ГАТИН Г. Н.
О ЗАКОНЕ ЦИПФА

УДК 81.32, ВАК 05.13.01, ГРНТИ 20.00.00

О законе Ципфа

Г. Н. Гатин

Ухтинский государственный технический университет, г. Ухта,

Автор производит проверку литературных, программных и случайных текстов на соответствие закону Ципфа. Показывает, что существуют тексты, не отвечающие этому закону. Обращает внимание, что использование в Internet Закона Ципфа для проверки естественности текста бесполезно.

Ключевые слова: *закон Ципфа, распределение, текст, язык, частота, слово.*

About Zipf's Law

G. N. Gatin

Ukhta state technical university, Ukhta

The author checks literary, programmatic and random texts for compliance with Zipf's law. Shows that there are texts that do not meet this law. It is noteworthy that the use of the Zipf Law on the Internet to verify the naturalness of the text is useless.

Keywords: *Zipf's law, text, distribution, language, frequency, word.*

Введение

Обычно, известие о том, что из всего словарного запаса языка мы используем где-то ~2000 слов, первый раз звучит достаточно удивительно. Выполняя краткий анализ этого утверждения, а именно, удалив из этих ~2000 слов предлоги, союзы, частицы или одно-двухбуквенные (а иногда более) слова, мы понимаем, что наш словарный запас грозит сравняться со словарём знаменитой Элочки - людоедки – это уже шокирует.

Тогда-то мы и ищем обоснование всего этого. Оказывается это закон Ципфа.

Например, в книге В. А. Лапшина "Лекции по математической лингвистике" [1]. В. А. Лапшин приводит его в форме:

$$P_n \approx \frac{1}{n^a} \quad (1)$$

В. А. Лапшин приводит этот закон для такого объекта как "мешок слов русского языка". Другими словами, закону Ципфа подчиняется весь язык, весь словарный запас языка. Но в печатной литературе вы не найдёте обоснования или доказательства закона Ципфа – это просто эмпирическая закономерность.

Теоретический анализ

Как профессиональный программист я привык все утверждения литературы по программированию не принимать на веру, а проверять. Не факт, что скопированная программа или методика "пойдут" на вашей машине или в вашем окружении.

Формулировка же закона Ципфа предельно проста, что подталкивает вас к решению проверить закон. К тому же, до 2005 года (см. ниже) это была недоказанная наблюдаемая статистическая закономерность.

Но, конечно, прежде всего выполняем обзор.

Открыл закон, оказывается, французский стенографист Жан Батист Эсту (фр. Jean-Daptiste Estoup) в 1908 году, что говорит в пользу применения закона ко всему словарному запасу языка. Википедия [2] приводит следующий график частот слов:

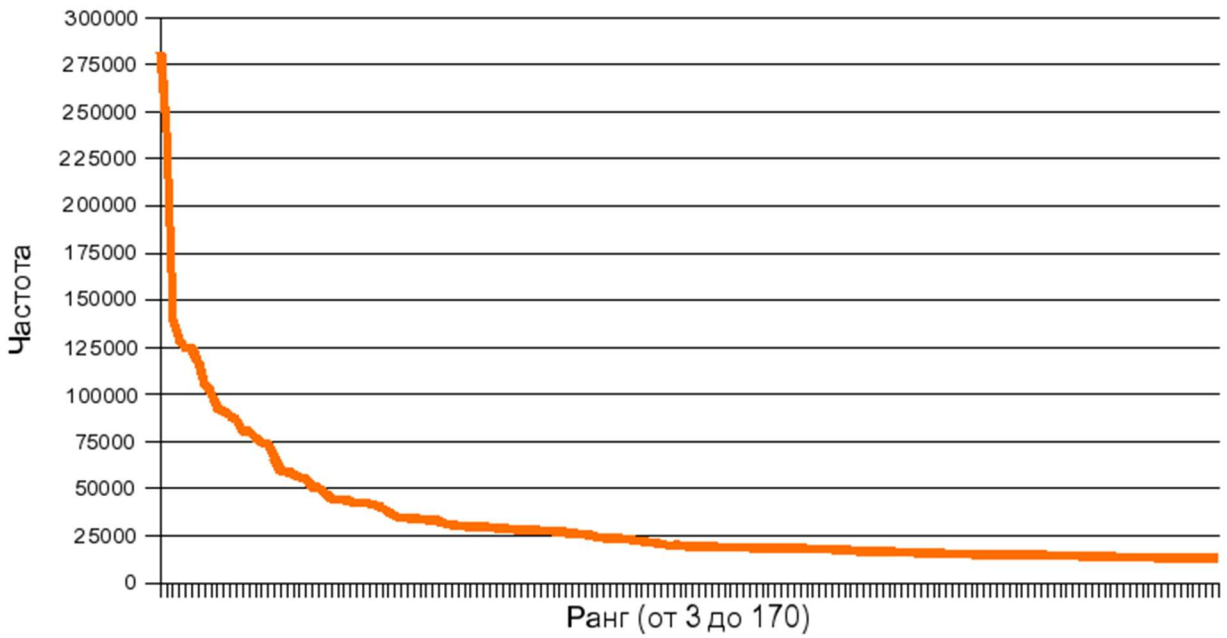

Рисунок 1. Закон Ципфа. График для для частот слов из статей русской Википедии с рангами от 3 до 170. [200]

Форма закона здесь немного иная, но, впрочем, та же:

$$Pi = P1 / i \quad (2)$$

Интересно, что этому закону подчиняется население городов: "город с самым большим населением в любой стране в два раза больше, чем следующий по размеру город, и так далее" [2]; доходы населения, ну и, конечно же, тексты.

Для городов в 1999 году экономист Ксавье Габэ показал, что если города будут расти случайным образом с одинаковым среднеквадратичным отклонением, то в пределе распределение будет сходиться к закону Ципфа. [2]

Но, самое главное, "Объяснение закона Ципфа, основанное на корреляционных свойствах аддитивных марковских цепей (со ступенчатой функцией памяти) было дано в 2005 году" [2].

Это утверждение, выпускает дух из вашего желания как-то проверить закон Ципфа. Но дальнейшая работа в Internete снова вас вдохновляет.

Оказывается, что часть сайтов проверяет тексты на естественность, считая, что чем более текст соответствует закону Ципфа, тем он естественней. С другой стороны, существует статья американского специалиста по биоинформатике [3], в которой он утверждает, что закону Ципфа подчиняются чисто случайные тексты. И если это утверждение верно, то вышеуказанные проверки бесполезны.

Безусловно, проверить живую речь на закон Ципфа достаточно трудоёмкое занятие, хотя и интересное. Ну, например, сколько слов мы говорим в сутки? В сутках 1620 мин. Исключаем 8 часов сна и один час на гигиену, остаётся 1080 мин. Если мы произносим, хотя бы десять слов в минуту, то за сутки обеспечено 10 800 слов, что позволяет выполнять проверку закона уже после первых суток. Я бы снизил число слов в четыре раза, что даёт в общей сложности время набора слов примерно от четырёх до пяти - семи дней. Основная проблема набрать статистически значимую выборку. Принципиально всё это считается: для массива из N слов (≈250 000 слов русского языка), выборка объёмом n слов с вероятностью P даст отклонение не боле k слов.

В общем-то, мы и так знаем, что в обыденной своей речи используем где-то 2000 слов, но подчиняется ли эта выборка закону Ципфа? Каких-либо данных на ту тему я не обнаружил. С другой стороны, кажется, что если бы мы чаще всего использовали некое значимое достаточно длинное слово, то это слово мы бы знали. Но таких слов никто не вспомнит!

Интуитивно понятно, что наиболее часто мы используем вспомогательные одно- и двухбуквенные слова (ну если быть педантичным, это требует проверки).

Другими словами, если вы собираетесь заниматься модной сегодня обработкой текста, то у вас должны быть все эти данные. Это веский аргумент в пользу проверки закона Ципфа.

Кажется, чего автор кипятится? Возьмите частотный словарь русского языка. Какие проблемы? А проблем-то достаточно. Вы не найдёте общедоступного словаря русского языка, который можно подключить к вашей программе и использовать его как источник слов, не данных даже, а просто слов. Автор имеет опыт обработки словаря из Internet. Некоторые слова набраны латиницей, ошибки в написании, пропущенные пробелы и т.п.

Вот почему мы проанализировали некоторое количество текстов.

Опять-таки, прежде чем заниматься, некоторой рутинной работой, необходимо обдумать: «А что, собственно, мы собираемся получить?»

Начнём с вопроса: Существуют ли тексты, не подчиняющиеся закону Ципфа? Закон сформулирован для языка, а значит, таких текстов не существует.

Тем не менее, такие тексты есть! В любом орфографическом словаре распределение слов равномерное. С толковыми и двуязычными словарями сложнее. Распределение слов, скорее всего, будет стремиться в закону Ципфа, но как велико отклонение?

Интуитивно ясно, что закону Ципфа не подчиняются: Простые словари, Каталоги, Сметы, Нормативы.

Со случайными текстами сложнее: всё сильно зависит от алгоритма генерации текста, результаты исследований представлены ниже в таблицах.

Экспериментальная часть

Для проверки были выбраны: две статьи автора, как короткие тексты, несколько литературных произведений, библия и толковый словарь русского языка (таблица 1). Кроме того, были добавлены авторские исходные модули на языке "C++", как смесь искусственного языка и имён, задаваемых программистом.

Таблица 1. Коэффициенты регрессии для выбранных текстов

	Всего слов	b_1	b_0	угол наклона	R^2	F
Статья автора "К критике компетенций"	2617	-0.58376	3.9258	120.275	0.890809	8.15825
Статья автора "Философия языка 'C' "	4117	-0.63439	4.4943	122.391	0.902867	9.2951
Сборник рассказов Габриэль Гарсия Маркес	81930	-0.8110	7.6787	129.043	0.93384	14.1162
Русские народные сказки	117170	-1.11948	10.593	138.226	0.949559	18.8253
М. А. Булгаков Мастер и Маргарита	118154	-0.85071	8.2974	130.388	0.948383	18.3734
Сказки 1000 и одна ночь	248974	-1.07892	10.653	137.174	0.974394	38.0526
Библия	714243	-1.1572	12.292	139.168	0,978777	46.1183
Толковый словарь русского языка (Ожегов С.И. Шведова Н.Ю.)	1140220	-0.9206	10.651	132.633	0.96212	25.403
Программы на языке C++	9434	-1.1225	7.9024	138.305	0.97085	33.3034

В таблице представлены прямые регресии, подбираемые методом наименьших квадратов, где:

- ось ординат – логарифмы частот слов;
- ось абсцисс – ранги (номеров слов, отсортированных по убыванию) – n,

уравнение подбиралось в виде:

$$chastota = b_1 * n + b_0;$$

- R^2 – (квадрат) множественный коэффициент корреляции показывает процент отклонения от среднего (Y-ср) и объясняется уравнением регрессии;
- при анализе - критерия Фишера F выявлено, что F > 1, поскольку число степеней свободы чрезвычайно велико, то является достаточным.

Написанное автором ПО, считывало текст, подсчитывало число слов и вычисляло частоту встречаемости слов. Затем методом наименьших квадратов определялись коэффициенты b_1 и b_0 регрессионной прямой зависимости логарифма частоты от ранга слова. Адекватность формулы проверялась по множественному коэффициенту корреляции (R^2) [3] и критерию Фишера (F). Результаты приведены в таблице 1. Результаты однозначно подтверждают истинность закона Ципфа для литературных текстов и алгоритмического языка "C++".

Была выполнена проверка на закон Ципфа случайных текстов. Однако, что считать "чисто случайным текстом"? Можно генерировать "случайный" текст, так чтобы сохранять частоту встречаемости букв русского языка, а слова случайные последовательности этих букв. Можно выбирать случайным образом слова из словаря русского языка (в этом случае, скорее всего, гипотеза о соответствии текста закону Ципфа не подтвердится, поскольку все слова будут распределены равномерно).

Автор ориентировался на "чисто случайный текст". Сначала генерировалась длина слова, затем буквы, входящие в слово; буквы и длины слов распределены равномерно. Результаты приведены в таблицах 2 – 5. Видим, что случайный текст закону Ципфа не соответствует. Таким образом, сообщение Вэньтянь Ли не подтверждается, с оговоркой – "на нашей модели случайного текста".

Таблица 2. Коэффициенты регрессии для случайных текстов

Всего найдено слов	Коэффициенты прямой регрессии	Угол наклона	Степеней свободы для критерия Фишера	Коэффициент корреляции	Критерий Фишера
40000	$b_1 = -0.0646027$ $b_0 = 0.632828$	93.6963	39998	$R^2 = 0.20381$	$F = 0.255981$

Таблица 3. Частоты встречаемости слов в зависимости от длины

длина слова	1	2	3	4	5	6	7	8	9	10
число слов	1382	1404	1445	1353	1414	1397	1398	1379	1407	1432
длина слова	11	12	13	14	15	16	17	18	19	20
число слов	1428	1373	1460	1433	1353	1358	1359	1475	1401	1389
длина слова	21	22	23	24	25	26	27	28	29	
число слов	1398	1442	1357	1370	1410	1406	1428	1423	1344	

Таблица 4. Коэффициенты прямой регрессии для случайных текстов

Коэффициенты прямой регрессии	Угол наклона	Степеней свободы для критерия Фишера	Коэффициент корреляции	Критерий Фишера
$b_1 = -3.78197$ $b_0 = 1476.12$	165.189	28	$R^2 = -0.0227175$	$F = 0.000520331$

Таблица 5. Массив слов, сортированный по частотам

№	Количество
1	63
2	60
3	53
4	52
5	51
6	45

Заключение

Очевидно, что если в случайном тексте длины слов распределены равномерно, то никакого соответствия текста любому другому распределению не будет. С другой стороны, можно генерировать длины слов под заранее заданное распределение, что может выполнить любой робот. А тогда, какую проверку естественности происхождения текста можно выполнить по распределению?

На сегодня никто не доказал, что все языки подчиняются закону Ципфа, хотя возникает идея поверки этой гипотезы для всех языков представленных в Internete.

Опять-таки, а соответствует ли закону Ципфа текст написанный иероглифами? Я, например, очень в этом сомневаюсь; к тому же есть ещё и слоговое письмо.

В завершении заметим, что закон Ципфа не что иное, как распределение Пуассона при $\lambda = 1$.

Список использованных источников и литературы

1. Лапшин В. А. Лекции по математической лингвистике. – М. : Научный мир, 2010.

2. Википедия [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org (Дата обращения: 07.02.2020).

3. Вэньтянь Ли, Закон Ципфа работает и для случайных текстов [Электронный ресурс]. – Режим доступа: https://santafe.edu (Дата обращения: 07.02.2020)

4. Норман Р. Дрейпер, Гарри Смит. Прикладной регрессионный анализ (третье издание). – М. : Диалектика, 2007.

List of references

1. Lapshin V. A. Lectures on mathematical linguistics, M, Scientific world, 2010.

2. Wikipedia, https://ru.wikipedia.org, accessed February 07, 2020.

3. Wentian Li, Zipf's Law also works for random texts, https://santafe.edu, accessed February 07, 2020.

4. Norman R. Draper, Harry Smith, Applied Regression Analysis (Third Edition). - M.: Dialectics, 2007.