Раздел Информационные процессы
Section Information processes

Научная статья / Research article
https://doi.org/ 10.31432/1994-2443-2023-18-4-21-37
УДК 004.77(051.2) (100)

Информационный потенциал корпуса научных текстов

Валентин Николаевич Комарица

ООО «Научно-исследовательский институт трубопроводного транспорта»
(ООО «НИИ Транснефть), Москва, Россия
KomaritsaVN@niitnn.transneft.ru

Аннотация. В статье рассматриваются общедоступные корпуса текстов, представленные в сети Интернет, дается характеристика и рассматривается потенциал корпусной лингвистики для анализа развития научных трендов, дискурса и изменений в области терминологии.
Представлен набор данных, подготовленный на основе корпуса текстов научных статей в отраслевом научном журнале по транспорту нефти и корпуса Google Books Corpus. Набор данных позволяет рассмотреть изменения в частотности применимости терминов с 1940 по 2019 гг.
Представлены результаты анализа частот использования терминов, сделано сопоставление изменений в технологической отрасли с развитием ключевой лексики.
Результаты показывают, что исследования, сделанные с использованием данных корпусов научно-технических текстов, имеют хороший потенциал для понимания трендов технологического развития и динамики изменений в промышленности и терминоведении.

Ключевые слова: ключевые слова, термины, корпуса текста, Google ngram, технологические тренды

Для цитирования: Комарица В.Н. Информационный потенциал корпуса научных текстов // Информация и инновации. 2023. Т.18, № 4. С. 21 - 37. https://doi.org/ 10.31432/1994-2443-2023-18-4-21-37.

Information Potential of a Corpus of Scientific Texts

Valentin N. Komaritsa

LLC Research Institute of Pipeline Transport (LLC NII Transneft), Moscow, Russian

Abstract. The article considers publicly available corpus of texts presented in the internet, characterises and considers the potential of corpus linguistics for analysing the development of scientific trends, discourse and changes in the field of terminology.
A dataset based on a corpus of texts of scientific articles in a petroleum transport trade journal and the Google Books Corpus is presented. The dataset allows us to examine changes in term usage frequencies from 1940 to 2019.
The results of analyses of term usage frequencies are presented, and a comparison is made between changes in the technology industry and the development of key vocabulary.

The results show that studies made using data from corpuses of scientific and technical texts have good potential for understanding trends in technological development and the dynamics of change in industry and terminology.

***Key words*:** keywords, terms, text corpora, Google ngram, technological trends

For citation: Komaritsa V. N. The information potential of a corpus of scientific texts. *Information and Innovations.* 2023;18(4):21-37. (In Russ.). https://doi.org/10.31432/1994-2443-2023-18-4-21-37.

Введение

Определение изменений, происходящих в процессе технологического развития, является важной стратегической задачей, актуальной как для государственных организаций и служб, так и для научных учреждений и бизнеса. Знание трендов развития позволяет строить прогнозы, принимать решения о дальнейшем развитии государства, общества, компании, выделять финансовые, материальные и интеллектуальные ресурсы [1, 2]. Одним из значимых источников данных для анализа и прогнозирования, являются тексты научных статей, в которых публикуются результаты научной деятельности создающей основы технологического развития. Достоверность и репрезентативность данных, представленных в научных статьях, подтверждается наличием системы рецензирования, востребованностью и цитируемостью научных статей [3, 4].

Научный текст — это совершенно новый тип больших данных. В основном большие данные являются большими, но «короткими» в рамках определенного периода времени — это записи, фиксирующие недавние события, катализируемые Интернетом. Корпус текстов созданный в проекте Google books corpus, это совершенно новый тип больших данных, позволяющий проводить оценку изменений, происходящих в науке во времени [5, 6]. По оценкам Google, в мире насчитывается более 129 млн печатных изданий, в рамках проекта Google Books было оцифровано более 40 млн произведений начиная с 16 века и до наших дней [7]. Это означает, что с помощью цифровых методов можно извлечь информацию из всех этих книг, охватывающих период почти в пять столетий [8]. Проект HathiTrust (hathitrust.org) — это крупнейший репозиторий цифрового контента библиотек, включая контент, оцифрованный проектами Internet Archive и Google Books, а также оцифрованный самостоятельно отдельными библиотеками.

Большие данные, это не те данные, которые получены в результате эксперимента — экспериментальные данные, это данные которые получены по заранее подготовленному плану эксперимента с определенной заданной точностью и которые впоследствии могут быть воспроизведены. Большие же данные часто сопровождаются своей неразберихой. Массив больших данных представляет собой смесь фактов и измерений, сделанных без какой-либо научной цели и с использованием неуниверсальных процедур. Он изобилует ошибками и огромным количеством пробелов и недостающих элементов информации. Исследования с помощью больших данных — это исследования без гипотезы, потому что никогда не будет известно в начале работы, что будет найдено в ее конце [5].

Корпус письменного или устного текста — подобранная совокупность текстов и обработанная по определённым правилам, является базой больших данных. Для исследования корпуса текста применяются специально созданные алгоритмы и компьютерные программы, с помощью которых удобно обрабатывать, быстро выполнять манипуляции с корпусом и получать точные готовые результаты [9, 10]. Размер корпуса может быть как от нескольких до сотен тысяч, так и до миллиона и миллиарда словоформ [7].

Общие корпуса текстов не подходят для изучения определенных предметных областей в силу их большого объема, разнообразного материала, а также отсутствия специальной терминологии, для описания терминологии применяемой в определенной предметной области соз-

даются корпуса узкоспециальных текстов [3]. Корпус — это не только целый массив текста, но и его фрагменты, например, корпус созданный из подрисуночных надписей, тоже является корпусом текста [11].

Классификация корпусов текста осуществляется на основе различных признаков. Пример такой классификации, составленный на основе материалов [12], представлен в табл. 1.

Таблица 1 / Table 1

Классификация корпусов по различным признакам / Classification of enclosures on various grounds

Вид классификации	Признаки, примеры
Тематическая классификация	Признаки: слова или термины, связанные с определенной темой. Пример: новостные статьи, обзоры продуктов, медицинские статьи, юридические документы и т. д.
Эмоциональная классификация	Признаки: слова или фразы, выражающие эмоциональную окраску (положительную, отрицательную, нейтральную). Пример: отзывы, комментарии в социальных сетях.
Языковая классификация	Признаки: синтаксические, морфологические и лексические особенности конкретного языка. Пример: классификация текстов на разных языках.
Уровень сложности	Признаки: длина предложений, сложность слов, использование сложных терминов. Пример: простые тексты, средние, сложные.
Источник	Признаки: автор, стиль письма, тип источника. Пример: тексты из новостных источников, блогов, академических статей.
Формат	Признаки: структура текста, наличие заголовков, списков, цитат. Пример: тексты в виде статей, книг, блогов, твитов и т. д.
Медиа-связанные признаки	Признаки: наличие медиаконтента, описание, теги. Пример: тексты с изображениями, аудиофайлами, видеороликами.
Классификация по времени	Признаки: дата, временные метки в тексте. Пример: новости за определенный период, исторические тексты и т. д.
Социальные признаки	Признаки: упоминания пользователей, хэштеги, ссылки. Пример: тексты из социальных сетей, форумов.
Классификация по качеству	Признаки: наличие подтвержденных фактов, качество источника. Пример: проверенные и непроверенные источники, фейковые новости.

Источник: составлено автором на основе [12]
Source: compiled by the author using [12]

Корпус текста имеет довольно обширный функционал, для того, чтобы использовать его в полном объеме, необходимо тексты в корпусе представить с дополнительной служебной информацией, которая называется разметкой. Разметка может содержать информацию о словах (терминах) — их исходных формах, которые употребляются в тексте, с указанием какой частью речи они является в предложении. Благодаря разметке, для исследования Корпуса можно применять специальные алгоритмы и получать более полную информацию — пример фрагмента кода разметки корпуса научного текста (1). В статье [13] предложено решение по разметке корпуса, основанное на структурных особенностях научно-технических текстов. Текст статьи предлагается представить в виде графа, вершинами и ребрами которого являются полноценные структурные элементы научной статьи.

```
<?xml version="1.0" encoding="utf-8" ?>                                    (1)
<corpus>
<se>
<w><ana lex="контрольный" gr="PRAEDIC"></ana>контроль`ный</w>
<w>< ana lex="трубопровод" gr="V,ipf,intr,act=n,sg,prate,indic"></ana>трубопровод`ный</w>
<w><ana lex="дефект" gr="CONJ"></ana>дефект`ный</w>
<w><ana lex="авария" gr="V,ipf,intr,act-inf"></ana>авари`йный</w>
<w><ana lex="нефть" gr="CONJ"></ana> нефть</w>
<w><ana lex="коррозия" gr="V,ipf,intr,act=inf"></ana>коррози`онное</w> ,
<w><ana lex="сварка" gr="CONJ"></ana>свар`ной</w>
<w><ano lex="напряжение" gr="V,pf,intr,med-inf"></ana>напряжен`ный</w>
 <w>< ano lex="магистральный" gr="PR"></ana>магистральный</w>
```

При создании текстовых корпусов применяются следующие подходы: 1) размер корпуса увеличивается со временем и содержит разнообразные данные; 2) корпус представляет собой определенный «языковой срез» в данный период времени; 3) корпусы не имеют строгой основы для построения выборки. Данные собираются для решения конкретной задачи.

Корпуса можно создавать самим или использовать готовые корпуса, представленные в сети Интернет, вот некоторые из таких корпусов:

1. Национальный корпус русского языка — информационно-справочная система по русскому языку на основе представительного электронного собрания текстов на русском языке, общим объемом более 2 млрд слов, оснащенная лингвистической разметкой и инструментами поиска [14].

2. Генеральный интернет-корпус русского языка — мегакорпус (около 20 млрд слов), созданный при помощи автоматической технологии сбора и разметки текстов из интернета и основанный на современных достижениях компьютерной лингвистики. Корпус включает в себя текстовые материалы из блогосферы, социальных сетей, с крупнейших новостных ресурсов и из литературных журналов [15].

3. Корпус биографических текстов — Russian corpus of biographical texts — корпус содержит биографии личностей, чья основная деятельность связана с наукой, техникой и образованием [16].

4. Корпус русских учебных текстов — коллекция текстов на русском языке,

написанных студентами разных вузов. Общий объем корпуса составляет около 3,1 млн слов. Тексты сопровождаются несколькими типами разметки: метатекстовой, морфологической и разметкой по ошибкам [17].

5. Clarin corpora of academic texts — корпус академических текстов содержит научные статьи, такие как исследовательские работы, эссе и рефераты, опубликованные в академических журналах, материалы конференций, научные монографии. Инфраструктура предоставляет доступ к 22 подкорпусам академических текстов, 2 из которых многоязычные и 20 одноязычные. Корпуса содержат научные тексты на 11 языках, в том числе на русском. Представлено более 15 различных научных дисциплин [18].

6. The Google books corpus — корпус составлен на массиве печатных источников, опубликованных с 16 века и собранных в сервис Google books [19].

Более полную информацию об имеющихся корпусах текстов можно получить на странице «Список текстовых корпусов» в энциклопедии Wikipedia.

В исследовании использовались данные корпуса Google books, корпус содержит около 500 млрд слов и представляет собой изменения с течением времени начиная с 16 века и до наших дней. По данным Google всего в мире было написано 129 млн книг, считается, что, исследуя данные оцифрованных книг, изучая статистику текстов, можно добраться до практической сути многомилионного книжного пространства, увидеть, как со временем меняется значимость тех или иных слов, угасает интерес общества к прошлому, происходят изменения в культуре, проявляется влияние пропаганды, социальную значимость тех или иных профессий, неожиданные и довольно абстрактные закономерности [7].

При работе с текстовыми корпусами как с цифровой моделью применяется кодировка текста в виде векторного представления слов. Методы векторного кодирования Word2vec, Pullenti, Skip-gram моделируют текст с учетом его контекста, сочетаемости и расположением по 2-5 слов по тексту, преобразуя слова в цифровой вид векторного типа [3]. Количественные методы исследования корпуса текста используются при проведении подсчетов и измерениях текстовых единиц любого уровня и основаны на использовании методов математической статистики [20,21].

Цель, задачи, методы исследования

Цель настоящего исследования — рассмотреть различные виды текстовых корпусов, определить информационный потенциал корпуса научного текста как базы больших данных и выявить область использования в терминологических исследованиях.

Задачи исследования:

1. Подготовить обзор публикаций в области корпусной лингвистики.

2. Сформировать корпус научного текста по отраслевой тематике и разметить его по тематическим разделам; определить словосочетания в размере 2–3 слов с участием значимых ключевых слов; выполнить автоматизированные подсчеты частот и построить распределение значимых словосочетаний во времени; сделать сопоставление технологических изменений с временным распределением частот словосочетаний в научном тексте.

3. Рассчитать по закону Ципфа показатели частотности корпуса текста по транспорту нефти и оценить естественность его текстового содержания.

4. Выполнить диахроническое исследование терминосодержащих N-грамм и сопоставить результаты с технологическим и историческими периодами развития отрасли.

Методы исследования: в работе были использованы теоретико-аналитический метод, включающий обзор и анализ научной литературы; метод сплошной выборки, сопоставительного и структурного анализа.

Анализ данных и результат

Данные для исследования

1. Корпус текстов научных статей по тематике трубопроводного транспорта нефти: содержит 916 275 словоупотреблений; разделен на 17 тематических разделов; выделены авторские и расчетные ключевые слова [22]. Частотный список распределения ключевых слов в корпусе и его разделах представлен в табл. 1 и табл. 2. Одиночные словоформы полностью не отображают смысловое содержание всего текста, двухсловные и трехсловные словосочетания с участием терминов и слов, сочетающихся с терминами и находящиеся в тексте рядом, являются более точными индикаторами содержания текста. Подготовленные для исследования N-граммы представлены в табл. 2.

2. Русский и английский подкорпуса корпуса Google books, содержащие данные о частотности слов, смежных последовательностей — N-грамм, n слов с n = 2 – 5. Корпус содержит тексты, распределенные во времени, с 16 века и до наших дней. Для исследования использовались данные за период 1940 — 2019 гг.

Таблица 2 / Table 2

Список распределения ключевых слов по частотности в основном корпусе / A list of keyword frequency distribution in the main corpus

Авторские	Расчетные
нефть, трубопровод, система, резервуар магистральный	нефть, трубопровод, система, грунт, покрытие

Таблица 3 / Table 3

Частотный список распределения ключевых слов по тематическим разделам корпуса / Frequency list of keyword distributions across thematic sections of the corpus

Раздел	Распределение ключевых слов по частотности
R(1)	состояние, трубопровод, изгиб, напряжение, разрушение
R(2)	модель, нефть, трубопровод, грунт, резервуар
R(3)	вертикаль, контроль, сварной, неразрушающий, трубопровод
R(4)	ремонт, надежность, трубопровод, капитальный, дефект
R(5)	защита, покрытие, коррозия, электрохимическая, катодная

Раздел	Распределение ключевых слов по частотности
R(6)	обеспечение, автоматизированная, программное, информационный, контур
R(7)	режим, насос, нефть, оптимизация, гидравлический
R(8)	диагностика, система, магистральный, сертификация, стандартизация
R(9)	оценка, качество, соответствие, арматура, сертификация
R(10)	нефть, качество, перекачка, измерение, свойство
R(11)	безопасность, нефть, авария, ликвидация, опасный
R(12)	система, управление, сеть, анализ, коэффициент
R(13)	разлив, ликвидация, нефтепродукт, испарение, деградация
R(14)	автоматизированные, управление, информационный, угрозы, кибератаки
R(15)	системы, моделирование, расчет, проектирование, цифровой
R(16)	резервуар, железобетонный, научно-технический, отрасль, Главтранснефть
R(17)	монополия, характеристики, регулирование, экономика, отрасль

Считается, что, диахронические исследования терминов являются одними из наиболее перспективных в области терминоведения [23]. Данные извлеченные из корпуса Google books за период 1940–2019 гг. использовались в данном исследовании, отображающем временные изменения в текстовом поле корпуса.

Частотность распределения ключевых слов

Частотное распределение слов в любом тексте подчинено эмпирической закономерности Ципфа (2): «Если все слова определенного текста расположить по убыванию частоты их использования, то частота r-го слова в таком списке окажется обратно пропорциональной его порядковому номеру r — рангу этого слова» [24].

$$f(w) = \frac{c}{r(w)}, \qquad (2)$$

где *f (w)* — количество слов в тексте;

r(w) — ранг слова;

c — постоянная величина.

Расчетные значения частот слов, полученные для исследуемого корпуса текстов научных статей по тематике трубопроводного транспорта нефти и соответствие модели закона Ципфа, графически показаны на рис. 1.

На графике частот (рис. 1) видно, что исследуемый корпус имеет избыточность ключевых слов в пределах до 30 ранга, но в целом соответствие расчетного значения частотности текста к значению, полученному по закономерности (2) составляет 59% — хорошим уровнем естественности текста считается соответствие в 50% и выше.

Временные изменения частотности ключевых слов

Ngram viewer [25, 26] строит графики временных изменений частотности N-грамм в тексте, где n — есть количество слов, расположенных в одной последовательности. В данном исследовании, для построения графиков:

- использовались русский и английский корпуса текстов;
- выбран временной период 1940-2019 гг.;

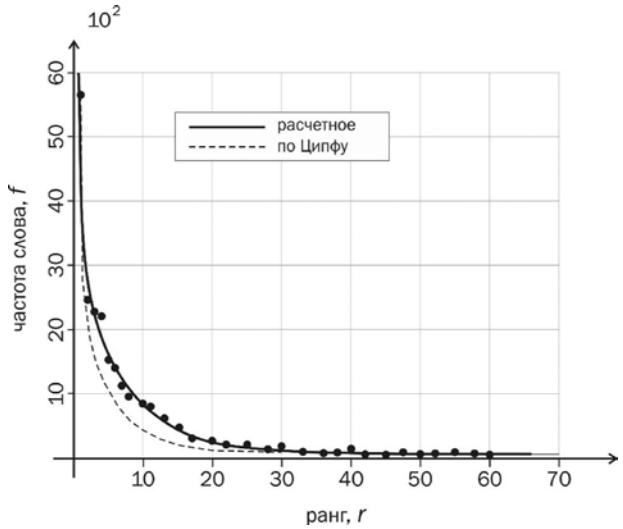

Рис. 1 Распределение частотности слов в корпусе текстов научных статей по тематике трубопроводного транспорта нефти
Fig. 1. Distribution of word frequency in the corpus of texts of scientific articles on the subject of oil pipeline transport

– установлено значение параметра Case-Insensitive (без учета регистра);
– применялись уровни сглаживания — 3, чтобы сделать графики более четкими с коэффициентом сглаживания, равный трем — количество слов для любого данного года является средним значением слов этого года и трех лет до и после него;
– для одной и той же фразы проверялось несколько терминов, вариантов написания, сначала делался первоначальный поиск возможных словоупотреблений (путем задания в начале или в конце слова (фразы) символа звездочка) и окончательно выбран наиболее часто используемый вариант.

В результате получены следующие данные для N-грамм по нефтегазовой тематике:

– графики частотности, показаны на рис. 2;
– данные об экстремальных значениях частот в корпусе, сведены в табл. 4;
– временные последовательности характеризующие периоды экстремальных частот, показаны на рис. 3.

Периоды с максимальными значениями N-грамм приходятся на 60-е и начало 80-х годов прошлого века (рис. 3), это время наиболее интенсивного развития нефте-газотранспортной отрасли: в Советском Союзе и за рубежом отмечается рост объемов добычи углеводородов и ввода в эксплуатацию сверхдальних магистральных трубопроводов. 90-ые годы отмечаются спадом производства, а начало 2000 подъёмом — вводятся в строй новые нефте и газотранспортные магистральные трубопроводы.

Популярные ключевые N-граммы

Расширение сферы применения цифровых технологий и искусственного интеллекта приводит к широкому обсуждению проблематики исследований и внедрения новых технологий в научных публикациях. Цифровая трансформация и искусственный интеллект оказывают-

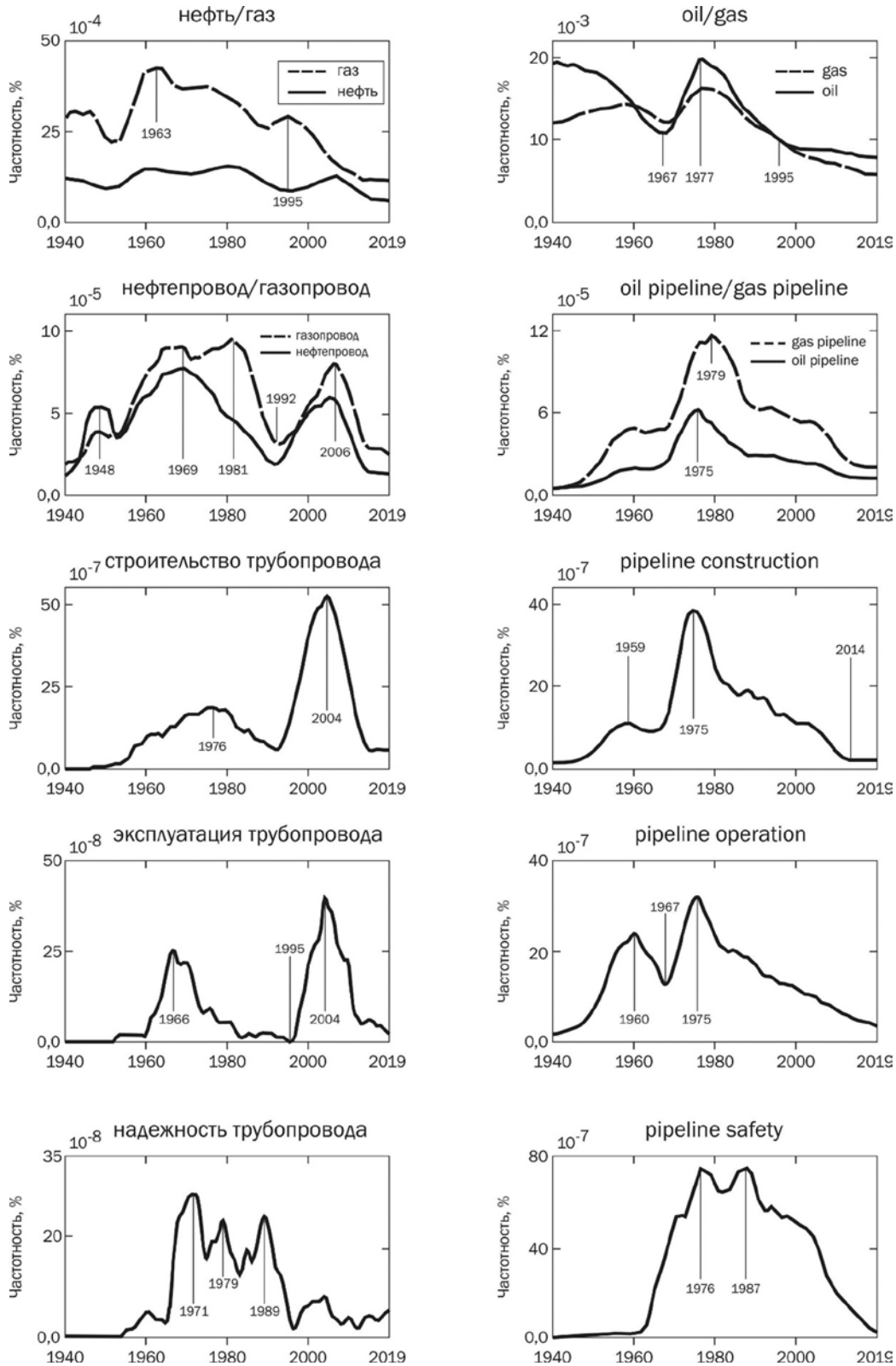

Рис. 2. Выборка частот терминов за временной период 1940–2019гг. по русскоязычным и англоязычным подкорпусам Google books corpus
Fig. 2. Sampling of term frequencies for the time period 1940-2019 by Russian- and English-language Google books corpus subcorps

Таблица 4 / Table 4

Нефтегазовые N-граммы и даты экстремального значения частот / Oil and gas N-grams and dates of extreme frequency values

Нефть Газ	1995 1963	Oil Gas	1967 1977 1995 1957 1967 1977
Нефтепровод Газопровод	1948 1969 1992 2005 1948 1969 1981 1992 2006	Oil pipeline Gas pipeline	1975 1979
Строительство трубопровода	1968 1979 1993 2004 2014	Pipeline construction	1959 1975 2014
Эксплуатация трубопровода	1966 1995 2004	Pipeline operation	1960 1967 1975
Надежность трубопровода	1971 1979 1989	Pipeline safety	1976 1987

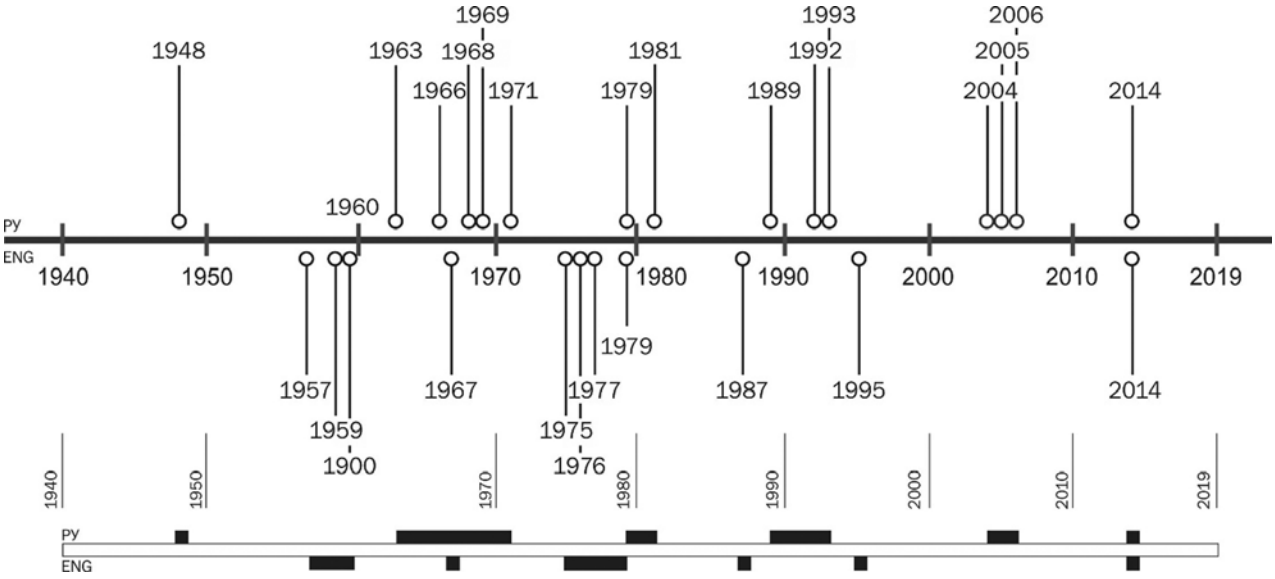

Рис. 3. Схема распределение дат, когда частотность употребления нефтегазовых N-грамм имеет максимальные и минимальные значения

Fig. 3. Scheme of distribution of dates when the frequency of use of oil and gas N-grams has maximum and minimum values

ся одними из самых частотных N-грамм в текстах по проблематике и обсуждению возможностей, которые создают новые технологии. Словосочетания «цифровые технологии и искусственный интеллект» также активно включаются в семантическое поле характеризующее научную и производственную деятельность в отрасли трубопроводного транспорта нефти.

На рис. 4 показано распределение частотности N-грамм «искусственный

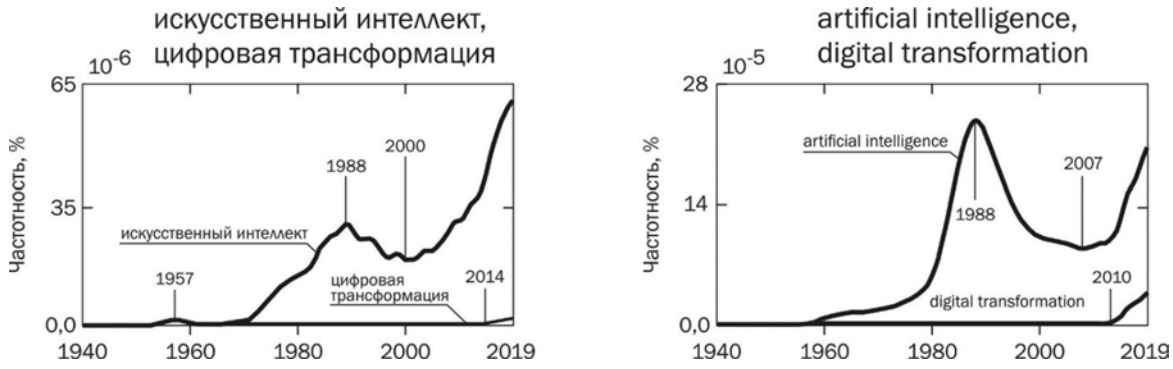

Рис. 4. Пример частотности N-грамм «искусственный интеллект, цифровая трансформация, artificial intelligence, digital transformation»
Fig. 4. An example of N-grams frequency is «artificial intelligence, digital transformation, artificial intelligence, digital transformation»

интеллект, цифровая трансформация, artificial intelligence, digital transformation» построенных на данных русскоязычного и англоязычного подкорпусов текстов основного корпуса Google books corpus. На графиках видно, что частотность практически совпадает в обоих подкорпусах. Интенсивный рост частотности прослеживается в периоды с 1960-х годов и достигает своего максимума в 1988 году, в последующем на протяжении более 10 лет наблюдается спад и в начале 2000-х начинается интенсивный рост.

Успешный способ сварки

Известно, что внедрение новых изобретений и технологий является цикличным процессом, имеющим вид S-образной кривой, так называемой «Кривой Гартнера». До достижения своей зрелости процесс проходит фазы интенсивного роста, спада и последующего подъёма до достижения плато продуктивности. По графикам N-грамм просматриваются частоты употребления названия изобретения с течением времени и косвенно графики могут быть использованы для выявления периодов создания, внедрения и дальнейшего распространения изобретения.

Созданный в Советском Союзе способ автоматической сварки под слоем флюса впервые был продемонстрирован в июле 1940 года и в последующем стал одним из основных способов для выполнения сварочных работ в промышленности и строительстве. За рубежом первый патент на способ submerged arc welding (SAW) был получен в 1935 году. В основном свое распространение сварка под слоем флюса получила уже в 1940-ые и последующие годы.

По графику частот N-грамм (рис. 5) построенного по тескам русскоязычного подкорпуса видно, что наибольшее упоминание в литературе способ получил в 1950–1960 годы, в последующем на протяжении 30 лет шло постепенное снижение влияния и с 1994 года практические нигде данный способ сварки не упоминался и уже в начале 2000-х годов видно, что способ начинает больше упоминаться в текстах. Аналогичная ситуация просматривается и в англоязычной части корпуса, здесь также отмечается два периода значительного интереса в технической литературе к данному способу, это 1959 и 1979 годы, и также видим последующий спад до 2016 года с дальнейшим постепенным подъемом. Это свидетельствует о том, что способ сварки продолжает ис-

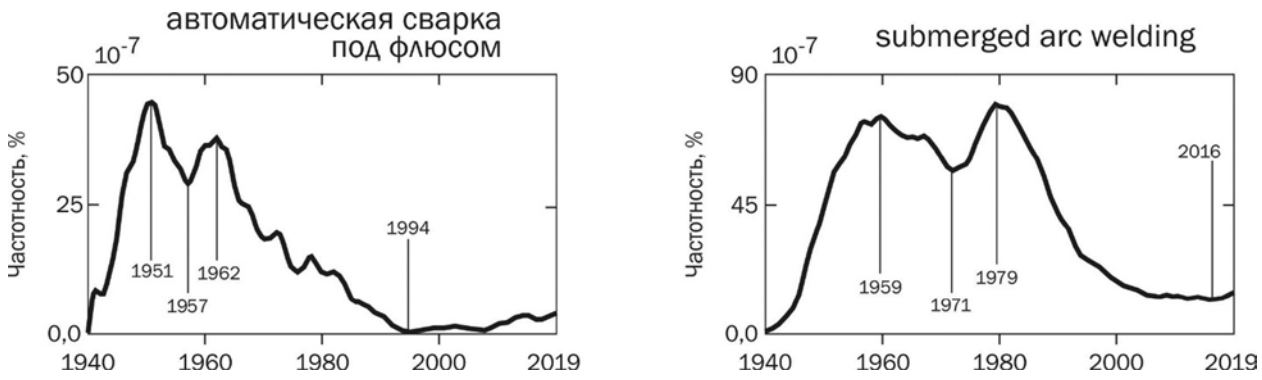

Рис. 5. Распределение частотности N-грамм «автоматическая сварка под флюсом, submerged arc welding»

Fig. 5. Frequency distribution of N-grams — «automatic submerged arc welding, submerged arc welding»

пользоваться в промышленности и обеспечивает решение задач повышения производительности и качества сварки.

Практическая значимость результатов

Результаты исследования могут быть использованы в методиках научно-технического прогнозирования, в области распознавания речи и машинного перевода, где текстовые корпуса применяются для создания моделей маркирования частей речи. Корпусы и частотные словари применяются в обучении и настройки языковых моделей с искусственным интеллектом создающих тексты и поддерживающих запросы на естественных языках [27]. N-граммы, выделенные из корпуса, используются при определении следующего слова, если известны предыдущие, для поиска и коррекции ошибок в тексте.

Выводы

По результатам исследования сделаны следующие выводы:

1. Наличие текстов в электронной форме существенно упрощает создание больших и представительных корпусов текста, по различным тематикам и на разных языках, объемами в десятки и сотни миллионов слов, но остается проблема с авторскими правами и приведением текстов к единому формату. Корпуса текстов обеспечивают решение задач как лексико-грамматического, так и информационного анализа в аспекте эволюционно-исторического развития явлений, происходящих в науке и технике. Хотя экспериментирование с N-граммами фиксирует только те изменения, которые уже произошли в прошлом, и не дает предсказаний о развитии в будущем, но как показывают результаты исследований по корпусу текста можно выявить тенденции и закономерности, которые уже проявились с течением времен и в будущем возможно сохранят свой инерционный след. Такой инерционный сценарий сбывается в 80-85% случаях.

2. Анализ показал, что частотность ключевых слов в исследуемом корпусе имеет избыточность до 30 ранга, при этом в целом по всему корпусу частотность, рассчитанная по формуле (2) составляет 59%, что в целом считается хорошим уровнем естественности текста и не требует дополнительных усовершенствований в части корректировки употребления слов с высокой или низкой частотностью.

3. Экстремальные значения в графиках распределения частотностей терминосодержащих N-грамм коррелируются с историческими периодами развития отрасли. На примере внедрения нового способа сварки по графику распределения частот N-грамм выделены временные периоды и изменения — с какой скоростью происходит внедрение изобретения и как быстро происходит технологическое обновление на производстве.

4. Вычислительные и графические эксперименты с данными корпуса показывают перераспределение частот отраслевых терминов в различные периоды времени. В настоящее время в терминоведении признается приоритетность диахронических исследований, которые позволяет выявить закономерности не только лингвистического, но и исторического значения: диахронические исследования помогают выявить изменения, вариативность и тенденции в развитии терминоведения. В целом следует отметить, что корпуса текста имеют лингвистический и фактологический потенциал для осуществления научных исследований.

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов

Conflict of Interests

The author declares no conflict of interests

ИСТОЧНИКИ / REFERENCES

1. Микова Н.С., Соколова А.В. Мониторинг глобальных технологических трендов: теоретические основы и лучшие практики // ФОРСАЙТ. 2014. Т. 8. № 4.

Mikova N.S., Sokolova A.V. Monitoring global`ny`x texnologicheskix trendov: teoreticheskie osnovy` i luchshie praktiki // FORSAJT. 2014. T. 8. № 4.

2. Нгуен Тхань Вьет, Кравец А.Г. Новый метод прогнозирования технологических трендов на основе анализа научных статей и патентов. International Journal of Open Information Technologies ISSN: 2307-8162 vol. 10, no. 10, 2022.

Nguen Txan` V`et, Kravecz A.G. Novy`j metod prognozirovaniya texnologicheskix trendov na osnove analiza nauchny`x statej i patentov. International Journal of Open Information Technologies ISSN: 2307-8162 vol. 10, no. 10, 2022.

3. Башков А.С., Соломенцев Я.К. Использование векторных методов представления слов в задачах выявления трендов // Вестник Российского нового университета. Серия «Сложные системы модели, анализ и управление». 2019. Выпуск 2. С. 80-88.

Bashkov A.S., Solomencev Ya.K. Ispol`zovanie vektorny`x metodov predstavleniya slov v zadachax vy`yavleniya trendov // Vestnik Rossijskogo novogo universiteta. Seriya «Slozhny`e sistemy` modeli, analiz i upravlenie». 2019. Vy`pusk 2, P. 80-88.

4. Сощенко А. Е., Комарица В.Н. Анализ зависимости между числом публикаций и количеством цитирования статей в научной периодике трубопроводного транспорта углеводородов // Наука и технологии трубопроводного транспорта нефти и нефтепродуктов. — 2015. — № 3(19). — С. 108-115.

Soshhenko A. E., Komaricza V.N. Analiz zavisimosti mezhdu chislom publikacij i kolichestvom citirovaniya statej v nauchnoj periodike truboprovodnogo transporta uglevodorodov // Nauka i texnologii truboprovodnogo transporta nefti i nefteproduktov. — 2015. — № 3(19). — P. 108-115.

5. Эрец Эйден. Неизведанная территория: как «большие данные» помогают раскрывать тайны прошлого и предсказывать будущее нашей культуры: / Эрец

Эйден, Жан-Батист Мишель. — Москва. Изд-во АСТ. 2016. — 350 с.

E`recz E`jden. Neizvedannaya territoriya: kak «bol`shie danny`e» pomogayut raskry`vat` tajny` proshlogo i predskazy`vat` budushhee nashej kul`tury`: / E`recz E`jden, Zhan-Batist Mishel`. — Moskva. Izd-vo AST. 2016. — 350 p.

6. Stop Hyping Big Data and Start Paying Attention to Long Data. URL: http://goo.gl/X7oEC (data dostupa 01.08.2023).

7. Google Books. URL: https://ru.wikipedia.org/wiki/Google_Книги, (data dostupa 08.08.2023).

Google Books. URL: https://ru.wikipedia.org/wiki/Google_Knigi, (data dostupa 08.08.2023).

8. Jean-Baptiste Michel, Erez Lieberman Aiden: What we learned from 5 million books. URL: https://www.ted.com/, (data dostupa 08.08.2023).

9. Котов Ю.А., Коломец Н.В. Элементы системы TextLab для частотного анализа текста. Современные тенденции развития науки и технологий. Сборник научных трудов по материалам Международной научно-практической конференции. В 5-ти частях. Часть II. Под общей редакцией Ж.А. Шаповал. 2017.

Kotov Yu.A., Kolomecz N.V. E`lementy` sistemy` TextLab dlya chastotnogo analiza teksta. Sovremenny`e tendencii razvitiya nauki i texnologij. Sbornik nauchny`x trudov po materialam Mezhdunarodnoj nauchno-prakticheskoj konferencii. V 5-ti chastyax. Chast` II. Pod obshhej redakciej Zh.A. Shapoval. 2017.

10. McEnery Tony, Wilson Andrew. Corpus Linguistics: An Introduction. 2nd edition. — Edinburgh University Press, 2001. — 235 p.

11. Zhongquan Du, Feng Jiang, Luda Liu. Profiling figure legends in scientific research articles: A corpus-driven approach, Journal of English for Academic Purposes, Volume 54, 2021, 101054, ISSN 1475-1585, URL: https://doi.org/10.1016/j.jeap.2021.101054.

12. Мордовин А. Ю. Лингвистическая идеология корпусов текстов / Иркутский гос. лингвистический ун-т. — Иркутск: 2014. – 190 с.

Mordovin A. Yu. Lingvisticheskaya ideologiya korpusov tekstov / Irkutskij gos. lingvisticheskij un-t. — Irkutsk: 2014. — 190 p.

13. Бутенко Ю.И. Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов. Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2022. Т. 20. № 3. С. 5-13.

Butenko Yu.I. Model` teksta nauchno-texnicheskoj stat`i dlya razmetki v korpuse nauchno-texnicheskix tekstov. Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informacionny`e texnologii. 2022. T. 20. № 3. P. 5-13.

14. Плунгян В.А., Резникова Т.И., Сичинава Д.В. Национальный корпус русского языка: общая характеристика. Научно-техническая информация. Серия 2: Информационные процессы и системы. 2005. № 3. С. 9-13.

Plungyan V.A., Reznikova T.I., Sichinava D.V. Nacional`ny`j korpus russkogo yazy`ka: obshhaya xarakteristika. Nauchno-texnicheskaya informaciya. Seriya 2: Informacionny`e processy` i sistemy`. 2005. № 3. P. 9-13.

15. Генеральный интернет-корпус русского языка. URL: http://www.webcorpora.ru/, (дата доступа 01.08.2023 г.).

General`ny`j internet-korpus russkogo yazy`ka. URL: http://www.webcorpora.ru/, (data dostupa 01.08.2023).

16. Корпус биографических текстов — Russian Corpus of Biographical Texts. URL:

https://sites.google.com/site/utcorpus (дата доступа 29.08.2023 г.).

Korpus biograficheskix tekstov — Russian Corpus of Biographical Texts. URL: https://sites.google.com/site/utcorpus (data dostupa 29.08.2023).

17. Корпус русских учебных текстов. URL: http://web-corpora.net/learner_corpus (дата доступа 07.09.2023 г.).

Korpus russkix uchebny`x tekstov. URL: http://web-corpora.net/learner_corpus (data dostupa 07.09.2023).

18. Corpora of Academic Texts. URL: https://www.clarin.eu/resource-families/corpora-academic-texts (data dostupa 07.09.2023).

19. Davies M. 2011. Google Books Corpus (155 billion words, 1810-2009). URL: http://googlebooks.byu.edu/, (data dostupa 01.08.2023).

20. Глазкова А.В. Автоматический поиск фрагментов, содержащих биографическую информацию, в тексте на естественном языке // Труды Института системного программирования РАН. 2018. Том 30. № 6. С. 221-236. DOI: 10.15514/ISPRAS-2018-30(6)-12.

Glazkova A.V. Avtomaticheskij poisk fragmentov, soderzhashhix biograficheskuyu informaciyu, v tekste na estestvennom yazy`ke // Trudy` Instituta sistemnogo programmirovaniya RAN. 2018. Tom 30. № 6. P. 221-236. DOI: 10.15514/ISPRAS-2018-30(6)-12.

21. Андреев Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении / АН СССР. Ин-т языкознания. — Ленинград: Наука. Ленингр. отд-ние, 1967. — 403 с.

Andreev N.D. Statistiko-kombinatorny`e metody` v teoreticheskom i prikladnom yazy`kovedenii / AN SSSR. In-t yazy`koznaniya. – Leningrad: Nauka. Leningr. otd-nie, 1967. — 403 p.

22. Комарица В.Н. Анализ ключевых слов в научных статьях. Научно-техническая информация. Серия 1. Организация и методика информационной работы. 2023. № 9. С. 9 — 15.

Komaricza V.N. Analiz klyuchevy`x slov v nauchny`x stat`yax. Nauchno-texnicheskaya informaciya. Seriya 1. Organizaciya i metodika informacionnoj raboty`. 2023. № 9. Pp. 9 — 15.

23. Гринев-Гриневич С.В., Сорокина Э.А. Перспективные направления развития терминологических исследований // Вестник Московского государственного областного университета. Серия: Лингвистика. 2018. № 5. С. 18–28.

Grinev-Grinevich S.V., Sorokina E`.A. Perspektivny`e napravleniya razvitiya terminologicheskix issledovanij // Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Lingvistika. 2018. № 5. Pp. 18–28.

24. Маслов В.П. О законе Ципфа и ранговых распределениях в лингвистике и семиотике / В.П. Маслов., Т.В. Маслова // Математические заметки. — 2006. — Т. 80. – N. 5 — С. 718-732.

Maslov V.P. O zakone Cipfa i rangovy`x raspredeleniyax v lingvistike i semiotike / V.P. Maslov., T.V. Maslova // Matematicheskie zametki. — 2006. — T. 80. — N. 5 — Pp. 718-732.

25. Google Books Ngram Viewer. URL: https://books.google.com/ngrams/, (data dostupa 21.08.2023).

26. 15 years of Google Books. Blog Google. URL: https://blog.google/products/search/15-years-google-books/, (data dostupa 15.08.2023).

27. ChatGPT. URL: http:// ru.wikipedia.org/, (data dostupa 04.08.2023).

Информация об авторе

Валентин Николаевич Комарица — кандидат технических наук, заместитель начальника отдела издательских проектов и медиакоммуникаций, ООО «Научно-исследовательский институт трубопроводного транспорта» (ООО «НИИ Транснефть), Москва, KomaritsaVN@niitnn.transneft.ru

Information about the author

Valentin Nikolaevich Komaritsa — Cand. Sci. (Eng.), deputy head of the department of Publishing Projects and Media Communications, LLC Research Institute of Pipeline Transport (LLC NII Transneft), Moscow, KomaritsaVN@niitnn.transneft.ru

■