлингтона, Ричард из Ставенесби и Гуго из Кройдона дополнили конкорданс Сен-Шера цитатами из параллельных текстов, а монах-францисканец Арлотто из Прато (конец XIII века) расширил его введением несклоняемых частиц. В 1310 году Конрад из Хальберштадта сократил конкорданс, оставив только основные слова цитаты, однако именно его версия получила наибольшее распространение из-за более удобной формы [Casson, 1960].

В 1470 году в Страсбурге появился первый печатный конкорданс, измененный и дополненный в версиях 1475, 1485 и 1496 года. Изданный в Базеле в 1496 году конкорданс Себастьяна Бранта послужил основой для конкорданса Роберта Этьенна (1555 год). Р. Этьенн добавил в корпус имена собственные, восполнил существующие пропуски, разместил и склоняемые, и несклоняемые слова в алфавитном порядке и дал указания ко всем

цитатам по стихам и главам, что в значительной мере приблизило его работу к современной модели лингвистических корпусов.

Также известны конкордансы к переводам Библии на европейские языки: немецкий — Ф. Ланкиш (Лейпциг, 1677); Г. Бюхнер (Йена, 1757), Ф. И. Бернгард (Лейпциг, 1850); английский — Т. Гибсон (Лондон, 1535), Дж. Марбек (Лондон, 1550), Дж. Даунэйм (Лондон, 1630), А. Круден (Лондон, 1737), Р. Янг (Эдинбург, 1879—1884); русский — Антиох Кантемир (Санкт-Петербург, 1727); Иван Ильинский (Санкт-Петербург, 1737); Андрей Богданов (Санкт-Петербург, 1737), холмогорский епископ Парфений (Москва, 1823). В целом, все вышеописанные корпусы составлялись по единому алгоритму. Для упрощения последующего поиска цитат в Ветхом и Новом Заветах авторы конкордансов вручную индексировали слова в каждой строчке Библии.

Помимо религиозных текстов, конкордансы в доэлектронный период составлялись и для текстов светского характера. В частности, работа немецкого филолога И. Кадинга (1897), в которой он рассчитал частоту распределения и последовательность появления букв в немецком языке с использованием корпуса из почти 11 миллионов слов [Assunção et al., 2019]; исследование Э. Итон (1940), где она сравнивала частоту значений слов в английском, французском, немецком и испанском языках [Eaton, 1940]. В рамках педагогики иностранных языков следует отметить работы Э. Л. Торндайка (1921), Г. Палмера (1933), Ч. К. Фриза и А. А. Тревер (1940), Х. Бонгерса (1947) и т. д. Эти ученые и ряд других вручную составили и использовали английские речевые корпуса различных размеров и форм для определения закономерностей в расширении словарного запаса у учащихся [Dash et al., 2018, p. 180—182].

Первым протоэлектронным конкордансом считается Index Thomisticus, составление которого началось в 1940-х годах. Под руководством Роберто Бузы проект за более чем 30 лет проиндексировал 10 631 980 слов из текстов Фомы Аквинского, первоначально на перфокартах, затем в автоматическом режиме [Mcgillivray et al., 2009]. В 1993 году Index Thomisticus был назван второй по величине печатной работой XX столетия, превзойти которую смогла лишь Британская энциклопедия [Guietti, 1993]. Некоторые исследователи считают, что авторы проекта Index Thomisticus не только впервые разработали методы работы с неструктурированным языком, но и положили начало цифровым гуманитарным наукам [Rockwell et al., 2019].

Однако очевидно, что изучение корпуса вышло на совершенно новый уровень в связи с применением компьютерных технологий в лингвистике. В частности, благодаря появлению персональных компьютеров в 1980-х годах были преодолены трудности сбора данных на мэйнфреймовых ком-

пьютерах. Это способствовало новому витку популярности исследований, основанных на корпусах. Так, новаторское партнерство Бирмингемского университета и издательства Collins привело к созданию первого словаря, построенного на принципах корпусной лингвистики, — Cobuild English Dictionary (1987). За основу проекта COBUILD (Collins Birmingham University International Language Database) был взят Бирмингемский корпус текстов, включающий 20 млн словоупотреблений, из которых 7,3 млн составили основной корпус, а 13 млн — резервный. При этом корпус на 75 % состоял из образцов письменной речи (преимущественно художественной прозы) и на 25 % — из записей устной речи, относящихся к периоду с 1960-х до 1982 года. По словам руководителя проекта Дж. Синклера, существует несколько факторов, отличающих COBUILD от других словарей английского языка. Во-первых, он отражает современное состояние английского языка. Во-вторых, по форме представления записей он представляет собой радикальный отход от существующих принципов лексикографии. В-третьих, при его составлении использовались передовые компьютерные технологии [Sinclair, 1987]. С. Йохансон отмечает, что корпус COBUILD можно считать настоящим прорывом для своего времени благодаря тому, что его объем превышал 20 млн словоупотреблений. При этом корпус отличался высокой репрезентативностью, так как включал тексты как устной, так и письменной речи разнообразных жанров. Также особенностью корпуса было то, что источниками словоупотреблений служили полные тексты, а не короткие фрагменты [Johansson, 2009].

Развитие корпусной лингвистики в 1990-е годы ознаменовалось появлением национальных корпусов, отражавших специфику различных языков. Например, созданный в 1994 году British National Corpus (BNC) вплоть до настоящего времени является одним из крупнейших корпусов английского языка. Именно данный корпус первым получил статус «национального». По мнению В. А. Плунгяна, первоначально слово *национальный* в названии корпуса являлось синонимом квалификации «британский вариант английского языка» с целью противопоставления другим вариантам — американскому, канадскому, австралийскому и др. Впоследствии же *национальными* стали называть объемные корпуса с репрезентативными выборками текстов различных жанров, отражающих специфику того или иного языка в определенный исторический период [Плунгян, 2005, с. 7]. В это же время появляется Corpus of Historical American English (400 млн слов американского варианта английского языка), Corpus of Contemporary American English (450 млн слов), корпус Берлинской Бранденбургской академии наук (1,8 млрд словоупотреблений), Мангеймский корпус немецкого языка (5,5 млрд слов) и Национальный корпус русского языка (500 млн

слов). Как можно заметить, в этот период наметилась тенденция к увеличению объемов корпусов. По меткому замечанию некоторых исследователей, создатели корпусов руководствовались девизом «чем больше, тем лучше» [Kuebler et al., 2015, p. 10]. Вместе с тем огромные массивы текстовых фрагментов, включенных в национальные корпуса, сделали возможным масштабные исследования частотности лексических единиц и коллокаций, состоящих из нескольких слов (так называемых *кластеров*) [Hyland, 2008].

Кроме того, в конце 1990-х — начале 2000-х годов создаются корпуса для документирования и изучения языков малочисленных народов, многие из которых находятся под угрозой исчезновения. В частности, стоит отметить Корпус айнского языка (AINU Corpus, Япония), Корпус языков коренных народов Латинской Америки (AILLA, США), Корпус языков коренных народов Австралии (PARADISEC, Австралия), Корпус языков коренных народов Сибири (INTAS, Россия), Корпус саамских языков (GiellaLT, Норвегия) и др.

В XXI веке корпусная лингвистика вышла на новый уровень благодаря созданию специализированных медицинских, юридических и научных корпусов. Например, PubMed Corpus содержит миллионы научных статей по медицине, что позволяет исследователям изучать терминологию и дискурс в этой области [Stefchov et al., 2018; Dernoncourt et al., 2017; Doğan et al., 2012]. Еще одним направлением стало создание параллельных корпусов, содержащих тексты на нескольких языках. В частности, Europarl Corpus включает тексты Европейского парламента на 21 языке, что делает его ценным ресурсом для изучения перевода и межъязыковых соответствий. А развитие мультимодальных корпусов, включающих аудиовизуальные данные, открыло новые возможности для изучения языка в его многообразии [Allwood, 2009].

Основная цель корпусных исследований заключается в выявлении закономерностей использования языковых единиц в различных контекстах. Для этого применятся целый ряд методов, включая частотный анализ, изучение коллокаций, конкордансов и ключевых слов. Так, частотный анализ позволяет определить, насколько часто те или иные слова или конструкции встречаются в корпусе. Данный метод помогает выявить наиболее употребительные элементы языка и их распределение в различных типах текстов. Так, исследование О. Н. Ляшевской и С. А. Шарова демонстрирует, что в русском языке частотность слов варьируется в зависимости от жанра текста. Например, ими было выявлено, что слова *ну, да, вот, угу, ага* и др. появляются в спонтанной устной речи в десятки раз чаще, чем в подготовленной письменной речи [Ляшевская и др., 2009].

Изучение коллокаций помогает понять паттерны взаимодействия слов в языке. По мнению Дж. Хилла и М. Льюиса, коллокации играют важную

роль в обеспечении связности речи, способствуют достижению беглости, повышают уровень понятности и предсказуемости языковых высказываний [Hill et al., 1997, p. 1] («one of the most powerful forces in making language coherent, fluent, comprehensible, and predictable» [Pawley et al., 1983, p. 205]), так как в большинстве случаев они вспоминаются и воспроизводятся целиком. Проведенный Дж. Синклером анализ коллокаций в British National Corpus выявил огромное количество устойчивых сочетаний в английском языке (*make a mistake, strong criticism, bitterly cold)* [Sinclair, 1991].

Конкордансы представляют собой списки контекстов употребления слова или фразы. Они позволяют исследовать семантику и синтаксические особенности языковых единиц. Например, конкордансы используются для изучения многозначных слов и их значений в различных контекстах [McEnery et al., 2012].

Что касается ключевых слов, то их анализ помогает выявить наиболее значимые слова в тексте или корпусе по сравнению с другим корпусом. Данный метод часто используется в дискурс-анализе и стилометрии [Scott et al., 2006]. Например, анализ корпуса политических речей позволяет выявить основные темы выступлений лидеров наций и риторические приемы, нацеленные на определенное воздействие на аудиторию [Partington et al., 2015].

Помимо корпусной педагогики, дискурс-анализа и стилометрии, методы корпусной лингвистики активно применяются для анализа грамматических конструкций в их естественном употреблении. В частности, исследования на основе корпуса COBUILD показали, что грамматические правила, описанные в учебниках, зачастую не соответствуют реальному употреблению [Hunston et al., 2000]. Анализ четырех совершенно разных типов макродискурса (разговорная речь, художественная литература, публицистика и академическая проза) позволил скорректировать общую грамматику современного английского языка [Grammar …, 1999]. Наиболее авторитетные словари, например, Oxford English Dictionary, Cambridge Dictionary и Macmillan English Dictionary for Advanced Learners, активно используют корпусы для обновления своих статей [Atkins et al., 2008]. Такие специализированные собрания, как корпус медицинских текстов или юридических документов, применяются для изучения терминологии. В настоящее время выявлены особенности употребления медицинских терминов в профессиональной коммуникации [Lei et al., 2016], создан список часто употребляемых академических слов и фраз, относящихся к сельскому хозяйству [Martínez et al., 2009], химии [Valipouri et al., 2013], защите окружающей среды [Liu et al., 2015], машиностроению [Chang, 2023] и т. д. Однако, несмотря на активное использование корпусной лингвистики для изучения

терминологии различных областей знания, до настоящего момента в отечественной и зарубежной лингвистике отсутствуют комплексные исследования, посвященные применению корпусных методов для анализа терминологии ИИ. Существующие работы в основном сосредоточены на теоретическом описании терминов ИИ или их лексикографической фиксации [Козловская и др., 2023; Термины …, 2024; Шалимова, 2024; Sabahuddin, 2024; Resslerová, 2024; A global …, 2023], тогда как корпусный подход, позволяющий изучать термины в их естественном контексте употребления, остается практически невостребованным.

2. Материал, методы, обзор = Material, methods, review

Цель исследования заключается в определении ключевых моделей образования терминов, а также в изучении того, как складываются орфографические и стилистические нормы использования терминологических единиц ИИ в русском языке. Ведется разработка алгоритма исследования корпусных данных; осуществляется выделение ядерных терминов на основе частотности использования в исследуемом материале; ставится задача установить наиболее характерные коллокации с этими терминами.

В настоящей работе объектом исследования выступают терминологические единицы ИИ в русском языке, поскольку анализ этого пласта лексики в академическом письменном дискурсе позволяет более точно определить значение и контекст использования терминов, что особенно важно в междисциплинарных исследованиях [Петрова и др., 2022]; выявить основные тенденции и изменения в использовании терминов, которые могут указывать на развитие научной мысли или появление новых направлений исследования в области ИИ [Suleimanova et al., 2024].

Источником фактического материала послужили 12 сборников аналитических материалов, посвященных отрасли ИИ в России и мире (альманах «Искусственный интеллект»). Для обеспечения репрезентативности выборки были выбраны 150 статей, охватывающих широкий спектр областей искусственного интеллекта, таких как обучение с подкреплением, ИИ в здравоохранении, нормативно-правовая база, аппаратное обеспечение, предиктивная аналитика, системы поддержки принятия решений, компьютерное зрение и обработка естественного языка. Для обработки текстов использовался корпусный менеджер AntConc (версия 4.3.1), функция частотного анализа которого позволила выделить наиболее часто встречающиеся в текстах статей ядерные термины-слова и термины-словосочетания, образованные на их основе. Контент-анализ, предполагающий содержательную интерпретацию данных, способствовал выявлению особенностей образования терминов-словосочетаний, вариативности орфографических

и стилистических норм, а также определению контекстуального использования анализируемых терминов в зависимости от тематики статей.

С развитием корпусной лингвистики и увеличением доступности электронных текстовых данных возникла необходимость в специализированных инструментах для анализа больших массивов текстов, так называемых ***корпусных менеджерах*** [WordSmith Tools, SketchEngine, tlCorpus, TextSTAT Google Books Ngram Viewer и др.]. Некоторые из этих инструментов являются платными, в то время как другие требуют определенных технических знаний для успешного использования. AntConc выделяется как мощный бесплатный инструмент для анализа больших массивов текстовых данных, разработанный профессором Л. Антони, директором Центра обучения английскому языку в области науки и техники Школы науки и техники университета Васеда (Япония). Среди его основных преимуществ — доступность для пользователей с различным уровнем подготовки, интуитивно понятный интерфейс, широкий спектр функций для текстового анализа (включая конкорданс, коллокации, частотный анализ, кластеризацию, создание частотных словарей и выделение ключевых слов), поддержка различных текстовых форматов, возможность гибкой настройки параметров анализа и кросс-платформенная совместимость [Anthony, 2011; Сулейманова и др., 2022].

Методика работы с платформой AntConc для изучения основных характеристик терминов состояла из 6 этапов:

Этап 1. Сбор эмпирического материала в цифровом формате, состоящего из 150 статей альманаха «Искусственный интеллект».

Этап 2. Подготовка корпуса к исследованию (очистка текста от метаданных и изображений; сохранение текстов в формате .txt, кодирование в UTF-8 для корректного отображения символов).

Этап 3. Импортирование данных в AntConc:

1) загрузка текстов в программу посредством инструмента Corpus Manager для дальнейшего использования в качестве основного корпуса (Target Corpus);

2) импортирование эталонного корпуса (Reference Corpus), включающего 1540 терминологических единиц русскоязычной части онлайн-глоссария по ИИ [aiterms.ru]. Онлайн-глоссарий создан на основе толковых, этимологических и терминологических словарей, национальных и международных стандартов по ИИ, тематических интернет-ресурсов.

На рисунке представлены подготовленные для проведения исследования основной и эталонный корпусы (рис. 1).

Этап 4. Сравнение эталонного корпуса (Reference Corpus) с основным (Targer Corpus) при помощи инструмента Keyword List (рис. 2). Данный

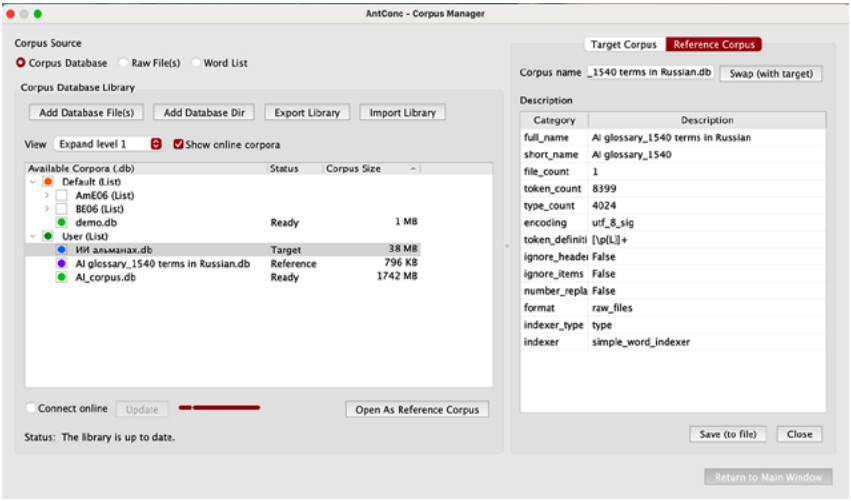

Рис. 1. Создание основного и эталонного корпусов в AntConc

инструмент позволяет рассчитать, какие слова в исследуемом корпусе обладают наибольшей частотностью по сравнению с эталонным корпусом. Поскольку в качестве эталонного корпуса используются термины ИИ, то на выходе мы получаем список высокочастотных терминологических единиц, которые можно охарактеризовать как ядерные термины или содержащие наиболее важные значения и являющиеся ключевыми для понимания темы или области знания [Винокурова, 2016]. Они служат основой для дальнейшего сужения и конкретизации значений, связанных с этой темой, а также сопровождаются производными терминами, которые могут иметь дополнительные значения или специфические контексты [Кондратюкова, 2012].

Этап 5. Отбор n-грамм посредством инструмента Clusters для определения словосочетаний с ядерными терминами справа и слева от искомого слова. На рисунке представлены коллокации, образованные на основе ядерного термина *данные* (рис. 3). Чтобы учесть все словоформы ядерного термина, варьируемая часть слова заменена на звездочку (*). Сортировка представлена по частотности употребления; количество элементов в кластере равно 3, так как средняя длина термина ИИ равна 2 или 3 терминоэлементам [Винокурова, 2016].

Следует отметить, что на этапах 4 и 5 цифровых статистических исследований текста требуется интеграция качественного контент-анализа, поскольку не все отобранные программой слова и словосочетания будут относиться к терминологической лексике в области ИИ.

[Научный диалог = Nauchnyi dialog = Nauchnyy dialog, 14(7), 2025]
[ISSN 2225-756X, eISSN 2227-1295]

Этап 6. Интерпретация результатов исследования.

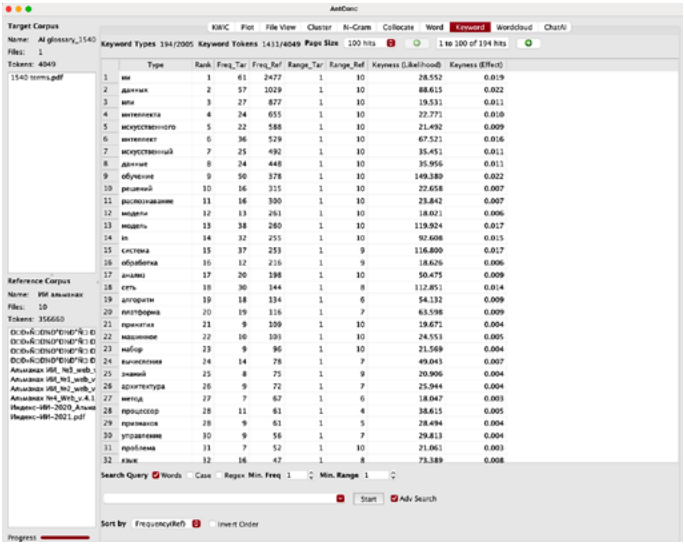

Рис. 2. Результат сравнения основного корпуса исследования с эталонным

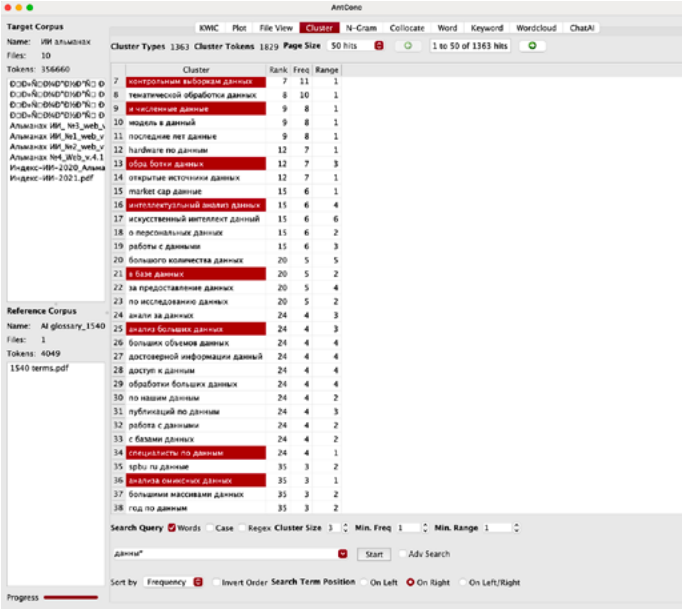

Рис. 3. Коллокации на основе ядерного термина *данные*

3. Результаты и обсуждение = Results and Discussion

В результате исследования, проведенного при помощи инструмента Keyword сервиса AntConc (версия 4.3.1) и последующего качественного контент-анализа, было выявлено 100 ядерных терминов. Наиболее высокочастотные приведены в таблице (табл. 1).

Таблица 1

Ядерные термины ИИ в русском языке
с высокой частотностью употребления

Термин	Частотность употребления	Термин	Частотность употребления
данные	1829	метод	489
система	1695	платформа	339
интеллект	1503	цифровой	303
обучение	1271	управление	300
модель	852	устройство	288
решение	833	процессор	280
сеть	770	функция	275
алгоритм	696	архитектура	245
анализ	692	вычисления	225
распознавание	592	классификация	190
обработка	550	набор	187
язык	530	признак	113
зрение	511	библиотека	99

На основании анализа представленных терминов можно выделить несколько ключевых направлений искусственного интеллекта, демонстрирующих наиболее динамичное развитие: обработка естественного языка (Natural Language Processing, NLP), машинное обучение (Machine Learning), генеративные модели, рапознавание образов и распознавание голоса, эмоциональный интеллект, обработка и защита персональных данных, классификация и кластеризация.

Большинство исследуемых терминов-словосочетаний имеют двухкомпонентную структуру и формируются в соответствии со следующими моделями:

1) прил. + сущ.: *тензорная сеть, сиамская сеть, сверточная сеть, рекуррентная сеть, глубокая сеть; открытая платформа, аппаратная платформа, геномная платформа, облачная платформа, цифровая платформа, краудсорсинговая платформа; коллективный интеллект, окружающий интеллект, поэтапное обучение, инкрементный алгоритм, нейро-*

OPEN ACCESS [Научный диалог = Nauchnyi dialog = Nauchnyy dialog, 14(7), 2025]
[ISSN 2225-756X, eISSN 2227-1295]

сетевая модель, скоринговая модель, компьютерное зрение, адаптивное управление, эталонная архитектура;

2) сущ. + сущ.: *анализ данных, хост-система, алгоритм оптимизации, алгоритм обучения, алгоритм word2vec, обучение модели, модель языка, распознавание речи, распознавание лиц, распознавание образов, распознавание текста, управление жестами, управление дронами, управление диалогом, управление рисками, валидация данных, архитектура агента, архитектура фон Неймана, лес решений;*

3) сущ. + предлог + сущ.: *обучение без учителя, обучение с учителем, обучение по сходству, поиск по графу, пересечение по объединению, площадь под кривой, сжатие без потерь, текст в речь;*

4) аббревиатура + сущ. (с предлогом или без): *ИИ-модель, ИИ-агент, МО с подкреплением, система ИИ, IOT-устройство, NLP-устройство, библиотека cuDNN, библиотека ZenDNN, свойство NP, алгоритм BLEU.*

Сужение и конкретизация значения исходного двухкомпонентного термина происходят за счет добавления к этому термину прилагательного или существительного, что приводит к уточнению области применения или аспекта, связанного с исходным термином:

1) прил. + двухкомпонентный термин-словосочетание: ***объяснимый*** *искусственный интеллект,* ***субсимвольный*** *искусственный интеллект, автономная система управления,* ***рекуррентная*** *нейронная сеть, долгая краткосрочная память,* ***сквозные*** *цифровые технологии,* ***случайный*** *лес решений;*

2) сущ. + двухкомпонентный термин-словосочетание: ***анализ*** *больших данных,* ***алгоритм*** *машинного обучения,* ***устройство*** *распознавания речи,* ***технология*** *распознавания голоса,* ***архитектура*** *вычислительной машины,* ***оптимизация*** *множетсвенного роя,* ***система*** *распознавания речи.*

Рассмотрим термин *большие данные* на примерах сужения фокуса на конкретном аспекте работы с данными (они выступают объектом номинации):

— *большие данные* — общий термин, обозначающий огромные объемы структурированных и неструктурированных данных, которые трудно обрабатывать традиционными методами. Этот термин охватывает сами данные, их характеристики (объем, скорость, разнообразие) и технологии их хранения;

— *анализ больших данных* представляет собой специализированный процесс, направленный на исследование, обработку и интерпретацию значительных объемов информации с целью извлечения ценной информации, выявления закономерностей и поддержки принятия обоснованных решений. В данном контексте внимание сосредотачивается на методах и

инструментах, применяемых для анализа, включая машинное обучение, статистические подходы и визуализацию данных;

— *обработка больших данных*, в свою очередь, акцентирует внимание на процедурах сбора, очистки, преобразования массивов данных и управления ими с целью их подготовки к последующему анализу или использованию. Здесь основное внимание уделяется техническим аспектам, таким как использование распределенных систем (например, Hadoop и Spark), алгоритмов параллельной обработки и методов оптимизации, которые обеспечивают эффективное управление большими объемами информации.

Следовательно, применение терминов *анализ* или *обработка* позволяет акцентировать внимание на специфических аспектах взаимодействия с большими данными. Такая терминологическая конкретизация способствует более точному описанию процессов, связанных с технической подготовкой данных, и их четкому разграничению с другими этапами работы, такими как визуализация, хранение, управление данными, разработка алгоритмов машинного обучения и принятие решений на основе полученных результатов.

Наряду с полным названием термина *искусственный интеллект* (1305 употреблений) в текстах используется его аббревиатура *ИИ* (2477 употреблений); аббревиатура *МО* встречается в текстах 40 раз, а *машинное обучение* — 320 раз. Такое распределение употреблений свидетельствует о том, что аббревиатуры активно используются в профессиональной и научной коммуникации, особенно в случаях, когда термин упоминается многократно. Однако полные названия сохраняют свою значимость для обеспечения ясности и точности, особенно в случаях, когда важно избежать двусмысленности или термин вводится впервые. Это также отражает тенденцию к балансу между лаконичностью и полнотой выражения в научных и технических текстах.

Русскоязычный термин или аббревиатура может сопровождаться данным в скобках термином и / или аббревиатурой на английском языке: обработка естественного языка (NLP), МО (machine learning), мета-обучение с подкреплением (meta-RL), большие данные (big data), графический процессор (GPU), сверточные нейронные сети (Convolutional Neural Network, CNN), генеративно-состязательные сети (GAN). Такая практика не только способствует однозначности и ясности изложения, обеспечивает интеграцию русскоязычной терминологии в международный научный контекст, но и помогает избежать терминологической путаницы, связанной с различиями в языковых и профессиональных традициях.

Гибридное написание также способствует обеспечению универсальности и однозначности терминов. Термины могут иметь английские

вкрапления: *8-слойные LSTM с механизмом внимания*, *модель word2vec, Q-обучение, подход end2end.* Как правило, аббревиатуры, часто используемые в профессиональной коммуникации, в составе сложного термина сохраняют английский вариант написания: *NLP-алгоритм, NLP технологии, NLP модель; модель LaMDA, процессор для ML, проектировщик GPU, сервер CPU/GPU, метрика BLEU, ЦП (CPU).*

Ряд терминологических единиц не имеют переводного аналога в русском языке и пишутся латиницей: речевые помощники *Siri, Alexa, Cortana, Amazon Echo*, библиотеки *Python, PyTorch, NumPy*, нейросети *AlexNet, DALL-E* и др. Использование в оригинальной форме имен собственных, названий продуктов, технологий или программных инструментов, которые изначально были созданы и названы на английском языке, обусловлено необходимостью сохранения узнаваемости, а также избежания искажения смысла или функциональной принадлежности.

Встречаются случаи транслитерации при наличии аналога в русском языке: *dataset — датасет / набор данных, autoencoder — автоэнкодер / автокодировщик; crowdsourcing — краудсорсинг / распределение задач, blockchain — блокчейн / технологии распределенного реестра.* Такая вариативность в использовании терминов отражает влияние англоязычной терминологии на русскоязычный научный и профессиональный дискурс, а также стремление к сохранению оригинального звучания терминов, что способствует их узнаваемости и унификации в международном контексте. В то же время наличие русскоязычных аналогов демонстрирует попытки адаптации и локализации терминологии.

Некоторые термины, несмотря на наличие русскоязычных аналогов, чаще используются в их английском варианте: *data scientist* и *data analyst* встречаются в текстах 19 раз, а их русские аналоги *специалист по данным* и *дата аналитик / аналитик-данных* — 1 и 5 раз соответственно. Кроме того, вместо русскоязычного термина может употребляться английская аббревиатура. Так, термин *генеративно-состязательные сети* часто замещается аббревиатурой *GAN*, нередко с добавлением русского окончания множественного числа: *Разумеется, такую задачу, как Style Transfer, не обошли стороной современные и нынче популярные* ***GAN'ы*** (Д. Нехаев, И. Лаптев «Computer Vision / Технологии», 2019).

Подчеркнем, что главной чертой терминосистемы ИИ в русском языке является наличие большого количества англицизмов [Брейтер, 1997, с. 132], которые заимствуются в связи с отсутствием соответствующего названия в языке-реципиенте. Однако тексты альманаха нельзя считать перенасыщенными англицизмами, написанными латинским шрифтом. Соотношение терминов, представленных кириллицей и латиницей, состав-

ляет 3 : 1; тем не менее существуют случаи, когда при наличии аналогов в русском языке используется исключительно английский вариант (*fuzzy systems*, *data mining*, *cognitive science*, *transfer learning*): *Основные работы по* ***transfer learning*** *сейчас ведутся в области донастройки и оптимизации (finetuning) обученных моделей таким образом, чтобы они могли быть перенесены в новые среды без потерь, а также «обогащении» обучающих выборок* (Н. Гутенева «RL — обзор основных технологий», 2020).

Проблема выбора между дефисным и раздельным написанием слов остается актуальной и не имеет четко установленного решения в ряде случаев. В ходе анализа было выявлено шесть различных вариантов написания термина *чатбот* в статьях различных авторов: *chat бот, chat-бот, чатбот, чат-бот, chatbot, chat bot, chat-bot*. Наиболее распространенной вариацией является *чат-бот*.

Аналогична ситуация с терминами, в составе которых присутствует аббревиатура ML (машинное обучение). В исследуемом корпусе текстов обнаружены различные варианты написания таких сложных слов, как *ML-specialist, ML Specialist и ML-специалист*.

В процессе анализа особенностей грамматического освоения заимствованных слов в русском языке было сделано интересное наблюдение. Прилагательные, как правило, образуются посредством гибридного присоединения русских аффиксов к транслитерированным английским терминам [Ермакова, 2001]. Например, английское существительное *software* находит в русском языке такие эквиваленты, как *программное обеспечение*, *ПО*, *софт*, от последней основы образуются прилагательные *софтовый* и *софтверный*; существительное *hardware* представлено в русском языке как *оборудование* или *железо* (в жаргонном употреблении), с соответствующим образованием прилагательного *хардверный*; от существительного *хакатон* (*hackathon)* образовано прилагательное *хакатонный.*

Для текстов альманаха характерен ряд стилистических особенностей. Во-первых, преобладание терминологической лексики свидетельствует о технической или специализированной направленности контекста, что предполагает целевую аудиторию, свободно владеющую соответствующей терминологией. Во-вторых, наличие небольшого процента жаргонизмов, таких как *железо, железка* и *фича*, с одной стороны, придает тексту неформальный и разговорный оттенок, что может быть интерпретировано как стремление авторов к установлению более близкого контакта с читателем, а с другой — создает эффект принадлежности к профессиональной среде, ориентированной на «своих». В-третьих, смешение различных стилистических регистров создает контраст, который способствует динамичности текста и делает его более живым. Такое сочетание формального

и неформального стилей отражает стремление к балансу между научной строгостью и доступностью изложения, что характерно для современных текстов, ориентированных на профессиональную, но широкую аудиторию. Приведем несколько примеров: *NVIDIA, почувствовав, что трон зашатался, вкладывает фантастические усилия в библиотеки ускорения нейросетей и новое* ***железо*** (Д. Ватолин «Аппаратное ускорение глубоких нейросетей в 2021», 2021); *Все это означает появление железок, обучение на которых будет относительно быстрым, но стоить которые будут дорого, что естественным образом приводит к идее разделять время использования этой дорогой* ***железки*** *между исследователями* (Д. Ватолин «Аппаратное ускорение глубоких нейросетей в 2021», 2021); *Однако просто брать нужные* ***фичи*** *и делать дообучение с датасета на датасет может быть недостаточно* [И. Захаркин «Вижу, значит существую: обзор Deep Learning в Computer Vision», 2019].

4. Заключение = Conclusions

Корпусная лингвистика обладает значительным потенциалом для анализа и изучения языковой системы, особенно в контексте быстро развивающейся области искусственного интеллекта (ИИ). Применение инструментов корпусной лингвистики, таких как AntConc, позволило выявить ключевые терминологические единицы, отражающие актуальные тренды и новшества в данной области. В проанализированных текстах был представлен всесторонний обзор терминологии ИИ в русском языке, выявлены основные словообразовательные модели, а также оценены стилистические и орфографические нормы употребления терминов.

Результаты исследования показывают, что терминология, связанная с искусственным интеллектом, продолжает эволюционировать и адаптироваться к требованиям современного научного дискурса. Обнаружено, что многие термины и словосочетания, заимствованные из английского языка, еще не имеют устоявшихся норм употребления в русском языке, что создает определенные трудности для исследователей и практиков. Это свидетельствует о необходимости дальнейшей работы по стандартизации терминов для улучшения понимания и коммуникации в профессиональной среде.

Кроме того, установлено, что англицизмы активно внедряются в русский язык, что указывает на глобализацию научной терминологии и влияние английского языка как lingua franca. Проблема, связанная с использованием различных форм написания терминов (латиница и кириллица), а также наличие гибридных форм требуют внимания и более четкой регламентации на пути к гармонизации научного языка.

Однако, несмотря на значительные достижения в области корпусной лингвистики, обнаружен ряд ограничений, связанных с использованием платформы AntConc. Высокое содержание «шума» в результатах, необходимость ручной фильтрации и ограниченные возможности по сведению словоформ к единой лемме ставят перед исследователями дополнительные вызовы. Будущие разработки в области корпусных технологий, вероятно, позволят преодолеть эти трудности и сделать анализ данных более эффективным и точным.

Важным направлением дальнейших исследований является углубленное изучение функционально-стилистической дифференциации терминов, что позволит установить более четкие нормы их употребления в контексте научного и профессионального дискурса. Также актуальной остается задача создания специализированных корпусов, которые будут отражать не только терминологию, но и контексты использования терминов, что поможет выявить закономерности и тенденции в их функционировании.

Заявленный вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.	**Contribution of the authors:** the authors contributed equally to this article.
Авторы заявляют об отсутствии конфликта интересов.	**The authors declare no conflicts of interests.**

Литература

1. *Архангельский Т. А.* Интернет-корпуса финно-угорских языков России / Т. А. Архангельский // Ежегодник финно-угорских исследований. — 2019. — Т. 13. — № 3. — С. 528—537. — DOI: 10.35634/2224-9443-2019-13-3-528-537.

2. *Брейтер М. А.* Англицизмы в русском языке : история и перспективы : пособие для иностр. студентов-русистов / М. А. Брейтер. — Москва : Диалог-МГУ, 1997. — 156 с.

3. *Винокурова Т. Н.* Структурные особенности терминологии искусственного интеллекта в английском языке / Т. Н. Винокурова // Международный научно-исследовательский журнал. — 2016. — № 10—3 (52). — С. 14—23. — DOI:10.18454/IRJ.2016.52.024.

4. *Ермакова О. И.* Особенности компьютерного жаргона как специфической подсистемы русского языка / О. И. Ермакова // Диалог. — 2001. — С. 173.

5. *Захаров В. П.* Корпусная лингвистика / В. П. Захаров. — Санкт-Петербург : Санкт-Петербургский государственный университет, 2005. — 48 с. — ISBN 978-5-288-05997-1.

6. *Козлова Н. В.* Лингвистические корпуса : определение основных понятий и типология / Н. В. Козлова // Вестник НГУ. Лингвистика и межкультурная коммуникация. — 2013. — № 1. — С. 79—88.

7. *Козловская Н. В.* Транстерминологизация в сфере искусственного интеллекта : к постановке вопроса о субтерминологии / Н. В. Козловская, А. С. Мусаева, Ю. В. Сложеникина // Art Logos. — 2023. — № 3 (24). — С. 98—118. — DOI: 10.24224/2227-1295-2025-14-4-9-37.

8. *Кондратюкова Л. К.* Заимствования и интернационализмы в терминологии английской компьютерной техники / Л. К. Кондратюкова // Динамика систем, механизмов и машин. — 2012. — № 4. — С. 155—158.

9. *Кононенко А. П.* Лингвистический потенциал компьютерных технологий в современной филологии / А. П. Кононенко, Л. А. Недосека // Гуманитарные и социальные науки. — 2023. — Т. 97. — № 2. — С. 50—54. — DOI: 10.18522/2070-1403-2023-97-2-50-54.

10. *Ляшевская О. Н.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) / О. Н. Ляшевская, С. А. Шаров. — Москва : Азбуковник, 2009. — 1090 с. — ISBN 978-5-91172-024-7.

11. *Петрова И. М.* Современные цифровые технологии в лингвистических исследованиях : учеб. пособие для обучающихся по направлению «Лингвистика» / И. М. Петрова, А. М. Иванова, В. В. Никитина. — Москва : Языки Народов Мира, 2022. — 259 с. — ISBN 978-5-6048046-8-1.

12. *Плунгян В. А.* Зачем нужен Национальный корпус русского языка? Неформальное введение / В. А. Плунгян // Национальный корпус русского языка : 2003—2005. — Москва : Индрик, 2005. — С. 6—20.

13. *Сулейманова О. А.* Методика лингвистического исследования как актуальный раздел современной научной публикации / О. А. Сулейманова, А. Б. Гулиянц // Вестник МГПУ. Серия : Филология. Теория языка. Языковое образование. — 2022. — № 4 (48). — С. 89—101. — DOI: 10.25688/2076-913X.2022.48.4.07.

14. *Термины* и понятия искусственного интеллекта в лингвистическом освещении / А. С. Мусаева, Ю. В. Сложеникина, Л. М. Гареева. — Москва : Спутник+, 2024. — 193 с. — ISBN 978-5-9973-6887-6.

15. *Шалимова П. А.* К вопросу о терминах и неологизмах в сфере искусственного интеллекта и нейросетей / П. А. Шалимова // Общество, экономика, культура : стратегии развития. Материалы XV Всероссийской научно-практической конференции. — 2024. — С. 218—223.

16. *A global* taxonomy of interpretable AI : unifying the terminology for the technical and social sciences / M. Graziani, L. Dutkiewicz, D. Calvaresi // Artificial Intelligence Review. — 2023. — Vol. 56. — № 4. — Pp. 347—3504. — DOI: 10.1007/s10462-022-10256-8.

17. *Aarts J.* Corpus Linguistics / J. Aarts, W. Meij. — Amsterdam : Rodopi, 1984. — 229 p.

18. *Abercrombie D.* Studies in Phonetics and Linguistics / D. Abercrombie — London : Oxford University Press, 1965. — 151 p.

19. *Corpus* Linguistics and Corpus-Based Research and Its Implication in Applied Linguistics : A Systematic Review / A. M. S. Al-Hamzi, A. Gougui, Y. Sari Amalia, T. Suhardijanto // PAROLE : Journal of Linguistics and Education. — 2020. — Vol. 10. — № 2. — Pp. 176—181.

20. *Allwood J.* Multimodal corpora / J. Allwood // Corpus Linguistics. An International Handbook. — Berlin : de Gruyter, 2009. — Pp. 207—225.

21. *Anthony L.* AntConc : A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit / L. Anthony // IWLeL 2004 : An Interactive Workshop on Language e-Learning. — 2011. — Pp. 7—13.

22. *Assunção C.* Entries on the History of Corpus Linguistics / C. Assunção, C. S. Araújo // Linha D Água. — 2019. — Vol. 32. — № 1. — Pp. 39—57. — DOI: 10.11606/issn.2236-4242.v32i1p39-57.

23. *Atkins B. T. S.* The Oxford guide to practical lexicography / B. T. S. Atkins, M. Rundell. — Oxford : Oxford university press, 2008. — 540 p.

24. *Bataillon L. J.* Hugues de Saint-Cher († 1263), bibliste et théologien / L. J. Bataillon, G. Dahan, P.-M. Gy. — Turnhout : Brepols, 2004. — 520 p.

25. *Biber D.* On the exploitation of computerized corpora in variation studies / D. Biber, E. Finegan // English corpus linguistics : Studies in honour of Jan Svartvik. — London : Longman, 1991. — Pp. 204—220.

26. *Boas F.* Handbook of American Indian Languages / F. Boas. — Cambridge : Cambridge University Press, 2013. — 570 p.

27. *Boulton A.* Using Corpora in Language Teaching, Learning and Use / A. Boulton, C. Landure // Recherche et pratiques pédagogiques en langues de spécialité. — 2016. — Vol. 35. — № 2. — Pp. 67—72. — DOI: 10.4000/apliut.5433.

28. *Casson L. F.* A Fourteenth Century Concordance to the Vulgate / L. F. Casson // Libri. — 1960. — Vol. 10. — № 2. — Pp. 111—128. — DOI: 10.1515/libr.1960.10.2.111.

29. *Chang L.* A Corpus-Based Mechanical Engineering Academic Word List / L. Chang // International Journal of TESOL Studies. — 2023. — Vol. 5. — № 3. — Pp. 126—142. — DOI: 10.58304/ijts.20230310.

30. *Chomsky N.* Quine's empirical assumptions / N. Chomsky // Synthese. — 1968. — Vol. 19. — Pp. 53—68. — DOI: 10.1007/BF00568049.

31. *Dash N. S.* History, Features, and Typology of Language Corpora / N. S. Dash, S. Arulmozi. — Springer : [b. i.], 2018. — 311 p. — DOI: 10.1007/978-981-10-7458-5_15.

32. *Dernoncourt F.* PubMed 200k RCT : a Dataset for Sequential Sentence Classification in Medical Abstracts / F. Dernoncourt, J. Y. Lee // Proceedings of the 8th International Joint Conference on Natural Language Processing. — Taipei : IEEE Signal Processing Society. — 2017. — Pp. 308—313.

33. *Doğan R. I.* An improved corpus of disease mentions in PubMed citations / R. I. Doğan, Z. Lu // Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012). — Montreal : Association for Computational Linguistics. — 2012. — Pp. 91—99.

34. *Eaton H.* Semantic frequency list for English, French, German, and Spanish ; a correlation of the first six thousand words in four single-language frequency lists / H. Eaton. — Chicago : Chicago University Press, 1940. — 440 p.

35. *Francis W. N.* Brown Corpus Manual : Manual of information to accompany. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers / W. N. Francis, H. Kucera. — Providence : Brown University, 1964. — 467 p.

36. *Grammar* of Spoken and Written English / D. Biber, S. Johansson, G. Leech, S. Conrad, E. Finegan. — Longman Harlow : Pearson Education Limited, 1999. — 1204 p.

37. *Guietti P.* Hermeneutic of Aquinas's Texts : Notes on the Index Thomisticus / P. Guietti // The Thomist : A Speculative Quarterly Review. — 1993. — Vol. 57. — № 4. — Pp. 667—686. — DOI: 10.1353/tho.1993.0006.

38. *Harris Z. S.* Structural Linguistics / Z. S. Harris. — Chicago : University Of Chicago Press, 1960. — 384 p.

39. *Hill J.* LTP Dictionary of Selected Collocations / J. Hill, M. Lewis. — Hove : Language Teaching Publications, 1997. — 288 p.

40. *Hunston S.* Pattern Grammar / S. Hunston, G. Francis. — Amsterdam : John Benjamins Publishing, 2000. — 288 p.

41. *Hyland K.* As it can be seen : Lexical bundles and disciplinary variation / K. Hyland // English for Specific Purposes. — 2008. — Vol. 27. — Pp. 4—21. — DOI: 10.1016/j.esp.2007.06.00.

42. *Johansson S.* Some aspects of the development of corpus linguistics in the 1970-s and 1980-s / S. Johansson // Corpus Linguistics: An International Handbook. — Berlin : De Gruyter, 2009. — Pp. 33—53.

43. *Kuebler S.* Corpus Linguistics and Linguistically Annotated Corpora / S. Kuebler, H. Zinsmeister. — London : Bloomsbury Publishing, 2015. — 320 p.

44. *Lei L.* A new medical academic word list : A corpus-based study with enhanced methodology / L. Lei, D. Liu // Journal of English for Academic Purposes. — 2016. — Vol. 22. — Pp. 42—53. — DOI: 10.1016/j.jeap.2016.01.008.

45. *Liu J.* A corpus-based environmental academic word list building and its validity test / J. Liu, L. Han // English for Specific Purposes. — 2015. — Vol. 39. — № 1. — Pp. 1—11. — DOI: 10.1016/j.esp.2015.03.001.

46. *Martínez I. A.* Academic vocabulary in agriculture research articles : a corpus-based study / I. A. Martínez, S. C. Beck, C. B. Panza // English for Specific Purposes. — 2009. — Vol. 28. — № 3. — Pp. 183—198. — DOI: 10.1016/j.esp.2009.04.003.

47. *McEnery T.* Corpus Linguistics : Method, Theory and Practice / T. McEnery, A. Hardie. — Cambridge : Cambridge University Press, 2012. — 312 p.

48. *Mcgillivray B.* The Index Thomisticus Treebank Project : Annotation, Parsing and Valency Lexicon / B. Mcgillivray, M. Passarotti, P. Ruffolo // Traitement Automatique des Langues. — 2009. — Vol. 50. — № 2. — Pp. 103—127.

49. *O'Keeffe A.* Routledge handbook of corpus linguistics / A. O'Keeffe, M. McCarthy. — London : Routledge, 2010. — 682 p.

50. *Partington A. U*sing corpora in discourse analysis / A. Partington, A. Marchi // The Cambridge Handbook of English Corpus Linguistics. — Cambridge : Cambridge University Press, 2015. — Pp. 216—234.

51. *Pawley A.* Two puzzles for linguistic theory : Nativelike selection and nativelike frequency / A. Pawley, F. H. Syder // Language and Communication. — London : Longman. — 1983. — Pp. 191—226.

52. *Resslerová V.* La terminologie du domaine de l'intelligence artificielle : néologie et pluridisciplinarité / V. Resslerová // Studia Romanistica. — 2024. — Vol. 24. — № 2. — Pp. 59—71. — DOI: 10.15452/SR.2024.24.0012.

53. *Rockwell G.* The Index Thomisticus as a Digital Humanities Big Data Project / G. Rockwell, M. Passarotti // Umanistica Digitale. — 2019. — № 5. — Pp. 13—34. — DOI: 10.6092/issn.2532-8816/8575.

54. *Sabahuddin A.* AI Lexica : Exploring the Vocabulary of Artificial Intelligence / A. Sabahuddin // Journal of Emerging Technologies and Innovative Research. — 2024. — Vol. 11. — Issue 4. — Pp. 123—137.

55. *Scott M.* Textual Patterns : Key words and corpus analysis in language education / M. Scott, C. Tribble. — Amsterdam : John Benjamins Publishing, 2006. — 203 p.

56. *Selivan L.* Corpus Linguistics and Vocabulary Teaching / L. Selivan // Demystifying Corpus Linguistics for English Language Teaching. — Springer. — 2023. — Pp. 139—161. — DOI: 10.1007/978-3-031-11220-1_8.

57. *Sinclair J.* Looking up : an account of the COBUILD Project in lexical computing / J. Sinclair. — London and Glasgow : Collins ELT, 1987. — 182 p.

58. *Sinclair J.* Corpus, Concordance, Collocation / J. Sinclair. — Oxford : University of Oxford, 1991. — 179 p.

59. *Stefanowitsch A.* Corpus linguistics : A guide to the methodology / A. Stefanowitsch. — Berlin : Language Science Press, 2020. — 510 p.

60. *Stefchov E.* Towards Constructing a Corpus for Studying the Effects of Treatments and Substances Reported in PubMed Abstracts / E. Stefchov, G. Angelova, P. Nakov // Lecture Notes in Computer Science. — 2018. — Vol. 11089. — Pp. 115—125. — DOI: 10.1007/978-3-319-99344-7_11.

61. *Suleimanova O. A.* Anthropocentrical Turn in Linguistics Through the Digital Lens : Evidence from Analyses of Russian Mnemonic Verbs / O. A. Suleimanova, I. V. Tivyaeva // Journal of Siberian Federal University. Humanities and Social Sciences. — 2024. — Vol. 17. — № 5. — Pp. 847—861.

62. *Valipouri L.* A corpus-based study of academic vocabulary in chemistry research articles / L. Valipouri, H. Nassaji // Journal of English for Academic Purposes. — 2013. — Vol. 12. — № 4. — Pp. 248—263. — DOI: 10.1016/j.jeap.2013.07.001.

Статья поступила в редакцию 20.02.2025,
одобрена после рецензирования 29.07.2025,
подготовлена к публикации 19.08.2025.

References

A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial Intelligence Review, 56 (4):* 347—3504. DOI: 10.1007/s10462-022-10256-8.

Aarts, J., Meij, W. (1984). *Corpus Linguistics.* Amsterdam: Rodopi. 229 p.

Abercrombie, D. (1965). *Studies in Phonetics and Linguistics.* London: Oxford University Press. 151 p.

Allwood, J. (2009). Multimodal corpora. In: *Corpus Linguistics. An International Handbook.* Berlin: de Gruyter. 207—225.

Anthony, L. (2011). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. *IWLeL 2004: An Interactive Workshop on Language e-Learning.* 7—13.

Arkhangelsky, T. A. (2019). Internet corpus of the Finno-Ugric languages of Russia. *Yearbook of Finno-Ugric studies, 13 (3):* 528—537. DOI: https://doi.org/10.35634/2224-9443-2019-13-3-528-537. (In Russ.).

Assunção, C. (2019). Entries on the History of Corpus Linguistics. *Linha D Água, 32 (1):* 39—57. DOI: 10.11606/issn.2236-4242.v32i1p39-57.

Atkins, B. T. S., Rundell, M. (2008). *The Oxford guide to practical lexicography.* Oxford: Oxford university press. 540 p.

Bataillon, L. J., Dahan, G. (2004). *Hugues de Saint-Cher († 1263), bibliste et théologien.* Turnhout: Brepols. 520 p.

Biber, D. (1991). On the exploitation of computerized corpora in variation studies. In: *English corpus linguistics: Studies in honour of Jan Svartvik.* London: Longman. 204—220.

Boas, F. (2013). *Handbook of American Indian Languages.* Cambridge: Cambridge University Press. 570 p.

Boulton, A., Landure, C. (2016). Using Corpora in Language Teaching, Learning and Use. *Recherche et pratiques pédagogiques en langues de spécialité, 35 (2):* 67—72. DOI: 10.4000/apliut.5433.

Breiter, M. A. (1997). *Anglicisms in the Russian language: history and prospects: a handbook for foreigners. students of Russian studies.* Moscow: Dialog-MSU. 156 p. (In Russ.).

Casson, L. F. (1960). A Fourteenth Century Concordance to the Vulgate. *Libri, 10 (2):* 111—128. DOI: 10.1515/libr.1960.10.2.111.

Chang, L. (2023). A Corpus-Based Mechanical Engineering Academic Word List. *International Journal of TESOL Studies, 5 (3):* 126—142. DOI: 10.58304/ijts.20230310.

Chomsky, N. (1968). Quine's empirical assumptions. *Synthese, 19:* 53—68. DOI: 10.1007/BF00568049.

Corpus Linguistics and Corpus-Based Research and Its Implication in Applied Linguistics: A Systematic Review. *PAROLE: Journal of Linguistics and Education, 10 (2):* 176—181.

Dernoncourt, F. (2017). PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. In: *Proceedings of the 8th International Joint Conference on Natural Language Processing.* Taipei: IEEE Signal Processing Society. 308—313.

Doğan, R. I., Lu, Z. (2012). An improved corpus of disease mentions in PubMed citations. In: *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012).* Montreal: Association for Computational Linguistics. 91—99.

Eaton, H. (1940). *Semantic frequency list for English, French, German, and Spanish; a correlation of the first six thousand words in four single-language frequency lists.* Chicago: Chicago University Press. 440 p.

Ermakova, O. I. (2001). Features of computer jargon as a specific subsystem of the Russian language. *Dialog.* P. 173. (In Russ.).

Francis, W. N. (1964). *Brown Corpus Manual: Manual of information to accompany. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers.* Providence: Brown University. 467 p.

Grammar of Spoken and Written English. (1999). Longman Harlow: Pearson Education Limited. 1204 p.

Guietti, P. (1993). Hermeneutic of Aquinas's Texts: Notes on the Index Thomisticus. *The Thomist: A Speculative Quarterly Review, 57 (4):* 667—686. DOI: 10.1353/tho.1993.0006.

Harris, Z. S. (1960). Structural Linguistics. Chicago: University of Chicago Press. 384 p.

Hill, J. (1997). *LTP Dictionary of Selected Collocations.* Hove: Language Teaching Publications. 288 p.

Hunston, S. (2000). *Pattern Grammar.* Amsterdam: John Benjamins Publishing. 288 p.

Hyland, K. (2008). As it can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27:* 4—21. DOI: 10.1016/j.esp.2007.06.00.

Johansson, S. (2009). Some aspects of the development of corpus linguistics in the 1970-s and 1980-s. In: *Corpus Linguistics: An International Handbook.* Berlin: De Gruyter. 33—53.

Kondratyukova, L. K. (2012). Borrowings and internationalisms in the terminology of English computer technology. *Dynamics of systems, mechanisms and machines, 4:* 155—158. (In Russ.).

Kononenko, A. P. (2023). Linguistic potential of computer technologies in modern philology. *Humanities and social sciences, 97 (2):* 50—54. DOI: 10.18522/2070-1403-2023-97-2-50-54. (In Russ.).

Kozlova, N. V. (2013). Linguistic corpus: definition of basic concepts and typology. *Bulletin of the NSU. Linguistics and intercultural communication, 1:* 79—88. (In Russ.).

Kozlovskaya, N. V., Musayeva, A. S., Sumenikina, Yu. V. (2023). Transterminologization in the field of artificial intelligence: towards raising the question of subterminology.

Art Logos, 3 (24): 98—118. DOI: https://doi.org/10.24224/2227-1295-2025-14-4-9-37. (In Russ.).

Kuebler, S. (2015). *Corpus Linguistics and Linguistically Annotated Corpora.* London: Bloomsbury Publishing. 320 p.

Lei, L., Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes, 22:* 42—53. DOI: 10.1016/j.jeap.2016.01.008.

Liu, J. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes, 39 (1):* 1—11. DOI: 10.1016/j.esp.2015.03.001.

Lyashevskaya, O. N., Sharov, S. A. (2009). *Frequency dictionary of the modern Russian language (based on the materials of the National Corpus of the Russian language).* Moscow: Azbukovnik. 1090 p. ISBN 978-5-91172-024-7. (In Russ.).

Martínez, I. A. (2009). Academic vocabulary in agriculture research articles: a corpus-based study. *English for Specific Purposes, 28 (3):* 183—198. DOI: 10.1016/j.esp.2009.04.003.

McEnery, T. (2012). *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press. 312 p.

Mcgillivray, B., Passarotti, M., Ruffolo, P. (2009). The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon. *Traitement Automatique des Langues, 50 (2):* 103—127.

O'Keeffe, A. (2010). *Routledge handbook of corpus linguistics.* London: Routledge. 682 p.

Partington, A., Marchi, A. (2015). Using corpora in discourse analysis. In: *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press. 216—234.

Pawley, A. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike frequency. In: *Language and Communication.* London: Longman. 191—226.

Petrova, I. M., Ivanova, A. M. (2022). *Modern digital technologies in linguistic research: textbook. handbook for students in the field of Linguistics.* Moscow: Languages of the Peoples of the World. 259 p. ISBN 978-5-6048046-8-1. (In Russ.).

Plungyan, V. A. (2005). Why do we need a National corpus of the Russian language? Informal introduction. In: *National corpus of the Russian language: 2003—2005.* Moscow: Indrik. 6—20. (In Russ.).

Resslerová, V. (2024). La terminologie du domaine de l'intelligence artificielle: néologie et pluridisciplinarité. *Studia Romanistica, 24 (2):* 59—71. DOI: 10.15452/SR.2024.24.0012.

Rockwell, G. (2019). The Index Thomisticus as a Digital Humanities Big Data Project. *Umanistica Digitale, 5:* 13—34. DOI: 10.6092/issn.2532-8816/8575.

Sabahuddin, A. (2024). AI Lexica: Exploring the Vocabulary of Artificial Intelligence. *Journal of Emerging Technologies and Innovative Research, 11 (4):* 123—137.

Scott, M., Tribble, C. (2006). *Textual Patterns: Key words and corpus analysis in language education.* Amsterdam: John Benjamins Publishing. 203 p.

Selivan, L. (2023). Corpus Linguistics and Vocabulary Teaching. *Demystifying Corpus Linguistics for English Language Teaching. Springer.* 139—161. DOI: 10.1007/978-3-031-11220-1_8.

Shalimova, P. A. (2024). On the question of terms and neologisms in the field of artificial intelligence and neural networks. *Society, economics, culture: development strategies. Materials of the XV All-Russian Scientific and Practical Conference.* 218—223. (In Russ.).

Sinclair, J. (1987). *Looking up: an account of the COBUILD Project in lexical computing.* London and Glasgow: Collins ELT. 182 p.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: University of Oxford. 179 p.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology.* Berlin: Language Science Press. 510 p.

Stefchov, E., Angelova, G. (2018). Towards Constructing a Corpus for Studying the Effects of Treatments and Substances Reported in PubMed Abstracts. *Lecture Notes in Computer Science, 11089:* 115—125. DOI: 10.1007/978-3-319-99344-7_11.

Suleimanova, O. A. (2024). Anthropocentrical Turn in Linguistics Through the Digital Lens: Evidence from Analyses of Russian Mnemonic Verbs. *Journal of Siberian Federal University. Humanities and Social Sciences, 17 (5):* 847—861.

Suleymanova, O. A., Guliyants, A. B. (2022). Methodology of linguistic research as an actual section of modern scientific publication. *Bulletin of the Moscow State Pedagogical University. Series: Philology. Theory of language. Language education, 4 (48):* 89—101. DOI: 10.25688/2076-913X.2022.48.4.07. (In Russ.).

Terms and concepts of artificial intelligence in linguistic illumination. (2024). Moscow: Sputnik+. 193 p. ISBN 978-5-9973-6887-6. (In Russ.).

Valipouri, L., Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes, 12 (4):* 248—263. DOI: 10.1016/j.jeap.2013.07.001.

Vinokurova, T. N. (2016). Structural features of artificial intelligence terminology in English. *International Scientific Research Journal, 10—3 (52):* 14—23. DOI:10.18454/IRJ.2016.52.024. (In Russ.).

Zakharov, V. P. (2005). *Corpus linguistics.* Saint Petersburg: Saint Petersburg State University. 48 p. ISBN 978-5-288-05997-1. (In Russ.).

The article was submitted 20.02.2025;
approved after reviewing 29.07.2025;
accepted for publication 19.08.2025.