КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 517.97 : 532.526 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2021.1/3378
Поступила в редакцию 24.02.2021
Подписана в печать 26.04.2021

ВОЗМОЖНЫЙ АЛГОРИТМ ВЫЧИСЛЕНИЯ ПРЕДЕЛЬНОГО РАЗМЕРА СЛОВАРЯ ПИСАТЕЛЯ

© 2021 А. А. Кретов✉, М. В. Ломец, И. П. Половинкин

Воронежский государственный университет
Университетская пл., 1, 394018 Воронеж, Российская Федерация

Аннотация. В работе предлагается метод оценивания предельного размера словаря писателя с помощью экстраполяции эмпирически задаваемой функции, выражающей зависимость коэффициента лексического разнообразия от объема текстового корпуса. Обсуждаются возникающие проблемы адекватности выбираемого способа экстраполяции. На примере творчества Л. Н. Толстого произведены расчеты с помощью логарифмических базисных функций для аппроксимации и экстраполяции.
Ключевые слова: коэффициент лексического разнообразия, закон Ципфа, экстраполяция, лемматизированный частотный словарь, предельный размер словаря.

Улучшенный вариант формулы Ципфа позволял количественно оценить и ранжировать богатство словарного запаса любого человека: высокое значение — богатый лексикон; низкое значение — бедный. Имея такую шкалу, можно измерять различия по словарному запасу между текстами или говорящими. Появляется возможность количественно оценить эрудицию.

(Бенуа Мандельброт «(Не)послушные рынки: фрактальная революция в финансах» [1])

ВВЕДЕНИЕ

В основе методологии данного исследования лежит понятие *Коэффициента лексического разнообразия* (КЛР, англ. *lexical diversity*, LD) — количество лемм делённое на длину текста в словоупотреблениях [2] и TTR (англ. *type/token ratio* — отношение словаря к тексту — ОСТ) — простейший способ вычисления КЛР, а также — в русистике — идеи Ф. Паппа [3] и А. Е. Супруна [4–6] об исследовании прироста новых слов по мере прироста текста. Суть их в том, что при разбиении текста на равные отрезки, например, длиной по 100 или по 1000 слов исследуется прирост новых (ранее не встречавшихся в тексте) слов во второй, третьей, четвёртой и т. д. сотне. Введение в научный обиход понятия коэффициента лексического разнообразия Википедия [7], хотя и «предположительно», приписывает М. Темплину [8]. Для такой осторожности: в 1961 году Ф. Папп [2] писал о TTR: «Этот метод применяется в лингвистике в течение не менее чем двадцати лет», ссылаясь при этом на работы [9–11]. Как показали исследования Ф. Паппа и А. Е. Супруна, скорость прироста новых слов в тексте с каждой последующей сотней снижается, а при превышении текстом размера в 500.000 словоупотреблений появление новых слов прекращается. «Существует определенная зависимость между объемом текста в словах T и объемом его словника C. $C/T<1$, причем в принципе чем больше текст, тем меньше этот показатель лексического разнообразия текста. Лексиче-

✉ Кретов Алексей Александрович
e-mail: kretov@rgph.vsu.ru

ское богатство текста должно, следовательно, определяться с учетом этой закономерности». «Дальнейшее замедление прироста новых слов ведет к тому, что усредненное ожидание новых слов в тексте объемом свыше полумиллиона словоупотреблений доходит до величины, приближающейся к нулю. Новые слова в таких текстах как бы не ожидаются: лексическая тема исчерпывает себя. Стоит напомнить, что именно полмиллиона словоупотреблений — это объем крупнейших романов, таких, как «Война и мир» Л. Н. Толстого или «Сага о Форсайтах» Дж. Голсуорси» [12].

Мы живём в эпоху лингвистических корпусов (множеств текстов, особым образом оформленных и организованных), поэтому было бы интересно исследовать, как эта закономерность проявляется не в тексте, а в *идиолекте* — языке отдельного человека, представленного корпусом всех созданных им текстов (в идеале — всем сказанным и написанным им).

Текст является единицей культуры. Идиолект индивида находит отражение в постоянно растущей совокупности создаваемых им текстов. При этом известно, что размер словаря зависит от размера корпуса текстов: чем больше человек написал за свою жизнь, тем обычно больше разных слов встречается в его текстах.

В текстах А. С. Пушкина длиной 544.777 словоупотреблений содержится чуть менее 23.000 разных слов [13], но Пушкин прожил 36 лет, не дожив до своего 37-летия. М. Ю. Лермонтов прожил на 10 лет меньше. Соответственно, общий объём написанных им текстов, а следовательно, и количество разных слов, употреблённых им, также меньше: в текстах М. Ю. Лермонтова зафиксировано «14.939 слов с общим числом словоупотреблений 342.269» [14].

Вопрос о том, чей словарь богаче: Пушкина или Лермонтова, не может быть даже поставлен. Тем более — вопрос о сравнении их словарей со словарём Л. Н. Толстого или Ф. И. Достоевского.

При такой постановке вопроса богатство словаря измеряется количеством разных слов в дошедших до нас текстах: у кого их больше, у того словарь и богаче. А как быть с Н. В. Гоголем, имевшим привычку сжигать свои произведения (не только 2-й том «Мёртвых душ»)?

Как известно, у каждого человека есть активный словарь — те слова, которые он употребляет, и пассивный словарь — те слова, которые он понимает. Об измерении объёма пассивного словаря речи не идёт, поскольку вряд ли эта величина может быть как-либо верифицирована. Когда говорят о богатстве словаря того или иного писателя, всегда имеют в виду (в полном согласии с позитивистским идеалом научности) активный словарь, а его оценивают исключительно по тем словам, которые употреблены в созданных автором в доступных подсчётам текстах. Но в своих текстах А. С. Пушкин употребил (в разных формах) слова *тётка*, *тёткин*, *тётушка* и не разу не употребил слово *тётя*. Значит ли это, что слово *тётя* не принадлежит его активному словарю? Едва ли. Другой пример. Люди, знавшие Пушкина, утверждают, что однажды он употребил слово *виносос*. Этот факт зафиксирован в их воспоминаниях. Значит ли, что слово *виносос* принадлежит активному словарю Пушкина? Во всяком случае у нас есть основания в этом усомниться.

Можно, конечно, взять за основу размер корпуса текстов М. Ю. Лермонтова и из наследия всех остальных русских писателей взять начальную часть их наследия такого же объёма. Потом посчитать число разных слов и объявить, что самый богатый словарь у того писателя, который употребил больше разных слов на «лермонтовской» дистанции своего творчества. В таком случае в «лермонтовский» корпус Л. Н. Толстого не попали бы ни «Война и мир», ни «Анна Каренина», ни «Воскресение», ни «Хаджи-Мурат». Имело бы смысл в таком случае утверждение о богатстве или бедности словаря Л. Н. Толстого? Вряд ли.

Что же можно предложить взамен? Упрощённо говоря, сравнивать такие размеры лексиконов писателей, каких они достигли бы, проживи те столько, сколько Л. Н. Толстой, и напиши столько же.

Отправная точка исследования состоит в том, что активный словарь любого человека —

величина конечная (как, впрочем, и величина корпуса созданных им текстов — хотя бы потому, что человек смертен) и при написании им корпуса текстов определённого (достаточно большого) объёма прирост новых слов в текстах данного автора прекращается. Размер этого словаря можно определить, если знать закон роста количества новых слов при росте размера корпуса. Величина прироста словаря постепенно уменьшается и на каком-то этапе становится равной нулю. Гипотеза исследования предполагает как возможность определения этой скорости, так и нахождение по ней предельного для данного индивида объёма активного словаря. Если при достижении нулевого прироста размера словаря корпус продолжает пополняться, коэффициент лексического разнообразия будет стремиться к нулю. Один из способов определить предельный размер словаря состоит в том, чтобы увязать этот предельный размер с таким соотношением размеров словаря и корпуса, при котором величина КЛР пренебрежимо мала.

Наиболее удобным для такого исследования объектом является творческое наследие Л. Н. Толстого, охватывающее более полувека. Немаловажным является то обстоятельство, что художественные тексты Л. Н. Толстого доступны в электронном виде.

Цель данного исследования — установить предельный объём словаря Л. Н. Толстого на основании корпуса из 20 художественных текстов. Видимо, при этом мы должны сделать оговорку, что речь пойдёт об объёме активного словаря художественных произведений Л. Н. Толстого: ведь мы не берём, ни драму, ни публицистику, ни дневники, ни эпистолярное наследие писателя.

1. МАТЕРИАЛЫ И МЕТОДЫ

Чтобы сделать **объект** исследования одновременно представительным и осуществимым, мы взяли 20 произведений разного объема, охватывающие более-менее равномерно отрезок времени в 52 года: 1) 1852 Детство; 2) 1854 Отрочество; 3) 1855 Севастопольские рассказы; 4) 1856 Два гусара; 5) 1856 Утро помещика; 6) 1857 Юность; 7) 1858 Альберт; 8) 1862 Поликушка; 9) 1863 Казаки; 10) 1869 Война и мир; 11) 1877 Анна Каренина; 12) 1884 Записки сумасшедшего; 13) 1886 Смерть Ивана Ильича; 14) 1889 Крейцерова соната; 15) 1890 Дьявол; 16) 1891 Мать; 17) 1895 Хозяин и работник; 18) 1898 Отец Сергий; 19) 1899 Воскресение; 20) 1904 Хаджи-Мурат.

Общая длина этих 20-ти текстов — 1.239.183 словоупотреблений. При этом сознательно брались тексты разного размера, чтобы иметь дело с наиболее сложным случаем прироста новых слов.

2. ХОД ИССЛЕДОВАНИЯ

Обработка каждого из 20 текстов Л. Н. Толстого состояла из реализации следующей последовательности шагов, приводившей текст в формате.txt на входе к *лемматизированному частотному словарю* данного текста на выходе.

Шаг 1. Текст разбивался на текстовые слова: последовательности букв между пробелами или знаками препинания.

Шаг 2. Из текста устранялись знаки препинания и цифры.

Шаг 3. Подсчитывалась частота употребления в тексте полученных последовательностей букв (словоформ), что давало *Частотный словарь словоформ* текста.

Шаг 4. С помощью размещенного в свободном доступе морфологического анализатора русского языка *MyStem*, разработанного Ильёй Сегаловичем в компании «Яндекс», осуществлялась лемматизация Частотного словаря словоформ.

Шаг 5. С использованием возможностей электронных таблиц MS-Excel осуществлялось превращение Частотного словаря словоформ в *Частотный словарь лемм* (словарных форм). (Иногда ***лемму*** мы называем *словом* — исключительно для простоты понимания, различая при этом *слово*, представленное ***леммой***, и *словоформу*, в количественном аспекте именуемую *словоупотреблением — сл/уп*: например, длина данного текста — 100.000 словоупотреблений; данный текст содержит 20.000 словоупотреблений).

Дальнейшая работа состояла в исследовании прироста новых слов (лемм) по мере роста корпуса текстов.

Тексты были расположены в хронологическом порядке. Каждый текст получил номер и сокращенное (1–3 буквы) обозначение — по первым буквам названия (исключение составило сокращение текста «Дьявол» — ***Дьв***, начинающегося с буквы ***Д***, используемой для обозначения повести «Детство»).

Исследование прироста новых слов по мере наращивания корпуса из 20 текстов потребовало 19 шагов и осуществлялось с использованием возможностей электронных таблиц MS-Excel.

Первый шаг состоял в сравнении частотного словаря лемм (далее Частотного словника — ЧС) текста-2 «Отрочество» с текстом-1 «Детство».

Это осуществлялось следующим образом.

1) На одну страницу таблиц Excel помещались ЧС теста-1 и текста-2 в формате, представленном в табл. 1:

Таблица 1
[Table 1]

Год	Текст	Лемма	ЧаЛем	Шаг-1

В столбце «Год» указывался год завершения работы над произведением. Исключение сделано для «Севастопольских рассказов», т. к. в 1856 году и так было создано 2 произведения: «Два гусара» и «Утро помещика». Остальные произведения Л. Н. Толстого завершены в разные годы.

В столбце «Текст» помещалось условное обозначение текста: **Д** или **О**.

В столбце «Лемма» помещалась словарная форма слова, как она была определена программой MyStem. Иногда её проходилось корректировать. Например, для прилагательного *белый* программа создавала несколько лемм, вероятно, трактуя эти прилагательные как субстантивы: *белые* (‘белогвардейцы’ или ‘боровики’) или фамилии: *белый*, *белая* (*Саша Белый* и *Маша Белая*). А словоформу *готов* программа соотносила с леммой существительного *гот*, а не с прилагательным *готовый*. Если бы в русских текстах ставились ударения, такого бы не случалось: отождествить *гОтов* и *готОв* было бы невозможно.

В столбце «ЧаЛем» (Частота лемм) указывалась частота слова, равная сумме частот всех его словоформ, отмеченных в исследуемом тексте. Фрагмент полученной при этом таблицы представлен в табл. 2.

В столбце «Шаг-1» использовались результаты работы функции СЧЁТЕСЛИ, применённой к каждой из ячеек столбца «Лемма».

Если в столбце «Шаг-1» стояла цифра 2, это означало, что слово не является уникальным для данного текста. *Новыми* признавались те леммы, для которых выполнялось условие: число 1 в столбце «ЧаЛем» и название текста-2 (на первом шаге это «О») в столбце «Текст», представлены в той же строке.

На всякий случай, мы подсчитали и число уникальных слов в тексте «Детство» (оно и указано в табл.3). Формально же — *все* леммы текста «Детство» — *новые*, поскольку встречаются впервые — на Шаге-0.

На Шаге-2 указанная процедура была применена к текстам «Детство + Отрочество + Севастопольские рассказы», при этом учитывались только слова уникальные для «Севастопольских рассказов».

На Шаге-3 — та же процедура была применена к корпусу текстов «Детство + Отрочество + Севастопольские рассказы» + «Два

Таблица 2
[Table 2]

Словоформа (сл/ф)	Лемма	Чсл/ф	ЧаЛем
бабушка	бабушка	32	59
бабушке	бабушка	6	
бабушки	бабушка	18	
бабушкой	бабушка	2	
бабушку	бабушка	1	

гусара» и т. д. — до последнего Шага-19, на котором к первым 19-ти текстам для прибавлен текст «Хаджи-Мурат». Результаты всех 19-ти шагов представлены в табл.3.

Наличие столбца «ЧаЛем» давало возможность не только определить число новых слов, но и подсчитать *размер текста*, покрываемый этими словами — простым суммированием значений в той части столбца «ЧаЛем», которая содержит частоты уникальных для *нового* (т. е. прибавленного на последнем шаге) *текста*.

В табл. 3 столбец «Длина» содержит число текстовых слов (словоформ, словоупотреблений) в данном тексте. Столбец «Длина» указывает число лемм (разных слов) в данном тексте. Столбец С/Д содержит результат деления числа лемм (столбец «Слов») на сумму их употреблений, которой и измеряется длина текста (столбец «Длина»). Эта величина носит название «**Коэффициент лексического разнообразия**» (КЛР, англ. *lexical diversity*, *LD*) — количественная характеристика текста, отражающая степень богатства словаря при построении текста заданной длины. В основе показателя лежит соотношение числа отдельных лексических единиц (*лемм*, англ. *types*) и количества их употреблений в тексте (*текстоформ*, англ. *tokens*)» [https://ru.wikipedia.org/wiki/Коэффициент лексического разнообразия].

Таблица 3. Прирост новых слов и покрываемого ими текста
[Table 3. The growth of new words and the text covered by them]

№	Год	Текст	Сокр	Длина	Слов	С/Д	Нов. Слов	Нов. Сл/уп	Доля НСлов	Доля Нсл/уп
1	1852	Детство	**Д**	30326	4253	0,14	2106	3473	0,50	0,11
2	1854	Отрочество	**О**	23020	3599	0,16	1452	2299	0,40	0,10
3	1855	Севастопольские рассказы	**СР**	36041	4276	0,13	2117	5258	0,45	0,15
4	1856	Два гусара	**ДГ**	17219	3179	0,18	844	1495	0,27	0,09
5	1856	Утро помещика	**УП**	15669	2952	0,19	751	1219	0,25	0,08
6	1857	Юность	**Ю**	49939	4964	0,10	1208	2085	0,24	0,04
7	1858	Альберт	**А**	7927	1660	0,21	147	382	0,09	0,05
8	1862	Поликушка	**П**	16879	3146	0,19	725	1338	0,23	0,08
9	1863	Казаки	**К**	46002	5268	0,11	1391	4451	0,26	0,10
10	1869	Война и мир	**ВиМ**	459672	15620	0,03	7212	28193	0,46	0,06
11	1877	Анна Каренина	**АК**	270110	11473	0,04	2702	10458	0,24	0,04
12	1884	Записки сумасшедшего	**ЗС**	3729	988	0,26	32	42	0,03	0,01
13	1886	Смерть Ивана Ильича	**СИИ**	17716	2804	0,16	190	280	0,07	0,02
14	1889	Крейцерова соната	**КС**	25434	3297	0,13	270	363	0,08	0,01
15	1890	Дьявол	**Дьв**	14246	2321	0,16	395	555	0,17	0,04
16	1891	Мать	**М**	3597	1058	0,29	48	71	0,05	0,02
17	1895	Хозяин и работник	**ХиР**	14270	2534	0,18	268	440	0,11	0,03
18	1898	Отец Сергий	**ОС**	13706	2603	0,19	153	283	0,06	0,02
19	1899	Воскресение	**В**	137305	9490	0,07	1591	4029	0,17	0,03
20	1904	Хаджи-Мурат	**ХМ**	36376	4852	0,13	481	2606	0,10	0,07

Столбец «Нов слов» указывает на количество «новых» – уникальных для «нового» (прибавленного к корпусу на данном шаге) текста слов (лемм). Столбец «Нов Сл/уп» содержит сумму частот «новых слов», указывая тем самым на размер покрываемого ими текста.

Столбец «Доля Нслов» содержит частное от деления числа «новых слов» на общее количество слов в данном («новом») тексте.

Столбец «Доля Нсл/уп» содержат частное от деления доли текста, покрываемого «новыми словами» (столбец «Нов Сл/уп») на длину «нового текста» в словоупотреблениях (столбец «Длина»).

Следующим этапом исследования является анализ данных, содержащихся в табл. 3. Результат этого анализа представлена табл. 4.

В столбце «ПрирС» указан рост словника (числа разных слов) корпуса текстов нарастающим итогом.

В столбце «ПрирТ» (прирост текста) указан пошаговый прирост длины корпуса текстов.

В столбце «КЛР» указаны значения коэффициента лексического разнообразия (частное от деления значений в столбце «ПрирС» на значения в столбце «ПрирТ») на каждом из шагов роста корпуса, включая нулевой (строка № 1).

Содержание столбца «КЛР» можно трактовать как скорость прироста словаря. Динамика этого параметра представлена на рис. 1.

Рис. 1 показывает, что вначале скорость прироста словаря Л. Н. Толстого последовательно убывает, но, начиная с «Анны Карениной» практически приближается к прямой, стабилизируясь на величине 0,23, а, начиная с «Воскресенья», — на величине 0,21. **Логарифмическая** зависимость, формула которой приведена на рис. 1 вверху справа, позволяет аппроксимировать полученную кривую с коэффициентом правдоподобия 0,96.

Таблица 4. Динамика КЛР в нарастающем корпусе текстов
[Table 4. The dynamics of LD in the growing corpus of texts]

№	Год	Текст	Сокр	НовСл	ДлинаТ	ПрирСл	ПрирТ	КЛР
1	1852	Детство	**Д**	4253	30326	4253	30326	0,1402
2	1854	Отрочество	**О**	1452	23020	5705	53346	0,1069
3	1855	Севастопольские рассказы	**Ср**	2117	36041	7822	89387	0,0875
4	1856	Два гусара	**Дг**	844	17219	8666	106606	0,0813
5	1856	Утро помещика	**Уп**	751	15669	9417	122275	0,0770
6	1857	Юность	**Ю**	1208	49939	10625	172214	0,0617
7	1858	Альберт	**А**	147	7927	10772	180141	0,0598
8	1862	Поликушка	**П**	725	16879	11497	197020	0,0584
9	1863	Казаки	**К**	1391	46002	12888	243022	0,0530
10	1869	Война и мир	**ВиМ**	7212	459672	20100	702694	0,0286
11	1877	Анна Каренина	**АК**	2702	270110	22802	972804	0,0234
12	1884	Записки сумасшедшего	**Зс**	32	3729	22834	976533	0,0234
13	1886	Смерть Ивана Ильича	**СИИ**	190	17716	23024	994249	0,0232
14	1889	Крейцерова соната	**Кс**	270	25434	23294	1019683	0,0228
15	1890	Дьявол	**Дьв**	395	14246	23689	1033929	0,0229
16	1891	Мать	**М**	48	3597	23737	1037526	0,0229
17	1895	Хозяин и работник	**ХиР**	268	14270	24005	1051796	0,0228
18	1898	Отец Сергий	**ОС**	153	13706	24158	1065502	0,0227
19	1899	Воскресение	**В**	1591	137305	25749	1202807	0,0214
20	1904	Хаджи-Мурат	**ХМ**	481	36376	26230	1239183	0,0212

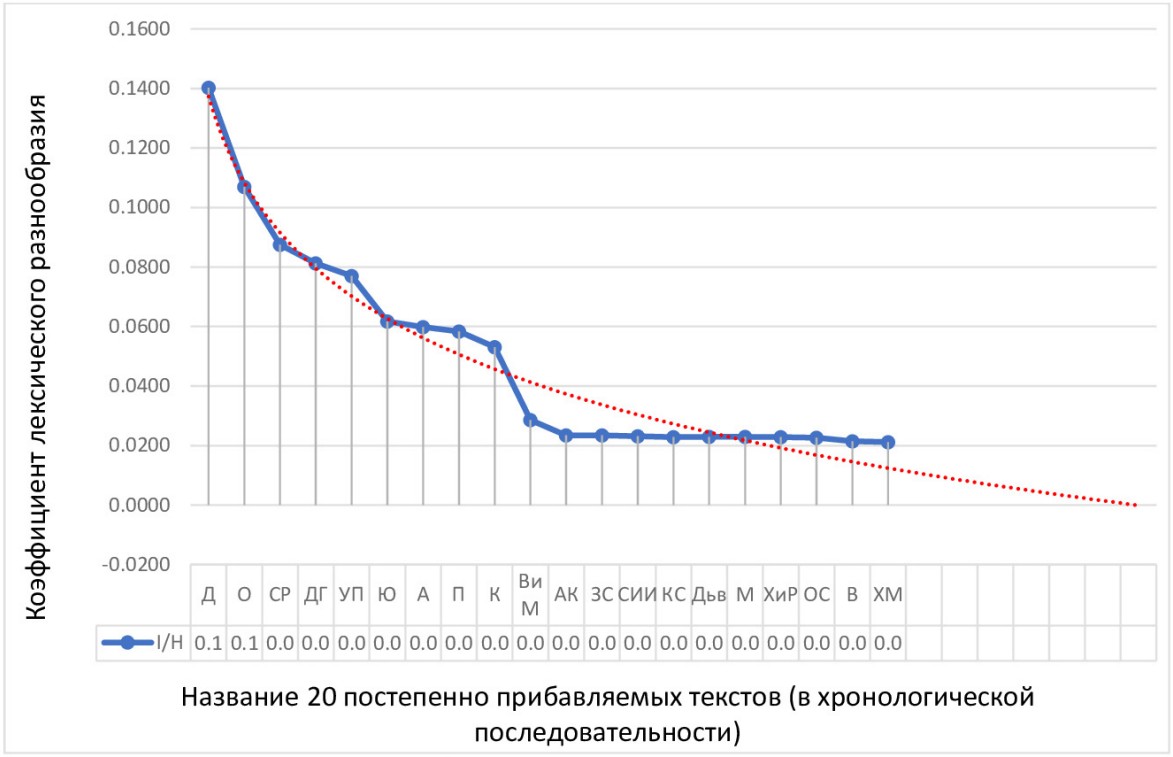

Рис. 1. Динамика КЛР в корпусе художественной прозы Л. Н. Толстого при присоединении к корпусу новых произведений

[Fig. 1. LD dynamics in the corpus of fiction L. N. Tolstoy when joining the corpus new works]

Прогнозирование поведения тренда показывает, что КЛР как функция от размера корпуса стремится к нулю. Это означает, что размер словаря Л. Н. Толстого (и любого человека вообще) конечен, а это, в свою очередь, позволяет нам вычислить *предельный размер словаря* Л. Н. Толстого, а также размер корпуса текстов, при достижении которого рост словаря прекратится и дальнейшее наращение корпуса не будет сопровождаться ростом словаря. Ср. наблюдение Ф. Паппа: «Все это говорит о том, что значения КОЕС (Кумулятивного отношения словарных единиц к словам) будут постепенно, очень медленно и все медленнее уменьшаться по мере удлинения протяженности испытуемого текста» [2:99].

Предельный размер словаря (ПРС) и есть, на наш взгляд, та величина, по которой корректно сравнивать богатство языка разных писателей и любых индивидов вообще.

Возникает естественный вопрос об адекватном моделировании тренда изменения КЛР с увеличением корпуса произведений писателя. Выбор средств моделирования, как известно, зависит от целей моделирования. В нашем случае этой целью является определение предельного размера словаря. Здесь возникает еще одна задача: указать формализованные признаки достижения предельного размера словаря. В качестве таковых можно предложить близость к нулю приращения словаря при включении в корпус текста очередного произведения или близость к нулю КЛР. Совершенно ясно, что величина КРЛ должна стремиться к нулю при неограниченном увеличении корпуса, но принимать нулевое значение не может, поскольку величина размера словаря всегда положительна. В связи с этим требуется уточнить, что понимается под «малостью» как приращения словаря, так и КЛР. Здесь возникает и проблема увязать это понятие малости с выбором модели тренда и как следствие способа экстраполяции тренда.

Имеются многочисленные попытки построения эмпирических формул для выражения зависимости объема словаря от объема текста, как и зависимости КРЛ от объема текста. Наиболее подходящей в агрегированном

смысле считается аппроксимация по степенному закону Ципфа, известному также как закон «аллометрического» или «постоянного относительного роста»:

$$КЛР = C\,x^{\gamma},$$

где $\gamma < 0$, x — накопленный размер текста корпуса. При таком моделировании тренда мы, конечно, не получим нулевого значения КЛР, что соответствует реальности. Поэтому мы можем считать, что рост словаря пренебрежимо мал, когда КЛР пренебрежимо мал. Что это означает, подчеркиваем, подлежит уточнению. Кроме проблемы уточнения «малости» есть еще одна проблема. Согласно Ю. А. Тулдаве [15, с. 99] при больших размерах текста прогнозирование тренда КЛР с помощью закона Ципфа дает значительные погрешности (завышенные оценки).

Мы предлагаем несколько иной путь. Выберем в качестве линии тренда логарифмическую зависимость (см. рис. 2). Более точно, мы выбираем логарифмические и постоянные функции в качестве базисных, а функцию зависимости КЛР от объема текста ищем в виде линейной комбинации базисных функций. Коэффициент правдоподобия в таком случае тоже очень высок. Зато такая функция имеет нуль. Значение размера текста при этом мы можем считать соответствующим предельному размеру словаря. Приравняем нулю функцию тренда и решим уравнение

$$0,408 - 0,028 \ln x = 0.$$

Пусть x_0 — корень этого уравнения. Легко видеть, что

$$\ln x_0 = 0,408 / 0,028 \approx 14,57,$$

$$x_0 \approx e^{14,57} \approx 2\,129\,565.$$

Итак, исходя из выбранного способа моделирования, мы заключаем, что размер текста корпуса, при котором достигается предельный размер словаря Л. Н. Толстого, составляет 2129565 слов. Ясно, что это некоторая приближенная оценка.

3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Теперь мы должны найти предельный размер словаря. Пойдем тем же путем. В качестве линии тренда выберем логарифмическую зависимость (см. рис.3) и приравняем нулю функцию тренда (на сей раз x означает размер словаря, а x — по-прежнему КЛР). Пусть x_1 – корень уравнения

$$0,6301 - 0,0604 \ln x = 0.$$

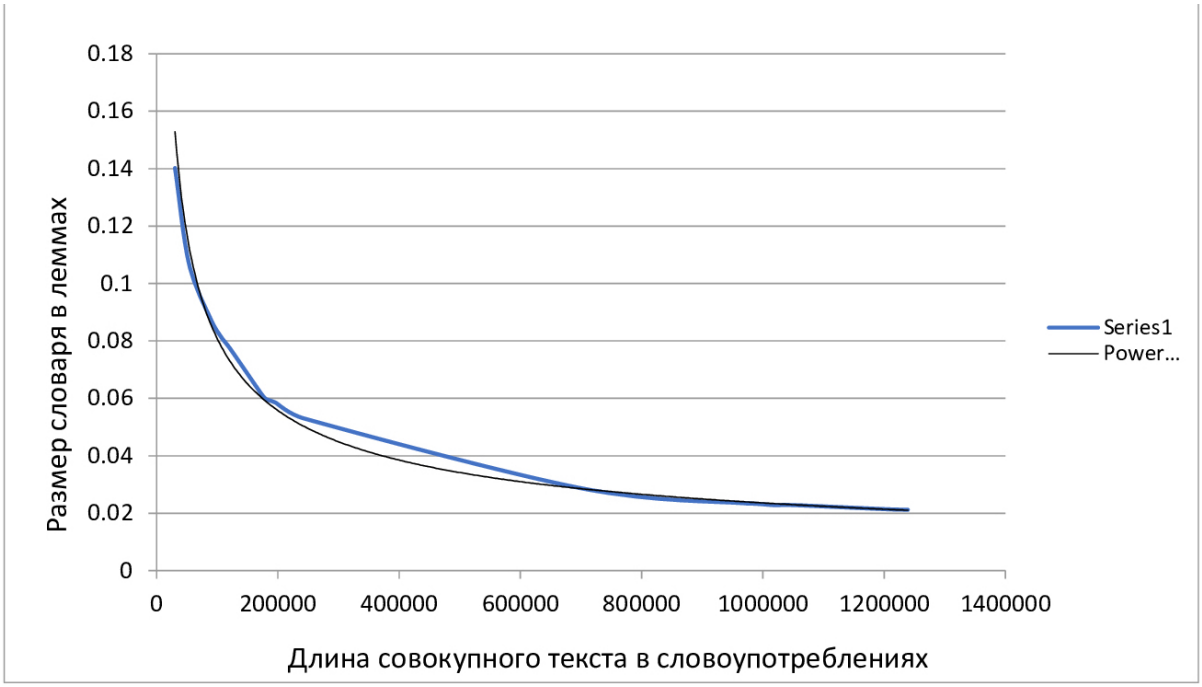

Рис. 2. Динамика КЛР в корпусе художественной прозы Л. Н. Толстого в зависимости от роста размера текста
[Fig. 2. LD dynamics in the corpus of fiction L. N. Tolstoy, depending on the growth of text size]

Тогда, очевидно,

$$\ln x_1 = 0,6301 / 0,0604 \approx 10,43,$$

$$x_1 \approx e^{10,43} \approx 33\,932.$$

Итак, оценка предельного размера словаря Л. Н. Толстого (с необходимым замечанием об учете выбранного метода моделирования) составляет примерно 33932 слова.

Есть еще одна проблема — проблема проверки достоверности полученных прогнозов. Классический способ сравнения приближенного решения с точным решением или с экспериментальными данными применен быть не может по причине отсутствия оных. Здесь нам доступны лишь косвенные способы проверки. Все же попробуем воспользоваться вариантом закона Ципфа, но теперь для описания зависимости размера словаря от размера текста:

$$Y = A\,X^{\beta},$$

где Y — размер словаря, X — размер текста, $0 < \beta < 1$. По данным табл. 2 устанавливается степенная зависимость вида (см. рис. 4):

$$Y = 38,069\,X^{0,4653}.$$

Подставив в эту формулу значение $x_0 \approx e^{14,57} \approx 2\,129\,565$, мы получим значение 33.506 слов как оценку для предельного размера словаря. Это значение отличается от полученного ранее как нуля логарифмической функции тренда КЛР. Однако относительная погрешность составляет

$$\frac{33932 - 33506}{33506} \times 100\% \approx 1,27\%,$$

что, на наш взгляд, вполне приемлемо. Осталось только принять окончательное решение о прогнозе предельного размера словаря и размера соответствующего размера корпуса. Произведя традиционные округления, приходим к следующим прогнозам:

– предельный размер словаря Л. Н. Толстого составляет 33.500–34.000 слов,

– размер текста, при котором достигается предельный размер словаря Л. Н. Толстого, составляет 2.129.500–2.130.000 слов.

ЗАКЛЮЧЕНИЕ

Итак, цель исследования достигнута и искомые значения получены.

Однако главный результат исследования видится в постановке проблемы поиска подходящего алгоритма, позволяющего вычис-

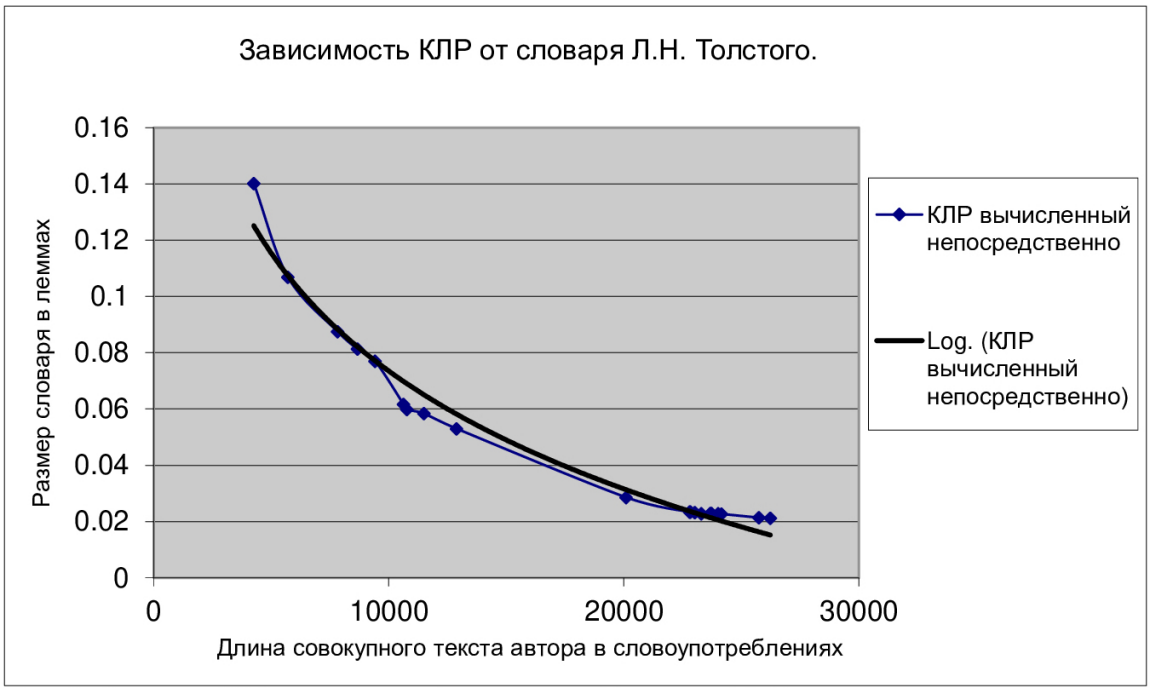

Рис. 3. Динамика КЛР в корпусе художественной прозы Л. Н. Толстого в зависимости от роста размера словаря
[Fig. 3. LD dynamics in the corpus of fiction L.N. Tolstoy, depending on the growth of the size of the dictionary]

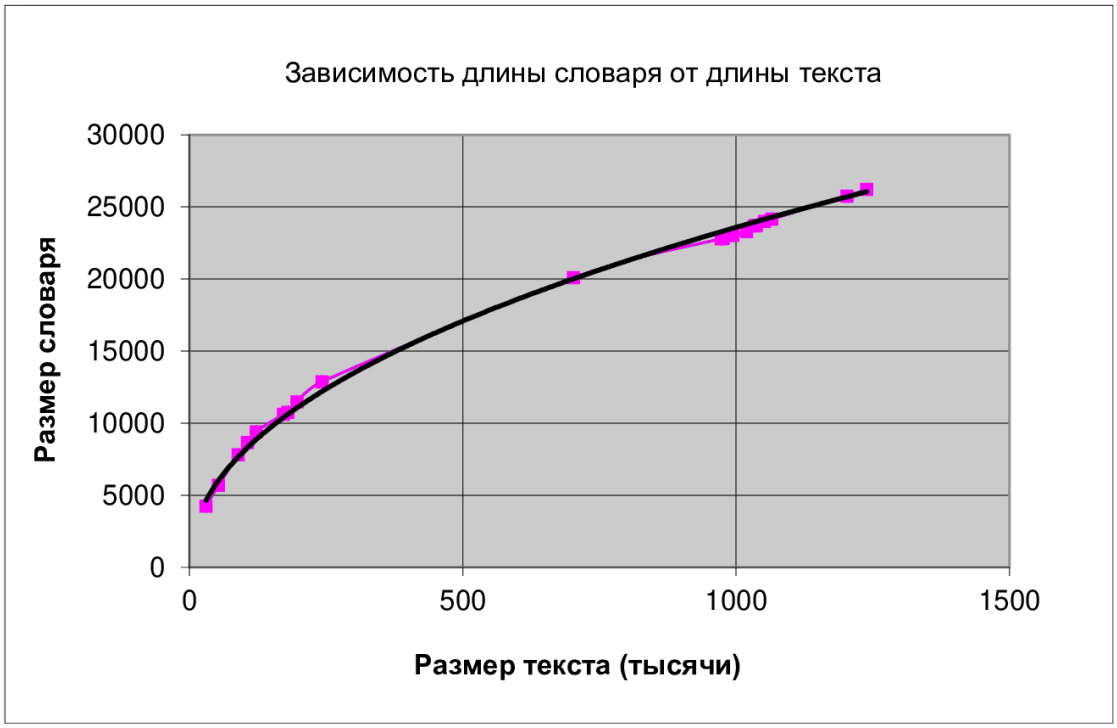

Рис. 4. Динамика размера словаря в корпусе художественной прозы Л. Н. Толстого в зависимости от роста размера текста
[Fig. 4. The dynamics of the size of the dictionary in the corpus of fiction L. N. Tolstoy, depending on the growth of text size]

лять предельный размер словаря для любого автора, представленного *минимально достаточным корпусом текстов* (МДКТ). Значение МДКТ ещё предстоит установить, но вряд ли он может быть меньше «ципфовского размера» текста в 100.000 словоупотреблений.

Авторы допускают мысль, что близость полученных двумя разными способами результатов может быть случайной. В связи с этим предложенный алгоритм определения предельного размера словаря должен быть проверен на текстах других авторов, а также на текстах Л. Н. Толстого другими способами.

Ближайшей перспективой исследования является получение данных о предельном размере словаря А. С. Пушкина, М. Ю. Лермонтова, Н. В. Гоголя, Ф. М. Достоевского, И. С. Тургенева, И. А. Гончарова, М. Е. Салтыкова-Щедрина, А. П. Чехова, М. Горького, И. А. Бунина, М. А. Шолохова и других деятелей русской литературы.

Отдалённой перспективой исследования является оценка предельного размера словаря у иноязычных авторов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Мандельброт Б.* (Не)послушные рынки: фрактальная революция в финансах / Б. Мандельброт, Р.Л. Хадсон – М. : Издательский дом «Вильямс», 2006. – 400 с.

2. *McKee G., Malvern D. & Richards B.* Measuring Vocabulary Diversity Using Dedicated Software // Literary and Linguistic Computing. – 2000. – № 15(3). – 323–337. doi: 10.1093/llc/15.3.323.

3. *Папп Ф.* Количественный анализ словарной структуры некоторых русских текстов / Ф. Папп // Вопросы языкознания. – 1961. – № 6. – С. 93–100.

4. *Супрун А. Е.* К количественной оценке лексического богатства текста / А. Е. Супрун // Филологические науки. – 1979. – № 1.

5. *Супрун А. Е.* Повтор в лексической структуре текста / А. Е. Супрун // Язык — система. Язык — текст. Язык — способность. К 60-летию члена-корреспондента РАН Ю. Н. Караулова. – М., 1995. – С. 133–141.

6. *Супрун А. Е.* Лекции по теории речевой деятельности [Текст] : Пособие для студентов филолог. фак. вузов / А. Е. Супрун. – Минск : Белорусский фонд Сороса, 1996. – 287 с.

7. https://ru.wikipedia.org/wiki /Коэффициент_лексического_разнообразия.

8. *Templin M.* Certain language skills in children. – Minneapolis : University of Minnesota Press, 1957. doi: 10.1044/jshd.1703.280.

9. *Johnson W.* Language and speech hygiene, an application of general semantics, Ann Arbor, 1944.

10. *Chotlos J. W.* Studies in language-behavior, IV – A statistical and comparative analysis of individual written language samples. – 1944. – P. 75–111. («Psychologie monographie»). doi: 10.1037/h0093511.

11. *Miller G. A.* Language and communication / G. A. Miller. – New York, 1951.

12. *Супрун А. Е.* Повтор в лексической структуре текста / А. Е. Супрун // Исследования по лингвистике текста: сборник статей / А. Е. Супрун. – Минск, 2001. – С. 108–117.

13. *Кретов А. А., Матыцина Л. Н.* Морфемно-морфонологический словарь языка А. С. Пушкина: Ок. 23.000 слов. / А. А. Кретов, Л. Н. Матыцина. – Воронеж : Центрально-Черноземное книжное издательство, 1999. – 208 с.

14. Частотный словарь языка М. Ю. Лермонтова / Под ред. В. В. Бородина, А. Я. Шайкевича; Сост. Авдеева А. А., Бородин В. В., Быкова Н. Я., Козокина С. М., Гордеева Н. А., Макарова Л. А., Шайкевич А. Я. // Лермонтовская энциклопедия / АН СССР. Ин-т рус. лит. (Пушкин. Дом); Науч.-ред. совет изд-ва «Сов. Энцикл.». – М. : Сов. Энцикл., 1981. – С. 717–774.

15. *Тулдава Ю. А.* Проблемы и методы квантитативно-системного исследования лексики / Ю.А. Тулдава. – Таллин : Валгус, 1987. – 204 с.

Кретов Алексей Александрович — д-р филол. наук, проф., профессор кафедры теоретической и прикладной лингвистики Воронежского государственного университета.
E-mail: kretov@rgph.vsu.ru
ORCID iD: https://orcid.org/0000-0002-1474-3177

Ломец Мария Викторовна — студентка кафедры теоретической и прикладной лингвистики факультета Романо-германской филологии Воронежского государственного университета.
E-mail: marusya.lomets@gmail.com.
ORCID iD: https://orcid.org/0000-0003-0885-4740

Половинкин Игорь Петрович — д-р физ.-матем. наук, профессор кафедры математического и прикладного анализа, доцент кафедры теоретической и прикладной лингвистики Воронежского государственного университета.
E-mail: polovinkin@yandex.ru
ORCID iD: https://orcid.org/0000-0002-4308-8909

DOI: https://doi.org/10.17308/sait.2021.1/3378 ISSN 1995-5499
Received 24.02.2021
Accepted 26.04.2021

POSSIBLE ALGORITHM FOR CALCULATING THE LIMIT SIZE OF A WRITER'S DICTIONARY

© 2021 A. A. Kretov✉, M. V. Lomets, I. P. Polovinkin

Voronezh State University
1, Universitetskaya Square, 394018 Voronezh, Russian Federation

Annotation. The article is proposed a method for estimating the maximum size of a writer's dictionary by extrapolating an empirically defined function expressing the dependence of the lexical diversity rate on the size of the text corpus. The problems of the approximation of the chosen extrapolation method are discussed. Calculations were made on the example of Leo Tolstoy using the logarithmic basis functions for approximation and extrapolation.
Keywords: lexical diversity rate, Zipf's law, extrapolation, lemmatized frequency dictionary, limit size of the glossary.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Mandelbrot B. B. & Hudson R. L.* (2004) The (mis)Behavior of Markets: A Fractal View of Risk, Ruin, and Reward. New York: Basic Books.

2. *McKee G., Malvern D., & Richards B.* (2000) Measuring Vocabulary Diversity Using Dedicated Software. Literary and Linguistic Computing. № 15(3). P. 323–337. doi: 10.1093/llc/15.3.323.

3. *Papp F.* (1961) Quantitative analysis of the vocabulary structure of some Russian texts. Problems of Linguistics. (6). P. 93–100. (In Russian).

4. *Suprun A. E.* (1979) To the quantitative assessment of the lexical richness of the text. Philological Sciences. 1. (In Russian).

5. *Suprun A. E.* (1995) Repeat in the lexical structure of the text. Language — a system. Language is text. Language is an ability. On the occasion of the 60th anniversary of Corresponding Member of the RAS Yu. N. Karaulov. Moscow. P. 133–141. (In Russian)

6. *Suprun A. E.* (1996) Lectures on the theory of speech activity. A manual for students philologist fak. university. Minsk. Belorusskij fond Sorosa. P. 287. (In Russian)

7. https://en.wikipedia.org/wiki/Lexical_diversity.

8. *Templin M.* (1957) Certain language skills in children. – Minneapolis: University of Minnesota Press. doi: 10.1044/jshd.1703.280.

9. *Johnson W.* (1944) Language and speech hygiene, an application of general semantics, Ann Arbor.

10. *Chotlos J. W.* (1944) Studies in language-behavior, IV — A statistical and comparative analysis of individual written language samples. P. 75–111. («Psychologie monographie»). doi: 10.1037/h0093511.

11. *Miller G. A.* (1951) Language and communication. New York.

12. *Suprun A. E.* (2001) Povtor v leksicheskoi strukture teksta [Repetition in the lexical structure of the text]. In: Research on text linguistics: a collection of articles. Minsk. P. 108–117.

13. *Kretov A. A., Matycina L. N.* (1999) Morpheme-morphological dictionary of the language A. S. Pushkin: approx. 23,000 words. Voronezh. Central Chernozemnoe book publishing house. 208 p. (In Russian).

14. *Avdeeva A. A., Borodin V. V., Bykova N. Ya., Kozokina S. M., Gordeeva N. A., Makarova L. A.,*

✉ Kretov Alexey A.
e-mail: kretov@rgph.vsu.ru

SHajkevich A. Ya. (1981) Frequency Dictionary of the Language M. Yu. Lermontov. Lermontov Encyclopedia / USSR Academy of Sciences. Inst. Rus. lit. (Pushkin. House); Scientific Ed. Council of the publishing house "Sov. Encycl". Moscow. Sov. Encikl. P. 717–774. (In Russian).

15. *Tuldava Yu.* (1987) Problems and Methods of the Quantitative Systemic Study of Vocabulary, Tallin. Valgus. (In Russian)

Kretov Alexey A. — doctor of philology, professor of the department of theoretical and applied linguistics of Voronezh State University.
Email: kretov@rgph.vsu.ru
ORCID iD: https://orcid.org/0000-0002-1474-3177

Lomets Maria V. — student of the department of theoretical and applied linguistics, faculty of Romano-Germanic philology, Voronezh State University.
Email: marusya.lomets@gmail.com
ORCID iD: https://orcid.org/0000-0003-0885-4740

Polovinkin Igor P. — doctor of physical and mathematical sciences, professor of the department of mathematical and applied analysis, docent of the department of theoretical and applied linguistics of Voronezh State University.
E-mail: polovinkin@yandex.ru
ORCID iD: https://orcid.org/0000-0002-4308-8909