ИЗДАТЕЛЬСТВО грамота

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2026. Том 19. Выпуск 3 | 2026. Volume 19. Issue 3
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Параметризация англоязычного научного дискурса (корпусное исследование)

Горожанов А. И., Гусейнова И. А., Денисова Г. В.

Аннотация. Цель проводимого исследования заключается в установлении параметров научного дискурса (жанр – исследовательская статья), транслируемого на английском языке в экспертном сообществе, на основе корпусного подхода. Научная новизна заключается в том, что в работе впервые предлагается и апробируется процедура общего (базового) корпусного анализа текстов научного дискурса, а также тестируются новые функции инновационного отечественного программного продукта «Генератор сбалансированного лингвистического корпуса и корпусный менеджер». В результате были получены статистические параметры текстов на синтаксическом уровне (средняя длина предложения в словах, доля сложноподчиненных и сложносочиненных предложений); определено лексическое разнообразие в научных статьях на основе количества уникальных словоформ; проведен анализ частотности сочетаний токенов для установления степени клишированности текстов статей и составлен список наиболее типичных словосочетаний.

EN

Parametrization of English-language scientific discourse (corpus-based study)

A. I. Gorozhanov, I. A. Guseynova, G. V. Denissova

Abstract. The aim of this study is to establish the parameters of scientific discourse (the research article genre) transferred in English within the expert community using a corpus-based approach. The study's novelty lies in its first proposal and validation of a procedure for a general (basic) corpus analysis of scientific discourse texts, as well as in testing new features of the innovative Russian software product "Balanced Linguistic Corpus Generator and Corpus Manager". As a result, statistical parameters of the texts were obtained at the syntactic level (average sentence length in words, proportion of complex and compound sentences); lexical diversity in scientific articles was determined based on the number of unique word forms; the frequency of token combinations was analyzed to determine the degree of cliché in the article texts; and a list of the most typical phrases was compiled.

Введение

Научный дискурс как важная составная часть академического дискурса реализуется в различных устных и письменных жанрах: научная статья, монография, диссертация, научный доклад, выступление на конференции, стендовый доклад, научно-технический отчет, рецензия, реферат, аннотация, тезисы (Карасик, 2002, с. 163). Среди указанных жанров научная статья занимает одно из центральных мест и «может рассматриваться в качестве отдельного этапа масштабной дискуссии, … позволяющей приблизиться к решению конкретной научной проблемы» (Малюга, 2019, с. 54).

Тексты научных статей характеризуются высокой степенью интертекстуальности (Слышкин, 2000). Их структура и содержание подчиняются не только объективным (со стороны научного сообщества) и субъективным (со стороны конкретной редакции) требованиям, но и нормам закона (Винокуров, 2024; Волков, 2024). Кроме того, размещение статей на специализированных веб-платформах может накладывать на автора и редакцию дополнительные обязательства к оформлению основного текста и метаданных (Кожемякин, Дубровская, 2024; Петров, Филиппова, 2024). Вместе с тем главной задачей научного текста остается «выявление структуры мысли, достижение прозрачного соотнесения языка и реальности, четкое разграничение значимых и пустых выражений и фраз, донесение самой сути примарной когнитивной информации» (Аликаев, Бредихин, 2015, с. 122).

Научная статья как форма публикации и как жанр является сегодня актуальным объектом обсуждения в связи с резким скачком развития технологий генеративного искусственного интеллекта. Исследователи

Научная статья (original research article) | https://doi.org/10.30853/phil20260146

поднимают проблему применения подобного программного обеспечения при подготовке статей к опубликованию (Валькова, 2024; Иванова, 2024). С указанной проблемой тесно связан вопрос о некачественных заимствованиях и уровне академической экспертизы (Чернявская, 2011; 2024).

Другая популярная область исследований относится к педагогике; здесь ставится задача научить создавать правильный научный текст на родном и иностранных (например, на английском) языках (Завьялова, 2023; Басик, Купалов, 2024; Мальцева, Черемохина, 2024). Упомянем также проблему перевода (Слободянюк, 2025), поскольку «в условиях роста публикационной активности ученых и международного обмена знаниями растут и требования, предъявляемые к качеству переводов научных статей» (Серова, 2025, с. 61).

Такой высокий интерес к научному дискурсу как объекту современных научных изысканий со стороны профессионального сообщества подтверждает актуальность нашей работы.

Материалом настоящего исследования являются опубликованные в период 2024-2025 гг. научные статьи из семи журналов издательства “Frontiers” (https://www.frontiersin.org/), которые были преобразованы в сбалансированные лингвистические корпусы по каждому журналу и отдельный сводный лингвистический корпус объемом около 4 млн токенов. Под «сбалансированностью» здесь подразумевается тот факт, что корпусы дают представление о части научного дискурса, а не о научном дискурсе в целом:

– “Frontiers in Artificial Intelligence” (176 статей, 22 282 предложения / 622 070 токенов);
– “Frontiers in Chemistry” (180 статей, 10 372 предложения / 331 584 токена);
– “Frontiers in Communication” (228 статей, 40 308 предложений / 1 257 008 токенов);
– “Frontiers in Education” (176 статей, 16 233 предложения / 483 117 токенов);
– “Frontiers in Language Sciences” (79 статей, 15 113 предложений / 492 347 токенов);
– “Frontiers in Physics” (108 статей, 5758 предложений / 152 913 токенов);
– “Frontiers in Water” (156 статей, 19 845 предложений / 620 189 токенов).

Преобразование текстов статей и их объединение в корпусы проводилось с помощью оригинального программного обеспечения «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» (Степанова, 2023).

Преследуя цель выявить черты современного англоязычного научного дискурса (жанра исследовательской статьи) и действуя в предметном поле корпусной лингвистики, мы ставим перед собой следующие задачи:

1) получить статистические параметры текстов на синтаксическом уровне (средняя длина предложения в словах, доля сложноподчиненных и сложносочиненных предложений);

2) определить лексическое разнообразие в статьях на основе количества уникальных словоформ;

3) провести анализ частотности сочетаний токенов для установления степени клишированности текстов статей (Кравцова, 2012, с. 131), составив список наиболее типичных словосочетаний.

Исследование опирается на общенаучные методы, включая анализ (современной предметно-специальной литературы по проблеме), синтез (формулирование выводов) и эксперимент (получение размеров частотных последовательностей), а также на ряд специальных методов, в числе которых обозначим обработку естественного языка и создание подкорпусов (для каждого журнала), метод эксперимента (тестирования) для вывода контекстов (по другой терминологии – конкордансов), содержащих заданные единицы и явления, статистический и синтаксический анализ, что в совокупности мы назовем *общим (базовым) корпусным анализом*, который направлен на получение данных о текстах определенного типа дискурса и жанра автоматическим или автоматизированным путем (с минимальной долей ручного труда исследователя).

Высокая степень автоматизации в нашем случае становится возможной благодаря применению входящего в состав упомянутого выше программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» корпусного менеджера – системы управления базой данных. Комплекс является локальным приложением, которое принадлежит Московскому государственному лингвистическому университету и в настоящее время не распространяется для стороннего пользования.

Метод эксперимента упоминается в том смысле, что мы впервые обращаемся к указанному программному комплексу для получения результатов запросов к корпусу текстов научных статей и тем самым тестируем его программные функции.

Теоретическую базу работы формируют труды ученых, посвященные характеристике научного дискурса (Аликаев, Бредихин, 2015; Карасик, 2002; Малюга, 2019; Слышкин, 2000), специфике языка научной статьи (Абросимова, Богданова, 2024; Ватина, 2024; Садовникова, 2024; Богинская, 2025), а также прикладной и корпусной лингвистике (Баранов, Добровольский, 2022; Баранов, 2024; Рахилина, Плунгян, 2025; Се, 2025). Безусловно, при работе с корпусом текстов неизбежно возникает вопрос о разработке общегуманитарных основ проводимого нами исследования (King, 2004; Krämer, 2008). В этой связи нам представляется целесообразным обратить особое внимание на концептуальные труды отечественных и зарубежных ученых, чье внимание сосредоточено на цифровых трансформациях, а также на их последствиях (Bauriedl, Strüver, 2018; Keith, 2011; Махлина, 2000). Не секрет, что создание корпусов текстов, посвященных определенной теме или проблематике, предполагает и построение определенной архитектуры в виртуальном пространстве, в которой потребуется выделение профессиональных и профессионально ориентированных жанров, а также жанров, направленных на популяризацию знаниевого компонента и, кроме того, на определение жанра или типа текста, чей универсальный характер мог бы обеспечить его требуемую трансформацию путем выполнения определенных автоматизированных действий. В этом свете актуальной представляется проблема диверсификации

целевых социокоммуникативных групп, для каждой из которых необходимой представляется разработка ее общих социолингвистических параметров (пол, средний возраст, социальный статус, вкусовые предпочтения, духовно-материальные запросы и т. п.). Одновременно возникает проблема сопряжения жанров научного дискурса с учетом характеристик выделенных опытным путем социокоммуникативных групп. На наш взгляд, упомянутый выше «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» во многом призван служить выполнению описываемых нами научно-практических задач. Очерченный нами круг проблем требует своего решения, которое не может быть выполнено в рамках одной статьи. Это обстоятельство свидетельствует о перспективности проводимых нами изысканий и вносит теоретико-методологический и одновременно практический вклад в развитие больших баз данных и инструментов для работы с ними.

Практическая значимость исследования заключается в том, что его результаты могут быть использованы в прикладной плоскости – для анализа текстов научного дискурса на английском и других языках, а также в педагогической области – в ходе чтения лекционных курсов и проведения семинарских занятий по широкому кругу проблем цифровой лингвистики.

Обсуждение и результаты

Прежде всего, замерим необходимые статистические параметры текстов. Средняя длина предложения в словах определяется делением количества токенов на количество предложений. Таким образом, мы получаем следующий результат (Таблица 1):

Таблица 1. *Данные о средней длине предложений*

Название журнала	Средняя длина предложения (в токенах)
“Frontiers in Language Sciences”	32,58
“Frontiers in Chemistry”	31,97
“Frontiers in Water”	31,25
“Frontiers in Communication”	31,19
“Frontiers in Education”	29,76
“Frontiers in Artificial Intelligence”	27,92
“Frontiers in Physics”	26,56

Данные о количестве токенов и предложений оператор корпусного менеджера получает сразу после загрузки лингвистического корпуса.

При средней длине предложения равной 30,48 слова гуманитарные и естественнонаучные журналы не показывают принципиальной разницы, хотя по общепринятым представлениям во вторых мысли должны излагаться более лаконично.

Журнал по физике имеет минимальное значение не в последнюю очередь благодаря наличию небольших простых предложений, например (числовой параметр в начале примера показывает его номер в корпусе):

1003 Finally, the used loss functions are discussed. / Наконец, обсуждаются используемые функции потерь (здесь и далее – перевод авторов статьи. – *А. Г., И. Г., Г. Д.*).

Достаточно краткими являются описания рисунков, например:

4010 Figure 6A shows the transient accelerating voltage in the main accelerating section. / На рисунке 6A показано переходное ускоряющее напряжение в основной ускоряющей секции.

Журнал по лингвистике нередко использует предложения длиной более 50 слов, например:

1502 Conversely, asking students to be seated at desks for hours with only a short break, over focusing on accuracy, drilling explicit grammar rules, or using course books for Polish children in Poland and thus focusing on concepts relevant to that population but often less engaging for children living abroad may be counterproductive. / И наоборот, просить учеников сидеть за партами часами с коротким перерывом, чрезмерно сосредотачиваясь на точности, заучивая подробные правила грамматики или используя учебники для польских детей в Польше и, таким образом, сосредотачиваясь на концепциях, актуальных для этой группы населения, но часто менее интересных для детей, живущих за границей, может быть контрпродуктивным.

Педагогическая тематика в приведенном примере не должна удивлять, поскольку в зарубежной трактовке методика преподавания иностранных языков относится к области (прикладной) лингвистики.

Журнал по химии также часто демонстрирует объемные высказывания, описывая экспериментальные процессы, например:

1007 In this paper, 3-amino-1,2,4-triazole (D) was synthesised using thiourea as a starting material, and finally the coupling end product E of triazole and Schiff base was obtained by aldolamine condensation reaction, and the structures of all the compounds were determined by spectroscopic analysis. / В данном исследовании 3-амино-1,2,4-триазол (D) был синтезирован с использованием тиомочевины в качестве исходного вещества. Конечный продукт, представляющий собой производное триазола, содержащее фрагмент основания Шиффа, был получен аминированием. Структуры всех соединений были подтверждены спектроскопическими методами.

На наш взгляд, данные примеры интересны как аутентичный материал для обучения профессиональному подъязыку или для формирования умений создания научного текста, однако сама по себе такая статистика

характеризует жанр научной статьи только в самых общих чертах, хотя и является значимой ступенью для основательной интерпретации текста (Степанова, Яновская, 2024; Писарик, 2024). Однако согласно поставленной задаче, рассчитаем долю сложноподчиненных и сложносочиненных предложений. Для этого мы воспользуемся функцией корпусного менеджера для поиска сочинительных и подчинительных союзов (Таблица 2).

Таблица 2. *Доли предложений с сочинительными и подчинительными союзами*

Название журнала	Доля предложений с сочинительными союзами (%)	Доля предложений с подчинительными союзами (%)
“Frontiers in Artificial Intelligence”	59,1	27,36
“Frontiers in Chemistry”	65,61	21
“Frontiers in Communication”	64,4	37,25
“Frontiers in Education”	65,11	34,36
“Frontiers in Language Sciences”	59,51	41,88
“Frontiers in Physics”	54,78	23,57
“Frontiers in Water”	64,01	26,72

Поиск сочинительных и подчинительных союзов осуществляется благодаря частичной разметке корпусов, при которой эти союзы классифицируются как отдельные «части речи».

Данная методика расчета имеет небольшую погрешность, однако хорошо работает при сопоставлении различных журналов. Заметим также, что показатель не отражает точной доли сложносочиненных предложений, поскольку сочинительный союз может объединять два дополнения или другие члены предложения, но придаточные предложения с союзами определяются довольно точно.

Все же по концентрации сочинительных союзов журналы показывают минимальный разброс параметров, тогда как тот же коэффициент для подчинительных союзов может различаться почти в два раза (для химии и для лингвистики).

Для того чтобы определить, могут ли полученные данные являться характеристиками научного дискурса жанра научной статьи, рассчитаем те же показатели для новостного дискурса. Итак, для корпуса новостных статей CNN объемом 982 107 токенов, собранного нами ранее в рамках исследований лаборатории фундаментальных и прикладных проблем виртуального образования (Бондарчук, Грачев, 2024), доли предложений с сочинительными и подчинительными союзами составили, соответственно, 52,11% и 35,63%. Если соотнести эти результаты со средним арифметическим по научным статьям – 61,79% и 30,31% – то можно заключить, что сочинительная связь скорее более характерна для текстов научного дискурса, чем новостного. Например, нами было выявлено множественное употребление союза “and” при парном перечислении цитируемых ученых, например: “Shuming Zhang et al., and Zhe Wu et al.”

Представим эти результаты в виде диаграммы (Диаграмма 1).

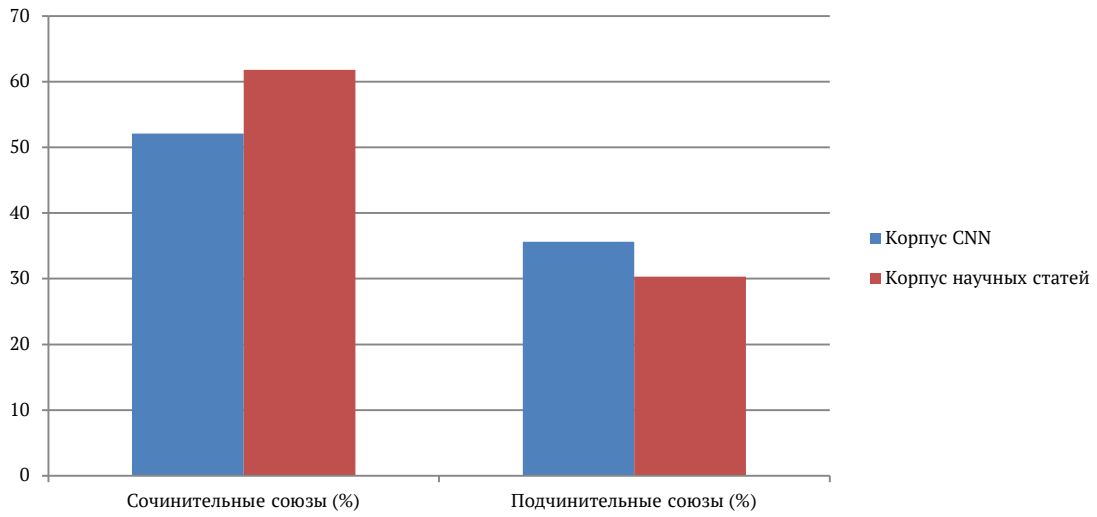

Диаграмма 1. *Доля предложений с сочинительными и подчинительными союзами в корпусе CNN и корпусе научных статей*

На следующем этапе работы было изучено лексическое разнообразие текстов статей, под которым мы здесь, в узком смысле, понимаем долю уникальных существительных и прилагательных. Более широкое понимание может потребовать привлечение также других частей речи, однако это выходит за рамки настоящего исследования и может быть предпринято в перспективе.

Для получения статистических данных и общего списка частотности уникальных лемм, а также списка частотности по заданным частям речи были применены функции корпусного менеджера «Частотный список» и «Частотный список 2» соответственно (Таблица 3).

Таблица 3. *Данные о лексическом разнообразии текстов*

Название журнала	Уникальные существительные (%)	Уникальные прилагательные (%)	Всего уникальных лемм (%)
“Frontiers in Artificial Intelligence”	3,94	5,39	3,44
“Frontiers in Chemistry”	**7,81**	**9,11**	5,45
“Frontiers in Communication”	3,34	4,24	**2,72**
“Frontiers in Education”	3,64	5,53	3,25
“Frontiers in Language Sciences”	4,38	5,42	3,35
“Frontiers in Physics”	**8,01**	**10,49**	5,48
“Frontiers in Water”	4,25	5,80	3,94

Поясним методику расчета. Общая доля уникальных лемм определяется как отношение уникальных лемм к общему количеству лемм в корпусе, выраженному в процентах. Например, в журнале об искусственном интеллекте выявлено 21 398 уникальных лемм при общем количестве токенов 622 070. При этом учитываются также и служебные слова (артикли, прилагательные и т. п.).

Чем выше этот показатель, тем выше доля уникальных лемм в тексте, т. е. если бы ни одно слово в тексте не повторялось, то показатель составил бы 100%.

Говоря об уникальных существительных, параметр определяется как доля уникальных имен существительных к их общему количеству в процентах. Так, для того же журнала об искусственном интеллекте получены данные, из которых мы приведем первые десять строк:

1. model: 4573 (количество употреблений в корпусе)
2. datum: 2595
3. study: 1658
4. dataset: 1371
5. method: 1295
6. system: 1253
7. performance: 1249
8. learning: 1210
9. feature: 1197
10. approach: 1195

То же действительно и для прилагательных, среди которых для этого же журнала самыми частотными являются:

1. high: 1172
2. different: 844
3. specific: 585
4. medical: 567
5. various: 565
6. real: 564
7. significant: 553
8. deep: 523
9. l ow: 490
10. human (в качестве определения): 471

Заметим, что приведенные выше данные позволяют заключить: в исследуемом журнале затрагиваются, в первую очередь, вопросы медицины, что свидетельствует о нивелировании в современных реалиях некогда остро стоящей в прикладной лингвистике проблемы реферирования текста с целью вычленения его смысловой доминанты.

В этом контексте интересны показатели по самым частотным существительным сводного корпуса по всем семи журналам:

1. study: 11 598
2. model: 9449
3. water: 6855
4. datum: 6601
5. student: 5949
6. language: 5636
7. %: 5628
8. research: 5348
9. time: 5114
10. result: 4984
11. system: 4410
12. effect: 4353
13. participant: 4336
14. information: 4307
15. analysis: 4280

16. level: 4148
17. process: 4123
18. medium: 3974
19. group: 3889
20. approach: 3755
21. method: 3655
22. value: 3645
23. use: 3580
24. learning: 3472
25. performance: 3037

В приведенном списке мы намеренно оставили погрешность в п. 7, где знак процента был принят за существительное. Заметим, что мы приводим этот список скорее как демонстрацию возможностей программного обеспечения, не осуществляя его тщательную интерпретацию. Однако необходимо заметить, что оценивать вес каждой приведенной леммы необходимо с учетом разницы в долях присутствия каждого отдельного журнала в сводном корпусе.

В Таблице 3 наиболее ярко видны следующие отклоняющиеся от общего среднего значения: уникальные существительные для химии и физики, уникальные прилагательные в тех же журналах, наконец, для коммуникаций – минимальная доля уникальных лемм.

Такие показатели мы можем объяснить тем, что в случае с физикой и химией имеет место описание сложных экспериментальных процессов, имеющих свои собственные характеристики. Это наталкивает нас намысль о том, что терминологический аппарат отдельной научной отрасли не обязательно может быть представлен исключительно существительными – важно и сочетание «прилагательное + существительное». Также в качестве существительных часто были определены формульные обозначения, что повысило общее разнообразие текстов. И противоположно этому – преимущественно гуманитарный журнал на тему коммуникаций позволил выявить относительно «бедное» разнообразие в средствах выражения, оперирование общими (спекулятивными) рассуждениями и фактическое отсутствие формульных обозначений. Таким образом, мы можем выдвинуть гипотезу о том, что именно описание реальных экспериментальных процессов лексически разнообразит текст научной статьи и, возможно, *a priori* делает ее более привлекательной для читателя. Так, в статье по физике предложение № 3689 (описание изображения) содержит три формульных токена и термины (“Quantum Stirling”, “entanglement”, “antisymmetric exchange”).

3689: In Figure 2, we plot the efficiency of the Quantum Stirling heat engine in terms of the ratio between the antisymmetric exchange parameters D1/D2, fixing the parameters {c1=4c2, J2=2J1} for different values of temperature (2 a) and {Th=4Tc, J2=2J1} (2 b), for different values of entanglement. / На рисунке 2 представлен график термодинамической эффективности квантовой тепловой машины Стирлинга в зависимости от отношения параметров антисимметричного обмена D1 и D2 при фиксированных значениях параметров {c1=4c2, J2=2J1}. График (a) соответствует различным значениям температуры, а график (b) – различным значениям запутанности при условии {Th=4Tc, J2=2J1}.

Исходя из принципов работы используемого нами токенизатора, который определяет в качестве токенов слова, числа и знаки препинания, данное предложение содержит 60 токенов, из которых 39 рассматриваются как уникальные.

Для наглядности нами была построена диаграмма (Диаграмма 2), демонстрирующая зависимость количества уникальных токенов и их частотности для журналов по физике (красный столбик) и коммуникациям (зеленый столбик).

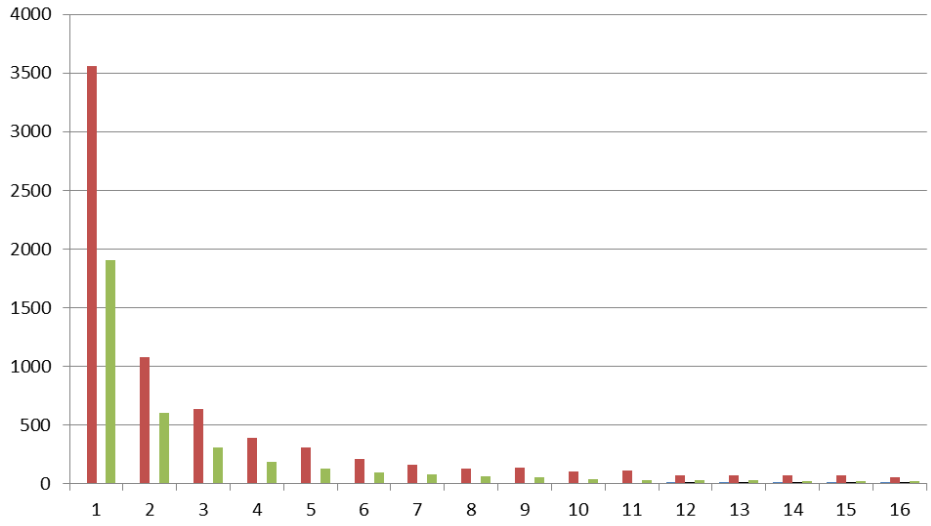

Диаграмма 2. *Сравнение количества и частотности токенов для журналов по физике и коммуникациям*

Для расчета была произведена выборка токенов, которые употребляются в каждом корпусе от одного до 16 раз (ось X). На оси Y показано количество таких токенов (первый столбик – для журнала по физике, второй – для журнала по коммуникациям).

Поскольку объем корпуса второго журнала больше в 8,22 раза, данные были нормализованы для того, чтобы было возможно работать с абсолютными значениями и, следовательно, повысить качество визуализации.

Например, в журнале по физике имеется 3557 токенов, которые встречаются в корпусе только один раз. В журнале по коммуникациям таких токенов насчитывается 1905. Для дважды упоминающихся токенов эти значения равны 1706 и 602 единицы соответственно. При этом значении столбик на диаграмме должен быть заметно выше метки 1500, а на диаграмме, в нормализованном варианте, он чуть выше 1000.

Таким образом, мы показали возможности определения лексического разнообразия в исследуемых текстах, привлекая объективные количественные данные, полученные в ходе работы нашего программного обеспечения.

Для решения третьей задачи исследования была впервые применена функция корпусного менеджера «Частотный список n ТНК», что вносит дополнительный вклад в новизну нашей работы. Функция получает от оператора два числовых параметра: длину последовательности и ее минимальную частотность. Поиск производился во всех семи корпусах, при этом были заданы последовательности от трех до семи токенов и минимальная частотность 20. Методика расчета была такова, что в число токенов входили знаки препинания, цифры, различные небуквенные символы (Таблица 4).

Таблица 4. *Количество последовательностей токенов (3-7) с частотностью от 20 употреблений*

Название журнала	3/20	4/20	5/20	6/20	7/20
“Frontiers in Artificial Intelligence”	987	108	9	0	0
“Frontiers in Chemistry”	531	155	110	100	93
“Frontiers in Communication”	2192	272	24	6	0
“Frontiers in Education”	694	80	14	4	0
“Frontiers in Language Sciences”	644	63	8	1	0
“Frontiers in Physics”	235	92	79	74	69
“Frontiers in Water”	955	106	15	2	0

При анализе полученных в ходе эксперимента данных следует принимать во внимание различные размеры корпусов. Значительное количество повторяющихся последовательностей в журналах по химии и физике вызвано тем, что в текст статей обязательно включается стандартный формальный параграф о правах, личном вкладе авторов, отсутствии сгенерированного текста и пр.

К слову сказать, этот параграф скорее можно отнести к метаданным, однако он был формально включен в основной текст статей.

На наш взгляд, наиболее показательными являются столбцы «5/20» и «6/20». Например, для журнала по искусственному интеллекту были выявлены следующие частотные сочетания из пяти токенов:

1. on the other hand,: 74
2. in this study, we: 54
3. , it is important to: 29
4. at the same time,: 26
5. , as well as the: 25
6. precision, recall, and: 24
7. %, due to the: 22
8. in this paper, we: 21
9. it should be noted that: 20

В журнале о водных ресурсах четко прослеживается тематическая специфика. Так, для «5/0» имеем:

1. on the other hand,: 71
2. % in the study area: 53
3. , as well as the: 46
4. % in the context of: 33
5. as a result, the: 23
6. , energy, and food: 22
7. the other hand, the: 21
8. at the same time,: 21
9. water, energy, and: 21
10. in the absence of a: 21
11. in this study, the: 20
12. in this study, we: 20
13. in the absence of any: 20
14. % of the study area: 20
15. the impacts of climate change: 20

Для «6/20»:

1. on the other hand, the: 21
2. water, energy, and food: 21

Анализ полученных результатов по всем семи журналам позволяет сделать вывод о наиболее распространенных сочетаниях токенов в исследуемых научных статьях вне зависимости от их тематического направления. Ниже мы приводим некоторые самые частотные сочетания различной длины (приведено к написанию со строчной буквы ввиду особенностей программного обеспечения):

as a result;
as well as;
at the same time;
children who are hard of hearing;
in the absence of a;
in the context of;
in the current / present study, we;
in the private / public sector;
in this study / paper, we;
it is important to note that;
it should be noted that;
of the fourth industrial revolution;
on the one / other hand;
the impacts of climate change;
the perceived utility of ai;
to adopt and use ai;
utility of ai integration in.

Детальное рассмотрение каждого из приведенных выше пунктов, расположенных в алфавитном порядке, требует более обширного исследования, хотя программный комплекс позволяет изучить контексты каждого из этих словосочетаний.

Говоря о степени клишированности изучаемых текстов, упомянем стандартные фрагменты в журналах по физике и химии. Самым же частотным сочетанием длиной свыше трех токенов для всех тематических направлений явилось “on the one/other hand”. Здесь мы не описываем другие частотные сочетания и отдельные токены, которые употребляются в научном тексте (“in conclusion”, “finally” и др.).

Заключение

Суммируя сказанное выше, мы приходим к выводу, что изучение текстов жанра англоязычной исследовательской статьи научного дискурса на основе корпусного подхода позволяет не только получить исчерпывающие статистические данные, которые могут служить стимулом для построения и проверки тех или иных лингвистических гипотез, но и предоставить ценный лингводидактический материал.

Исследование показало, что для текстов жанра научной статьи характерна четко выраженная сочинительная связь, технические статьи более разнообразны по лексическому составу (включая формульный компонент), клишированные сочетания имеют не только общенаучный характер, но и обусловлены актуальной тематикой описываемых исследований.

Безусловно, к ограничивающим факторам относится то, что материал работы охватывал определенный круг журналов одного издательства, а также некоторые погрешности в работе программного комплекса, которые, однако, не касались системных явлений и не оказали значительного влияния на полученные результаты.

Кроме решения поставленных задач в статье нами была предпринята попытка осуществить общий корпусный анализ как совокупность процедур по составлению запросов к корпусам и интерпретации полученных данных.

Важно, что исследование осуществлялось с опорой на авторское программное обеспечение (собственно, тестирование функций и составляет суть заявленного эксперимента), что позволяет не зависеть от сторонних (зарубежных) программных продуктов и повышает степень технологического суверенитета отечественной науки.

Уместным считаем подробно остановиться на перспективе нашего исследования. Особенность программного комплекса «Генератор сбалансированного лингвистического корпуса и корпусный менеджер» состоит в том, что с его помощью возможно создавать (и расширять ранее созданные) лингвистические корпусы из заданных текстов полностью автоматическим путем, причем без подключения к сети Интернет, поскольку программная система является оффлайн-приложением, что повышает устойчивость ее работы в условиях кибератак и прочих вредоносных вмешательств.

Предложения, полученные в результате запросов к базам данных лингвистических корпусов возможно изучить более детально, запрашивая для них более широкий контекст. Выполненная параметризация фокусировалась преимущественно на формальных (*количественных*) характеристиках, важность которых не стоит недооценивать. В будущем для получения широкого спектра *качественных* параметров научного дискурса представляется продуктивным сфокусироваться на семантике отдельных токенов и словосочетаний, проблеме описания идиостиля, а также на вопросе автоматизированного составления дидактических материалов для обучения академическому письму.

Источники | References

1. Абросимова Л. С., Богданова М. А. О возможности красноречия в жанрах научной прозы // Известия Южного федерального университета. Филологические науки. 2024. Т. 28. № 3. https://doi.org/10.18522/1995-0640-2024-3-21-31
2. Аликаев Р. С., Бредихин С. Н. «Схемы действования» как маркер дискурсивности научного текста: формальная логика vs. герменевтика // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2015. № 2 (26). https://doi.org/10.15688/jvolsu2.2015.2.17
3. Баранов В. А. Исторический корпус средневековых славянских рукописей «Манускрипт» как исследовательский интернет-ресурс (количественно-статистические характеристики существительных брань и рать) // Scriptorium slavicum. 2024. № 1. https://doi.org/10.20913/script-2024-1-06
4. Баранов А. Н., Добровольский Д. О. Понятие корпусной модели идиостиля // Труды института русского языка им. В. В. Виноградова. 2022. № 1. https://doi.org/10.31912/pvrli-2022.1.2
5. Басик Н. Ю., Купалов Г. С. Проектирование и разработка научной публикации в высшей школе: реалии и устремления // Современная высшая школа: инновационный аспект. 2024. Т. 16. № 2 (64).
6. Богинская О. А. Языковые актуализаторы хеджирования в научных статьях по гуманитарным наукам // Вестник Новосибирского государственного университета. Серия: История, филология. 2025. Т. 24. № 2. https://doi.org/10.25205/1818-7919-2025-24-2-19-30
7. Бондарчук Г. Г., Грачев Г. В. Функционирование английских наименований питания в современном общественно-политическом дискурсе // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 13 (894).
8. Валькова Ю. Е. Использование технологий искусственного интеллекта для подготовки и написания научных статей // Информатика и образование. 2024. Т. 39. № 6. https://doi.org/10.32517/0234-0453-2024-39-6-38-52
9. Ватина А. Е. Бисубстантивные предложения в научном дискурсе: виды оценок в научных статьях // Известия Саратовского университета. Новая серия. Серия: Филология. Журналистика. 2024. Т. 24. № 2. https://doi.org/10.18500/1817-7115-2024-24-2-139-146
10. Винокуров В. А. Научные публикации как произведения науки: вопросы оформления // Право интеллектуальной собственности. 2024. № 4 (78). https://doi.org/10.55291/1999-480X-2024-4-12-15
11. Волков В. И. Научная статья: основные ее элементы, требования к изложению содержания и рекомендации по написанию // Научный вестник оборонно-промышленного комплекса России. 2024. № 4.
12. Завьялова М. С. Обучение работе над научной статьей на материале дисциплины «Иностранный язык в профессиональной деятельности» // Вестник Тверского государственного университета. Серия: Педагогика и психология. 2023. № 1 (62). https://doi.org/10.26456/vtpsyped/2023.1.193
13. Иванова Л. А. Искусственный интеллект при написании научных статей – положительный или вредоносный фактор? // Crede Experto: транспорт, общество, образование, язык. 2024. № 4. https://doi.org/10.51955/2312-1327_2024_4_6
14. Карасик В. И. Языковой круг: личность, концепты, дискурс. Волгоград: Перемена, 2002.
15. Кожемякин Е. А., Дубровская Т. В. Жанр научной статьи в Интернете: структурно-семиотическая трансформация // Жанры речи. 2024. Т. 19. № 1 (41). https://doi.org/10.18500/2311-0740-2024-19-1-41-66-78
16. Кравцова Е. В. Научный дискурс как вид институционального типа дискурса // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2012. № 25.
17. Мальцева Г. Ю., Черемохина Д. А. Практика организации и проведения вузовского курса «академическое письмо (на русском языке)» // Вопросы методики преподавания в вузе. 2024. Т. 13. № 2. https://doi.org/10.57769/2227-8591.12.2.09
18. Малюга Е. Н. Новые тенденции англоязычного научного дискурса: вопросы актуальности исследования и языковой идентичности // Вестник Томского государственного университета. Филология. 2019. № 58. https://doi.org/10.17223/19986645/58/4
19. Махлина С. Т. Семиотика в системе современного научного знания // Философия XX века: школы и концепции: материалы научной конференции (г. Санкт-Петербург, 23-25 ноября 2000 г.). СПб.: Санкт-Петербургское философское общество, 2000.
20. Петров С. А., Филиппова С. Д. Проектирование веб-приложения для сопровождения процесса публикации научных статей в электронном журнале // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2024. № 4. https://doi.org/10.18137/RNU.V9187.24.04.P.114
21. Писарик О. И. Репрезентация сферы «культура и искусство» в корпусе современных средств массовой информации ФРГ // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 11 (892).
22. Рахилина Е. В., Плунгян В. А. О цифровой лексикографии // Труды института русского языка им. В. В. Виноградова. 2025. № 1. https://doi.org/10.31912/pvrli-2025.1.32
23. Садовникова Е. В. Участие модальных глаголов в реализации авторской интенции (на материале немецкоязычной научной статьи) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 7 (888).

24. Се Ж. Высокочастотные слова в спонтанных монологах-описаниях: методика создания частотного списка для лексического анализа // Вестник Донецкого национального университета. Серия Д: Филология и психология. 2025. № 3. https://doi.org/10.5281/zenodo.15301555
25. Серова О. У. Автоматизированное исследование индикаторов переводческого языка на материале переводов научных статей // Terra Linguistica. 2025. Т. 16. № 1. https://doi.org/10.18721/JHSS.16105
26. Слободянюк В. В. Проблемы передачи терминологии при переводе заглавия и аннотации научной статьи с русского на английский язык // Вестник Пермского национального исследовательского политехнического университета. Проблемы языкознания и педагогики. 2025. № 1. https://doi.org/10.15593/2224-9389/2025.1.5
27. Слышкин Г. Г. От текста к символу: лингвокультурные концепты прецедентных текстов в сознании и дискурсе. М.: Academia, 2000.
28. Степанова Д. В. Программный комплекс для генерации динамического корпуса текстов СМИ // Вестник Минского государственного лингвистического университета. Серия 1: Филология. 2023. № 6 (127).
29. Степанова Д. В., Яновская А. С. Языковые средства описания внешности литературного героя: сопоставительный анализ на базе корпуса параллельных текстов // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2024. № 13 (894).
30. Чернявская В. Е. Культура отказа в научной коммуникации: семантика и прагматика отрицательной оценки в экспертизе // Вестник Санкт-Петербургского университета. Язык и литература. 2024. Т. 21. № 4. https://doi.org/10.21638/spbu09.2024.411
31. Чернявская В. Е. Плагиат как социокультурный феномен // Известия Санкт-Петербургского университета экономики и финансов. 2011. № 3 (69).
32. Bauriedl S., Strüver A. Smart City. Kritische Perspektiven auf die Digitalisierung in Städten. Bielefeld: Transcript, 2018.
33. Keith D. City Branding. Theory and Cases. L.: Macmillan, 2011.
34. King A. Spaces of Global Cultures: Architecture, Urbanism, Identity. L. – N. Y.: Routledge, 2004.
35. Krämer S. Medium, Karte, Übertragung. Kleine Metaphysik der Medialität. Frankfurt a. M.: Suhrkamp, 2008.

Информация об авторах | Author information

RU

Горожанов Алексей Иванович[1], д. филол. н., доц.
Гусейнова Иннара Алиевна[2], д. филол. н., проф.
Денисова Галина Валерьевна[3], д. культ., проф.
[1, 2] Московский государственный лингвистический университет
[3] Московский государственный университет имени М. В. Ломоносова

EN

Alexey Ivanovich Gorozhanov[1], Dr
Innara Alievna Guseynova[2], Dr
Galina Valerievna Denissova[3], Dr
[1, 2] Moscow State Linguistic University
[3] Lomonosov Moscow State University

[1] *a.gorozhanov@linguanet.ru,* [2] *ginnap@mail.ru,* [3] *denissovagv@my.msu.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 05.03.2026; опубликовано online (published online): 24.03.2026.

Ключевые слова (keywords): параметризация; научный дискурс; жанр научной исследовательской статьи; корпусная лингвистика; устойчивое словосочетание; parametrization; scientific discourse; research article genre; corpus linguistics; set phrase.