Terra Linguistica. 2023, Том 14, № 3. С. 80–93.
Terra Linguistica. 2023, Vol. 14, No. 3. Pp. 80–93.

Научная статья
УДК 004.62
DOI: https://doi.org/10.18721/JHSS.14307

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ЛЕКСИКОГРАФИЧЕСКИХ ИССЛЕДОВАНИЯХ: ПРЕДСТАВЛЕНИЕ ЧАСТОТНОЙ ЛЕКСИКИ

М.В. Хохлова ✉

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ m.khokhlova@spbu.ru

Аннотация. Статистические методы используются в лингвистике на протяжении долгого времени. Однако в последнее время в связи с развитием информационных технологий статистический аппарат получил свое второе развитие и стал более активно применяться для решения прикладных задач, в том числе при обработке и представлении текстовых данных. Целью работы является описание ряда статистических метрик, которые используются в лексикографических исследованиях, на примере частотного словаря русского языка, корпусов текстов и баз данных, в которых содержится информация сочетаемости лексических единиц. Данные показатели используются для дифференциации лексики по разным основаниям, представления высоко- и низкочастотных единиц, выделения слов и устойчивых словосочетаний, характерных для текстов определенного стиля или темы. Также в статье содержится краткий исторический обзор применения количественных методов к анализу текстов и обсуждаются вопросы, связанные со статистической лексикографией.

Ключевые слова: статистические методы, корпусы текстов, частотные словари, устойчивые словосочетания, базы данных, русский язык.

Финансирование: Исследование выполнено за счет гранта Российского научного фонда («Структура и функционирование устойчивых неоднословных единиц русской повседневной речи», проект № 22-18-00189).

Для цитирования: Хохлова М.В. Статистические методы в лексикографических исследованиях: представление частотной лексики // Terra Linguistica. 2023. Т. 14. № 3. С. 80–93. DOI: 10.18721/JHSS.14307

Research article
DOI: https://doi.org/10.18721/JHSS.14307

STATISTICAL METHODS IN LEXICOGRAPHIC RESEARCH: REPRESENTING FREQUENCY VOCABULARY

M.V. Khokhlova ✉

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ m.khokhlova@spbu.ru

Abstract. Statistical methods have been used in linguistics for a long time. However, recently, information technologies have boosted the development of statistical tools, which are now more actively used for applied tasks, including processing and presentation of text data. The purpose of the work is to describe a number of statistical metrics used in lexicographic studies, involving a frequency dictionary of the Russian language, text corpora and databases that present information about lexical collocability. These measures are implemented to differentiate vocabulary on different grounds, highlighting key words and phrases characteristic of texts of a certain style or topic. The paper also provides a brief historical overview of the application of quantitative methods to text analysis.

Keywords: statistical methods, text corpora, frequency dictionaries, collocations, databases, Russian language.

Acknowledgements: The study was supported by a grant from the Russian Science Foundation ("Structure and functioning of stable non-single-word units of Russian everyday speech", project No. 22-18-00189).

Citation: Khokhlova M.V., Statistical methods in lexicographic research: representing frequency vocabulary, Terra Linguistica, 14 (3) (2023) 80–93. DOI: 10.18721/JHSS.14307

Введение

В связи с автоматизацией, проникшей во многие области, которые связаны с анализом языкового материала, статистический анализ получил свое второе развитие. Обработка естественного языка (*англ*. natural language processing) стала важным этапом при проектировании разнообразных систем, для создания которых необходимо обработать большие текстовые массивы, а также обучить на них языковые модели. Вычисление различных квантитативных характеристик существенно упростилось благодаря появлению компьютерных инструментов для статистической обработки данных. В статье мы дадим краткий обзор истории статистических методов в филологии и лингвистике и опишем тот статистический аппарат, который используется в словарях и корпусах текстов и связан с представлением информации о лексике, в том числе с ее дифференциации на основе квантитативных показателей.

Статистические методы в исследовании текстов: краткий обзор

К первой попытке применить количественные методы в филологии можно отнести работу «Опыт о русском стихосложении» видного русского филолога А.Х. Востокова, в которой он попытался осуществить статистический анализ стиха (в частности, автором обсуждался вопрос частотности метрических форм русского стиха) [1]. Позднее математиком В.Я. Буняковским был опубликован первый том «Лексикона чистой и прикладной математики», в котором была описана математическая терминология [2]. В словарных статьях для терминов на французском языке были даны переводы на русский язык, которые сопровождались подробными толкованиями. Также исследователем затрагивались вопросы теории вероятностей и применения статистических методов на новом материале. Известный русский поэт, писатель и ученый А. Белый

провозгласил эстетику как точную науку, а сама методика исследования была заимствована им у естественных наук [3]. Им были введены количественные методы при исследовании стиха и получены статистические результаты на большом объеме проанализированных вручную стихов. Это стремление к точности было продолжено русским и советским филологом Б.И. Ярхо в его фундаментальной монографии «Методология точного литературоведения» [4]. В работе развиваются идеи анализа текстов с привлечением формальных и квантитативных методов.

Широко известные попытки применить статистические методы на языковом материале были предприняты в начале XX века в работах российских исследователей Н.А. Морозова и А.А. Маркова. Н.А. Морозов рассматривал вопросы, связанные с изучением языка разных авторов, и стремился вывести общие стилеметрические законы [5]. Он выделял служебные частицы как маркеры особенностей стиля писателя, предлагая вычислять частоту той или иной единицы в первой тысяче слов изучаемого текста. Данный метод был проиллюстрирован на примерах произведений А.С. Пушкина, Н.В. Гоголя, И.С. Тургенева, Л.Н. Толстого, Н.М. Карамзина и М.Н. Загоскина. Таким образом частотные распределения, или спектры, использовались для различения подлинных текстов авторов и текстов, которые могли быть им ошибочно приписаны. Результаты подверглись критике в работе видного русского математика А.А. Маркова, который указывал, что найденные закономерности могут относиться только к тем отрывками, на материале которых проводилось исследование, и могут быть характерны не для всех текстов, а только для указанных русских писателей. Исследователь использовал материал поэмы «Евгений Онегин» для демонстрации случайных процессов, получивших название «марковских» [6]. В упомянутых работах были заложены принципы целого направления, которое впоследствии стало активно развиваться и связано с решением задачи установлении авторства и определения подлинности текста, — стилеметрии. Позднее академик В.В. Виноградов указывал на необходимость исследовать частоты употребления разных типов слов в текстах разных стилей и тем самым выявлять их различия [7, с. 155–156]. Отметим важность работы группы под руководством выдающегося русского и советского математика А.Н. Колмогорова, которая достигла значительных успехов в изучении русского стиха: были выявлены метрические законы, дана классификация и получены статистические данные о ритмических вариациях метра [8].

Направление, посвященное сопряжению статистических методов и традиционной лингвистики, получило новый импульс в связи с появлением технических возможностей, позволивших привлечь вычислительные средства к решению ряда задач, в том числе получить новые данные и проверить гипотезы на обработанном текстовом материале. Во второй половине XX века в Ленинграде рабочая группа «Статистика речи» под руководством Р. Г. Пиотровского занималась изучением языковых структур при помощи статистических методов. Основной целью проекта была разработка систем машинного перевода. Результаты исследований публиковались в одноименных сборниках и были связаны, например, с квантитативным исследованием текстов на лексическом и морфологическом уровнях [9], анализом частотных словарей, принципов их составления [10]. Интерес к квантитативному анализу текстов привел к появлению работ, которые демонстрировали многообразие статистических методов и знакомили с ними гуманитариев [11–13]. Статистические методы нашли активное применение в задачах, связанных с атрибуцией текстов [14], с определением стилистических характеристик текстов [15, 16], при составлении словарей языка авторов [17] и частотных словарей [18]. В качестве иных направлений прикладной лингвистики, которые используют квантитативные методы, можно отметить квантитативную типологию текста, стилистическую диагностику, реконструкцию древних текстов и др.

Корпусная лингвистика стала еще одним направлением в прикладном языкознании, где активно стали использоваться компьютерные технологии. Результаты, выдаваемые системами в ответ на разнообразные запросы, снабжены статистической информацией. Она может включать данные о частотах слов или грамматических категорий, об оценке устойчивости выделен-

ных словосочетаний, о продуктивности словообразовательных моделей или синтаксических конструкций и многое другое. Стремительный рост информационных технологий в начале 2000-х дал новый толчок к развитию прикладной и, как следствие, компьютерной лингвистики и пополнил ее арсенал методами обработки больших массивов данных (*англ.* big data): в частности, машинное обучение стало широко применяться при решении лингвистических задач.

Ранг и частота

Для описания квантитативных характеристик речевых единиц могут использоваться абсолютные и относительные частоты. В словарях и корпусах наряду с абсолютной частотой приводятся данные в единицах ipm (*англ.* instances per million), т.е. в пересчете на 1 миллион словоупотреблений, что позволяет сравнивать между собой частоты слов в корпусах разного объема:

$$\text{ipm} = \frac{f * 10^6}{N}, \tag{1}$$

где f – абсолютная частота слова в корпусе; N – объем корпуса.

Рангом называют порядковый номер слова в списке, упорядоченном по частоте. Ранг 1 имеет самая частотная единица, ранг 2 — вторая по частотности, ранг 3 — третья и т.д. Самый низкий ранг будет приписан лексеме, находящей на последней позиции. Если слова имеют одинаковую частоту, то их ранги совпадают. В словарях это может быть отмечено диапазоном, например, нескольким словам будет присвоен ранг 100–105.

Распределение частот слов в языке неравномерно: некоторые слова используются очень часто, в то время как другие употребляются редко. Первые попытки проанализировать частотную структуру языковых данных были связаны с созданием стенографических систем. В начале XX века французским стенографистом Жаном-Батистом Эсту была выведена экспериментальным путем, закономерность, лежащая в основе распределения слов в тексте:

$$f * r = const, \tag{2}$$

где f – абсолютная частота слова, r – ранг слова.

Спустя некоторое время американским лингвистом Дж. Ципфом был предложен эмпирический закон, согласно которому частота слова в частотном словаре обратно пропорциональна его рангу (цитируется по [19]):

$$k_r = \frac{k_{\max}}{r^{\gamma}}, \tag{3}$$

где r – ранг слова, k_r – частота слова ранга r, $k_{\max}$ – частота самого частого слова, γ – коэффициент, характеризующий неравномерность распределения частот. Значения $k_{\max}$ и γ сильно варьируются в зависимости от языков и жанров.

Во второй половине XX века с развитием теории информации возрос интерес к закону Ципфа: стали появляться работы, связанные с его последующим уточнением и внесением поправок, а также вычислением констант для разных языков. Наибольший вклад был внесен математиком Б. Мандельбротом, в связи с чем наиболее часто в настоящее время закон упоминается как *закон Ципфа–Мандельброта*. Та же закономерность прослеживается не только в текстах, но и на ином материале. Например, численность населения, уровень дохода и др.

Распределение числа уникальных слов в документе как функция от его длины описывается *законом Хипса*:

$$V = \alpha * n^{\beta}, \tag{4}$$

где V – число уникальных слов в документе объемом n; α, β – параметры, задаваемые эмпирически.

Существуют и иные формулы, которые могут быть использованы для дифференциации лексики. Средняя уменьшенная частота ARF (*англ.* average reduced frequency) была введена в Чешском национальном корпусе его разработчиками [20] и призвана решить проблему слов, которые часто встречаются только в определенных документах. Ее значение указывается на странице с результатами выдачи вместе с ipm. Ниже приводится формула для вычисления ARF [21]:

$$\mathrm{ARF} = \frac{1}{\nu} \sum_{i=1}^{f} \min\{d_i; \nu\}, \tag{5}$$

где $\nu = N/f$; d_i – расстояние между двумя позициями употребления слова в корпусе.

Корпус объемом N единиц делится на f сегментов, где f – частота рассматриваемого слова. Всего может быть N/f способов разбиения корпуса. Далее подсчитывается количество сегментов, в которых встретилось слово (за счет этого уменьшается влияние того, что слово может встречаться часто в одном и том же документе). Вычисляется сумма частот, которая усредняется по количеству разбиений. Мера ARF может принимать значения от 1 (для слов, у которых абсолютная частота равна 1) до f (для слов, которые равномерно распределены в корпусе).

Статистический аппарат в частотных словарях и корпусах текстов

Статистическая лексикография занимается созданием словарей, в которых перечислены частотные характеристики лексики (см. подробнее, например, [22]). Частоты отдельных лексических элементов отличаются в зависимости от разных факторов: например, жанра произведения или времени его создания. При рассмотрении частот можно учитывать лексемы или словоформы, что также влияет на результат.

В частотном словаре лексические единицы характеризуются количественными характеристиками. При создании подобных словарей необходимо ответить на следующие вопросы: какие слова необходимо включать (например, как рассматривать глагольные формы, сложные союзы и предлоги), а также как оценивать частотность. Очень важен вопрос репрезентативности отбираемого материала (какие тексты наиболее точно и полно могут представить язык в его многообразии). Также можно отдельно выделить значимую лексику, которая характерна для определенного функционального стиля.

Во второй половине XX в. печатались серии специализированных словарей по электронике, машиностроению, судостроению и т.д. Издавались словари общей лексики, а также словари писателей и отдельных произведений. Частотные словари востребованы при преподавании языков, создании новых словарей, разработке приложений компьютерной лингвистики и решении других прикладных задач.

Первый словарь, который можно было бы назвать частотным, был составлен для немецкого языка стенографом Ф. Кедингом в 1898 г. [23] и был основан на корпусе текстов объемом около 11 млн слов. Словарь под редакцией Э.А. Штейнфельд был ориентирован на преподавание русского языка как иностранного [24]. Далее в 1977 г. на кафедре математической лингвистики ЛГУ под ред. Л.Н. Засориной был издан ставший известным «Частотный словарь русского языка» [25], в котором были представлены 40 тыс. наиболее употребительных слов современного русского языка. Общий объем выборки составил 1 млн словоупотреблений, что явилось своего рода прорывом, поскольку существовавшие до того времени проекты основывались на коллекциях текстов в 400–500 тыс. словоупотреблений.

В 2009 г. вышел из печати «Частотный словарь современного русского языка», который был составлен на материале Национального корпуса русского языка [18]. Объем выборки, в которую вошли тексты за период 1950–2007 гг., был равен 92 млн словоупотреблений. В словаре представлены следующие разделы: алфавитный и частотный списки лемм, распределение лемм по функциональным стилям (художественная литература, публицистика, другая нехудожественная литература и устная речь). Отдельно выделяются частотные списки разных частей речи (имен существительных и прилагательных, глаголов, наречий и предикатов, местоимений и служебных частей речи). В табл. 1 представлен пример 10 наиболее частотных слов из словаря.

Таблица 1. Наиболее частотные слова согласно «Частотному словарю русского языка»
Table 1. The most frequent words according to the "Frequency dictionary of the Russian language"

	Лемма	Часть речи	Частота (ipm)	Худ.лит. 1950-60-е	1970-80-е	1990-2000-е	Публ. 1950-60-е	1970-80-е	1990-2000-е
1	и	conj	35801.8	37546.0	39187.7	36514.7	36703.6	38352.1	35539.5
2	в	pr	31374.2	23943.5	25005.7	26185.7	35295.3	33264.3	36372.5
3	не	part	18028.0	21154.3	22921.5	22349.1	16127.4	18796.9	16853.2
4	на	pr	15867.3	17169.9	16411.7	16663.0	15249.8	15961.9	16675.2
5	я	spro	12684.4	17022.7	17090.5	18257.0	13881.5	16386.7	10447.1
6	быть	v	12160.7	12467.9	13492.8	12778.1	14458.2	15091.8	12646.0
7	он	spro	11791.1	19708.0	18587.1	16882.0	11721.6	12799.3	8811.7
8	с	pr	11311.9	11186.1	10974.4	11655.3	11060.5	11662.5	11462.7
9	что	conj	8354.0	8155.2	8313.9	8938.8	7634.5	9702.4	8743.1
10	а	conj	8198.0	10298.0	10772.2	9897.5	5748.9	7643.3	7298.8

Дополнительно к традиционным таблицам «Частотный словарь современного русского языка» также содержит списки значимой лексики — наиболее часто употребляемых слов, характерных для разных функциональных стилей. Данные списки были составлены на основе сравнения частот лемм в заданном подкорпусе текстов и в остальном корпусе при помощи коэффициента логарифмического правдоподобия (*англ.* log-likelihood, LL-score). Формула основана на сравнении наблюдаемых и ожидаемых частот [26: viii]:

$$\text{LL-score} = 2\left(a\ln\left(\frac{a}{E1}\right)+b\ln\left(\frac{b}{E2}\right)\right), \quad (6)$$

где $E1 = c\frac{a+b}{c+d}$, $E2 = d\frac{a+b}{c+d}$, a – наблюдаемая частота в рассматриваемом подкорпусе; b – наблюдаемая частота в других текстах; c – объем рассматриваемого подкорпуса; d – объем других текстов.

Таким образом сопоставляются частоты в текстах определенных стилей или тематики с частотами в остальных текстах. В случае если значение коэффициента LL-score превышает 15,31, разница между частотами признается статистически значимой, то есть слово или словосочетания является характерным для данного подкорпуса текстов.

Например, для устной речи к наиболее значимой лексике (см. табл. 2) можно отнести *ну, да, вот, там, ты*, в то время как публицистике в большей степени соответствуют следующие лексемы: *президент, театр, год, спектакль, правительство*. Это может быть объяснено тем фактом, что статьи в газетах посвящены политическим и культурным событиям.

Таблица 2. Значимая лексика устной речи согласно «Частотному словарю русского языка»
Table 2. Significant vocabulary of oral speech according to the "Frequency dictionary of the Russian language"

	Лемма	Часть речи	Частота в корпусе (ipm)	Частота в подкорпусе (ipm)	LL-score
1	ну	part	1114.6	17208.0	39648
2	да	part	787.5	11847.0	26881
3	вот	part	1785.1	15698.6	24493
4	там	advpro	1128.1	10531.7	17241
5	ты	spro	3171.2	13503.8	9491
6	угу	intj	24.6	2068.8	9324
7	я	spro	12684.4	33686.4	9186
8	нет	part	589.2	4618.9	6532
9	а	conj	8198.0	20593.3	4783
10	вообще	adv	417.6	2989.8	3890

Наряду с показателем ipm в словаре используются специальные метрики, которые учитывают не только частоту слова, но и число документов, в которых данное слово встречается, что может помочь определить, является ли слово характерным для корпуса в целом или только для текстов определенных стиля или тематики. Речь может идти о вычислении некоторой усредненной частоты. В качестве такого показателя могут использоваться широко известные статистические меры, например, дисперсия или стандартное отклонение. Также примером может служить показатель R (*англ*. range), который используется в [18] и равен количеству тех сегментов, в которых встретилось слово. В Британском национальном корпусе [27], Новом частотном словаре русской лексики [18] и ряде других проектов приводятся значения коэффициента D Жуйана (*англ*. Juilland's D) [28], который отражает распределения частот в разных сегментах корпуса [26: vi]:

$$D = 100 \times \left(1 - \frac{\sigma}{\mu\sqrt{n-1}}\right), \quad (7)$$

где n – количество сегментов, на которые разбит корпус; μ – средняя частота слова в корпусе; σ — среднее квадратичное отклонение средней частоты μ в корпусе.

Для вычисления данного коэффициента в [18] тексты корпуса были предварительно упорядочены по функциональным стилям, а далее разбиты на 100 равных частей (т.е. n равно 100). Значение D близко к 100, если слово встречается почти во всех сегментах корпуса, и близко к 0, если слово характерно только для небольшого числа документов.

Оба коэффициента R и D позволяют дифференцировать слова, которые могут иметь одинаковые частоты в корпусе, по тому, насколько они являются общеупотребительными. Так, глагол *возвращать* и прилагательное *вирусный* имеют одинаковую среднюю частоту 21 ipm, однако в первом случае коэффициент R равен 100 и коэффициент D равен 96, в то время как для второй лексемы 61 и 65 соответственно. Следовательно, первое слово встречается во всех сегментах и является в них частотным, в то время как второе значимо для определенного круга предметных областей.

Метрика DP (*англ*. deviation of proportions) была предложена в работе [29], автор которой обратил внимание на недостатки других статистик. В частности, в ней могут быть учтены сегменты корпуса разной длины. Она более проста в вычислении и менее чувствительна к выбросам. Мера принимает значения в диапазоне от 0 до 1 и основана на сравнении ожидаемых и наблюдаемых частотностей слова, измеренных относительно его появления в сегментах корпуса [30]:

$$DP = 0,5\sum_{i=1}^{n}\left|\frac{v_i}{f} - s_i\right|, \quad (8)$$

где n – количество сегментов, на которые разбит корпус; v – частота появления слова в сегментах корпуса; s — размер сегмента корпуса.

В системе Sketch Engine [31, 32] был введен показатель *keyness score* для выявления ключевых слов и словосочетаний, которые характерны в большей степени для заданного тематического (фокусного) корпуса или подкорпуса, чем для нейтрального (референтного) корпуса или всего корпуса в целом [33]:

$$keyness\ \ score = \frac{fpm_{focus} + n}{fpm_{ref} + n}, \quad (9)$$

где fpm_{focus} – относительная частота в фокусном корпусе в единицах ipm; fpm_{ref} – относительная частота в референтном корпусе в единицах ipm; n – сглаживающий коэффициент.

Обычно в качестве референтного корпуса рассматривается корпус большого объема, который в идеале должен быть сбалансированным и не содержать искажений частот. По умолчанию n равно 1, однако у пользователя есть возможность выбрать иное значение.

Базы данных сочетаемости для русского языка со статистическим аппаратом

Как было отмечено выше, в некоторых корпусах текстов используются статистические метрики. Так, в Национальном корпусе русского языка есть возможность поиска коллокаций по ключу (то есть по ключевому слову, для которого будут показаны сочетания) и коллокату [34]. Наряду с указанием совместной частоты, частот ключа и коллоката для количественной оценки применяются такие широко известные показатели как t-score, log-likelihood, logDice и MI^3 [35, 36]. Дополнительно введена агрегированная мера, которая вычисляет геометрическое среднее t-score и MI^3.

Ряд специализированных проектов ориентирован на исследование сочетаемости ключевых слов на материале русского языка с привлечением статистического аппарата. Таким ресурсом является система CoCoCo (“Collocations, Colligations, Constructions”)[1] [37], которая разрабатывается под руководством М. Копотева в Хельсинкском университете, представляет информацию о многословных выражениях (*англ.* multi-word expressions, MWE) на основе Национального корпуса русского языка [38], а также корпусов Taiga [39] и ruWac [40]. Многословные выражения понимаются предельно широко: к ним относятся идиомы, составные единицы, коллокации и коллигации. В системе предусмотрен поиск по леммам или токенам (словоформам), который можно ограничить морфологическими параметрами. При выборе части речи есть возможность задать значения грамматических категорий (например, род, число или падеж для существительных или вид, время и залог для глаголов).

Результаты поиска демонстрируют не все возможные словосочетания, а только наиболее значимые — то есть те, на которые нужно обратить внимание при изучении русского языка (в том числе как иностранного). Подобная информация основывается на применении статистических метрик. Например, при поиске сочетания *глагол* + «*решение*» в корпусе «Тайга» наиболее важными оказываются следующие словоформы: *принял, приняло, приняли, приняла, обжаловать, вынес, отменил, принять, примет, оспорить*. Отдельно отмечены значимые граммемы, которые вносят наибольший «вес» в степень устойчивости словосочетания. Для решения этой задачи используется расстояние Кульбака–Лейблера (мера KLD), которое позволяет оценить, какие грамматические признаки оказываются наиболее употребительными для данной лексической

[1] https://cosyco.ru/cococo/

единицы [41]. При помощи алгоритма сравниваются частоты некоторой единицы во всем корпусе и в выборке, которая ограничена моделью запроса, заданной пользователем. Ниже приводится формула для меры KLD [там же]:

$$Div(C) = \sum_{i=1}^{N} P_i^{pattern} \times \log\left(\frac{P_i^{pattern}}{P_i^{corpus}}\right), \quad (10)$$

где C – морфологическая категория в рамках заданной модели, $P_i^{pattern}$ – относительная частота i-го значения категории, ограниченного моделью, P_i^{corpus} – частота того же значения во всем корпусе.

Мера KLD принимает наибольшее значение для категории падежа (по сравнению с родом, числом и одушевленностью) в модели *предлог + существительное* — таким образом именно данная категория оказывается наиболее важной [там же]. Чтобы выявить наиболее часто встречающееся значение рассматриваемой категории, используется отношение частот (*англ*. frequency ratio) по следующей формуле [41, 42]:

$$\text{frequency ratio} = \frac{P_i^{pattern}}{P_i^{corpus}}. \quad (11)$$

Если данное отношение больше 1, это означает, что значение категории является значимым для рассматриваемой модели. Например, для лексемы «рука» таковым является значение родительного падежа.

Еще одним ресурсом, в котором используется статистический аппарат для представления информации об устойчивых словосочетаниях, является «База данных коллокаций», разработанная для русского языка [43, 44]. В системе приведена информация из разнообразных лексикографических источников (толковые словари, словари сочетаемости, онлайн-словари), в которых представлены словосочетания разной степени устойчивости. В ней также доступны следующие возможности для поиска:

- поиск коллокаций по главному (опорному) слову;
- поиск коллокаций по коллокату;
- визуализация;
- просмотр каждой отдельной коллокации и её лингвостатистических характеристик;
- просмотр ссылок на печатные и электронные издания словарей;
- просмотр ссылок на корпусы текстов.

Для оценки того, насколько часто словосочетание встречается в разных источниках, был введен словарный индекс — количество словарей, в которых зафиксирована коллокация. Чем выше его значение, тем больше вероятность того, что словосочетание является воспроизводимым в речи и, как следствие, его необходимо запомнить (если речь идет об изучающем русский язык). Пример являются следующие коллокации, которые имеют индекс 5: *глубокая благодарность, нестерпимая боль, полная свобода, твердая уверенность, железный характер*. Каждая коллокация также сопровождается оценками согласно мерам ассоциации [35] на основе Araneum Russicum Maximum (MI, MI3, log-likelihood, logDice, t-score). Таким образом у пользователя есть возможность получить следующую информацию об устойчивом словосочетании [43]:

- информация о вхождении той или иной коллокации в словари (всего 9 словарей);
- общий словарный индекс коллокации;
- относительная частота в ipm на основе НКРЯ и корпуса Araneum Russicum Maximum;
- значения устойчивости словосочетания по разным метрикам.

Заключение

Статистический аппарат, который был проиллюстрирован в статье, не ограничивается приведенными примерами. Круг задач, с которыми сталкиваются филологи и лингвисты, постоянно расширяется, что заставляет специалистов искать новые способы для их решения, в том числе с помощью новых статистических методов. Приход больших данных в лингвистику, а также развитие нейронных сетей позволяют по-новому исследовать материал. Как было отмечено, возможности современных языков программирования облегчают использование в исследованиях уже готовых пакетов и функций и их применение, как для промежуточной обработки собственных данных, так и для представления результатов конечному пользователю. Также в работе не была затронута тема визуализации примеров, которая представляет интерес для лексикографических задач и заслуживает отдельного исследования.

СПИСОК ИСТОЧНИКОВ

1. **Востоков А.Х.** Опыт о русском стихосложении. СПб., 1817.
2. **Буняковский В.Я.** Лексикон чистой и прикладной математики. Т. 1.: A-D. СПб., 1839.
3. **Белый А.** Символизм. Книга статей. М., 1910.
4. **Ярхо Б.И.** Методология точного литературоведения. М., 2006.
5. **Морозов Н.А.** Лингвистические спектры: Средство для отличения плагиатов от истин. произведений того или др. известного авт. Петроград, 1916.
6. **Марков А.А.** Об одном применении статистического метода // Известия Императорской Академии Наук, серия VI, Т.X, N4, 1916. С. 239.
7. **Виноградов В.В.** Современный русский язык: в 2 т. М., 1938.
8. **Колмогоров А.Н.** Труды по стиховедению / ред.-сост. А.В. Прохоров. М.: МЦНМО, 2015.
9. **Пиотровский Р.Г.** Информационные измерения языка. Л., 1968.
10. **Алексеев П.М.** Статистическая лексикография. Л.: Изд-во ЛГПИ им. Герцена, 1975.
11. **Андреев Н.Д.** Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Л.: Наука, 1967.
12. **Головин Б.Н.** Язык и статистика. М.: Просвещение, 1970.
13. **Арапов М.В.** Квантитативная лингвистика. М., 1988.
14. **Марусенко М.А.** Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л., 1990.
15. **Мартыненко Г.Я.** Основы стилеметрии. Л.: Изд-во ЛГУ, 1988.
16. **Мартыненко Г.Я., Чебанов С.В.** Стилеметрия // Прикладное языкознание. Учебник / Отв. ред. А.С. Герд. СПб.: Изд-во СПбГУ, 1996. С. 420–435.
17. **Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А.** Статистический словарь языка Достоевского. М.: Языки славянских культур, 2003.
18. **Ляшевская О.Н., Шаров С.А.** Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. URL: http://dict.ruslang.ru/freq.php (дата обращения: 10.08.2023)
19. **Мартыненко Г.Я.** Методы математической лингвистики в стилистических исследованиях. СПб.: Нестор-История, 2019.
20. **Čermák F., Křen M.** New Generation Corpus-Based Frequency Dictionaries: The Case of Czech. In International Journal of Corpus Linguistics, 10 (4), 2005. Pp. 11–13.
21. **Hlaváčová J.** New Approach to Frequency Dictionaries – Czech Example. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy. European Language Resources Association (ELRA), 2006. URL: http://lrec.elra.info/proceedings/lrec2006/pdf/11_pdf.pdf (дата обращения: 10.08.2023)
22. **Алексеев П.М.** Частотные слова. Учебное пособие. СПб.: Изд-во СПбГУ, 2011.
23. **Kaeding F.W.** Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographiesysteme / Hrsg. von F.W. Kaeding. Steglik bei Berlin: Selbsverlag des Herausgebers, 1898.

24. **Штейнфельдт Э.А.** Частотный словарь современного русского литературного языка: Справочник для преподавателей рус. яз. М.: Прогресс, 1965.

25. Частотный словарь русского языка /под ред. Л.Н. Засориной. М.: Рус. яз., 1977. URL: http://project.phil.spbu.ru/lib/data/slovari/zasorina/zasorina.html (дата обращения 10.09.2023)

26. **Шаров С.А., Ляшевская О.Н.** Введение к частотному словарю современного русского языка. URL: http://dict.ruslang.ru/freq.pdf (дата обращения: 10.08.2023)

27. British National Corpus. URL: http://www.natcorp.ox.ac.uk/ (дата обращения: 10.08.2023)

28. **Juilland A., Dorothy B., Davidovitch C.** Frequency dictionary of French words. The Hague–Paris: Mouton, 1970.

29. **Gries S.Th.** Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 2008. № 13 (4). Pp. 403–437.

30. **Gries S.Th.** Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.). In A practical handbook of corpus linguistics, New York: Springer, 2020. Pp. 99–118.

31. **Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V.** The Sketch Engine: ten 2023s on. Lexicography, 2014. № 1. Pp. 7–36.

32. Sketch Engine. URL: https://www.sketchengine.eu (дата обращения 10.09.2023)

33. Statistics used in the Sketch Engine. URL: https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf (дата обращения 10.09.2023)

34. Поиск коллокаций. URL: https://ruscorpora.ru/page/tool-collocations/ (дата обращения: 10.08.2023)

35. **Evert S.** The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, 2004. URL: http://purl.org/stefan.evert/PUB/Evert2004phd.pdf (дата обращения: 10.08.2023)

36. **Хохлова М.В.** Экспериментальная проверка методов выделения коллокаций. // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы. Под ред. А. Мустайоки, М.В. Копотева, Л.А. Бирюлина, Е.Ю. Протасовой. Хельсинки, 2008. С. 343–357.

37. **Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R.** CoCoCo: Online Extraction of Russian Multiword Expressions. In The 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria). Sofia: INCOMA Ltd, 2015. Pp. 43–45.

38. Национальный корпус русского языка. 2003–2023. URL: http://ruscorpora.ru (дата обращения 10.09.2023)

39. **Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.** Corpus as language: from scalability to register variation. In Dialogue, Russian International Conference on Computational Linguistics, Bekasovo, 2013. URL: http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BelikovVI.pdf (дата обращения: 10.08.2023)

40. **Sharoff S., Nivre J.** The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge. In Russian International Conference on Computational Linguistics "Dialogue 2011", May 25-29, Bekasovo, Russia. Moscow, 2011. Pp. 591–604. URL: http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/58.pdf (дата обращения: 10.08.2023)

41. **Kopotev M., Pivovarova L., Kochetkova N., Yangarber R.** Automatic Detection of Stable Grammatical Features in n-gramms. In Papers from the 9th Workshop on Multiword Expressions at NAACL 2013, Atlanta, 2013. Pp. 73–81.

42. **Kormacheva D., Pivovarova L., Kopotev M.** Automatic Collocations Extraction and Classification of Automatically Obtained Bigrams. Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations, Tübingen, 2014. Pp. 27–33.

43. **Хохлова М.В., Мамаев И.Д.** Разработка базы данных коллокаций: обзор золотого стандарта на примере атрибутивных словосочетаний // Труды международной конференции «Корпусная лингвистика–2021». СПб.: Скифия-принт, 2021. С. 370–379.

44. **Хохлова М.В.** Атрибутивные коллокации в золотом стандарте сочетаемости русского языка и их представление в словарях и корпусах текстов // Вопр. лексикографии. 2021. № 21. С. 33–68.

REFERENCES

[1] **A.Kh. Vostokov,** Opyt o russkom stikhoslozhenii [On the Russian versification]. SPb., 1817.

[2] **V.Ya. Bunyakovskiy,** Leksikon chistoy i prikladnoy matematiki [Lexicon of pure and applied mathematics]. T. 1: A-D. SPb., 1839.

[3] **A. Belyy,** Simvolizm. Kniga statey [Symbolism. Book of articles] M., 1910.

[4] **B.I. Yarkho,** Metodologiya tochnogo literaturovedeniya [Methodology of precise literary criticism] M., 2006.

[5] **N.A. Morozov,** Lingvisticheskiye spektry: Sredstvo dlya otlicheniya plagiatov ot istin. proizvedeniy togo ili dr. izvestnogo avt. [Linguistic spectra: A means for distinguishing plagiarism from truth. works of one or another famous author] Petrograd, 1916.

[6] **A.A. Markov,** Ob odnom primenenii statisticheskogo metoda [On one application of the statistical method], Izvestiya Imperatorskoy Akademii Nauk [News of the Imperial Academy of Sciences], seriya VI, T.X, N4, 1916. P. 239.

[7] **V.V. Vinogradov,** Sovremennyy russkiy yazyk v 2 t. [Modern Russian language in two volumes] M., 1938.

[8] **A.N. Kolmogorov,** Trudy po stikhovedeniyu [Works on poetry] / red.-sost. A.V. Prokhorov. M.: MTsNMO, 2015.

[9] **R.G. Piotrovskiy,** Informatsionnyye izmereniya yazyka [Information dimensions of language] L., 1968.

[10] **P.M. Alekseyev,** Statisticheskaya leksikografiya [Statistical lexicography] L.: Izd-vo LGPI im. Gertsena, 1975.

[11] **N.D. Andreyev,** Statistiko-kombinatornyye metody v teoreticheskom i prikladnom yazykovedenii [Statistical-combinatorial methods in theoretical and applied linguistics] L.: Nauka, 1967.

[12] **B.N. Golovin,** Yazyk i statistika [Language and statistics] M.: Prosveshcheniye, 1970.

[13] **M.V. Arapov,** Kvantitativnaya lingvistika [Quantitative linguistics] M., 1988.

[14] **M.A. Marusenko,** Atributsiya anonimnykh i psevdonimnykh literaturnykh proizvedeniy metodami raspoznavaniya obrazov [Attribution of anonymous and pseudonymous literary works by image recognition methods] L., 1990.

[15] **G.Ya. Martynenko,** Osnovy stilemetrii [Fundamentals of stylometry] L.: Izd-vo LGU, 1988.

[16] **G.Ya. Martynenko, S.V. Chebanov,** Stilemetriya [Stylemetry] Prikladnoye yazykoznaniye. Uchebnik [Applied linguistics. Textbook] / Otv. red. A.S. Gerd. SPb.: Izd-vo SPbGU (1996) 420–435.

[17] **A.Ya. Shaykevich, V.M. Andryushchenko, N.A. Rebetskaya,** Statisticheskiy slovar yazyka Dostoyevskogo [Statistical Dictionary of Dostoevsky's Language] M.: Yazyki slavyanskikh kultur, 2003.

[18] **O.N. Lyashevskaya, S.A. Sharov,** Chastotnyy slovar sovremennogo russkogo yazyka (na materialakh Natsionalnogo korpusa russkogo yazyka) [Frequency Dictionary of the Modern Russian Language (based on materials from the National Corpus of the Russian Language)] M.: Azbukovnik, 2009. URL: http://dict.ruslang.ru/freq.php (accessed 10.08.2023)

[19] **G.Ya. Martynenko,** Metody matematicheskoy lingvistiki v stilisticheskikh issledovaniyakh [Methods of mathematical linguistics in stylistic research] SPb.: Nestor-Istoriya, 2019.

[20] **F. Čermák, M. Křen,** New Generation Corpus-Based Frequency Dictionaries: The Case of Czech. In International Journal of Corpus Linguistics, 10 (4) 2005 11–13.

[21] **J. Hlaváčová,** New Approach to Frequency Dictionaries – Czech Example. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy. European Language Resources Association (ELRA), 2006. Available at: http://lrec.elra.info/proceedings/lrec2006/pdf/11_pdf.pdf (accessed: 10.09.2023)

[22] **P.M. Alekseyev,** Chastotnyye slova. Uchebnoye posobiye. SPb.: Izd-vo SPbGU, 2011.

[23] **F.W. Kaeding,** Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographiesysteme / Hrsg. von F. W. Kaeding. Steglik bei Berlin: Selbsverlag des Herausgebers, 1898.

[24] **E.A. Shteynfeldt,** Chastotnyy slovar sovremennogo russkogo literaturnogo yazyka: Spravochnik dlya prepodavateley rus. yaz [Frequency Dictionary of the Modern Russian Literary Language: A Handbook for Russian Teachers] M.: Progress, 1965.

[25] Chastotnyy slovar russkogo yazyka [Frequency dictionary of the Russian language] /ed. by L.N. Zasorinoy. M.: Rus. yaz., 1977. Available at: http://project.phil.spbu.ru/lib/data/slovari/zasorina/zasorina.html (accessed 10.08.2023)

[26] **S.A. Sharov, O.N. Lyashevskaya,** Vvedeniye k chastotnomu slovaryu sovremennogo russkogo yazyka. Available at: http://dict.ruslang.ru/freq.pdf (accessed: 10.09.2023)

[27] British National Corpus. Available at: http://www.natcorp.ox.ac.uk/ (accessed 10.08.2023)

[28] **A. Juilland, B. Dorothy, C. Davidovitch,** Frequency dictionary of French words. The Hague–Paris: Mouton, 1970.

[29] **S.Th. Gries,** Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13 (4) (2008) 403–437.

[30] **S.Th. Gries,** Analyzing dispersion. In Magali Paquot & Stefan Th. Gries (eds.). In A practical handbook of corpus linguistics, New York: Springer, 2020. Pp. 99–118.

[31] **A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel,** The Sketch Engine: ten years on. Lexicography, 1 (2014) 7–36.

[32] Sketch Engine. Available at: https://www.sketchengine.eu (accessed 10.08.2023)

[33] Statistics used in the Sketch Engine. Available at: https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf (accessed 10.08.2023)

[34] Poisk kollokatsiy [Collocation search]. Available at: https://ruscorpora.ru/page/tool-collocations/ (accessed 10.08.2023)

[35] **S. Evert,** The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, 2004. Available at: http://purl.org/stefan.evert/PUB/Evert2004phd.pdf (accessed 10.08.2023)

[36] **M.V. Khokhlova,** Eksperimentalnaya proverka metodov vydeleniya kollokatsiy [Experimental study of methods for collocation extraction], Slavica Helsingiensia 34. Instrumentariy rusistiki: Korpusnyye podkhody [Tools for Russian studies: corpus approaches]. Ed. by A. Mustayoki, M.V. Kopoteva, L.A. Biryulina, Ye.Yu. Protasovoy. Khelsinki, 2008. Pp. 343–357.

[37] **M. Kopotev, L. Escoter, D. Kormacheva, M. Pierce, L. Pivovarova, R. Yangarber,** CoCoCo: Online Extraction of Russian Multiword Expressions. In The 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria). Sofia: INCOMA Ltd, 2015. Pp. 43–45.

[38] Natsionalnyy korpus russkogo yazyka [National Corpus of the Russian Language] 2003—2023. Available at: http://ruscorpora.ru (accessed 10.08.2023)

[39] **V. Belikov, N. Kopylov, A. Piperski, V. Selegey, S. Sharoff,** Corpus as language: from scalability to register variation. In Dialogue, Russian International Conference on Computational Linguistics, Bekasovo, 2013. Available at: http://www.dialog-21.ru/digests/dialog2013/materials/pdf/BelikovVI.pdf (accessed: 10.09.2023)

[40] **S. Sharoff, J. Nivre,** The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge. In Russian International Conference on Computational Linguistics "Dialogue 2011", May 25-29, Bekasovo, Russia. Moscow, 2011. Pp. 591–604. Available at: http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/58.pdf (accessed: 10.09.2023)

[41] **M. Kopotev, L. Pivovarova, N. Kochetkova, R. Yangarber,** Automatic Detection of Stable Grammatical Features in n-gramms. In Papers from the 9th Workshop on Multiword Expressions at NAACL 2013, Atlanta, 2013. Pp. 73–81.

[42] **D. Kormacheva, L. Pivovarova, M. Kopotev,** Automatic Collocations Extraction and Classification of Automatically Obtained Bigrams. Workshop on Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations, Tübingen, 2014. P. 27–33.

[43] **M.V. Khokhlova, I.D. Mamayev,** Razrabotka bazy dannykh kollokatsiy: obzor zolotogo standarta na primere atributivnykh slovosochetaniy Development of a collocation database: a review of the gold standard using the example of attributive phrases, Trudy mezhdunarodnoy konferentsii «Korpusnaya lingvistika–2021» [Proceedings of the international conference "Corpus Linguistics-2021"] SPb.: Skifiya-print, 2021. Pp. 370–379.

[44] **M.V. Khokhlova,** Attributive collocations in the gold standard of combinability of the Russian language and their representation in dictionaries and text corpora, Vopr. leksikografii [Questions of lexicography], 21 (2021) 33–68.

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT AUTHOR

Хохлова Мария Владимировна
Maria V. Khokhlova
E-mail: m.khokhlova@spbu.ru
https://orcid.org/0000-0001-9085-0284

Поступила: 10.08.2023; Одобрена: 25.09.2023; Принята: 27.09.2023.
Submitted: 10.08.2023; Approved: 25.09.2023; Accepted: 27.09.2023.