Закономерности роста словаря в 1-11 томах «Полного собрания сочинений» В. И. Ленина

В. А. Оганисян, email: lera.oganisyan@mail.ru
А. А. Кретов, email: kretov@rgph.vsu.ru

Воронежский государственный университет

Аннотация. *В данной статье исследуются закономерности роста словаря В. И. Ленина. Цель исследования – определить предельный объём активной лексики В. И. Ленина на основе корпуса, состоящего из 1-11 томов «Полного собрания сочинений» (1893-1905).*

Ключевые слова: *коэффициент лексического разнообразия, лемматизированный частотный словарь, предельный размер словаря, язык В. И. Ленина.*

Введение

Использование новых информационных технологий в лингвистических исследованиях является знамением нашего времени. Они особенно эффективны в обработке больших массивов данных, таких как авторский корпус текстов, в котором представлены работы только одного автора. Количественный анализ[1] текстов конкретного человека даёт возможность охарактеризовать его *идиолект* – индивидуальный вариант общенародного языка, находящий отражение во всем множестве текстов, порождённых индивидом. Чем больше корпус текстов, представляющих данный идиолект, тем полнее и богаче информация о зависимости между размером корпуса и богатством словаря индивида. В связи с этим, «Полное собрание сочинений» В. И. Ленина (т. 1-55, 5 изд., М., 1960-65; далее ПСС-5) даёт едва ли не уникальную возможность исследовать эту закономерность.

В работе использован метод моделирования и прогнозирования роста словаря индивида, предложенный в статьях А. А. Кретова, И. П. Половинкина и их соавторов [1-6].

Цель исследования – оценить влияние размера корпуса текстов на результаты прогнозирования роста словаря индивида – на материале ПСС-5 В. И. Ленина. При этом исследуется только русский словарь Ленина – слова, написанные кириллицей.

[1] Авторы приносят благодарность Игорю Петровичу Половинкину за консультации по математическим вопросам исследования.

Исследование проводится в пять шагов, на каждом из которых осуществляется прирост корпуса на примерно равную часть, сопровождаемый определением максимального размера активного словаря В. И. Ленина. На первом шаге исследуется корпус из 1-11 томов, на втором 1-22, на третьем 1-33, на четвёртом 1-44, на пятом 1-55. Такой подход позволит проследить влияние размера корпуса текстов на максимальный объём активной лексики В. И. Ленина. Данный доклад содержит описание первого шага исследования

История вопроса

Ранее реальный размер активного словаря Ленина по ППС-5 был определён в 37500 слов по алфавитно-частотному словоуказателю к «Полному собранию сочинений» В. И. Ленина [7]. Составление «Словаря языка В.И. Ленина» началось в 1972 году в Институте русского языка АН СССР. К моменту распада СССР по материалам картотеки словаря (около 2,3 млн. карточек-цитат) были опубликованы различные статьи [8, 9, 10], 3 докторские диссертации, более 10 кандидатских, «Фразеологический словарь языка В. И. Ленина» [11], подготовлен, но не издан первый том «Словаря языка В. И. Ленина» (буквы А-В, более 6 тыс. словарных статей, рукопись) из предполагавшихся пяти [12].

Попыток определить размер *предельного* (при котором прирост словаря пренебрежимо мал) активного словаря Ленина, насколько нам известно, не предпринималось.

1. Реальный и предельный размер словаря

Для определения реального и предельного словаря В. И. Ленина на первом шаге исследования взяты первые одиннадцать томов «Полного собрания сочинений» общей длиной 962.436 словоупотреблений.

В ходе исследования 11 томов было выполнено 10 последовательных шагов, позволяющих осуществить наращивание корпуса посредством конкатенации текста для получения его длины (в словоупотреблениях) и определения прироста реального размера словаря (в *леммах* – словарных формах, представляющих всю парадигму слова в словаре).

Лемматизация (приведение текстовых слов к лемме) осуществлялась с помощью морфологического анализатора русского языка MyStem, разработанного Ильей Сегаловичем в компании "Яндекс" и размещённого в свободном доступе. Благодаря этому появилась возможность превратить частотный словарь словоформ в частотный словарь лемм и получить необходимые для вычисления данные. При исследовании прироста новых слов по мере увеличения

корпуса были использованы возможности электронных таблиц MS-Excel, в которых производились расчёты и строились графики.

Важной характеристикой, позволяющей отследить прирост новых слов по мере наращивания метакниги (корпуса текстов), является "коэффициент лексического разнообразия" (КЛР, англ. lexical diversity, LD). Он представляет собой количественную характеристику текста, отражающую степень богатства словаря при построении текста заданной длины. В основе данного показателя лежит отношение количества лемм к количеству их употреблений в тексте.

Для получения реального и предельного размера словаря В.И. Ленина необходимы суммарные (кумулятивные) значения размера метакниги и словаря. Эти данные приведены в табл. 1.

Таблица. 1. Прирост новых слов и покрываемого ими текста

Том	Год	Длина текста ΔM	Кол-во слов ΔN	КЛР	Кум. длина текстов M	Кум. размер словаря N	КумКЛР Y_{TTR}
T01	1893-1894	108604	7092	0,0653	108604	7090	0,0653
T02	1895-1897	104156	7435	0,0714	212760	10124	0,0476
T03	1896-1900	104507	6499	0,0622	317267	12057	0,0380
T04	1898 - 1901 апр.	96831	7177	0,0741	414098	13716	0,0331
T05	1901 май - 1901 дек.	78779	7392	0,0938	492877	15307	0,0311
T06	1902 янв. - 1902 авг.	87138	7031	0,0807	580015	16455	0,0284
T07	1902 сент. -1903	67412	6358	0,0943	647427	17260	0,0267

	сент.						
T08	1903 сент. - 1904 сент.	91709	6419	0,0700	739136	18171	0,0246
T09	1904 июль - 1905 март	73290	6651	0,0907	812426	18996	0,0234
T10	1905 март - 1905 июнь	66348	6001	0,0904	878774	19560	0,0223
T11	1905 июль -1905 окт.	83662	6516	0,0779	962436	20191	0,0210

В табл. 1: N – текущее значение размера словаря; ΔN – приращение словаря, то есть количество новых уникальных слов при добавлении новых текстов в корпус; M – текущее значение размера корпуса; ΔM – приращение размера корпуса, то есть количество словоупотреблений в добавляемом в корпус тексте; YTTR – текущее значение КЛР.

2. Результаты исследования

Нашей задачей является определение *предельного* (при котором прирост словаря пренебрежимо мал) объёма Достижение предельного размера словаря активной лексики В. И. Ленина. достигается при приращении словаря близком к нулю [3], когда КЛР в процессе увеличения корпуса стремится к нулю, но не принимает нулевого значения, поскольку размер словаря – величина положительная.

Для исследования используется линия тренда – логарифмическая зависимость. (рис. 1) Логарифмические и постоянные функции выбираются в качестве базисных, а функцию зависимости КЛР от объема текста ищутся в виде линейной комбинации базисных функций. При этом значение суммарной длины текста можно считать соответствующим предельному размеру словаря. Полученную функцию тренда приравниваем нулю и решаем полученное нами уравнение.

$$-0,02 \ln M + 0,2893 = 0$$

Пусть M_0– корень этого уравнения.

$$\ln M_0 = 14,465$$

$$M_0 = e^{14,465} = 1914563$$

Таким образом, размер текста корпуса, при котором достигается предельный размер словаря В. И. Ленина, составляет 1 914 563 слова. Это приближенное значение.

Для того, чтобы найти предельный объем активной лексики, мы используем тот же метод (рис. 2). Выбираем в качестве линии тренда логарифмическую зависимость и приравняем полученное уравнение к нулю.

$$-0,041\ln N + 0,4241 = 0$$

Пусть N_0 – корень данного уравнения.

$$\ln N_0 = 10,3439$$

$$N_0 = e^{10,3439} = 31067$$

Итак, оценка предельного словаря В. И. Ленина «прогнозно» составляет 31 067 слов.

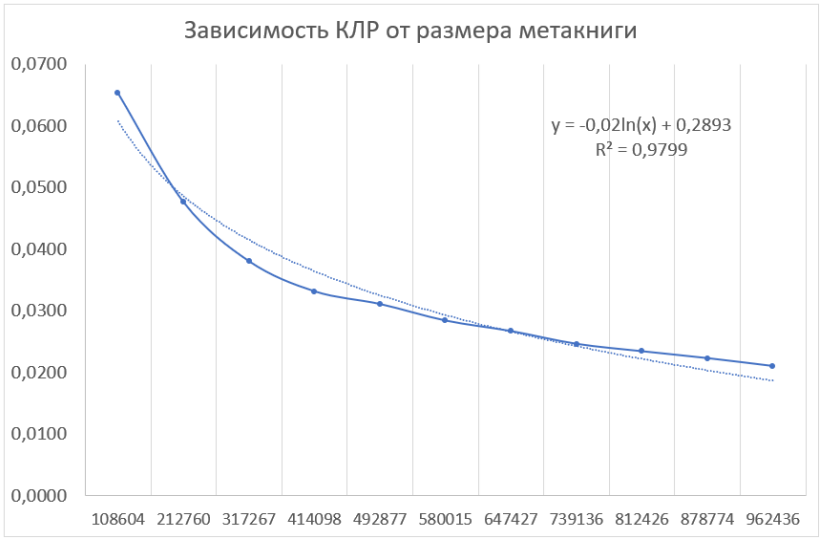

Рис. 1. Динамика КЛР в корпусе работ В. И. Ленина при присоединении к корпусу новых томов

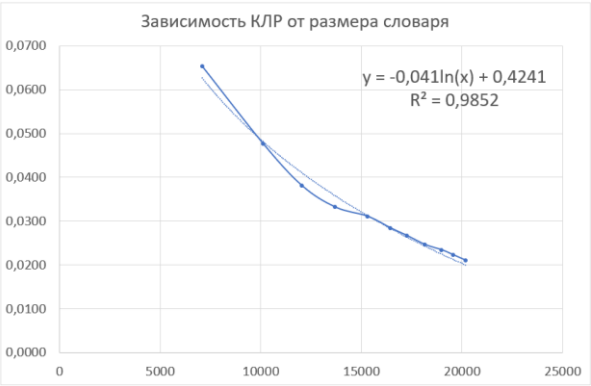

Рис. 2. Зависимость КЛР от размера словаря В. И. Ленина

Необходимо провести проверку полученных результатов, чтобы удостовериться в их правильности. Для проверки прогноза используется вариант закона Ципфа.

$$N = AM^{\beta},$$

где N – размер словаря, M – размер текста, $0 < \beta < 1$. По данным табл. устанавливается степенная зависимость вида:

$$N = 18.012M^{0,482}$$

В эту формулу подставляем $M_0 \approx e^{14,465} \approx 1914563$ и получаем 28 795 слов – новую оценку предельного размера словаря В. И. Ленина. Данное значение отличается от того, которое мы получили раннее при использовании КЛР.

Рис. 3. Зависимость размера словаря от размера корпуса текстов В. И. Ленина

Абсолютная погрешность равна разности полученных значений предельного словаря.

$$31067 - 28795 = 2272$$

Посмотрим, чему равна относительная погрешность.

$$\frac{31067 - 28795}{31067} \times 100\% \approx 7,3\%$$

Данный результат вполне приемлем.

Заключение

В ходе исследования определён предельный размер словаря В. И. Ленина на основе корпуса текстов 1-11 томов «Полного собрания сочинений» (1893-1905). Он находится в интервале 28.795-31.067, т.е. 29.931±1.136 слов, округлённо – 30.000±1.000 слов. Длина текста, при которой достигается предельный размер словаря, равна 1 914 563 словоупотреблениям.

Литература

1. Кретов, А. А., Ломец, М. В., Половинкин, И. П. Возможный алгоритм вычисления предельного размера словаря писателя./ Вестник ВГУ. Серия: Системный анализ и информационные технологии, 2021, 133-145.

2. Кретов А.А., Половинкина М.В., Половинкин И.П., Ломец М.В. О моделировании изменений языка. / Современные методы теории функций и смежные проблемы, 2021, 173-174.

3. Кретов А.А., Половинкина М.В., Половинкин И.П., Ломец М.В. О некоторых количественных характеристиках фрактальности в языке. / Информатика: проблемы, методы, технологии, 2020, 1627-1634.

4. Кретов А.А., Половинкин И.П., Ломец М.В. Абсолютное и относительное «богатство словаря» на примере произведений Л.Н. Толстого. / Математика и междисциплинарные исследования – 2020, 200-203.

5. Кретов А.А. и др. Лексическое богатство словаря В.В. Набокова / А.А. Кретов, И.П. Половинкин, Н.А. Касымова, М.В. Половинкина // Электронный научный журнал «Квантитативная филология», Смоленск: Смоленский Центр квантитативной филологии, 2021, № 1, С. 39-48. DOI 10.35785/0000-0000-2021-1-39-48.

6. Кретов А.А. и др. О предельном размере словаря и фрактальной размерности метакниги М.Е. Салтыкова-Щедрина / А.А. Кретов, М.В. Половинкина, И.П. Половинкин, Н.А. Касымова // Информатика: проблемы, методы, технологии: сборник материалов XXII международной научно-методической конференции / под редакцией

Д.Н. Борисова; Воронеж, Воронежский государственный университет, 10-12 февраля 2022 г. – Воронеж : «ВЭЛБОРН», 2022. – С. 1146-1154.

7. Словарь языка В. И. Ленина : [В 2 т.] / АН СССР, Ин-т рус. яз. ; Отв. ред. П. Н. Денисов. - М. : Наука, 1987. - 22 см.[Ч. 1]: А - Одолжение. - Москва : Наука. - 592 с.[Ч. 1]: А - Одолжение. - Москва : Наука. - 592 с. [Ч. 2]: Одряхлеть - Ящичек. - Москва : Наука. - С. 595-1189,[2].

8. Филин Ф. П. О словаре языка В.И. Ленина. // Вопросы языкознания, 1974, № 6, С. 3-10.

9. Даниленко В.П. и др. Словарь языка В.И. Ленина / В.П. Даниленко, В.Н. Хохлачева // Русская речь, 1975. № 2, 3-10.

10. Денисов П.Н. Богатство языка В.И. Ленина / П.Н. Денисов // Русская речь, 1983. № 2, 3-10.

11. Байрамова Л.К. Фразеологический словарь языка В.И. Ленина / Л.К. Байрамова, П.Н. Денисов – Казань: Казанский университет [КазГУ], 1991. - 349 с.

12. Петерс, Я. Как создавался «Словарь языка В.И. Ленина» // Regla, 2007, № 39; см. [Как создавался «Словарь языка В.И. Ленина» | Статьи | Главная | Научно-культурологический журнал (relga.ru)].