Воронежский государственный университет

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

А. А. Кретов, В. А. Оганисян

# 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 .

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

# 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*.

# Коэффициент лексического разнообразия (КЛР)

«Коэффициент лексического разнообразия" (КЛР, англ. lexical diversity, LD) – количественная характеристика текста, отражающая степень богатства словаря при построении текста заданной длины.

В основе данного показателя лежит отношение количества лемм к количеству их употреблений в тексте.

# Таблица « Прирост новых слов и покрываемого ими текста»

| Том | Год | Длина текста ΔM | Колич. слов ΔN | КЛР | Кумулят. длина текстов M | Кумулят. 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 апрель | 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 сент. | 67412 | 6358 | 0,0943 | 647427 | 17260 | 0,0267 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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; $\Delta N$ – приращение словаря, то есть количество новых уникальных слов при добавлении новых текстов в корпус; $M$ – текущее значение размера корпуса; $\Delta M$ – приращение размера корпуса, то есть количество словоупотреблений в добавляемом в корпус тексте; $Y_{TTR}$ – текущее значение КЛР.

# Динамика КЛР в нарастающем корпусе текстов В.И. Ленина

Для исследования используется линия тренда – логарифмическая зависимость. Полученную функцию тренда приравниваем нулю и решаем полученное нами уравнение.

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

Функция достигает нулевого значения в точке
Размер метакниги составляет 1 914 563 слова.

# Динамика КЛР в нарастающем корпусе текстов В.И. Ленина

Для того, чтобы найти предельный объем активной лексики, мы используем тот же метод.

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

Функция достигает нулевого значения в точке

Оценка предельного словаря В. И. Ленина «прогнозно» составляет 31 067 слов.

# Закон Ципфа

где – размер словаря, – размер текста, .

По данным таблицы устанавливается степенная зависимость вида:

В эту формулу подставляем и получаем 28 795 слов, которые оценивает в качестве предельного размера словаря В. И. Ленина

# Абсолютная и относительная погрешности

Абсолютная погрешность равна разности полученных значений предельного словаря.

Посмотрим, чему равна относительная погрешность.

$$\frac{\qquad\qquad\qquad\qquad}{}$$

Данный результат вполне приемлем.

Спасибо за внимание!