Deviations in the Zipf and Heaps laws in natural languages

Vladimir V. Bochkarev, Eduard Yu.Lerner, Anna V. Shevlyakova
Kazan Federal University, Kremlevskaya str.18, Kazan 420018, Russia

E-mail: vladimir.bochkarev@kpfu.ru

Abstract. This paper is devoted to verifying of the empirical Zipf and Heaps laws in natural languages using Google Books Ngram corpus data. The connection between the Zipf and Heaps law which predicts the power dependence of the vocabulary size on the text size is discussed. In fact, the Heaps exponent in this dependence varies with the increasing of the text corpus. To explain it, the obtained results are compared with the probability model of text generation. Quasi-periodic variations with characteristic time periods of 60-100 years were also found.

1. Introduction
Dependence of the vocabulary size (total number of different words) on the increasing text size is indistinct nowadays though great text corpora are available. Understanding of the processes governing the appearance and usage of new words and their dependence on the text size is essential. This problem is widely discussed in different linguistic and non-linguistic contexts. The empirical laws which focus on the distribution of word frequencies are known as Heaps and Zipf laws. Here we study the relation between the two laws analysing the evolution of written language. Statistical analysis is used to test the validity of these laws using an extremely large text corpus Google Books Ngram which provides a great set of written texts over a period of more than four centuries.

2. Connection between the Zipf and Heaps laws
According to the Zipf law, word usage frequency is defined as a power function \( p_r \sim r^{-\beta} \), where \( r \) is a word rank, in other words it is a word number in a list of words ranked according to their frequency [1]. The Zipf law is closely connected with the Heaps law which says that the size of lexicon \( N \) (the number of particular words in a text) in a text or set of texts of size \( L \) is determined by an expression \( N \sim L^{-k} \). Different probability models of text generation (if it’s assumed that the Zipf law is fulfilled) result in a simple relation between the exponents \( \beta \) and \( k \):

\[
\beta = k^{-1}
\]

Initially, the Zipf law was tested on a relatively small corpus of texts where the values of exponent \( \beta \) were close to 1. However, the Heaps law was initially derived while analysing news items. A \( k \) exponent was estimated as close to 0.5.

A creation of a large text corpus Google Books Ngram which contains data on more than 9 languages over a great time interval offered an opportunity to study static dependencies of words usage [2,3,4]. The Zipf law is tested in detail on the texts written in the main European languages in [4]. It was shown that the Zipf law in its classical form isn’t fulfilled but two power sections can be
seen in the frequency distribution plot. The exponent of the 1st section (for frequently used words) is close to 1 and the exponent of the second section (for non-frequent words) is much higher and varies from 1.7 to 2.5 for different languages. The last values correlate better with the power values of the Heaps law and expression (1). The typical word frequency distribution is shown in fig. 1.A.

Let’s consider the following simple model of text generation. Let us have a (finite or infinite) set of possible words, which usage probabilities at a subsequent stage is an a-priori known and equal to $p_i$. How many words will be used on average at least once in a text of $L$ length? To answer this question let us use the method of indicators. Consider the random variable which takes on value 1 if the $i$th word from our lexicon is used at least once and value 0 if it isn’t used at all. The probability that the $i$th word will not be used at least once is $(1 - p_i)^L$, consequently our random variable takes value 1 with the probability $1 - (1 - p_i)^L$ and its mathematical expectation is equal to this number. The mathematical expectation of such random variables is the average number of N words. This yields the following:

$$ N(L) = \sum_i \left(1 - (1 - p_i)^L\right) $$. 

(2)

Figure 1. A. Word frequency distribution, the common base of English words, year 2000. The solid line represents the empirical data, the dashed and dotted lines – the power approximation; B. The dependence of the lexicon size on the text size of the common base. The dots represent the empirical data, the dotted line – the power law approximation; The dashed and solid lines – the calculation according to the models (2) and (3)

3. The comparison of the modelled and empirical data

Fig 1.B shows the dependence of the lexicon size on the text size according to the Google Books Ngram corpus data for the common base of the English language. The number of different word forms used in the given year (the letter case wasn’t differentiated) and the total number of words contained in the text base were calculated. The word forms consisting only of the letters of the Latin alphabet were considered. The approximation of the empirical data using power law dependence is also shown in the figure. The power exponent was calculated using the least square criterion and was 0.5503. As it can be seen, the empirical data is poorly described by the exponential function. It should be noted that the correlation with the power law for the other languages presented in the Google Books Ngram is worse than for the English language.

The modelling of the expected lexicon size was performed using formula (2). It’s necessary to know word-usage frequencies to calculate the expected lexicon size using formula (2). The calculation of frequencies of rare words is relatively problematic, mostly of words used in earlier time periods because the number of texts in Google Books Ngram where they can be found is rather small. Empirical frequencies of words used in the year 2000 were analysed to estimate probabilities $p_i$ because the biggest number of texts is dated back to this year. Thus, we can estimate the frequencies of the greatest number ($3.97 \times 10^5$) of unique words. The counting results according to formula (2) are represented in the figure by the dashed line. It can be seen that the modelled dependence is close to the
empirical one but it lies slightly above it. To enable better match of the modelled and experimental data, the model was improved.

Traditionally, the words are divided into content and functional words. The latter indicate a grammatical relationship and their frequency directly depend on the sentence structure. However, the percentage of functional words can be relatively high. For example, the percentage of functional words in the English language changed from about 0.48 to 0.57 in 1800-2000 (the list of English words was used during the calculation which can be found in [5]). It can be assumed that all the function words will be used in a set of texts covering the wide range of topics and the number of content words will grow up with the increase of the text size. This results in the following modified model:

\[
N(L) = N_{serv} + \sum_{i \in I} \{1 - (1-p_i)^{\zeta L}\},
\]

where \(N_{serv}\) is the number of functional words, \(I\) – the set of content word numbers in a common list, \(\zeta\) – the part of content words in a text. In the modelling process, the parameter \(\zeta\) is defined for each year according to the Google Books Ngram data, after that the expected lexicon size is calculated using formula (3).

It can be seen that the model (3) yields the best approximation of the empirical data. The exponent obtained during the fitting procedure of the modelled curve using the power dependence (on the section from \(10^3\) to \(10^{10}\)) is 0.5674. It is close to the value mentioned above obtained during the fitting of the empirical data. At the same time, a serious divergence of the modelled and empirical curves can be seen. For example, variations in the large values domain can hardly be explained using a simple model.

Consequently, the observed divergence can be due to imperfect model or result from the dynamic processes in a language. The analysis of the data was performed for different time intervals to verify which of the possibilities is realised. The data selected using the 50 year sliding time window was fitted by power dependence. The dependence variations of the Heaps exponent with the time obtained for the English, Russian, German, and French languages are shown in fig. 2A. Two features of the obtained curves should be mentioned. Firstly, the descending trend is typical of all the languages. Secondly, quasi-periodic variations with characteristic time periods of 60-100 years can be seen in the plots.

![Figure 2. A. Change of the Heaps exponent with the time for the four European languages; B. Modelled values of the Heaps exponent at different corpus size. The solid line represents the empirical frequency distribution (English, 2000), the dashed, dot-and-dash and dotted lines represent the Zipf distribution with different exponents](image)

On average the number of texts in the Google Books Ngram base is much higher for later time periods. So the question is: can the observed variations of the Heaps exponent with the time be due to the increase of the average size of the corpus or are they do to some dynamic processes in the language? The modelling results of the expected Heaps exponent at different text size are shown in fig. 2B. The modelling was performed using empirical word-usage frequencies and the modelled frequencies correlated with the Zipf law. The dependence \(N(L)\) is approximated using the power law
\( N \sim L^{-k(L)} \) for each small section, where \( k(L) \) is assumed to be a slowly varying function. The required power exponent \( k(L) \) can be calculated by the following formula

\[
k(L) = -\frac{\partial \ln N}{\partial \ln L} \tag{4}
\]

The Heaps exponents were modelled by (2,4) using empirical word-usage frequencies of the common English base for the year 2000 (the solid line). A descending trend and the Heaps exponent tending to zero at \( L \to \infty \) can be observed in the plot. It’s not surprising because our model has the finite potential lexicon and the dependence \( N(L) \) should tend to saturation. Two steps are seen in the plot. The dashed and dotted lines in the fig. 2B represent the modelling results assumed that the word-usage frequencies correspond with the Zipf law \( p_k = A k^{-\beta} \) (with the exponent \( \beta \) equal to 1.077 and 1.698 correspondingly) with the finite number of \( 3.97 \times 10^6 \) words. The curve tends asymptotically to zero as in the previous case. At that, almost a horizontal section of the local Heaps exponent can be seen in the plot close to the value \( k(L) = 1/\beta \) as it was expected from the above mentioned thoughts about the connection of the exponents of the Zipf and Heaps laws. Thus, it can be expected that two flat sections of the first curve correspond with the two almost power sections in the word frequency plot (fig. 1A).

Comparison of fig. 2A and fig. 2B shows that the Heaps exponent decreases with time due to text size increase. But both model (2) and model (3) can (in contrast to the given in fig.3) give only monotonically decreasing dependencies for the Heaps exponent. Thus, the quasi-periodic variations of the Heaps exponent observed in fig. 2.A are most likely due to dynamic processes in the language.

4. Conclusion

Thus, the Heaps law is fulfilled restrictively for small texts and texts related to a short historical period. At that, the Heaps exponent varies with characteristic interval of 60—100 years which reflects the dynamic processes in a language.

This work was supported by the Russian Foundation for Basic Research (grant №12-06-00404а).

References

[1] Baayen, R.H.: Word Frequency Distributions. Springer, Berlin (2002)
[2] J.-B. Michel, Y.K. Shen, A.P. Aiden, et. al., Quantitative Analysis of Culture Using Millions of Digitized Books, Science 331, 176 (2011).
[3] A.M. Petersen, J.N. Tenenbaum, S. Havlin, H.E. Stanley, M. Perc, Languages cool as they expand: Allometric scaling and the decreasing need for new words, Sci. Rep. 2 (2012) 943.
[4] M. Gerlach and E.G. Altmann, Stochastic Model for the Vocabulary Growth in Natural Languages, Physical Review X, 3, 021006 (2013)
[5] J.M. Hughes, N.J. Foti, D.C. Krakauer, D.N. Rockmore, Quantitative patterns of stylistic influence in the evolution of literature, PNAS, 2012, vol. 109, no. 207682-7686.