Zipf's Law for LiveJournal

Nikita N. Trifonov

Kazan Federal University, Kazan, Russia
nikita-trif@yandex.ru

Abstract. The paper provides an overview of research of frequency of language units on the material of the LiveJournal corpus. The corpus includes texts on Russian language from 2002 to 2014 year, totaling more than 5 million words of articles written by 2 thousand authors. Research was held in the following main directions, represented in the present work: estimation of coefficients for the Zipf's law for different authors, estimation of coefficients for the Zipf's law for the total number of words in all the analyzed articles.

Keywords: Zipf's law, LiveJournal corpus, frequency of words, rank distribution.

1 Introduction

Perhaps the most famous statistical distribution in linguistics is Zipf's law: in any large enough text, the frequency ranks (starting from the highest) of wordforms or lemmas are inversely proportional to the corresponding frequencies [1]:

$$f(r) * r = c, \tag{1}$$

where $f(r)$ is the frequency of the unit (wordform or lemma) having the rank r and c is a constant. With Mandelbrots improvements to Zipf's law, the formula (1) has next form [2]:

$$f(r) = \frac{c}{r^{\gamma}}, \tag{2}$$

where γ is the exponent coefficient (near to 1). Zipf's law is most easily observed by plotting the data on a $log - log$ graph, with the axes being $log(rank\,order)$ and $log(frequency)$. After taking the logarithm of the formula (2) :

$$ln(f(r)) = C - \gamma ln(r), \tag{3}$$

The LiveJournal source chosen to collect the corpus because it makes it possible to explore articles as a whole and separately for each author.

The LiveJournal[1] (LJ) is a social network owned by SUP Media where Internet users can keep a blog, journal or diary, and is also the name of the free and open source server software which runs the LiveJournal website and online community.

[1] http://www.livejournal.com/

In order to collect corpus of LiveJournal, created a program that gets the text of articles written by one author, saves the text in a database and goes over to another author for further information gathering.

2 Experimental Results

The graph plotted (Fig. 1) using for Zipf's law the points: $x_r = log\, r$, $y_r = log\, f(r)$ where $r = 1 \dots n$, and n is the number of different units (wordforms or lemmas). The Ordinary Least Squares used to approximate such a graph by a straight line $y = ax + b$, where a and b correspond to γ and C for Zipf's law (the formula (3)) .

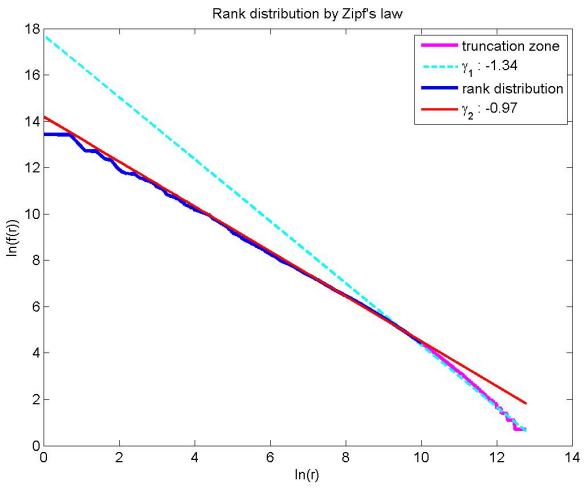

Fig. 1. The Zipf's law for the total number of words in all the analyzed articles.

The graph (Fig. 1) can be divided into three different parts. The first part is a nuclear zone consisting of the most frequently used words in the Russian language - prepositions, pronouns, etc. The central part of the graph very important for exploring. It is most accurately described by Zipf's law. The last part, called "zone of truncation " consists of words which do not carry meaning, rarely used terms and grammatical errors.

As the graph shows, the zone of truncation affects the result of the approximation, and γ coefficient in approximating line is differs from the expected. However, if we do not consider the zone of truncation, approximation line almost merges with the graph of frequency distribution, and γ coefficient satisfies the improvements of Mandelbrot for Zipf's law.

Ten authors, who written the highest number of letters in articles, were selected for further researches.

Table 1. List of 10 authors with the highest total number of words in articles

Author page on LiveJournal	The total number of words used by the author
http://eto_fake.livejournal.com/	584061
http://mzadornov.livejournal.com/	583071
http://cuamckuykot.livejournal.com/	347962
http://aillarionov.livejournal.com/	302165
http://mgsupgs.livejournal.com/	260479
http://matveychev_oleg.livejournal.com/	243579
http://steissd.livejournal.com/	240800
http://kak_eto_sdelano.livejournal.com/	234767
http://annatubten.livejournal.com/	225701
http://adamashek.livejournal.com/	214743

For each author from the list given in Table 1 made separate research. These researches have shown that all of graphs correspond to the Zipf's law. An example of this graph you can see in Figure 2.

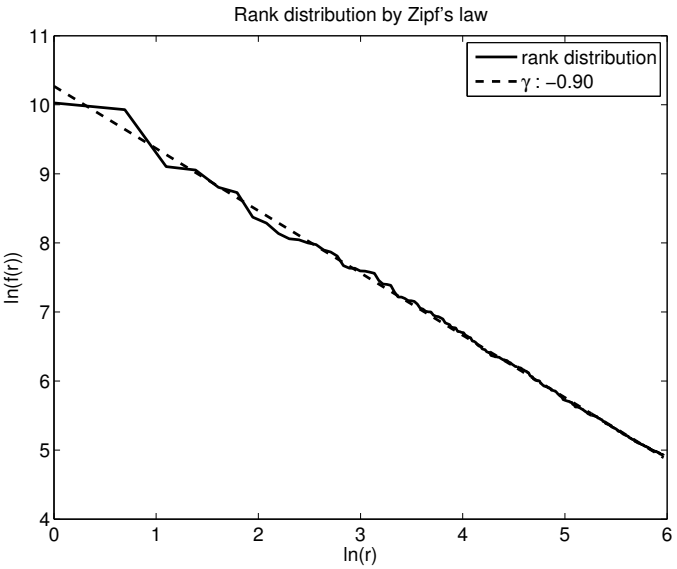

Fig. 2. Rank distribution by Zipf's law for http://eto_fake.livejournal.com/

For comparison, made the research of rank distribution of word frequencies of Zipfs law based on 4 volumes of books Leo Tolstoy's "War and Peace." (Fig. 3)

There are differences between the list of the most frequently encountered words of Leo Tolstoy's works and the list of the most frequently encountered words of contemporary authors represented on LiveJournal. These differences are

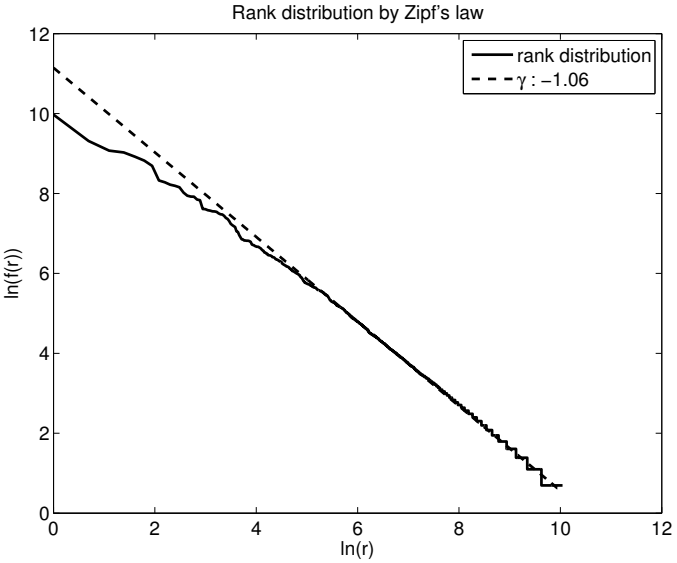

Fig. 3. Rank distribution by Zipf's law for L. N. Tolstoi

related with the difference between the vocabulary of Leo Tolstoy and modern vocabulary, as well as skills in writing texts.

3 Conclusion

Exponential coefficients of Zipf's law depend on text volume, genre of the text and author's style. The zone of truncation encountered in the research of large texts or texts written by different authors. Explanation of this phenomenon needs more investigation.

Acknowledgements. I would like to thank Valery Dmitrievich Solovyev, Eduard Yulyevich Lerner, and Vladimir Vladimirovich Bochkarev for helpful discussions.

References

1. *Zipf, G. K.* Human behavior and the principle of least effort. Cambridge, MA, Addison-Wesley, 1949, p. 36.
2. *Mandelbrot, B.* An informational theory of the statistical structure of languages, Communication Theory, ed. W. Jackson, Betterworth, 1953, pp. 486502
3. *Gelbukh, A., Sidorov, G.* Zipf and Heaps Laws' Coefficients Depend on Language. Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335.

Закон Ципфа для LiveJournal

Никита Н. Трифонов

Казанский (Приволжский) федеральный университет, Казань, Россия
nikita-trif@yandex.ru

Аннотация В статье представлен обзор исследований частоты языковых единиц на материале корпуса LiveJournal. Корпус включает тексты на русском языке, написанные в период с 2002 по 2014 год. Были исследованы статьи 2000 авторов, а так же более 5 000 000 словоформ из этих статей. Исследование было проведено в следующих основных направлениях, представленных в настоящей работе: расчет коэффициентов закона Ципфа по отдельным авторам, расчет коэффициентов закона Ципфа по всем проанализированным статьям без дифференциации по авторам.

Ключевые слова: закон Ципфа, корпус LiveJournal, частота встречаемости слов, ранговое распределение.