2025;10(2):XX-XX Научный редактор и издатель / Science Editor and Publisher

НАУЧНЫЕ КОММУНИКАЦИИ / SCIENTIFIC COMMUNICATIONS

Оригинальная статья / Original paper

https://doi.org/10.24069/SEP-251041

Авторские ключевые слова и редакторские термины в реферативной базе данных: статистический анализ различий

О. В. Федорец ✉, **Н. С. Солошенко**

Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН), г. Москва, Российская Федерация

✉ *ovf@viniti.ru*

Резюме. Авторские ключевые слова (КС), в отличие от терминов, присвоенных профессиональными индексаторами, не регулируются нормативными документами и не контролируются с помощью специальных словарей. Цель исследования — выявление статистических различий между массивами КС: присвоенных авторами, с одной стороны, и редакторами реферативной базы данных Всероссийского института научной и технической информации (ВИНИТИ) — с другой. Предполагается, что подтверждение и понимание этих различий может оказаться полезным для более рационального использования КС, полученных из различных источников, в поисковых системах и при библиометрических исследованиях. Впервые выполнен сравнительный анализ количественных показателей новизны и лексического разнообразия авторских и редакторских КС, а также сравнение мер включения авторских и редакторских КС в другие элементы метаданных на нескольких независимых тематических выборках. Методологической основой исследования является обобщение — выделение и количественный анализ общих признаков, присущих изучаемым массивам данных. Эмпирическую базу исследования составили пять независимых статистических выборок объемом от 10,40 до 18,97 тыс. статей. Тематика выборок соответствовала пяти рубрикам Государственного рубрикатора научно-технической информации (ГРНТИ): «52. Горное дело; 53. Металлургия; 55. Машиностроение; 61. Химическая технология. Химическая промышленность; 73. Транспорт». Отбирались русскоязычные статьи, загруженные в реферативную базу данных ВИНИТИ в 2021–2024 гг. и одновременно имеющие следующие непустые элементы метаданных: заглавие, авторские КС, авторскую аннотацию, редакторские КС и реферат, специально подготовленный для этой базы данных. Для каждой выборки и по отдельности для авторских и редакторских КС были получены точечные статистические оценки выделенных общих признаков: лексического разнообразия, новизны и включенности КС в другие элементы метаданных. Во всех пяти тематических коллекциях наблюдались одинаковые статистические различия авторских и редакторских КС: мера лексического разнообразия авторских КС выше, чем редакторских; коэффициент новизны у авторских КС выше, чем у редакторских КС, а у авторских аннотаций выше, чем у рефератов; мера включения авторских КС в заглавие статьи ниже, по сравнению с аналогичным показателем для редакторских терминов. Повторение выявленных различий в пяти независимых тематических выборках, соответствующих случайно выбранным областям знаний, позволяет говорить о статистической устойчивости этих различий. Лексика авторских КС более изменчива во времени по сравнению с более стабильной лексикой редакторских КС, что может быть полезно для оперативного выявления новой терминологии и фронтиров науки. В отличие от редакторских, авторские КС не могут самостоятельно выражать основные темы и понятия документа, так как являются дополнением к терминам, которые можно извлечь из заглавий публикаций.

Ключевые слова: научные журналы, авторские ключевые слова, реферативная база данных, редакторские термины, тематические коллекции статей, мешок слов, стемминг, сравнительный статистический анализ, мера включения, лексическое разнообразие, коэффициент новизны

Научный редактор и издатель / Science Editor and Publisher 2025;10(2):XX-XX

Для цитирования: Федорец О. В., Солошенко Н. С. Авторские ключевые слова и редакторские термины в реферативной базе данных: статистический анализ различий. *Научный редактор и издатель*. 2025;10(2):XX-XX. https://doi.org/10.24069/SEP-251041

Author keywords and editorial terms in the abstract database: a statistical analysis of differences

O. V. Fedorets ✉, **N. S. Soloshenko**
All-Russian Institute for Scientific and Technical Information, Russian Academy of Sciences (VINITI RAS), Moscow, Russian Federation
✉ *ovf@viniti.ru*

Abstract. Author keywords, unlike terms assigned by professional indexers, are not regulated by normative documents or controlled by special dictionaries. The aim of this study is to identify statistical differences between two sets of keywords (KWs): those assigned by authors, on the one hand, and those assigned by editors of abstract database of VINITI RAS, on the other. It is believed that confirming and understanding these differences may be useful for more rational use of keywords obtained from various sources. A comparative analysis of quantitative indicators of the novelty and lexical diversity of author and editorial KWs was conducted for the first time in this study. A comparison of the inclusion measures of author and editorial KWs in other metadata elements was conducted for the first time on several independent thematic samples. The methodological basis of the study is generalization—the identification and quantitative analysis of common features inherent in the studied data arrays. The empirical base of the study consisted of five independent statistical samples, the size of which varied from 10.40 thousand to 18.97 thousand articles. The topics of the samples corresponded to five headings of the State Rubricator of Scientific and Technical Information: 52. Mining; 53. Metallurgy; 55. Mechanical Engineering; 61. Chemical Technology. Chemical Industry; 73. Transport. We selected Russian-language articles uploaded to the VINITI abstract database in 2021–2024 and simultaneously containing the following non-empty metadata elements: title, author's keywords, author's abstract, editor's keywords, and an abstract specially prepared for the VINITI abstract database. For each sample and separately for author's and editor's KWs, point statistical estimates of the identified common features were obtained: lexical diversity, novelty, and inclusion of keywords in other metadata elements (title and abstract). Similar statistical differences of author's and editor's KWs were observed across all five thematic collections: the degree of lexical diversity in author-generated KWs is higher than that of editor-generated terms; the novelty coefficient of author-generated KWs is higher than that of editor-generated terms; the novelty coefficient of author-generated annotations is higher than that of abstracts; and the degree of inclusion of author-generated KWs in article titles is lower than the degree of inclusion of editor-generated terms. Replication of the identified differences across five independent thematic samples, corresponding to randomly selected fields of knowledge, suggests the statistical stability of these differences. The vocabulary of author KWs is more variable over time compared to the more stable vocabulary of editor-generated terms, which may be useful for the rapid identification of new terminology and scientific frontiers. Unlike editor-generated KWs, author-generated KWs cannot independently express the main themes and concepts of a document, as they supplement the terms that can be extracted from publication titles.

Keywords: scientific journals, author keywords, abstract database, editor's keywords, thematic collections of articles, bag of words, stemming, comparative statistical analysis, inclusion measure, lexical diversity, novelty coefficient

For citation: Fedorets O. V., Soloshenko N. S. Author keywords and editorial terms in the abstract database: a statistical analysis of differences. *Science Editor and Publisher*. 2025;10(2):XX-XX. https://doi.org/10.24069/SEP-251041

2025;10(2):XX-XX Научный редактор и издатель / Science Editor and Publisher

ВВЕДЕНИЕ

Неуклонный рост количества научных публикаций и соответствующее увеличение объемов библиографических и полнотекстовых баз данных (БД) повышают роль метаданных, среди которых основными содержательными элементами являются заглавие, аннотация (или реферат) и ключевые слова (КС). Поля метаданных публикаций в научных БД используются для различных целей — от поиска информации до наукометрических исследований. Традиционно основными источниками данных были Web of Science (WoS) и Scopus. Однако с появлением новых БД, таких как Dimensions, выбор электронных информационных ресурсов еще более расширился [1].

Обнаружение цифровых материалов учеными и специалистами в значительной степени зависит от эффективности метаданных, обеспечивающих включение публикаций в результаты поиска и приоритетность ранжирования в поисковых системах [2]. Выбирая подходящие КС и фразы при подготовке рукописи к публикации, авторы могут значительно способствовать представлению полной и релевантной информации о своих работах в библиографических БД [3]. Составленный автором перечень КС призван отражать тематику статей и является важным элементом метаданных. КС в основном относятся к предметной области исследования и отражают понимание авторами своей работы в тематическом контексте их исследовательских областей [4]. Таким образом, авторские КС являются достоверными индикаторами тематики статьи. Авторы выбирают КС произвольно, не ориентируясь на контролируемые словари по своим дисциплинам [5].

Авторы научных публикаций не единственные, кто формулирует ключевые слова и словосочетания, загружаемые в БД. В работе Б. Хьёрланда (B. Hjørland) [6] исследованы теоретические основы как ручного, так и автоматического индексирования и предложено выделять пять типов индексаторов, обладающих различными аспектами субъективности, которые влияют на качество процесса индексирования: 1) авторов как индексаторов; 2) индексаторов, прошедших обучение по библиотечно-информационным технологиям; 3) участников систем социальных тегов; 4) людей с высоким уровнем формальных знаний по предмету; 5) специалистов с высоким уровнем формальных знаний по предмету, прошедших обучение в качестве профессиональных индексаторов. К этому списку можно добавить шестой тип индексатора — программное обеспечение (ПО), выполняющее индексирование текста в автоматизированном или автоматическом режиме.

В условиях постоянно растущего объема документов все чаще библиотеки по всему миру изучают и внедряют автоматизированные подходы к предметному индексированию [7]. Широко известный пример результата автоматического индексирования в научных БД — КС плюс (Keywords Plus), которые генерирует ПО поисковой платформы WoS на основе анализа списков литературы в статьях. КС плюс используются совместно с авторскими КС для поиска и наукометрических исследований [8–10]. Если ментальные модели индексаторов не совпадают с моделями пользователей, возникает семантический разрыв между пользователями и индексаторами, который мешает пользователям находить ожидаемые информационные ресурсы [11].

Несоответствия между КС, присвоенными статьям, могут быть вызваны не только различиями в ментальных моделях индексаторов и пользователей, но и особенностями поведенческих моделей выбора КС, влияющими на результаты индексирования. Различия моделей поведения обычно оцениваются статистическими методами на основе распределения выбранных индексатором КС по разделам научной статьи [11–15]. Другим направлением исследований сходства и различий является сравнительный анализ КС, присвоенных статьям разными типами индексаторов [8; 11; 16; 17].

Несмотря на имеющиеся исследования поведенческих моделей авторов и сравнение авторских КС с автоматизированными индексами (например, WoS Keywords Plus), отсутствует масштабный воспроизводимый статистический анализ различий между КС, присвоенными авторами, и КС, назначенными профессиональными индексаторами, в контексте трех ключевых измерений: разнообразия, новизны и включенности в другие элементы метаданных.

На основании изучения рекомендаций и нормативных документов, которыми пользуются авторы статей и редакторы реферативной БД, авторами в данной статье была выдвинута гипотеза: различия в рекомендациях и нормативных документах, а также в целях индексирования неизбежно приводят к расхождениям в моделях поведения авторов и редакторов при выборе КС. Проявлением этих расхождений должно стать устойчивое различие некоторых количественных показателей массивов авторских и редакторских КС. Для подтверждения гипотезы мы собрали тематические коллекции метаданных научных статей по пяти обширным научным областям, относящимся к техническим наукам. Заглавия статьей, авторские КС и аннотации были отобраны из Научной электронной библиотеки eLIBRARY.RU[1].

[1] URL: https://elibrary.ru (дата обращения: 10.12.2025).

Редакторские КС и рефераты, назначенные тем же статьям, были взяты из БД Всероссийского института научной и технической информации Российской академии наук (ВИНИТИ) РАН[2].

Цель исследования — выявление статистических различий между двумя массивами КС: присвоенных авторами, с одной стороны, и редакторами реферативной БД ВИНИТИ — с другой. Для достижения цели решается задача подтвердить различия по количественным показателям: а) лексического разнообразия; б) новизны; в) включенности в другие элементы данных (заглавие и реферат). Предполагается, что выявление и анализ различий между этими массивами могут оказаться полезными в библиометрических и терминологических исследованиях для более рационального использования КС, полученных из различных источников.

ЛИТЕРАТУРНЫЙ ОБЗОР

Роль авторских КС в картировании науки

Картирование, являясь одним из приемов наглядного представления данных, подчиняется общим принципам визуализации, т.е. компактному представлению большого объема информации, делающему заметными закономерности, выявленные в ходе анализа. На практике чаще всего встречаются различные типы наукометрического анализа, позволяющие ответить на базовые вопросы: статистический анализ / построение профиля (Кто?); геопространственный анализ (Где?); темпоральный анализ (Когда?); тематический анализ (Что?); сетевой анализ (С кем?) [18].

Несмотря на присущие КС ограничения (недостаточное количество, проблема синонимии и унификации), их используют для тематического анализа — идентификации предмета исследований и визуализации изменения тематики публикаций [18]. Тематический анализ может сочетаться с другими типами анализа, перечисленными выше, чтобы визуализировать изменения тематики публикаций авторов из определенного региона мира или из определенных научных организаций в заданный интервал лет. Структура научной дисциплины обычно визуализируется в виде сети взаимосвязанных кластеров, при этом могут сравниваться результаты картирования, полученные различными методами. В качестве примера сочетания тематического анализа с сетевым можно привести работу [19], в которой интеллектуальная структура публикаций 2006–2015 гг. по информационным технологиям строится двумя методами: посредством сетевого анализа (по связям в сетях цитирования авторов) и тематического анализа связей между авторскими КС.

В наукометрических исследованиях авторские КС могут использоваться как единственный источник данных [20; 21] либо в комбинации с терминами, полученными из других элементов метаданных, например из авторских аннотаций [22] или из массива КС плюс, которые автоматически были присвоены статьям в БД WoS Core Collection (CC) [10; 9]. Для проведения наукометрических исследований широко используются специализированные пакеты прикладных программ, такие как CiteSpace и VOSviewer [23–25]. Эти пакеты позволяют строить кластеры КС по их совместной встречаемости (*co-occurrence clustering of keywords*), сети совместной встречаемости слов (*co-occurrence network of keywords*), выделяя КС с использованием цветового кодирования: категориального (разными цветами) для группировки по кластерам или темам и последовательного (градацией оттенков одного цвета) для отображения временно́й динамики, например года первого появления. VOSviewer позволяет загружать данные как из библиографических БД, так и из текстовых файлов в формате RIS. При этом пакет может не только обрабатывать авторские КС (тег KW в формате RIS), но и автоматически извлекать и обрабатывать КС из заглавий статей (Title), аннотации/реферата (Abstract) или из сочетания этих элементов метаданных (Title/Abstract), что позволяет следить за динамикой изменения основной терминологии в конкретной научной тематике и в разные хронологические периоды по частоте встречаемости термина и показателю его релевантности (*relevance score*)[3].

Авторские КС могут являться как источником данных для инструментов наукометрического анализа, так и самостоятельным объектом исследований, при этом предметы исследования могут быть различными. Далее мы подробнее остановимся на моделировании поведения авторов при выборе КС и на влиянии КС на цитируемость статей.

Поведенческие модели выбора КС и цитируемость статей

Поведение авторов при выборе ключевых терминов обычно моделируется статистическими методами — на основании распределения авторских КС по заглавию, аннотации, списку литературы и разделам статьи (введение, методы, результаты, обсуждение, заключение) [11–15]. Статистические методы так-

[2] URL: https://www.viniti.ru/database (дата обращения: 10.12.2025).

[3] van Eck N. J., Waltman L. Manual for VOSviewer, Version 1.6.19. 17 November 2025. URL: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.19.pdf (дата обращения: 10.12.2025).

2025;10(2):XX-XX Научный редактор и издатель / Science Editor and Publisher

же используются, чтобы исследовать, как зависит цитируемость статей от различных количественных показателей, характеризующих авторские КС и шаблоны поведения авторов в процессе выбора ключевых терминов [4; 13; 15].

Статистика позволяет определить разделы статьи и элементы метаданных, из которых автор выбирает КС чаще, и таким образом моделировать поведение авторов в процессе выбора КС. Например, В. Лу с соавт. (W. Lu et al.) [13] получили следующие эмпирические результаты: средний процент КС автора, встречающихся в заглавиях статей, аннотациях и в обоих элементах метаданных одновременно, составляет 31; 52,1 и 56,7% соответственно. При этом в ссылках обнаружено 41,6% КС. В результате корреляционного анализа авторы обнаружили отрицательную корреляцию между долей КС в заглавии и аннотации и последующим цитированием статьи. Иными словами, чем меньше КС встречается в заглавии и аннотации статьи, тем больше цитирований она получит. С другой стороны, процент КС автора, встречающихся в массиве высокочастотных КС, имеет положительную корреляцию с количеством цитирований статей. Высокочастотными КС здесь названы КС, частота встречаемости которых в БД растет в течение последних нескольких лет, что говорит о растущем интересе ученых к темам, представленным этими КС.

Поведение автора при выборе ключевых терминов и зависимость цитирования от распределения авторских КС по разделам статьи исследовали Д. Янг, З. Лю, К. Ченг (J. Yang, Z. Liu, X. Cheng) и Г. Йе (G. Ye) [15] на массиве документов из БД China National Knowledge Infrastructure (CNKI), относящемся к тематике *Library and Information Science*. Авторы выделили в полных текстах статей четыре сегмента: *Введение*, *Связанные работы*, *Методы*, *Результаты и обсуждение*, предложили количественный индикатор степени сбалансированности распределения КС по сегментам, выявили 24 типа моделей поведения автора при выборе КС и исследовали статистическую зависимость цитирования статьи от типа модели поведения. Степень сбалансированности положительно коррелирует как с цитированием публикации, так и с количеством ее загрузок. В частности, присутствие авторских КС в разделах *Введение* и *Результаты и обсуждение* оказывает существенное положительное влияние на цитирование.

Исследователи С. Уддин и А. Хан (S. Uddin, A. Khan) [4] изучили влияние на цитируемость статей четырех статистических показателей КС, выбранных автором (рост частоты, разнообразие, количество, а также процент новых КС). Корреляционно-регрессионный анализ показал, что все эти факторы имели значимую положительную корреляционную связь с количеством цитирований, за исключением доли новых КС, имевшей значимую отрицательную корреляцию. Авторы работы [4] предложили объяснение этой отрицательной корреляции: недавно появившийся термин может быть недостаточно хорошо известен научному сообществу или не принят им как тема для исследования, что приводит к сокращению числа читателей статьи и, соответственно, количества цитирований публикации.

В отличие от авторов, стремящихся повысить видимость и цитируемость своих статей, у читателей нет такой задачи, поэтому они назначают КС этим статьям с той же целью, что и редакторы реферативных служб,— для создания поискового образа, достаточно точно и полно отражающего смысловое содержание при последующей навигации в своих тематических коллекциях статей. Поэтому поведенческая модель читателя должна отличаться от поведенческой модели автора, что было подтверждено в исследовании Я.Н. Чена и Х.Р. Ке (Y.N. Chen, H.R. Ke) [11], в котором проведен сравнительный анализ КС, присвоенных авторами и читателями статей. Авторские КС были взяты из БД Library and Information Science Abstracts (LISA). Теги, присвоенные читателями этим же статьям, были взяты с сайта CiteULike, который на тот момент функционировал (прекратил работу в 2019 г.). CiteULike был бесплатным веб-сервисом для ученых, который позволял сохранять, организовывать ссылки на научные статьи и обмениваться этими ссылками, а также систематизировать их при помощи тегов. Около трети (31,97%) тегов CiteULike и только 18,38% дескрипторов LISA были идентичны соответствующим КС в заглавиях статей, что говорит о существенном различии поведения авторов и читателей в извлечении ключевых терминов из заглавий статей.

Если целью ПО, присваивающего КС документу, является не продвижение документа вверх в поисковых сервисах (т.е. поисковая оптимизация), а выражение его понятийного ядра, то поведенческая модель ПО должна отличаться от поведенческой модели авторов статьи. В работе М. Трипати с соавт. (M. Tripathi et al.) [8] одной из задач было сравнение двух массивов КС: авторских и КС плюс, автоматически присваиваемых в БД WoS CC. Основной задачей авторов было выявление тенденций исследований в области социальных и гуманитарных наук в Индии, они также анализировали исходные данные и обнаружили статистически значимую отрицательную корреляцию между количеством авторских КС и КС плюс в разных областях исследований. Сравнение КС с терминами в заглавиях статей показало, что автоматически присвоенные КС плюс значитель-

но чаще встречаются в заглавиях, чем авторские КС. Хотя бы одно КС плюс встречалось в заглавиях 53,4% статей. При этом хотя бы одно авторское КС встречалось в заглавиях лишь 34,5% публикаций.

Согласно исследованиям [8; 11], целеполагание и поведенческая модель выбора КС влияют на результаты координатного индексирования и, как следствие, порождают статистические различия между массивами КС, полученными из различных источников. Аналогичный результат был зафиксирован и авторами данной статьи: мера включения авторских КС в заглавие статьи оказалась существенно меньше по сравнению с редакторскими КС.

Сопоставление авторских КС с дескрипторами, присвоенными этим же статьям в четырех реферативных БД, выполнено в работе И. Гил-Лейва и А. Арройо Алонсо (I. Gil-Leiva, A. Alonso Arroyo) [16]. Материал был взят из следующих БД: INSPEC (Information Service for Physics, Engineering, and Computing), CAB (Current Agriculture Bibliography), ISTA (Information Science and Technology Abstracts), LISA (Library and Information Science Abstracts). Исследование показало: почти 25% всех авторских КС появились в той же форме, что и дескрипторы в БД, и, если бы они были нормализованы, в дескрипторах обнаружился бы еще 21%. Это означает, что в общей сложности до 46% авторских КС присутствовали в массиве дескрипторов — либо в идентичной форме, либо в нормализованном виде.

КС и автоматические предметные индексы в открытых библиографических ресурсах

Быстрое наполнение библиографических БД свободного доступа, в частности The Lens, Semantic Scholar и OpenAlex, открытыми метаданными, полученными из Medline, Microsoft Academic[4] и CrossRef, приводит к тому, что доля метаданных научных публикаций, имеющих авторские КС, в этих ресурсах резко сокращается. Так, по данным системы The Lens[5], в настоящее время из 291 млн публикаций только 21,6 млн (8,9%) имеют в наборе метаданных авторские КС. Результаты, полученные авторами настоящей работы при исследовании тематической выборки научных публикаций по перспективной композитной керамике, подтверждают неполноту тематических выборок КС в БД The Lens и OpenAlex. Например, в тематической выборке научных статей из БД The Lens на один документ приходится 0,46 КС в 2018 г. и 0,64 КС в 2024 г. Кроме того, выборки КС в БД The Lens преимущественно состояли из однократно встречаемых авторских слов и словосочетаний. В OpenAlex складывается аналогичная ситуация со сгенерированными системой ключевыми терминами — 0,55 КС на один документ в 2018 г. и 0,42 КС — в 2024 г. Доля метаданных статей с КС в этом ресурсе даже уменьшилась в 2024 г. по сравнению с 2018 г.

Восполнить отсутствие авторских КС помогает инструментарий открытых библиографических БД, который автоматически присваивает публикациям предметные индексы. Эти индексы можно использовать как функциональные эквиваленты авторских КС. Так, в БД The Lens каждому документу приписывается совокупность областей исследования (*Fields of Study*), выбранных из 698 тыс. предметных индексов, которые определяются на основе машинного обучения и парсинга всех доступных текстовых данных в описаниях статей. В открытой версии БД Dimensions каждая публикация сопровождается набором метаданных, выделяющих ее ключевые аспекты (*key highlights*): цели исследования, методы, характеристики, результаты, а также десять основных КС (*top keywords*) и резюме (*summarize*). В OpenAlex каждому документу присваивается иерархический набор предметных индексов: научное направление (*Domain*), научная область (*Field*), предметная категория (*Subfield*) и предметная рубрика (*Topic*).

По мере роста количества научных публикаций и прогресса искусственного интеллекта (ИИ) в области семантической обработки текстов неуклонно будут расти массивы КС, присвоенных научным публикациям автоматически. В этих условиях особую актуальность приобретают методики количественного сравнения массивов КС, присвоенных различными индексаторами, независимо от того, являются ли они людьми или автоматическими системами. Статистические различия между массивами КС отражают различия в ментальных или поведенческих моделях индексаторов, а результаты количественного сравнения массивов позволяют судить о степени расхождения между моделями или алгоритмами индексирования.

МАТЕРИАЛЫ И МЕТОДЫ

Обоснование адекватности используемых методов

Методы измерения лексического разнообразия текстов давно разрабатываются и исследуются в лексикологии, однако обзор этой темы выходит за рамки данной статьи. Отметим лишь, что в монографии [26] описано около десятка различных мер лексического разнообразия, изобретенных в XX в. В XXI в. к ним добавилась MTLD (англ. *The Measure of Textual Lexical Diversity* — «мера текстового лексического

[4] Проект Microsoft Academic закрыт в конце 2021 г., его данные были загружены в БД OpenAlex.

[5] The Lens. URL: https://www.lens.org/ (дата обращения: 20.09.2025).

разнообразия») [27]. Применительно к анализу КС может использоваться комбинация нескольких мер лексического разнообразия. Например, Д. Пауэлл с соавт. (J. Powell et al.) [28] с помощью четырех мер: *Type–Token Ratio* (TTR), *Hypergeometric Distribution Diversity* (HD-D), Maas index (MTLD) — оценили лексическое разнообразие терминов, автоматически извлеченных из заглавий статей.

Мы используем самую первую и простую меру лексического разнообразия — *Type–Token Ratio* (TTR). TTR рассчитывается путем деления количества уникальных слов (или типов) в тексте на количество токенов (или общее количество слов), которые он содержит:

$$\mathrm{TTR} = V / N,$$

где V — количество уникальных слов; N — количество всех слов.

У этой формулы имеется существенный недостаток: ее значение нелинейно зависит от размера измеряемого текста. Это происходит потому, что с увеличением длины текста количество уникальных слов растет медленнее, чем количество всех слов. Последующие меры лексического разнообразия были изобретены главным образом для того, чтобы избавиться от этого фундаментального недостатка и сделать сравнимым лексическое разнообразие текстов, существенно различающихся по длине. Как мы покажем далее, в нашем случае длина текстов (объемы массивов авторских и редакторских КС) различалась незначительно, поэтому выбор меры TTR для оценки лексического разнообразия был правомерным.

В отличие от измерения лексического разнообразия, оценка новизны КС представляет собой сравнительно новую и менее исследованную научную проблему. В нашей работе новизна измерялась методом, описанным в статье [4]:

$$Percentage\ of\ new\ keywords = \\ = \frac{Number\ of\ new\ keywords}{Total\ number\ of\ keywords} \times 100\%.$$

Однако в научной литературе можно встретить и иные подходы к анализу динамики КС. Например, в рамках другого исследования Ф. Песет с соавт. (F. Peset et al.) [5] также выявляли новые авторские КС, но затем фокусировались не на оценке самой новизны, а на анализе выживаемости этих новых слов в течение последующего 10-летнего периода при помощи кривых Каплана — Мейера (*Kaplan–Meier curves*). Этот метод широко используется в медицине для статистических исследований выживаемости пациентов. Анализ новых КС показал, что в течение первого года 65,3% КС исчезли и больше не использовались до конца анализируемого периода. С этого момента наблюдается плавное снижение кривой Каплана — Мейера, которое слегка усиливается к концу. Только 11,5% КС дожили до конца 10-летнего периода.

В свою очередь, Д. Янг с соавт. (J. Yang et al.) [29] применили подход, основанный на принципах экологии знаний, к выявлению новых тем исследований, используя при этом авторские КС, извлеченные из электронной библиотеки Ассоциации вычислительной техники (Association for Computing Machinery, ACM) ACM Digital Library за период с 1969 по 2018 г. Они оценивали новизну не КС, а темы исследования в году t по следующей формуле:

$$Novelty = \frac{1}{t - t_0 + 1} \cdot e^{-\frac{f_t}{\sum_{t_0}^{t} f_t}},$$

где t_0 — первый год с ненулевой частотой встречаемости темы исследования в списке лет, упорядоченном по убыванию; f_t — частота встречаемости исследовательской темы в году t. По нашему мнению, этот подход также можно применять к оценке новизны КС, если оценивать их новизну в каждом году статистического периода.

Для оценки меры включения применительно к анализу КС сложные математические модели пока не применяются. Обычно указывается средний процент КС автора, встречающихся в других элементах метаданных [13], или вычисляется коэффициент, являющийся аналогом одного из известных коэффициентов сходства или меры включения. Например, *Overlapping Index* (индекс перекрытия) и *Redundancy Index* (индекс избыточности), судя по формулам в [8], являются полным аналогом *Jaccard Index*[6] (коэффициента Жаккара) и меры включения, которые используются в этой статье.

В данном исследовании массив авторских КС сравнивается с результатами координатного индексирования статей при подготовке Реферативного журнала (РЖ) / БД ВИНИТИ. Координатное индексирование КС выполняется редакторами реферативной БД ВИНИТИ с целью создания предметных указателей к тематическим выпускам этого информационного ресурса. Редакторы являются профессиональными информационными работниками,

[6] Wikipedia contributors. 17 September 2025. Jaccard index. In: Wikipedia, The Free Encyclopedia. URL: https://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=1311865137 (дата обращения: 25.11.2025).

выполняющими индексирование и классификацию научных публикаций. Все они являются профильными специалистами.

Институциональные различия авторского и редакторского индексирования

Ответ на вопрос «Почему авторы и редакторы индексируют по-разному?» следует искать не столько в научных публикациях, сколько в нормативной базе, регламентирующей индексирование в библиотечно-информационных центрах. Если на модель поведения редакторов оказывают прямое влияние стандартизация и средства автоматизированного контроля лексики по словарю, то на модель поведения авторов могут воздействовать рекомендации, публикуемые на сайтах научных журналов.

Авторы обычно индексируют публикации КС, не проверяя свой выбор по словарям терминов и ограничиваясь орфографическим контролем. Профильные специалисты (информационные и библиотечные работники), как правило, используют для замены и контроля специальные словари — тезаурусы, которые содержат дескрипторы терминов и их синонимы. Согласно п. 2.62 стандарта ГОСТ Р 7.0.91-2015 (ИСО 25964-1:2011), тезаурусом называется «контрольный… структурированный словарь… в котором понятия… представлены терминами… организованными таким образом, что отношения между понятиями… представлены эксплицитно, и дескрипторы… снабжены указателями перехода от синонимов… и квазисинонимов… Задачей, решаемой тезаурусом, является обеспечение того, чтобы как индексатор, так и пользователь выбирали для представления данного содержания (предмета) один и тот же дескриптор или комбинацию дескрипторов. По этой причине тезаурус оптимизирован так, чтобы стать средством навигации и терминологического покрытия предметной области для человека»[7].

Индексирование документов и построение тезаурусов регламентируется стандартами в России и за рубежом, причем российские стандарты являются переводами международных стандартов ISO [30]. Редакторы тематических выпусков БД ВИНИТИ руководствуются нормативными документами, главным из которых является стандарт ГОСТ Р 7.0.66-2010 (ИСО 5963:1985). Процитируем фрагменты этого стандарта, важные для понимания методики индексирования, которой пользуются редакторы:

«6.1. Число характеристик и понятий, отраженных в ПОД [поисковом образе документа], определяет его полноту и является важнейшим показателем качества индексирования.

6.1.1. В ПОД необходимо отразить все понятия всех тем, которые могут иметь ценность для пользователей системы.

<…>

А.4. Если найденная лексическая единица — аскриптор, то заменить ее указанным в ссылке дескриптором (или комбинацией дескрипторов)»[8].

Таким образом, индексирование, согласно стандарту, должно обеспечить достаточную полноту отражения понятий и тем статьи, а синонимы дескрипторов (аскрипторы) должны быть заменены на дескрипторы. ГОСТ не содержит никаких рекомендаций или запретов включать термины из заглавия или аннотации в ПОД.

Исследование Е. В. Тихоновой и М. А. Косычевой показало, что авторы, как правило, подбирают КС, дополняющие основную тему статьи, и активно используют синонимы, родственные термины, а также аббревиатуры. При этом во избежание повторов авторы избегают дублирования слов, содержащихся в заглавии статьи [31].

Если редактор, следуя стандарту, стремится унифицировать терминологию и последовательно использовать предпочтительный синоним (дескриптор), то автор, напротив, может сознательно применять аскрипторы, в том числе аббревиатуры. Если редактор выбирает термины независимо от их наличия в других элементах метаданных (заглавии и аннотации), то автор может сознательно отказаться от использования термина, присутствующего в этих элементах, или заменить термин синонимом. Приведем два примера рекомендаций авторам, взятые с сайта, предлагающего услуги для авторов, и с сайта научного журнала. Эти рекомендации демонстрируют различие подходов к выбору КС. На сайте https://научныепереводы.рф в разделе «Аспиранту» в статье «Подбор ключевых слов» приведена следующая рекомендация: «Важно: не используйте слова из заголовка. Это неэффективно, так как поисковая система в любом случае его покажет. Ключевое слово должно дополнять, уточнять, расшифровывать термины из заглавия статьи, но не дублировать их»[9].

[7] ГОСТ Р 7.0.91–2015 (ИСО 25964–1:2011). Система стандартов по информации, библиотечному и издательскому делу. Тезаурусы для информационного поиска, с. 9. URL: https://ifap.ru/library/gost/70912015.pdf (дата обращения: 25.11.2025).

[8] ГОСТ Р 7.0.66–2010 (ИСО 5963:1985). Система стандартов по информации, библиотечному и издательскому делу. Индексирование документов. Общие требования к координатному индексированию. Москва: ФГУП «Стандартинформ», с. 3, 9. URL: https://ifap.ru/library/gost/70662010.pdf (дата обращения: 25.11.2025).

[9] Субачев Ю. В. Подбор ключевых слов. Научные переводы: [сайт]. 2024. URL: https://научныепереводы.рф/podbor-klyuchevyh-slov/ (дата обращения: 01.10.2025).

На сайте журнала «Российское право: образование, практика, наука» опубликована следующая рекомендация авторам: «Ключевые слова должны отражать основную терминологию по проблеме, раскрытой в статье. Старайтесь избегать общих слов и выражений, не раскрывающих специфику статьи (типа "система принципов"). Ключевые слова должны быть сформулированы максимально конкретно и в совокупности давать полное представление о тезаурусе статьи»[10].

В первом случае приоритетом является повышение видимости статьи в поисковых системах и КС рассматриваются как дополнение к остальным элементам метаданных. Во втором случае главная цель — обеспечить полноту индексирования, поэтому КС выступают как самостоятельный, независимый элемент метаданных, отражающий основные термины статьи. При таком различии в подходах авторов и редакторов к координатному индексированию следует ожидать подтверждения нашей гипотезы о существовании устойчивых статистических различий между двумя массивами КС (авторских и редакторских), присвоенных одной и той же выборке статей.

Наборы данных: происхождение и формирование

В настоящем исследовании представлены статистические различия, выявленные на пяти тематических коллекциях русскоязычных статей по научным областям, соответствующим пяти ветвям Государственного рубрикатора научно-технической информации (ГРНТИ):

52. Горное дело;
53. Металлургия;
55. Машиностроение;
61. Химическая технология. Химическая промышленность;
73. Транспорт.

Тематические коллекции для исследования аккумулированы из технологической БД, которая используется в процессе производства выпусков БД ВИНИТИ РАН, доступной внешним пользователям. Поскольку редакторы классифицируют документы кодами ГРНТИ и индексируют КС на русском языке, в статистические выборки были включены статьи из русскоязычных научных журналов, у которых заполнены следующие элементы метаданных (в скобках приведены обозначения элементов):

– код(ы) ГРНТИ (RUBR)[11];
– заглавие статьи (TITLE);
– авторские КС (AU.KW);
– авторская аннотация (AU.ABST);
– редакторские КС (ED.KW);
– реферат (ED.ABST).

Элементы TITLE, AU.KW и AU.ABST загружены в технологическую БД из доступных электронных ресурсов. Элементы RUBR, ED.KW и ED.ABST появились в технологической БД в результате обработки документов в ВИНИТИ РАН. Поскольку элемент TITLE не является уникальным (заглавия разных статей могут совпадать), у каждой статьи в технологической БД имеется уникальный идентификатор (ID).

Критерии включения документов в тематические коллекции

Процесс формирования тематических коллекций схематически представлен на рис. 1. Перечислим критерии отбора и включения статей в тематические коллекции.

Критерии отбора:

– источники: русскоязычные рецензируемые научные журналы, издающиеся в России и включенные в Российский индекс научного цитирования (РИНЦ);
– ретроспектива: описания статей загружены в реферативную БД ВИНИТИ в 2021–2024 гг.;
– полнота данных: непустые элементы данных TITLE, AU.ABST, AU.KW, RUBR, ED.ABST, ED.KW.

Критерием для включения статьи в тематическую коллекцию служат первые две цифры кода ГРНТИ, присвоенного статье в реферативной БД ВИНИТИ.

Поскольку статья может быть классифицирована несколькими кодами ГРНТИ, встречаются случаи, когда ей присвоено более одного кода ГРНТИ на первом уровне иерархии. Например, статья с кодами ГРНТИ, начинающимися на 55 и 73, будет включена в обе тематические коллекции: «55.* Машиностроение» и «73.* Транспорт». По этой причине тематические коллекции частично пересекаются.

Предобработка текста и нормализация слов

Для сравнения КС использовалась модель текста, которая называется «мешок слов» (*bag of words*). В этой модели текст упрощенно представлен как неупорядоченный набор слов без сведений о связях между ними. Список КС (т.е. слов и словосочетаний), присвоенных статье автором или редактором, был вначале представлен в виде «мешка словоформ», приведенных к нижнему регистру. Затем на основе «мешка словоформ» был сформирован «мешок

[10] Руководство для авторов. Рекомендации по оформлению аннотации и ключевых слов. Российское право: образование, практика, наука: [сайт]. 2024. URL: https://rospravojournal.usla.ru/index.php/rp/about/annotationandkeywords (дата обращения: 01.10.2025).

[11] Если документ политематический, он может быть классифицирован более чем одним кодом ГРНТИ.

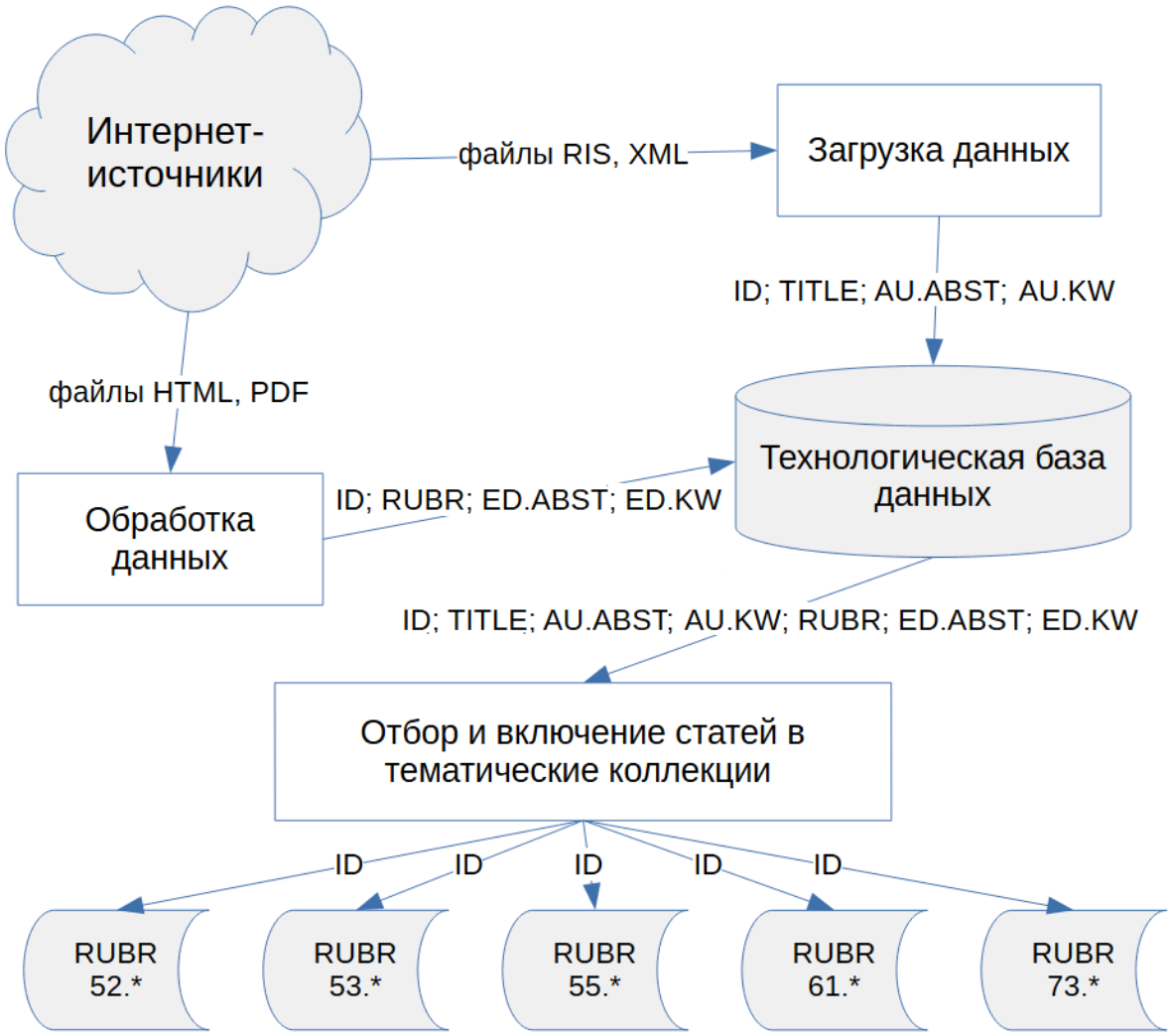

Рис. 1. Схема формирования тематических коллекций
Figure 1. Scheme of formation of thematic collections

основ слов» с применением стемминга (*stemming*) по алгоритму Портера для русского языка[12]. Алгоритм Портера выделяет основу слова путем удаления суффиксов и окончаний согласно набору правил без использования словарей. По сравнению с лемматизацией слов, которая выполняет более точный морфологический анализ с использованием словарей, стемминг гораздо быстрее работает, но может приводить к ошибкам, так как у разных слов могут быть одинаковые основы. К словам, записанным латинскими буквами или имеющим длину менее четырех символов, стемминг не применялся. Предлоги и союзы русского языка были помечены как стоп-слова и в «мешок основ слов» не включались (табл. 1).

[12] URL: http://snowball.tartarus.org/algorithms/russian/stemmer.html (дата обращения: 01.10.2025).

Описательные характеристики собранных коллекций

Все тематические коллекции статей были отобраны из библиографических записей, загруженных в БД ВИНИТИ в 2021–2024 гг. (табл. 2).

РЕЗУЛЬТАТЫ

Разнообразие КС

В качестве меры разнообразия использована мера TTR, вычисленная авторами как отношение количества уникальных основ слов к общему количеству основ слов (т.е. доля уникальных основ слов). Если бы каждая основа слова встречалась в выборке текстов один раз, то мера разнообразия приняла бы максимальное значение, равное единице. Чем чаще повторяются основы слов, тем меньше мера разнообразия.

Таблица 1. Преобразование списка КС в «мешок основ слов»
Table 1. Transforming a list of keywords into a bag of word stems

Этап нормализации	Авторские КС	Редакторские КС
1. Список КС	1-алкил-3-метилимидазолий хлорид; ионная жидкость; проточный микроканальный реактор; реакция	1-алкил-3-метил-, имидазолий, хлориды, получение; ионные жидкости; оптимизация; реакторы
2. Мешок словоформ	1-алкил-3-метилимидазолий; хлорид; ионная; жидкость; проточный; микроканальный; реактор; реакция	1-алкил-3-метил- имидазолий; хлориды; получение; ионные; жидкости; оптимизация; реакторы
3. Мешок основ слов	1-алкил-3-метилимидазол; хлорид; ион; жидкост; проточн; микроканальн; реактор; реакц	1-алкил-3-метил- имидазол; хлорид; получен; ион; жидкост; оптимизац; реактор

Таблица 2. Количественные показатели тематических коллекций статей
Table 2. Quantitative indicators of thematic collections of articles

Коллекция	Количество статей в коллекции	Общее количество словоформ*				
		TITLE	AU.ABST	ED.ABST	AU.KW	ED.KW
52	10 402	96 852	1 098 803	1 017 305	125 340	91 479
53	10 894	103 789	1 068 716	1 014 027	109 594	83 984
55	18 967	171 756	1 747 193	1 614 261	192 490	149 774
61	14 183	124 018	1 440 211	1 312 741	128 950	139 298
73	10 531	92 998	990 284	940 499	113 411	91 829

*Без учета стоп-слов.

Разница в количестве основ слов редакторских и авторских КС была незначительной: варьировалась от 7 до 37%. Поэтому мы сочли допустимым использование меры TTR для оценки лексического разнообразия. Хотя в дальнейших исследованиях желательно использовать другие меры, которые позволяют нивелировать разницу в длине текстов. Согласно данным табл. 3, только в коллекции «61» количество основ авторских КС оказалось на 7,4% меньше, в остальных коллекциях количество основ авторских КС было больше, при этом разница варьировалась от 23,5 до 37%. Другой способ получить более точную оценку лексического разнообразия — использовать лемматизацию слов вместо стемминга.

Согласно данным табл. 3, во всех пяти тематических коллекциях авторские КС имеют более разнообразную лексику по сравнению с редакторскими. Более того, в четырех коллекциях из пяти мера разнообразия авторских КС превысила меру разнообразия заглавий статей. Меры разнообразия лексики в авторской аннотации и реферате почти не различаются и существенно меньше, чем в заглавиях и КС. Сравнительно низкая мера разнообразия лексики аннотации и реферата объясняется значительным количеством общеупотребительных слов и общенаучных терминов, которые часто повторяются.

Новизна КС

В качестве меры новизны авторы использовали коэффициент новизны *Novelty* — показатель, отражающий степень обновления основ слов в элементе метаданных тематической коллекции:

$$Novelty = (b - c) / b,$$

где b — количество уникальных основ слов, выявленных в 2023–2024 гг.; c — количество уникальных основ слов, выявленных в оба периода: в 2021–2022 и 2023–2024 гг.

Для повышения интерпретируемости выявленных различий мы вычислили коэффициент выбытия основ слов:

$$Disposal = (a - c) / a,$$

где a — количество уникальных основ слов, которые встретились в 2021–2022 гг.; c — количество уникальных основ слов, которые встретились в оба периода: в 2021–2022 и 2023–2024 гг. (табл. 4).

Хотя во всех пяти тематических коллекциях коэффициент новизны авторских КС оказался выше, чем коэффициент новизны редакторских КС, разница незначительна и варьируется от 0,9% (в коллекции «61») до 7,4% (в коллекции «73»).

Таблица 3. Разнообразие лексики в элементах метаданных
Table 3. Lexical diversity in metadata elements

Элементы метаданных	Количество уникальных основ слов	Общее количество основ слов	Доля уникальных основ слов, %
53. Горное дело			
AU.ABST	40 184	1 098 803	3,66
ED.ABST	37 492	1 017 305	3,69
AU.KW	13 499	125 340	10,77
ED.KW	8 370	91 479	9,15
TITLE	10 079	96 852	10,41
53. Металлургия			
AU.ABST	51 002	1 068 716	4,8
ED.ABST	48 574	1 014 027	4,8
AU.KW	12 933	109 594	11,8
ED.KW	8 181	83 984	9,7
TITLE	12 429	103 789	12,0
55. Машиностроение			
AU.ABST	51 936	1 747 193	2,97
ED.ABST	47 579	1 614 261	2,95
AU.KW	18 163	192 490	9,44
ED.KW	11 699	149 774	7,81
TITLE	14 793	171 756	8,61
61. Химическая технология. Химическая промышленность			
AU.ABST	75 876	1 440 211	5,3
ED.ABST	70 267	1 312 741	5,4
AU.KW	19 786	128 950	15,3
ED.KW	17 556	139 298	12,6
TITLE	17 422	124 018	14,1
73. Транспорт			
AU.ABST	31 030	990 284	3,1
ED.ABST	29 478	940 499	3,1
AU.KW	12 122	113 411	10,7
ED.KW	7 606	91 829	8,3
TITLE	8 779	92 998	9,4

Таблица 4. Коэффициенты новизны и выбытия основ слов
Table 4. Coefficients of novelty and disposal of word stems

Элементы метаданных	*a*	*b*	*c*	Novelty, %	Disposal, %
52. Горное дело					
AU.ABST	28 128	25 791	13 735	46,7	51,17
ED.ABST	26 812	23 897	13 217	44,7	50,70
AU.KW	9 226	9 133	4 860	46,8	47,32
ED.KW	6 169	5 467	3 266	40,3	47,06
TITLE	7 117	6 653	3 691	44,5	48,14
53. Металлургия					
AU.ABST	33 883	32 104	14 985	53,3	55,77
ED.ABST	32 491	30 585	14 502	52,6	55,37
AU.KW	8 751	8 605	4 423	48,6	49,46
ED.KW	5 680	5 380	2 879	46,5	49,31
TITLE	8 533	7 984	4 088	48,8	52,09
55. Машиностроение					
AU.ABST	32 571	37 315	17 950	51,9	44,89
ED.ABST	30 550	34 014	16 985	50,1	44,40
AU.KW	11 607	13 250	6 694	49,5	42,33
ED.KW	7 922	8 592	4 815	44,0	39,22
TITLE	9 798	10 632	5 637	47,0	42,47
61. Химическая технология. Химическая промышленность					
AU.ABST	47 744	48 749	20 617	57,7	56,82
ED.ABST	45 396	44 405	19 534	56,0	56,97
AU.KW	12 366	13 591	6 171	54,6	50,10
ED.KW	11 295	11 656	5 395	53,7	52,24
TITLE	11 235	11 481	5 294	53,9	52,88
73. Транспорт					
AU.ABST	20 278	22 292	11 540	48,2	43,09
ED.ABST	19 402	21 364	11 288	47,2	41,82
AU.KW	7 541	8 782	4 201	52,2	44,29
ED.KW	5 129	5 532	3 055	44,8	40,44
TITLE	5 687	6 450	3 358	47,9	40,95

Таблица 5. Средние значения меры включения / коэффициента сходства КС, %
Table 5. Average measures of keywords inclusion and coefficient of similarity

A	*B*				
	AU.KW	ED.KW	TITLE	AU.ABST	ED.ABST
52. Горное дело					
AU.KW	100 / 100	39,2 / 28,6	31,6 / 20,6	65,8 / 10,2	64,4 / 10,4
ED.KW	49,0 / 28,6	100 / 100	40,2 / 25,4	73,7 / 9,8	72,7 / 10,1
53. Металлургия					
AU.KW	100 / 100	37,8 / 26,7	35,7 / 21,9	68,2 / 10,3	67,0 / 10,4
ED.KW	45,5 / 26,7	100 / 100	41,9 / 24,5	69,4 / 8,8	69,4 / 9,1
55. Машиностроение					
AU.KW	100 / 100	34,9 / 24,1	36,4 / 23,1	67,8 / 11,2	66,7 / 11,4
ED.KW	40,7 / 24,1	100 / 100	49,6 / 31,1	70,0 / 10,3	69,5 / 10,5
61. Химическая технология. Химическая промышленность					
AU.KW	100 / 100	36,2 / 21,3	34,6 / 20,9	67,3 / 9,1	65,9 / 9,3
ED.KW	33,5 / 21,3	100 / 100	48,8 / 35,0	67,0 / 10,4	66,1 / 10,8
73. Транспорт					
AU.KW	100 / 100	41,5 / 29,0	35,3 / 22,8	65,8 / 10,7	65,1 / 10,9
ED.KW	45,2 / 29,0	100 / 100	39,4 / 24,9	65,6 / 10,0	65,3 / 10,2

Во всех пяти тематических коллекциях коэффициент новизны авторской аннотации незначительно превысил коэффициент новизны реферата: разница коэффициентов варьируется от 0,7 до 2,0%. В четырех коллекциях из пяти коэффициент выбытия авторских КС незначительно превысил коэффициент выбытия редакторских КС. Между коэффициентами новизны и выбытия наблюдается заметная положительная корреляция: коэффициент корреляции Пирсона, оценивающий линейную статистическую связь между двумя переменными, равен 0,583.

Включение КС в другие элементы метаданных

Для вычисления меры включения множества A в множество B мы используем формулу

$$K(A; B) = n(A \cap B) / n(A),$$

где $n(A)$ — количество элементов множества A; $n(A \cap B)$ — количество элементов, принадлежащих обоим множествам, т.е. пересечению множеств A и B.

Данная мера включения показывает, какая доля элементов множества A является элементом множества B. Однако без симметричного коэффициента сходства трудно оценить пересечение и сравнимость массивов при различной длине списков слов, поэтому мы также вычислили коэффициент сходства Жаккара:

$$Kj(A; B) = n(A \cap B) / n(A \cup B).$$

Вначале меры включения и коэффициенты сходства были вычислены для каждой статьи в тематической коллекции. Затем внутри каждой коллекции вычислены средние арифметические значения мер включения и коэффициентов сходства (average($K(A; B)$), average($Kj(A; B)$)) (табл. 5).

Графа TITLE в табл. 5 подтверждает, что большинство авторов статей действительно придерживается рекомендации «не дублировать слова, содержащиеся в заглавии статьи». В то же время редакторы, как и предполагалось, индексируют статьи независимо от присутствия КС в заглавии статьи. Во всех пяти тематических коллекциях мера включения авторских КС в заглавие статьи оказалась ниже, чем редакторских. Разница мер включения варьируется от 4,1% (в коллекции «73») до 14,2% (в коллекции «61»).

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Настоящее исследование показало, что различия между авторскими и редакторскими КС являются **системными, статистически устойчивыми и воспроизводимыми** в разных областях технических наук. На пяти независимых тематических выборках

выявлены однонаправленные различия по трем измерениям — **лексическому разнообразию**, **новизне** и **включенности в другие элементы метаданных**. Эти различия могут быть интерпретированы в рамках более широкой исследовательской традиции, изучающей поведенческие и когнитивные модели индексаторов разных типов.

Во-первых, более высокая мера лексического разнообразия в массиве авторских КС последовательно наблюдалась во всех пяти коллекциях. Этот результат согласуется с описанными в литературе закономерностями: авторы стремятся расширять терминологическое поле публикации за счет синонимов, редких терминов, акронимов и узкоспециализированных обозначений, ориентированных на улучшение видимости статьи. Подобная стратегия находит подтверждение в исследовании С. Уддина и А. Хана [4], демонстрирующем положительную связь разнообразия авторских КС с цитируемостью, и в работе И. Гил-Лейва и А. Арройо Алонсо [16], где авторские списки оказались значительно менее нормализованными, чем дескрипторы БД. Редакторские КС, напротив, демонстрируют меньшую вариативность, что является прямым следствием их нормативной функции: индексатор обязан заменять аскрипторы на дескрипторы согласно стандартам и тезаурусу. Аналогичное сужение семантического поля отмечают Я.Н. Чен и Х.Р. Ке [11], которые показали, что профессиональные индексаторы систематически стремятся к терминологической унификации, в отличие от авторов и читателей.

Во-вторых, превосходство авторских КС над редакторскими по коэффициенту новизны согласуется с наблюдениями Ф. Песет с соавт. [5] и Д. Янг с соавт. [29], где «тонкие» терминологические сигналы и появление новых слов впервые фиксируются именно в авторских списках. Новизна редакторских КС закономерно ниже: официальные словари и тезаурусы обновляются с существенным временны́м лагом, о чем также сообщают М. Трипати с соавт. [8], анализируя различия между авторскими терминами и автоматически присвоенными WoS Keywords Plus.

Дополнительно в настоящем исследовании выявлено: новизна авторских аннотаций также стабильно превышает новизну рефератов в БД, что подтверждает нормирующую функцию последних.

В-третьих, значительно более низкая доля авторских КС, включенных в заглавия, подтверждает рекомендацию, распространенную в практике академического письма: не дублировать термины заглавий, а расширять семантическое поле метаданных. В противоположность этому редакторские КС во всех коллекциях демонстрируют бóльшую степень пересечения с заглавиями, что согласуется с наблюдениями М. Трипати с соавт. [8] и И. Гил-Лейва и А. Арройо Алонсо [16], согласно которым дескрипторы и автоматически генерируемые термины чаще совпадают с заголовками, чем авторские КС.

Таким образом, результаты исследования не только воспроизводят ранее обнаруженные закономерности, но и впервые демонстрируют **параллельное проявление этих закономерностей в трех статистических измерениях одновременно** и на масштабных выборках русскоязычных технических дисциплин.

Теоретические следствия

Совокупность выявленных эффектов позволяет интерпретировать различия между массивами в рамках **институциональной логики индексирования**: 1) авторские КС служат механизмом сигнализации о новизне, тематической чувствительности и оптимизации поиска; 2) редакторские термины формируют **нормализованное понятийное ядро**, минимизируя синонимию и обеспечивая согласованность предметной рубрикации.

Отмеченное различие согласуется с концепцией семантического разрыва (*semantic gap*) между пользователями и профессиональными индексаторами, которую описали Я.Н. Чен и Х.Р. Ке [11], и с моделями шаблонов поведения авторов, которые предложили Д. Янг с соавт. [15], С. Уддин и А. Хан [4].

Практическая значимость

Полученные результаты имеют прямое практическое значение:

1) для мониторинга научных трендов: предпочтительным является использование авторских КС, обладающих высокой обновляемостью и тематической чувствительностью;

2) для ретроспективного поиска и классификации: предпочтительнее применять редакторские термины, соответствующие тезаурусу и обеспечивающие стабильную рубрикацию;

3) для построения обучающих наборов данных: целесообразно интегрировать три источника — авторские КС + заглавия + редакторские дескрипторы. Такая комбинация обеспечивает необходимый баланс между нормированием и новизной терминологии. Как показывают Р. Вэй с соавт. [10] и М. Кобо с соавт. [9], подобные гибридные массивы терминов являются наилучшей основой для тематического анализа, автоматического индексирования и построения поисковых запросов к наукометрическим БД.

Ограничения исследования

Интерпретация результатов должна учитывать ряд факторов:

1) применение стемминга Портера может приводить к усечению форм и смешению морфологических вариантов;

2) модель «мешка слов» не различает многословные термины, что особенно критично для технических дисциплин;

3) использование меры TTR ограничено ее зависимостью от длины текста, хотя в данном исследовании разница в длине массивов была умеренной;

4) оценка новизны двухпериодной моделью снижает чувствительность к длинным трендам;

5) отсутствие статистических тестов значимости связано с малым числом агрегированных точек сравнения. Несмотря на это, повторяемость эффектов в пяти различных областях знаний существенно снижает вероятность случайного совпадения.

ЗАКЛЮЧЕНИЕ

Настоящее исследование продемонстрировало, что сопоставление авторских и редакторских КС может служить надежным инструментом выявления различий в практиках координатного индексирования, отражающих различающиеся цели и нормативные основания работы двух групп индексаторов. На основе пяти крупных тематических коллекций русскоязычных статей, относящихся к различным областям технических наук, была разработана и испытана методика количественного сравнения массивов КС по трем измерениям: лексическому разнообразию, новизне и включенности в другие элементы метаданных. Полученные статистические профили массивов позволили установить, что расхождения между авторскими и редакторскими списками носят не ситуативный, а воспроизводимый характер, проявляющийся на уровне крупных дисциплинарных массивов.

Методика исследования продемонстрировала применимость простых и интерпретируемых показателей к большому объему данных из технологической базы реферативной системы и показала, что даже точечные оценки позволяют выявлять устойчивые различия между типами индексирования. Этот результат подтверждает возможность использования КС не только как служебного элемента библиографических описаний, но и как самостоятельного объекта лингвистического, терминологического и наукометрического анализа. В условиях расширения массивов автоматически генерируемых терминов и появления новых источников метаданных возрастает значение методов, способных фиксировать расхождение поведенческих или алгоритмических моделей индексаторов — человека или инструмента.

Полученные результаты подчеркивают актуальность дифференцированного подхода к использованию КС в задачах информационного поиска, тематического анализа и построения обучающих наборов данных. Авторские КС, характеризующиеся более высокой изменчивостью и чувствительностью к появлению новых терминов, могут служить индикаторами терминообразовательной динамики. Редакторские КС, напротив, отражают стабильное понятийное ядро и обеспечивают унифицированное представление предметной области. Сочетание этих массивов открывает возможности для проектирования гибридных информационно-поисковых решений и улучшения качества аналитических инструментов.

Несмотря на методологические ограничения, связанные с использованием стемминга, модели «мешка слов» и точечных статистических оценок, исследование показало, что предложенный подход масштабируем и применим к выборкам различного объема и тематического охвата. Перспективы дальнейшей работы связаны с расширением корпуса, совершенствованием методов нормализации терминов, применением продвинутых мер разнообразия, а также с внедрением статистических тестов значимости различий. В совокупности эти направления позволят углубить понимание механизмов формирования терминологического пространства научных публикаций и усилят теоретическую и практическую ценность КС как объекта аналитики.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

CONFLICT OF INTERESTS

The authors declare no relevant conflict of interests.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Singh P., Singh V.K., Kanaujia A. Exploring the publication metadata fields in Web of Science, Scopus and Dimensions: Possibilities and ease of doing scientometric analysis. *Journal of Scientometric Research*. 2024;13(3):715-731. https://doi.org/10.5530/jscires.20041144
2. Stapleton S.C., Dinsmore C.S., Van Kleec D., Ma X. Computer-assisted indexing complements manual selection of subject terms for metadata in specialized collections. *College & Research Libraries.* 2021;82(6):792-807. https://doi.org/10.5860/crl.82.6.792

3. Gbur E.E., Trumbo B.E. Key words and phrases – the key to scholarly visibility and efficiency in an information explosion. *The American Statistician.* 1995;49(1):29-33. https://doi.org/10.1080/00031305.1995.10476108
4. Uddin S., Khan A. The impact of author-selected keywords on citation counts. *Journal of Informetrics*. 2016;10(4):1166-1177. https://doi.org/10.1016/j.joi.2016.10.004
5. Peset F., Garzón-Farinós F., González L.M., et al. Survival analysis of author keywords: An application to the library and information sciences area. *Journal of the Association for Information Science and Technology.* 2020;71(4):462-473. https://doi.org/10.1002/asi.24248
6. Hjørland B. Indexing: Concepts and theory. *Knowledge Organization*. 2018;45(7):609-639. https://doi.org/10.5771/0943-7444-2018-7-609
7. Golub K. Automated subject indexing: An overview. *Cataloging & Classification Quarterly.* 2021;59(8):702-719. https://doi.org/10.1080/01639374.2021.2012311
8. Tripathi M., Kumar S., Sonker S.K., Babbar P. Occurrence of author keywords and keywords plus in social sciences and humanities research: A preliminary study. *COLLNET Journal of Scientometrics and Information Management*. 2018;12(2):215-232. https://doi.org/10.1080/09737766.2018.1436951
9. Cobo M.J., López-Herrera A.G., Herrera-Viedma E., Herrera F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*. 2011;5(1):146-166. https://doi.org/10.1016/j.joi.2010.10.002
10. Wei R.-Z., Liu X.-Y., Lyu P.-H. Bibliometrics of public administration research hotspots: Topic keywords, author keywords, keywords plus analysis. *Heliyon.* 2024;10(21):e39352. https://doi.org/10.1016/j.heliyon.2024.e39352
11. Chen Y.-N., Ke H.-R. A study on mental models of taggers and experts for article indexing based on analysis of keyword usage. *Journal of the Association for Information Science and Technology*. 2014;65(8):1675-1694. https://doi.org/10.1002/asi.23077
12. Lu W., Li X., Liu Z., Cheng Q. How do author-selected keywords function semantically in scientific manuscripts? *Knowledge Organization*. 2019;46(6):403-418. https://doi.org/10.5771/0943-7444-2019-6-402
13. Lu W., Liu Z., Huang Y., et al. How do authors select keywords? A preliminary study of author keyword selection behavior. *Journal of Informetrics*. 2020;14(4):101066. https://doi.org/10.1016/j.joi.2020.101066
14. Дубинина Е.Ю. Выделение ключевых слов текста научной статьи в процессе создания автоматического реферата. *Вестник Воронежского государственного университета. Серия: Филология. Журналистика*. 2020;(1):26-28. URL: http://www.vestnik.vsu.ru/pdf/phylolog/2020/01/2020-01-06.pdf
Dubinina E.Yu. Selection of keywords in a scientific article in the process of creating an automatic abstract. *Proceedings of Voronezh State University. Series: Philology. Journalism.* 2020;(1):26-28. (In Russ.). URL: http://www.vestnik.vsu.ru/pdf/phylolog/2020/01/2020-01-06.pdf
15. Yang J., Liu Z., Cheng X., Ye G. Understanding the keyword adoption behavior patterns of researchers from a functional structure perspective. *Scientometrics.* 2024;129(6):3359-3384. https://doi.org/10.1007/s11192-024-05031-1
16. Gil-Leiva I., Alonso-Arroyo A. Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science and Technology*. 2007;58(8):1175-1187. https://doi.org/10.1002/asi.20595
17. Zhang J., Yu Q., Zheng F., et al. Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. *Journal of the Association for Information Science and Technology.* 2016;67(4):967-972. https://doi.org/10.1002/asi.23437
18. Акоев М.А. Картирование науки и технологии, прогноз развития. *Руководство по наукометрии: индикаторы развития науки и технологии*. Екатеринбург: Изд-во Уральского университета; 2014:164-184. https://doi.org/10.15826/B978-5-7996-1352-5.0007
Akoev M.A. Mapping science and technology, forecasting research and development. In: *Handbook for Scientometrics: Indicators of Science and Technology Development*. Ekaterinburg: Ural University Publ.; 2014:164-184. (In Russ.). https://doi.org/10.15826/B978-5-7996-1352-5.0007
19. Yang S., Han R., Wolfram D., Zhao Y. Visualizing the intellectual structure of information science (2006-2015): Introducing author keyword coupling analysis. *Journal of Informetrics*. 2016;10(1):132-150. https://doi.org/10.1016/j.joi.2015.12.003
20. Lu W., Huang S., Yang J., et al. Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management.* 2021;58(4):102594. https://doi.org/10.1016/j.ipm.2021.102594
21. González L.M., García-Massó X., Pardo-Ibañez A., Peset F., Devís-Devís J. An author keyword analysis for mapping Sport Sciences. *PloS ONE.* 2018;13(8):e0201435. https://doi.org/10.1371/journal.pone.0201435

22. Pearson W.S. Research topics in applied linguistics as keywords from authors and keywords from abstracts: A bibliometric study. In: Meihami H., Esfandiari R., eds. *A Scientometrics Research Perspective in Applied Linguistics*. Cham: Springer; 2024:113-134. https://doi.org/10.1007/978-3-031-51726-6_5
23. Gao J., Wang X. Exploring research hotspots and trends in the field of intelligent diagnosis and treatment from a bibliometric perspective: A comprehensive analysis of Citespace and VOSviewer. In: *Proc. 2024 5th Int. Symp. on Artificial Intelligence for Medicine Science (ISAIMS 2024). (Amsterdam, August 13-17, 2024)*. New York, NY: Association for Computing Machinery; 2024:108-113. https://doi.org/10.1145/3706890.3706908
24. Song C., Chen K., Jin Y., Chen L., Huang Z. Visual analysis of research hotspots and trends in traditional Chinese medicine for depression in the 21st century: A bibliometric study based on citespace and VOSviewer. *Heliyon*. 2025;11(1):e39785. https://doi.org/10.1016/j.heliyon.2024.e39785
25. Nabilah N., Nakamo S.J., Mwakapemba M.L. Mapping the evolution of research themes on ChatGPT integration in education: Thematic and novelty keywords. *Elementaria: Journal of Educational Research.* 2025;3(1):34-44. https://doi.org/10.61166/elm.v3i1.90
26. Malvern D., Richards B., Chipere N., Durán P. *Lexical Diversity and Language Development.* London: Palgrave Macmillan; 2004. 253 p. https://doi.org/10.1057/9780230511804
27. McCarthy P.M., Jarvis S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods.* 2010;42(2):381-392. https://doi.org/10.3758/BRM.42.2.381
28. Powell J., Klein M., Balakireva L. Combining keyphrase extraction and lexical diversity to characterize ideas in publication titles (3; Version 1). *arXiv.* 2022. https://doi.org/10.48550/ARXIV.2208.13978
29. Yang J., Lu W., Hu J., Huang S. A novel emerging topic detection method: A knowledge ecology perspective. *Information Processing & Management*. 2022;59(2):102843. https://doi.org/10.1016/j.ipm.2021.102843
30. Тимошенко И.В. Современные тенденции развития методов и нормативной базы индексирования библиотечных информационных ресурсов. *Научные и технические библиотеки*. 2024;(10):102-122. https://doi.org/10.33186/1027-3689-2024-10-102-122
Timoshenko I.V. The current trends in the development of methods and regulatory framework for indexing library information resources. *Scientific and Technical Libraries*. 2024;(10):102-122. (In Russ.). https://doi.org/10.33186/1027-3689-2024-10-102-122
31. Тихонова Е.В., Косычева М.А. Эффективные ключевые слова: стратегии формулирования. *Health, Food & Biotechnology*. 2021;3(4):7-15. https://doi.org/10.36107/hfb.2021.i4.s122
Tikhonova E.V., Kosycheva M.A. Effective keywords: Strategies for their formulation. *Health, Food & Biotechnology*. 2021;3(4):7-15. (In Russ.). https://doi.org/10.36107/hfb.2021.i4.s122

ИНФОРМАЦИЯ ОБ АВТОРАХ

Олег Владимирович Федорец, кандидат технических наук, заведующий лабораторией средств автоматизации, Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН), г. Москва, Российская Федерация; https://orcid.org/0009-0005-5149-5669; e-mail: ovf@viniti.ru
Наталия Сергеевна Солошенко, кандидат педагогических наук, заведующий отделом комплектования, Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН), г. Москва, Российская Федерация; https://orcid.org/0000-0002-3288-3755; e-mail: solns@viniti.ru

INFORMATION ABOUT THE AUTHORS

Oleg V. Fedorets, Cand. Sci. (Eng.), Head of Automation Tools Laboratory, All-Russian Institute for Scientific and Technical Information, Russian Academy of Sciences (VINITI RAS), Moscow, Russian Federation; https://orcid.org/0009-0005-5149-5669; e-mail: ovf@viniti.ru
Nataliya S. Soloshenko, Cand. Sci. (Educ.), Head of the Acquisitions Department, All-Russian Institute for Scientific and Technical Information, Russian Academy of Sciences (VINITI RAS), Moscow, Russian Federation; https://orcid.org/0000-0002-3288-3755; e-mail: solns@viniti.ru

Поступила в редакцию / Received 09.10.2025
Поступила после рецензирования / Revised 17.11.2025
Принята к публикации / Accepted 05.12.2025