Вестник Томского государственного университета. Филология. 2024. № 90. С. 5–25
Tomsk State University Journal of Philology. 2024. 90. pp. 5–25

ЛИНГВИСТИКА

Научная статья
УДК 81’32
doi: 10.17223/19986645/90/1

Оптимальный текст, частотные характеристики деформированных и авторских текстов в русле распределения Ципфа

Ольга Григорьевна Горина[1], Наталья Сергеевна Царакова[2], Михаил Владимирович Крайторов[3], Дарья Алексеевна Куганова[4], Игорь Дмитриевич Петров[5]

[1, 2, 3, 4, 5] Национальный исследовательский университет «Высшая школа экономики» – Санкт-Петербург, Санкт-Петербург, Россия
[1] gorina@bk.ru
[2] n.carakova@gmail.com
[3] mvkraytorov@edu.hse.ru
[4] dakuganova@edu.hse.ru
[5] idpetrov@edu.hse.ru

Аннотация. Исследуются ранговые распределения на лингвистическом материале, а именно анализируются частотные характеристики слов в текстах, написанных на английском языке; фиксируются расхождения между вычисляемыми по формуле Ципфа и реально наблюдаемыми частотами в зависимости от размера текста, таким образом, исследуется размер и устанавливаются причины существования оптимального текста; кроме того, экспериментально изучаются влияние целостности текста и его авторства на частотные характеристики.

Ключевые слова: ранговые распределения, теоретические и наблюдаемые частоты, оптимальный размер текста, частотные характеристики деформированных текстов, авторская константа

Благодарности: публикация подготовлена в ходе проведения исследования (№ 23-00-005, «Исследование частотных характеристик языка в рамках распределения Ципфа-Мандельброта; решение прикладных задач по отбору вокабуляра в русле корпусных процедур») в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2022–2023 гг.

Для цитирования: Горина О.Г., Царакова Н.С., Крайторов М.В., Куганова Д.А., Петров И.Д. Оптимальный текст, частотные характеристики деформированных и авторских текстов в русле распределения Ципфа // Вестник Томского государственного университета. Филология. 2024. № 90. С. 5–25. doi: 10.17223/19986645/90/1

Original article
doi: 10.17223/19986645/90/1

Zipf's distribution in language: Optimal text, frequency parameters of distorted texts and authorship ratio

Olga G. Gorina[1], Natalya S. Tsarakova[2], Michael V. Kraytorov[3], Daria A. Kuganova[4], Igor D. Petrov[5]

[1, 2, 3, 4, 5] *National Research University Higher School of Economics, Saint Petersburg, Russian Federation*
[1] *gorina@bk.ru*
[2] *n.carakova@gmail.com*
[3] *mvkraytorov@edu.hse.ru*
[4] *dakuganova@edu.hse.ru*
[5] *idpetrov@edu.hse.ru*

Abstract. The article focuses on a family of rank distributions. These are of key importance in the linguistic study of texts within the framework of quantitative and corpus linguistics. The efforts within this research are invested into the word frequency behavior study for texts written in English: we have fixed the discrepancies that inevitably occur between the frequencies calculated by Zipf's formula and the actually observed frequencies depending on the size of the text. Thus, not only the size per se but also the reasons behind the existence of an "optimal" text size are investigated; in addition, the influence of the integrity of the text and its authorship on frequency characteristics are experimentally studied. Firstly, within the framework of the goals stated, in a series of experiments, the "optimal" text size was determined. It has to be noted that the optimal text size was predicted by George Zipf himself; this size constitutes the minimal discrepancy between the formula-based or calculated (theoretical) and actually observed frequencies. Moreover, the emphasis is put on the size of the optimal text, which proves to be a key parameter in investigating distorted texts. Secondly, this article also touches upon a number of controversial provisions that used to be expressed in relation to incomplete or distorted texts. The frequency characteristics of distorted texts are studied to verify or defy the previously proposed hypotheses, expressed, in particular by the Russian mathematician Yury Orlov (1980). The assumption, which has been taken for granted, deals with the commonly held hypothesis that the distribution of word frequencies in incomplete texts might disagree with Zipf's Law. We sought empirical proof for this assumption and thoroughly explored the correlation between observed frequencies and text's completeness or its entirety. Contrary to expectations, the results prove that only the size of the text is crucial: the distribution remains Zipfian even for fragments of the text and randomly selected words from the text, provided that they collectively make up the text of the optimal size. Finally, this study discovers that authorship lends itself to being investigated with frequency derivatives. There is the so-called author's ratio, defined as the relative frequency of the most frequent word, which proves insensitive to whether the texts of a given author are incomplete, distorted or even fragmentary. It remains remarkably stable throughout both complete texts and random fragments such as sentences written by any given author.

Keywords: rank distributions, theoretical and observed frequencies, optimal text size, frequency characteristics of deformed texts, author's ratio

Acknowledgements: The publication was prepared within the framework of the Academic Fund Program at HSE University in 2022–2023 (grant No. 23-00-005, Research

on Language Frequency Characteristics within the Zipf-Mandelbrot Distribution; Addressing Applied Issues for Vocabulary Selection Using Corpus-based Methods).

For citation: Gorina, O.G., Tsarakova, N.S., Kraytorov, M.V., Kuganova, D.A. & Petrov, I.D. (2024) Zipf's distribution in language: Optimal text, frequency parameters of distorted texts and authorship ratio. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology.* 90. pp. 5–25. (In Russian). doi: 10.17223/19986645/90/1

Введение

Рассмотрение частотных характеристик слов начнём с обзора основных исследований в области ранговых распределений. Становление соответствующей области лингвистических изысканий связано с именем Джорджа Ципфа, в частности с одноименным законом, который описывает закономерности частотных распределений в естественном языке [1]. Реферативная часть статьи охватывает основные направления в исследовании ранговых распределений, в том числе на лингвистическом материале, а также включает ссылки на близкие по духу работы ученых-математиков и физиков, работавших на лингвистическом материале. Практическая часть работы посвящена эмпирической проверке выполнимости закона Ципфа в зависимости от объема текста; также анализировалось влияние целостности произведения (текста) на выполнимость закона Ципфа. Статистическая часть работы сначала выполнялась с помощью корпусных инструментов, предлагаемых программным продуктом WordSmith 6,0. Впоследствии методология эксперимента потребовала разработки собственных программных модулей для (i) автоматического сравнения наблюдаемых и теоретических частот; (ii) деформации текстов в рамках наблюдения за корреляцией между целостностью текстов и выполнимостью распределения Ципфа.

Проведенные эксперименты сосредоточены на некоторых особенностях распределений Ципфа, рассмотрение которых продиктовано рядом соображений практического и теоретического характера. Во-первых, существует ряд примечательных взаимосвязей, связанных с наличием оптимального текста. Во-вторых, в соответствии с широко распространенным предположением о влиянии целостности текста на ранжирование распределения мы провели ряд экспериментов для проверки корреляций между наблюдаемыми частотами и целостностью текста. Мы также пытались ответить на вопрос, в какой степени нарушение целостности текста искажает распределение Ципфа. В-третьих, в ходе эксперимента было выявлено, что авторство текста частично выражается через его производные частотные характеристики. Относительная частота самого частотного слова в тексте является постоянной величиной для каждого конкретного автора, практически не зависящей от рассматриваемого произведения этого автора и от размера текста. Таким образом, существует величина, являющаяся производной от частотных характеристик, которая сохраняет свое значение в рамках произведений одного и того же автора. В статье мы предлагаем называть данную величину

авторской константой, которая отличается устойчивостью не только в рамках полных текстов авторов, но и в отдельных фрагментах текстов каждого конкретного автора.

Актуальность работы. Обращение к тематике ципфовских распределений не ново, но не утратило релевантности и по сей день. «Сегодня понятие рангового распределения в информатике стало вполне привычным. Идея «распределения» информационных потоков по закону Ципфа – Мандельброта (иначе Брэдфорда, Лотка и т.д.) принята общественным мнением, является теоретической основой изучения этих потоков. Формы соответствующих распределений несколько варьируются, но понятно, что речь идет о некоем едином типе ранговых распределений. Тот же самый тип распределений известен в лингвистике, в биологии, в экономике и в социологии» [2. С. 1]. В настоящее время выводы об оптимальном размере и количестве городских агломераций страны в связи с ее экономической успешностью могут строиться на законе Ципфа [3]. Более того, проблематика может изучаться с разных ракурсов; до сих пор проводится анализ самого механизма, вызывающего появление подобных распределений. Так, при изучении происхождения закона Ципфа путем создания длинных связных сообщений в процессе коммуникации на основе стохастической динамической модели генерации текста был сделан вывод, что эта модель неизбежно приводит к ранговым распределениям в количественном согласии с эмпирическими данными; иными словами, результаты подтверждают неизбежность проявления и значимость закона Ципфа в человеческом языке [4].

Таким образом, тематика еще не исчерпала себя полностью особенно для междисциплинарных, или так называемых «дефисных» направлений, поскольку существуют определенные лакуны в знании, осмыслении и трактовках; в частности, мотивом к написанию этой статьи послужили ранее сделанные выводы о неципфовском характере распределений в несвязных или неполных текстах. Мы выработали специальную процедуру по деформации текстов и в данной работе называем подвергшиеся такому изменению тексты деформированными. Заключения о частотных характеристиках неполных текстов были сделаны в «докорпусную» эпоху, когда проверка утверждений была очень трудоемкой. Со временем эти утверждения стали восприниматься как устоявшиеся, в то время как многие из них не проверялись с помощью современных, более эффективных корпусных инструментов. Поэтому мы вновь обратились к ципфовским распределениям. Кроме того, в процессе экспериментального исследования выявилась уже упомянутая авторская константа, что потребовало проведения дополнительных экспериментов для проверки ее существования. В ходе данных экспериментов и сравнения частотных характеристик как полных текстов, так и их фрагментов мы получили результаты, которые будут изложены в заключительных разделах статьи.

В выводах, сделанных на основании более чем 10 000 экспериментальных вычислений, обсуждаются причины рассогласованности вычисляемых (теоретических) и реально наблюдаемых частот. В выводах данной работы

рассматриваются также парадоксы, связанные с ранговыми распределениями, разъясняются границы применимости понятия математического ожидания в контексте появления слов в тексте. В качестве побочного продукта вычислений была обнаружена авторская характеристика, которая требует тщательной проверки на более обширной выборке произведений различных авторов.

Уникальной особенностью авторского стиля является наличие авторской константы – относительной частоты наиболее частотного слова, которая остается неизменной в следующих случаях: (i) в завершенных работах одного автора, (ii) в отдельных фрагментах текстов этого автора, (iii) в целых предложениях и в выборке случайно взятых слов из произведений данного автора.

Таким образом, данная статья представляет собой теоретический и практический вклад в понимание ранжирования и частотных распределений в естественном языке. Результаты экспериментов, демонстрирующих выполнимость закона Ципфа вне зависимости от целостности текста, указывают на важность и необходимость исследований в данной области. Внимание, уделенное авторской константе, может способствовать развитию перспективных исследований в данной области.

Проблематика исследования ранговых распределений в различных областях. Взаимосвязь ранговых распределений и языка

Ранговые распределения описывают сложные системы, включая язык, и другие явления человеческой жизни, характеризующиеся существенной неравномерностью поведения, которое напоминает лавину. Часто употребительные слова становятся еще более употребительными подобно тому, как большие города укрупняются быстрее, а «успех порождает успех» (success breeds success). Американский социолог Р. Мертон назвал этот феномен эффектом Матфея [5, 6]. Первым, кто обратил внимание на то, что произведение частоты слова на его порядковый номер в частотном словаре есть величина постоянная, был журналист Жан Батист Эсту [7]; позже зависимость частоты от ранга была сформулирована как закон Ципфа [1].

Выбор математических соотношений, описывающих подобные явления, зиждется на степенной зависимости типа $Y = AX^{\alpha}$, где A и α – константы. При отрицательном показателе степени мы получаем гиперболическую зависимость. Один из первых результатов, описываемых таким распределением, – кривую распределения доходов на базе статистических данных о подоходном налоге – получил Вильфредо Парето. Еще одним примером гиперболического распределения является закон распределения научной продуктивности (frequency distributions of scientific performance) Альфреда Лотки. Таким образом, теория ранговых распределений может манифестироваться во многих направлениях исследований и изучаться на различных эмпирических базах данных, включающих экономические, лингвистические, публикационные и другие варианты данных.

Как уже упоминалось, эмпирический анализ гиперболических распределений *ранговым методом* был впервые предложен Дж. Ципфом, исследовавшим лингвистический материал [8]. В рамках данного метода для конкретно взятого текста последовательность всех различных слов ранжируется в порядке убывания частотности и сопоставляется с его местом, или рангом.

По мнению многих исследователей, с математической точки зрения законы Ципфа – Парето обнаруживают некоторые неожиданные свойства, вступающие в противоречие с традиционным для нас «гауссовским» представлением о вероятностной природе многих из встречающихся нам в окружении явлений [6]. Известно, что гауссово распределение изучаемых параметров очень широко распространено, так как мы выходим на это распределение каждый раз, когда результат зависит от большого количества независимых случайных величин, каждая из которых может иметь свое распределение. Главным условием для случайных величин, согласно известной центральной предельной теореме, будет наличие конечных математического ожидания и дисперсии (первого и второго моментов). В распределениях же типа Ципфа – Парето с ростом размера выборки растет среднее значение и дисперсия стремится к бесконечности. У данных распределений нет конечных первого и второго моментов, а наблюдаются эффекты так называемой «концентрации» и «рассевания» параметров. Это означает, что параметры с большими значениями концентрируются на слишком маленьком числе элементов и вместе с тем параметры с малыми значениями рассеяны на слишком большом количестве элементов [6]. Например, самое частотное слово в словаре обычно составляет примерно 5% от количества всех слов в тексте, а слова, встречающиеся в тексте всего один раз, составляют примерно половину всего словаря. Этот факт противоречит гауссовскому представлению, поскольку имеет слишком большой массив редких элементов (хвост) и медленную, по сравнению с гауссовским распределением, сходимость.

Если обратиться к лингвостатистике языка, то увидим, что в английском языке, как правило, 20% служебных слов (grammar words) выполняют 80% всей работы (hard-working core) по формированию предложений [9]. Если бы частотные характеристики языка соответствовали распределению Гаусса, то основное число слов в языке имело бы среднюю частотность с некоторым разбросом по краям, т.е. редкими выбросами как частотных, так и нечастотных слов [6].

В действительности это не так, и основное число слов «малопродуктивно»; более того, 50% слов любого корпуса текстов встречаются в нем лишь один раз, относясь к так называемым гапакс легоменонам (hapax legomena). Этот факт часто отмечается как эмпирически установленный и упоминается на страницах работ, посвященных корпусным исследованиям языка [10]. Несмотря на то, что этот факт легко обнаруживается генерированием частотного словаря, количественный объем слов с единичной частотностью может быть выведен и аналитически [11]. С утилитарной точки зрения ранговый характер распределения слов по частотности тесно связан

с показателем покрываемости текста минимальным списком; следовательно, при обучении иностранным языкам или РКИ отбор, например, лексических минимумов продвинутых уровней, в том числе профессиональных, может базироваться на более сложном частотном анализе низкочастотных единиц [12]. Таким образом, в основе законов, названных по имени изучавших их ученых – Ципфа, Парето, и Мандельброта, лежат негауссовы закономерности, что должно учитываться при рассмотрении лингвистического материала как с теоретических, так и с прикладных позиций [13].

Возвращаясь к вопросу о формульных аппроксимациях закона Ципфа, также можно заметить, что, во-первых, существование границы между редкими и частотными элементами может вызывать необходимость описывать их с помощью разных аппроксимирующих выражений, а во-вторых, существует неподтвержденное предположение, что внутри самого ядра, ограниченного лишь высокочастотными грамматическими словами, распределение может быть не ранговым.

Методы, материалы, инструменты

Для проверки гипотез наличия оптимального размера текста и частотных характеристик деформированных текстов, проиллюстрированных в таблицах, использовались тексты следующих произведений: (i) роман Джека Лондона «Мартин Иден» и набор коротких рассказов О. Генри, (ii) роман Джоан Роулинг «Гарри Поттер и философский камень», книга Кейт Фокс «Наблюдая за англичанами» и целый ряд других произведений, написанных на английском языке. Для проверки гипотезы об *авторской константе* изучались тексты большего количества авторов; мы просчитали величину авторской константы для более ста текстов, написанных различными авторами. На момент написания данной статьи коллективно написанные работы в научных журналах не проверялись на наличие авторской константы, тем не менее при использовании определенной методики такая проверка представляет интерес.

Первоначально, как мы уже упоминали, набор отобранных литературных и публицистических текстов исследовался с помощью инструмента, или корпусного менеджера WordSmith Tools в версии 6.0 [14]. Известно, что это набор программ для генерации ранжированных по частотности списков слов, ранжированных по частотности, конкордансов, а также вычисления ключевых слов и коллокаций. Впоследствии процедура анализа частотностей была усовершенствована, в результате был выработан наш собственный подход к анализу и сравнению частот. В частности, было разработано собственное программное обеспечение, которое позволяет разбивать исходный текст на некоторое количество равных последовательных частей (обозначим их буквой M), а затем зафиксировать, сколько вхождений каждого слова присутствует в каждой M-й части текста. Следовательно, программа работает не с таблицей частот, а с матрицей размером V на M, где V – количество слов в тексте (или размер словаря), а M – количество равных частей, на которые разбит исследуемый текст.

Таким образом, для сравнения частот был создан целевой программный продукт, сопоставляющий частоты на основе матрицы, размеры которой определяются объемом словаря и количеством отрывков текста. Очевидно, что в такой матрице для слов, встретившихся во всем тексте один раз, репрезентирующая эти слова строка матрицы будет содержать все нули и лишь одну единицу, соответствующую той части текста, где встретилось слово с единичной частотностью. Для высокочастотных слов соответствующая строка матрицы будет состоять из примерно одинаковых чисел, поскольку высокочастотные слова (а это всегда служебные слова) встречаются в тексте сравнительно равномерно.

В ходе проведения эксперимента по сравнению теоретических или расчетных частот с реально наблюдаемыми в текстах значениями, текстовый материал увеличивался и обрабатывался итерационно. Текст произведения разбивался на некоторое количество равных частей. Далее генерировался частотный словарь и вычислялись частотные параметры для первого фрагмента текста, затем для первого и второго фрагментов, далее рассчитывались параметры для совокупности первого, второго и третьего фрагментов. Таким образом, весь текстовый материал изучаемого произведения был разбит на увеличивающиеся фрагменты, частоты которых последовательно изучались и сводились в таблицу. Так, в табл. 1 приведены значения параметров для текста романа Джека Лондона «Мартин Иден», разбитого на 18 частей (по 8 тысяч словоупотреблений в каждой части), что соответствует 18 строкам таблицы, каждая из которых репрезентирует на 8 тысяч словоупотреблений больше, чем предыдущая.

Деформация текстов

Определенная часть исследования посвящена неполным текстам, поэтому была разработана процедура деформации текстового материала для получения текстов, частично потерявших связность и смысловую ясность. В данной части эксперимента мы осуществляли сокращение текстов путем случайного удаления слов из исходного текста; этот процесс является основной составляющей деформации текста. Отметим, что отдельные слова удаляются из текста, как мы уже упомянули, случайным образом, по специально разработанной процедуре. Чтобы достичь этой цели, мы начали с нумерации всех слов, а затем использовали генератор случайных чисел для выбора тех номеров из пронумерованного списка слов, которые должны были исключаться из текста. При анализе частотных характеристик отобранных произведений рассматривались как сами случайно выбранные слова, так и деформированные тексты.

Важно отметить, что порядок слов в тексте не оказывает влияния на его частотные характеристики.

Таблица 1

Размер текста (количество словоупотреблений, *N*) и словаря (*V*), относительная (P_1) и абсолютная (F_1) частоты, значения константы *K* и ошибки *E*

№	*N*	*V*	F_1	P_1	$K = 1/(Ln(V)+\gamma)$	$K = V/N$	*E*
1	8 000	2 048	450	0,05625	0,12192	0,25600	0,43377
2	16 000	3 123	831	0,05194	0,11596	0,19519	0,34724
3	24 000	3 885	1 229	0,05121	0,11310	0,16188	0,24994
4	32 000	4 658	1 588	0,04963	0,11082	0,14556	0,18688
5	40 000	5 262	1 993	0,04983	0,10934	0,13155	0,11468
6	48 000	5 907	2 390	0,04979	0,10798	0,12306	0,06627
7	56 000	6 507	2 824	0,05043	0,10686	0,11620	0,02407
8	64 000	6 956	3 157	0,04933	0,10611	0,10869	0,03056
9	72 000	7 508	3 560	0,04944	0,10525	0,10428	0,00451
10	80 000	7 931	4 001	0,05001	0,10465	0,09914	0,06123
11	88 000	8 427	4 486	0,05098	0,10399	0,09576	0,06848
12	96 000	8 824	4 912	0,05117	0,10349	0,09192	0,08539
13	104 000	9 281	5 322	0,05117	0,10296	0,08924	0,09947
14	112 000	9 759	5 767	0,05149	0,10243	0,08713	0,11072
15	120 000	10 135	6 250	0,05208	0,10203	0,08446	0,12604
16	128 000	10 451	6 725	0,05254	0,10171	0,08165	0,14294
17	136 000	10 688	7 137	0,05248	0,10148	0,07859	0,16444
18	142 259	10 887	7 462	0,05245	0,10129	0,07653	0,18003

В нашем эксперименте мы искажали тексты двумя способами: а) случайным удалением выбранных слов из изучаемого текста; б) составлением изучаемого текста на основе случайно выбранных слов из целостного текста.

Исследование

Для решения задач нашего исследования необходимо представить некоторые преобразования классической формулы закона Ципфа, а также определиться со способом сравнения частот. Для этого (i) были выполнены преобразования формулы Ципфа; (ii) было введено понятие ошибки как меры сравнения расчетных и теоретических частот.

Кроме того, прежде чем приступить к описанию эксперимента, необходимо определить понятие «оптимальный текст». Под оптимальным понимается текст такого размера (по количеству словоупотреблений), при котором наблюдается наилучшее совпадение между реально наблюдаемыми и теоретически вычисленными частотами по формуле Ципфа.

Сопоставляя различные терминологические традиции, следует сказать, что частотным словарем в российской науке принято называть набор слов с частотой их вхождения в данный текст, упорядоченный по убыванию частоты, в то время как в зарубежной литературе, в частности в продукте Word Smith 6.0, используется термин вордлист (wordlist).

Оптимальный текст и ошибка

Для выполнения преобразований формулы Ципфа обратимся к некоторым обозначениям. Рассмотрим множество всех слов, которым соответствуют словоформы, образующие текст. Каждому слову ~~(словоформе)~~ мы поставили в соответствие целое число ***n***, равное количеству словоформ в тексте. Далее полученное множество было упорядочено по убыванию ***n***, а для слов с одинаковым (совпадающим) значением ***n*** – в алфавитном порядке. Каждому слову в полученном упорядоченном множестве присваивается порядковый номер, начиная с 1, который в дальнейшем будем называть рангом. Число ***n*** именуется частотой слова, а полученный упорядоченный набор из трех параметров: ранг, слово, частота – принято называть частотным словарем текста. Построенный таким образом словарь упорядочен по возрастанию ранга и по убыванию частоты. Как несложно заметить, максимальный ранг слова равен количеству разных слов в словаре. Это значение составляет размер словаря; обозначим его символом ***V***. Вместе с тем сумма частот всех слов в словаре равна количеству словоформ в тексте и является размером (объемом) текста, который обозначен буквой ***N***. Как следует из принятых обозначений, ранг слова в словаре может принимать значения от 1 до *V*, т.е. ***i*** = 1, … ***V***, где ***i*** – ранг слова.

Абсолютную частоту слова с *i*-м рангом обозначим F_i, соответственно относительная частота слова определяется как частное $p_{i=}\frac{F_i}{N}$. Таким образом, формулу закона Ципфа представим следующим образом:

$$F_i = \frac{\widetilde{\mathrm{K}}}{i}, \qquad (1)$$

где $\widetilde{\mathrm{K}}$ – подбираемая для данного текста константа.

Вернемся к рассмотрению закона Ципфа и изложим некоторые соображения, связанные константой $\widetilde{\mathrm{K}}$ из формулы (1). Для простоты дальнейших рассуждений перепишем эту формулу для относительных частот, поделив обе части выражения на N:

$$p_i = \frac{K}{i}, \qquad (2)$$

где $K = \frac{\widetilde{K}}{N}$.

Учитывая правила, по которым создан частотный словарь, можно представить следующие выражения:

$$\sum_{i=1}^{V} F_i = N \text{ или } \sum_{i=1}^{V} p_i = 1.$$

С учетом последнего выражения соотношение (2) может быть представлено в виде

$$K \sum_{i=1}^{V} \frac{1}{i} = 1\,. \qquad (3)$$

Для больших значений V, используя формулу Эйлера для частичной суммы гармонического ряда (или $\boldsymbol{\nu}$-е гармоническое число, т.е. сумма обратных величин первых $\boldsymbol{\nu}$ последовательных чисел натурального ряда), представим последнее равенство (3) в виде

$$K = \frac{1}{(\ln(V)+\gamma)}\,, \tag{4}$$

где γ = 0,577…. – постоянная Эйлера – Маскероне. Таким образом, выражение (3) представляет собой не что иное, как одно из условий нормировки K (подбираемой константы K) для выражения (1). Два других условия нормировки для константы K выглядят следующим образом:

$$K = p_1\,. \tag{5}$$

$$K = \frac{V}{N}\,. \tag{6}$$

Выражение (6) было выведено с учетом того, что частота наименее частотного слова в частотном словаре всегда равна единице. Таким образом, соотношение (6) представляет собой хорошо известный корпусным лингвистам коэффициент лексического разнообразия (так называемый Type – Token Ratio, TTR).

Для проверки согласованности данных, основанных на теоретических или расчетных выражениях, и эмпирических значениях, полученных из аутентичных лингвистических материалов, мы провели эксперимент. Были исследованы текст полного собрания сочинений Джека Лондона под названием «Мартин Иден», а также тексты из 15 сборников рассказов О. Генри и другие текстовые материалы.

Несмотря на то, что общий вид закона Ципфа (1) подразумевает разные значения частот для различных рангов, на практике это не выполняется, и количество уникальных частот в частотном словаре значительно уступает количеству повторяющихся. Кроме того, когда речь идет об абсолютных частотах, имеет смысл говорить только о целочисленных значениях, и поэтому при вычислениях частот мы использовали округление. В качестве меры различия теоретической частоты и наблюдаемой для слова с рангом i из частотного словаря использовалась величина E_i, которую мы ввели выражением

$$E_i = \frac{|F_i - \hat{F}_i|}{\sqrt{F_i \hat{F}_i}}\,. \tag{7}$$

где F_i и $\hat{F}_i$ соответственно наблюдаемая и вычисляемая частоты. Суммарная ошибка, как количественная мера расхождения кривых для наблюдаемых и вычисляемых частот, считается по формуле, которая суммирует все ошибки E_i и затем делит сумму на значение V, т.е. размер словаря:

$$E = \frac{1}{V}\sum_{i=1}^{V} E_i\,. \tag{8}$$

Величину, вычисляемую по формуле (8), будем в дальнейшем называть **ошибкой**.

В ходе проведения эксперимента текстовый материал рассматриваемых произведений увеличивался и обрабатывался итерационно. Текст разбивался на некоторое количество равных частей. Далее генерировался частотный словарь и вычислялись частотные параметры для первого фрагмента текста, затем для первого и второго, потом для первого, второго и третьего. И так до конца произведения. В табл. 1 приведены значения параметров для текста романа Джека Лондона «Мартин Иден», разбитого на 18 частей по 8 тысяч слов в каждой части.

Результаты, представленные в табл. 1, приводят нас к следующим выводам:

1. Значение относительной частоты первого слова в частотном словаре (p_1) практически не меняется с увеличением размера текста, и как следствие, в качестве значения для нормировки K уравнение (5) подходит плохо. Проведенные эксперименты показывают, что при использовании уравнения (5) для вычисления значения K получается существенное расхождение значений наблюдаемых и вычисляемых частот.

2. Значения константы K для нормировочных уравнений, вычисляемые как $K = 1/((ln(V)+\gamma))$ и $K = V/N$, практически совпадают при размере текста порядка 72 000 словоформ.

3. Минимальное значение ошибка E принимает также и при размере текста приблизительно равном 72 000 словоформ. Иными словами, как показывает эксперимент, объем текста имеет принципиальное значение при оценке рассогласованности между теоретическими и эмпирическими данными в изучаемых ранговых распределениях.

Так, для текста объемом 72 000 словоформ, т.е. оптимального текста, расхождение между кривыми наблюдаемых и вычисляемых частот минимально. Для визуализации результатов были построены графики зависимости частот от ранга слова. На рис. 1, 2 сопоставлены кривые, построенные в соответствии со значениями наблюдаемых и вычисляемых частот слов в частотном словаре. Для наглядности кривые приведены в двойном логарифмическом масштабе. Для текста, состоящего из первых 72 000 словоупотреблений, кривые показаны на рис. 1. Для полного текста, состоящего из 142 259 словоупотреблений, кривые представлены на рис. 2.

Таким образом, очевидно, что распределение (1) точно выполняется при достаточно ограниченном объеме текста, который и признается оптимальным. Однако проведенные нами эксперименты показывают, что с ростом отклонения от оптимального объема текста по количеству словоупотреблений в большую или меньшую сторону все явственнее наблюдается рассогласованность между теоретически вычисляемыми (расчетными) и значениями частот в реальном тексте.

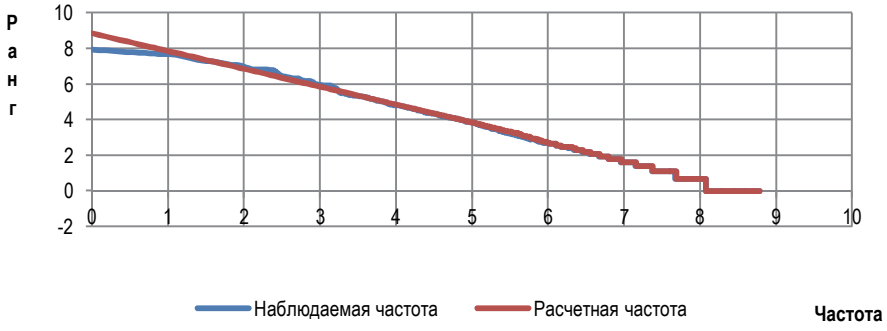

Рис. 1. График зависимости относительных частот от ранга для текста романа Джека Лондона «Мартин Иден» объемом 72 000 словоупотреблений в двойном логарифмическом масштабе

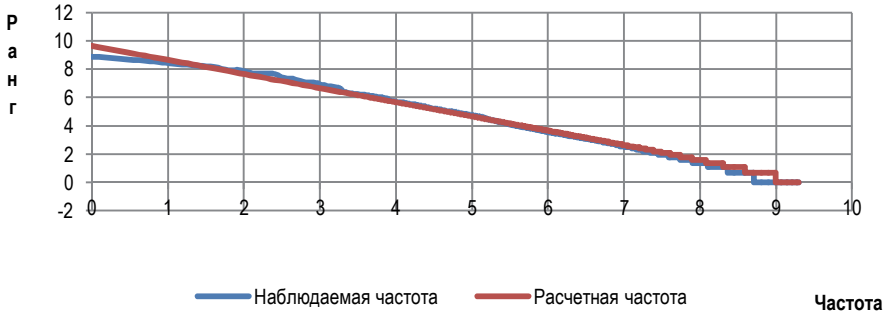

Рис. 2. График зависимости относительных частот от ранга для полного текста романа Джека Лондона «Мартин Иден» объемом 142 259 словоупотреблений в двойном логарифмическом масштабе

Примечательно, что этот факт не ускользнул от внимания самого Дж. Ципфа, который писал, что наилучшее совпадение теоретических и наблюдаемых частот будет иметь место на выборках так называемого «оптимального объема» или «размера» [1]. Вместе с тем ученый не привел развернутых объяснений этого факта. Вероятно, эмпирические проверки не проводились по причине отсутствия мощных компьютерных возможностей, которые доступны сейчас.

Исследование зависимости целостности текста и выполнимости распределения Ципфа – Мандельброта

В одной из своих работ российский математик Ю.К. Орлов [15] поднимает вопрос о том, что частоты слов в частотных словарях согласуются с вычисляемыми в распределении Ципфа–Мандельброта только в тех случаях, когда в качестве текста взято законченное литературное произведение. Автор также считал, что справедливо и следующее обобщение: указанное

частотное соответствие оговоренным распределениям заметно ухудшается, если в качестве текста взять лишь отрывок литературного произведения или текст нехудожественного жанра. Данное утверждение цитировалось во многих работах и со временем стало рассматриваться как устоявшийся факт.

Поскольку корпусные инструменты и сегодняшние вычислительные мощности позволяют проверить данное предположение, был разработан и проведен эксперимент для проверки высказанной гипотезы на материале художественного произведения. В качестве текстовой основы этой части исследования были взяты роман Джека Лондона «Мартин Иден» и роман Джоан Роулинг «Гарри Поттер и философский камень». Результаты по каждому из произведений помещены в отдельные таблицы. Так, в табл. 2 приведены результаты по тексту романа Джека Лондона «Мартин Иден». Как упоминалось в секции, описывающей методологию вычислений, целые тексты каждого произведения разбивались на равные последовательные отрывки объемом 8 тыс. словоупотреблений каждый. Соответственно, для каждого из отрывков текстов объемом 8 000, 16 000, 24 000 и т.д. словоупотреблений был составлен частотный словарь, посчитан размер словаря $V_{\text{набл}}$, вычислен расчетный размер словаря $V_{\text{расч}}$ с использованием формул (2) и (4), приведено значение относительной частоты самого частотного слова p_1 и значение ошибки E. Напомним, что в метрике данного исследования было введено понятие ошибки, которая рассчитывается по формуле (8) и дает возможность количественно оценить расхождение между наблюдаемыми и вычисляемыми частотами. Таким образом, все описанные результаты для целостных текстов приведены в левой половине таблицы.

Аналогичные вычисления были выполнены и для деформированных текстов. Результаты соответствующих расчетов приведены в правой половине табл. 2. Для эксперимента было важно, чтобы деформированный текст представлял собой бессмысленный набор слов, в котором искажены как связность, так и смысловое содержание и законченность. Вопреки рассуждениям о корреляции между целостностью текстов и их частотными характеристиками при сравнении полученных значений наблюдаемых $V_{\text{набл}}$ и вычисляемых частот $V_{\text{расч}}$, а также ошибки E и относительной частоты самого частного слова p_1 в исходных и деформированных текстах обнаруживается незначительное расхождение рассматриваемых параметров. Следует отметить две замечательные особенности. Во-первых, значение относительной частоты самого частотного слова p_1 сохраняет свое постоянство как для исходного, так и для деформированного текста и практически не зависит от размера фрагмента. Во-вторых, расхождение между наблюдаемыми и расчетными частотами, характеризующееся величиной ошибки E, достигает своего минимального значения при практически одинаковом размере как для исходного, так и для деформированного текста, равном 72 000 словоупотреблений.

Таблица 2

Параметры распределения Ципфа (2) для целостного текста в левой части таблицы и для деформированного текста в правой части таблицы (по тексту романа Джека Лондона «Мартин Иден»)

N	*V*набл	*V*расч	p_1	*ERR*	*N*	*V*набл	*V*расч	p_1	*ERR*
8 000	2 048	975	0,05625	0,43377	8 000	1 868	986	0,04888	0,35201
16 000	3 123	1 855	0,05194	0,34724	16 000	3 083	1 858	0,05150	0,30075
24 000	3 885	2 714	0,05121	0,24994	24 000	4 075	2 700	0,05046	0,25599
32 000	4 658	3 546	0,04963	0,18688	32 000	4 892	3 527	0,05113	0,20498
40 000	5 262	4 374	0,04983	0,11468	40 000	5 575	4 346	0,05168	0,15595
48 000	5 907	5 183	0,04979	0,06627	48 000	6 211	5 155	0,05171	0,10580
56 000	6 507	5 984	0,05043	0,02407	56 000	6 768	5 959	0,05250	0,06467
64 000	6 956	6 791	0,04933	0,03056	64 000	7 267	6 759	0,05248	0,02825
72 000	7 508	7 578	0,04944	0,00451	72 000	7 760	7 552	0,05289	0,00334
80 000	7 931	8 372	0,05001	0,06123	80 000	8 178	8 345	0,05356	0,04924
88 000	8 427	9 151	0,05098	0,06848	88 000	8 614	9 130	0,05363	0,06602
96 000	8 824	9 935	0,05117	0,08539	96 000	9 009	9 914	0,05373	0,08012
104 000	9 281	10 707	0,05117	0,09947	104 000	9 367	10 697	0,05368	0,09655
112 000	9 759	11 472	0,05149	0,11072	112 000	9 742	11 474	0,05372	0,11372
120 000	10 135	12 244	0,05208	0,12604	120 000	10 064	12 252	0,05358	0,13209
128 000	10 451	13 019	0,05254	0,14294	128 000	10 390	13 027	0,05328	0,14750
136 000	10 688	13 801	0,05248	0,16444	136 000	10 673	13 803	0,05303	0,16415
142 259	10 887	14 410	0,05245	0,18003	142 259	10 887	14 410	0,05245	0,18003

Таким образом, эти данные позволяют сделать вывод о наличии «оптимального» размера текста не только для целостных, но и для деформированных текстов. Очевидно, что, несмотря на потерю связности и деформацию, *оптимальный* объем практически совпадает у исходного и деформированного текстов. В табл. 3 приведены аналогичные расчеты для романа Джоан Роулинг «Гарри Поттер и философский камень». Для этого произведения, как видно из таблицы, «оптимальный» размер для исходного и деформированного текста равен 24 тыс. словоупотреблений. Тем же способом были проанализированы еще несколько произведений разных жанров, на материале которых подтвердился аналогичный результат. Отметим, что анализировались частоты как в самих усеченных за счет рандомно выброшенных слов текстах, так и в «текстах», состоящих исключительно из самих выброшенных слов. Результат исследования частот в этих текстовых материалах достаточно стабилен: основным параметром является объем или размер текста, именно близость к оптимальному размеру текста определяет то, насколько ципфовскими будут его частотные характеристики.

Таким образом, обратившись к совокупности результатов проведенных вычислений, можно сделать некоторые важные выводы. Во-первых, многочисленные вычисления доказывают, что степень соответствия наблюдаемых частотных характеристик текста соответствующим вычисляемым характеристикам в действительности не зависит от жанра произведения, его законченности и целостности.

Таблица 3

Параметры распределения Ципфа (2) для целостного текста в левой части таблицы и для деформированного текста в правой части таблицы (по тексту романа Джоан Роулинг «Гарри Поттер и философский камень»)

N	*V*набл	*V*расч	p_1	*ERR*	*N*	*V*набл	*V*расч	p_1	*ERR*
4 000	1 103	769	0,04450	0,52163	4 000	1 280	782	0,04050	0,56235
8 000	1 708	1 348	0,04663	0,43680	8 000	1 931	1 371	0,04213	0,49190
12 000	2 179	1 880	0,04900	0,34672	12 000	2 422	1 922	0,04250	0,42778
16 000	2 609	2 405	0,04813	0,28851	16 000	2 791	2 452	0,04231	0,35750
20 000	3 026	2 930	0,04585	0,24999	20 000	3 134	2 961	0,04265	0,29169
24 000	3 350	3 437	0,04471	0,00946	24 000	3 444	3 446	0,04392	0,02897
28 000	3 596	3 924	0,04468	0,02437	28 000	3 745	3 929	0,04425	0,06044
32 000	3 857	4 394	0,04528	0,06869	32 000	3 989	4 407	0,04438	0,06548
36 000	4 178	4 853	0,04614	0,06297	36 000	4 220	4 879	0,04433	0,06659
40 000	4 431	5 322	0,04583	0,06884	40 000	4 449	5 343	0,04445	0,06698
44 000	4 614	5 785	0,04557	0,07606	44 000	4 659	5 804	0,04445	0,07208
48 000	4 824	6 223	0,04654	0,09044	48 000	4 852	6 255	0,04473	0,08435
52 000	5 022	6 673	0,04652	0,10134	52 000	5 030	6 709	0,04458	0,09916
56 000	5 238	7 118	0,04654	0,11301	56 000	5 192	7 145	0,04514	0,11596
60 000	5 389	7 563	0,04640	0,13708	60 000	5 366	7 586	0,04530	0,13051
64 000	5 536	8 008	0,04614	0,15004	64 000	5 524	8 018	0,04569	0,14782
68 000	5 681	8 448	0,04597	0,16791	68 000	5 671	8 460	0,04546	0,16688
72 000	5 820	8 889	0,04569	0,18356	72 000	5 796	8 892	0,04557	0,18415
76 000	5 953	9 309	0,04617	0,20477	76 000	5 946	9 324	0,04554	0,20415
82 000	6 138	9 965	0,04563	0,23079	82 000	6 100	10 000	0,04546	0,21994
82 588	6 163	10 029	0,04559	0,23392	82 588	6 163	10 029	0,04559	0,23392

Во-вторых, наблюдаемые и вычисляемые частотные характеристики у целостных и деформированных текстов практически совпадают. Фактически вид вычисляемой кривой распределения частот слов зависит лишь от двух величин: относительной частоты самого употребляемого слова $\boldsymbol{p_1}$ и размера текста $\boldsymbol{N}$. Величина $\boldsymbol{p_1}$ является постоянной, практически не зависящей от размера и деформации текста, но вместе с тем ощутимо различающейся для каждого автора. В-третьих, оптимальным текстом является текст определенного размера, при котором наблюдаемые и вычисляемые частотные характеристики совпадают как для целостных текстов, так и для деформированных. Напомним, что существование оптимального текста было описано Дж. Ципфом и предполагает совпадение наблюдаемого размера частотного словаря с вычисляемым. Иными словами, для размера текста, при котором значение ошибки $\boldsymbol{E}$ минимально, расхождение между наблюдаемым и вычисляемым размерами словаря тоже будет минимальным. Для придания большей наглядности вышесказанному приведем графики зависимостей ошибки $\boldsymbol{E}$ от размера текста $\boldsymbol{N}$ для целостного и деформированного текстов книги Джоан Роулинг. Графики представлены на рис. 3.

Две кривые данного графика, соответствующие целостному и деформированному тексту, имеют практически равный по значению минимум при совпадающем значении текста (примерно 24 000 словоупотреблений). При

больших значениях размера текста, начиная приблизительно с 30 000 словоупотреблений, кривые практически совпадают. При малых значениях размера текста ошибка деформированного текста демонстрирует чуть большее значение по сравнению с ошибкой целостного текста.

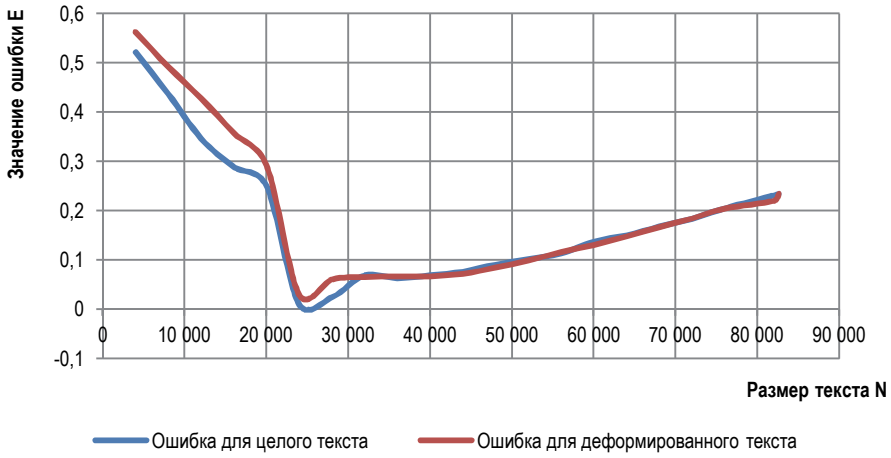

Рис. 3. Значение ошибки в зависимости от размера текста для целостного и деформированного текстов

Заключение и обсуждение результатов

В заключении представлено обсуждение результатов нашей работы, а также более общих наблюдений и выводов, касающихся распределения Ципфа–Парето, которые мы получили в ходе эксперимента.

Выводы нашего экспериментального исследования заключаются в следующем:

а) у каждого автора существует свой (уникальный) оптимальный объем текста, характеризующийся наилучшим совпадением вычисляемых и наблюдаемых частот. Однако по мере отклонения от оптимального размера текста, выполнение указанных распределений становится менее точным. Под вычисляемыми мы понимаем частоты, определенные по формулам распределения Ципфа;

б) выполнимость распределения Ципфа зависит фактически от единственного фактора, а именно объема текста. Отклонение от оптимального объема вызывает искажение частотных распределений, в то время как нарушение целостности текста не вызывает значимых расхождений между вычисленными и наблюдаемыми частотными характеристиками текста. Даже случайным образом отобранные из текста совокупности отдельных слов сохраняют ципфовские черты при условии, что их количество составляет оптимальный объем;

в) в ходе эксперимента также удалось обнаружить некую *константу автора,* равную относительной частотности самого частотного слова p_1. Эта величина проявляет существенную устойчивость к разрушению целостности текста и остается постоянной не только в законченных работах одного и того же автора, но и в текстах, составленных из отдельных фрагментов текстов автора.

Общие наблюдения по ранговым распределениям, к которым относится человеческий язык, касаются часто используемого понятия «вероятность», применяемого в отношении предсказания появления слова на страницах произведений. Мы нередко сталкиваемся с тем, что относительные частоты слов в частотном словаре ассоциируют с некими вероятностями, с которыми данные слова встречаются или в произведениях определенного жанра, или в языке вообще. На этом фоне мы считаем, что задача нашего исследования состоит в том, чтобы разобраться в причинах возникновения тех ошибок, которые стали нормой филологического и корпусно-лингвистического узуса. Многие авторы – по всей вероятности, в силу недостаточной осведомленности – зачастую искаженно представляют суть ранговых и гауссовых распределений, неправомерно расширяя (или, наоборот, сужая) область их применения, трактуя язык в терминах гауссовых распределений и пользуясь такими понятиями, как математическое ожидание, дисперсия, которые, вообще говоря, могут быть неприменимы в распределениях под названием «ранговые».

С нашей точки зрения, относительная частота слова в частотном словаре может быть интерпретирована как вероятность только в одной единственной трактовке: при случайном выборе слов из *определенного* текста, вероятность выпадения, или выбора *определенного* слова именно в этом тексте. будет соответствовать его относительной частоте в частотном словаре. Следует отметить, что речь идет об одном и том же тексте, и расширение понятия вероятности появления данного слова в других текстах неправомерно. Например, если взять два одинаковых по размеру фрагмента романа Джека Лондона «Мартин Иден» и составить для каждого из них частотный словарь, то частотные характеристики двух фрагментов будут совпадать, но слова, обладающие одинаковым рангом в двух различных словарях, будут разными. Таким образом, ранг, по существу, является оболочкой, в которую попадает слово, но в разных текстах в эту оболочку попадут разные слова. Можно метафорически представить ранг как должность, занимаемую сотрудниками в компании, например должность менеджера, старшего менеджера или директора компании. Эту позицию с неизбежностью кто-то занимает, однако в разных компаниях это будут разные люди.

Согласно результатам нашего эксперимента в каждом из двадцати фрагментов романа «Мартин Иден» Джека Лондона, состоящих из 8 тыс. слов, наблюдаются некоторые различия в ранжировании и составе самых частотных слов. Некоторые слова могут иметь разный ранг в разных фрагментах, а некоторые вообще могут не встречаться в других фрагментах. Тем не менее частота использования слов примерно одинакова для каждого ранга во

всех фрагментах, за исключением артикля «the», который почти всегда занимает первое место в частотных словарях.

Таким образом, при анализе содержания текста и составлении частотного словаря на основе прочитанных ста страниц не представляется возможным сделать вывод относительно того, какие слова и в каком объеме будут присутствовать на сто первой странице, за исключением, возможно, некоторого набора наиболее часто встречающихся слов.

Список источников

1. *Zipf G.K.* Human Behavior and the Principle of Least Effort // Science 110. 1949. № 2868. P. 669.

2. *Арапов М.В., Ефимова Е.Н., Шрейдер Ю.А.* О смысле ранговых распределений // НТИ. Сер. 2. 1975. № 1. С. 9–20.

3. *Bowen Cai, Zhenfeng Shao, Shenghui Fang, Xiao Huang, Yun Tang, Muchen Zheng, Hao Zhang.* The Evolution of urban agglomerations in China and how it deviates from Zipf's law // Geo-spatial Information Science. 2022. doi: 10.1080/10095020.2022.2083527

4. *Zanette D., Montemurro M.* Dynamics of Text Generation with Realistic Zipf's Distribution // Journal of Quantitative Linguistics. 2005. № 12:1. P. 29–40. doi: 10.1080/09296170500055293

5. *Merton R.K.* The Matthew Effect in Science // Science. 1968b. Vol. 5, № 159. P. 3810. Reprinted in: Merton, 1973.

6. *Петров В.М., Яблонский А.И.* Математика и социальные процессы: гиперболические распределения и их применение. М. : Знание, 1980. 64 с.

7. *Estoup J.B.* Gammes stenographiques: methode et exercices pour l'acquisition de la vitesse. 4e ed. rev. et aug. Parıs, 1916. 151 p.

8. *Zipf G.K.* Relative frequency as a determinant of phonetic change // Harvard studies in classical philology. 1929. № 40.

9. *O'Keeffe A., McCarthy M., Carter R.* From corpus to classroom: Language use and language teaching. Cambridge : Cambridge University Press, 2007.

10. *Scott M., Tribble C.* Textual Patterns: key words and corpus analysis in language education: Studies in Corpus Linguistics. Amsterdam/Philadelphia: John Benjamins, 2006. 200 p.

11. *Gorina O.G., Tsarakova N.S., Tsarakov S.K.* Study of Optimal Text Size Phenomenon in Zipf–Mandelbrot's Distribution on the Bases of Full and Distorted Texts. Author's Frequency Characteristics and derivation of Hapax Legomena // Journal of Quantitative Linguistics. 2020. № 27:2. P. 134–158. doi: 10.1080/09296174.2018.1559460

12. *Горина О.Г., Царакова Н.С.* Корпусные инструменты, маршруты и эксперименты в современной лингводидактике // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2021. Т. 19, № 2. С. 36–53. doi: 10.25205/1818-7935-2021-19-2-36-53

13. *Mandelbrot B.B.* The Fractal Geometry of Nature. New York : Freeman, 1983.

14. *Scott M.* Wordsmith Tools: Software. Oxford : Oxford University Press, 2012.

15. *Орлов Ю.К.* Невидимая гармония // Число и мысль. Вып. 3. М. : Знание, 1980. С. 70–106.

References

1. Zipf, G.K. (1949) Human Behavior and the Principle of Least Effort. *Science 110*. 2868. pp. 669.

2. Arapov, M.V., Efimova, E.N. & Shreyder, Yu.A. (1975) O smysle rangovykh raspredeleniy [On the meaning of rank distributions]. *NTI. Ser. 2*. 1. pp. 9–20.

3. Bowen Cai et al. (2022) The Evolution of urban agglomerations in China and how it deviates from Zipf's law. *Geo-spatial Information Science*. doi: 10.1080/10095020.2022.2083527

4. Zanette, D. & Montemurro, M. (2005) Dynamics of Text Generation with Realistic Zipf's Distribution. *Journal of Quantitative Linguistics.* 12:1. pp. 29–40. doi: 10.1080/09296170500055293

5. Merton, R.K. (1968) The Matthew Effect in Science. *Science*. 5 (159). p. 3810. Reprinted in: Merton, 1973.

6. Petrov, V.M. & Yablonskiy, A.I. (1980) *Matematika i sotsial'nye protsessy: giperbolicheskie raspredeleniya i ikh primenenie* [Mathematics and social processes: hyperbolic distributions and their applications]. Moscow: Znanie.

7. Estoup, J.B. (1916) *Gammes stenographiques: methode et exercices pour l'acquisition de la vitesse.* 4e ed. Parıs: Institut Sténographique.

8. Zipf, G.K. (1929) Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*. 40.

9. O'Keeffe, A., McCarthy, M. & Carter, R. (2007) *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

10. Scott, M. & Tribble, C. (2006) *Textual Patterns: key words and corpus analysis in language education: Studies in Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins.

11. Gorina, O.G., Tsarakova, N.S. & Tsarakov, S.K. (2020) Study of Optimal Text Size Phenomenon in Zipf–Mandelbrot's Distribution on the Bases of Full and Distorted Texts. Author's Frequency Characteristics and derivation of Hapax Legomena. *Journal of Quantitative Linguistics*. 27:2. pp. 134–158. doi: 10.1080/09296174.2018.1559460

12. Gorina, O.G. & Tsarakova, N.S. (2021) Corpus Routes and Experiments in Language Teaching. *Vestnik NGU. Seriya: Lingvistika i mezhkul'turnaya kommunikatsiya*. 19 (2). pp. 36–53. (In Russian). doi: 10.25205/1818-7935-2021-19-2-36-53

13. Mandelbrot, B.B. (1983) *The Fractal Geometry of Nature*. New York: Freeman.

14. Scott, M. (2012) *Wordsmith Tools: Software*. Oxford: Oxford University Press.

15. Orlov, Yu.K. (1980) Nevidimaya garmoniya [Invisible Harmony]. In: *Chislo i mysl'* [Number and Thought]. Vol. 3. Moscow: Znanie. pp. 70–106.

Информация об авторах:

Горина О.Г. – канд. пед. наук, доцент Департамента иностранных языков НИУ ВШЭ – Санкт-Петербург (Санкт-Петербург, Россия). E-mail: gorina@bk.ru

Царакова Н.С. – преподаватель Департамента иностранных языков НИУ ВШЭ – Санкт-Петербург (Санкт-Петербург, Россия). E-mail: n.carakova@gmail.com

Крайторов М.В. – студент Санкт-Петербургской школы экономики и менеджмента НИУ ВШЭ – Санкт-Петербург (Санкт-Петербург, Россия). E-mail: mvkraytorov@edu.hse.ru

Куганова Д.А. – студент Санкт-Петербургской школы экономики и менеджмента НИУ ВШЭ – Санкт-Петербург (Санкт-Петербург, Россия). E-mail: dakuganova@edu.hse.ru

Петров И.Д. – аспирант Санкт-Петербургской школы физико-математических и компьютерных наук НИУ ВШЭ – Санкт-Петербург (Санкт-Петербург, Россия). E-mail: idpetrov@edu.hse.ru

Авторы заявляют об отсутствии конфликта интересов.

Information about the authors:

O.G. Gorina, Cand. Sci. (Pedagogics), associate professor, National Research University Higher School of Economics (HSE University) (Saint Petersburg, Russian Federation). E-mail: gorina@bk.ru

N.S. Tsarakova, lecturer, National Research University Higher School of Economics (HSE University) (Saint Petersburg, Russian Federation). E-mail: n.carakova@gmail.com

M.V. Kraytorov, student, National Research University Higher School of Economics (HSE University) (Saint Petersburg, Russian Federation). E-mail: mvkraytorov@edu.hse.ru

D.A. Kuganova, student, National Research University Higher School of Economics (HSE University) (Saint Petersburg, Russian Federation). E-mail: dakuganova@edu.hse.ru
I.D. Petrov, postgraduate student, National Research University Higher School of Economics (HSE University) (Saint Petersburg, Russian Federation). E-mail: idpetrov@edu.hse.ru

The authors declare no conflicts of interests.

Статья поступила в редакцию 18.03.2023;
одобрена после рецензирования 21.11.2023; принята к публикации 12.07.2024.

The article was submitted 18.03.2023;
approved after reviewing 21.11.2023; accepted for publication 12.07.2024.