Analysis of the probability distribution of time series of word frequencies

Vladimir V. Bochkarev, Anna V. Shevlyakova
Kazan Federal University, Kazan, Russia
E-mail: vbochkarev@mail.ru, anna_ling@mail.ru

Abstract. Knowledge of probability distribution of frequency of words and phrases is important in solving many practical problems such as estimation of semantic similarity between words, detection of semantic changes and others. It is assumed by many researchers that frequencies of words obey the Poisson law. However, there is much evidence that the Poisson distribution describes empirical data unsatisfactorily. The analysis of the probability law in this case is greatly complicated by the fact that series of frequencies in most cases are non-stationary. This paper discusses the distribution law of time series of word frequencies based on the Google Books Ngram corpus data. It is shown that the correlation between the first moments of the frequencies differs from that expected in the assumption of the Poisson distribution. In particular, anomalously high values of frequency dispersion are observed for words with high frequency. To check the significance of deviations from the Poisson law, statistical modeling of frequency series was performed.

1. Introduction
Quantitative methods are widely used in linguistics for solving various problems such as automatic language processing, creation of machine translation systems, creation of electronic dictionaries, analysis of authors’ idiolect, studying semantic changes etc. Quantitative methods, for example, those based on word frequency, allow one to understand how words are distributed and behave in the text.

Often frequency distribution of words is understood as the dependence of frequency of a word on its rank (the number of a word in a list in frequency descending order). Such understanding is due to broad discussion of power laws in linguistics, in particular Zipf’s law. This paper studies the probability distribution of fluctuations of the sampled frequencies of words in a large corpus of texts.

It is necessary to know the probability distribution of word frequencies to draw statistically valid conclusions about various effects revealed during language evolution studies by the frequency method. In particular, it is required to know the distribution to reason rapidly developing methods of quantitative study of semantic changes, as well as to compare various Word Embedding methods.

Probability distribution of word frequencies has been discussed in many papers [1, 2]. It is assumed that word frequency fluctuations are described by the Poisson law:

$$P_n = \frac{\lambda^n}{n!} e^{-\lambda}$$  \hspace{1cm} (1)

Here, \(n\) is the absolute frequency (that is, the number of word occurrences in the text or corpus), \(\lambda\) is a distribution parameter equal to the expected frequency mean value. For example, the assumption of the Poisson law is used in [3] to estimate the number of different n-grams using large corpora data.
The assumptions that the frequency distribution obeys the lognormal law [2] or is described by a mixture of distributions [1] are also used. The assumption of the Poisson law is widely used since it is a universal law describing the distribution of the number of independent rare events.

This paper studies the applicability of the assumption of the Poisson law for word frequencies using data on annual word frequencies in the large diachronic corpus Google Books Ngram [4]. It is a large electronic library which contains millions of scanned books written in eight languages. Nowadays this library is widely used for corpus-based studies.

2. Method and Data
It is difficult to directly determine the probability distribution of frequencies since the series of frequencies are not stationary. Relative frequencies of words change over time in a complex way due to dynamic language processes, changes in genres and topics of the corpora texts and other reasons. The work objective is to check the assumption of probability distribution by analysing the correlation between the distribution moments. The characteristic range of frequency fluctuations for the Poisson law is proportional to the square root of the mean frequency value, more precisely, the standard deviation of frequency. This ratio can be checked by assuming that the relative frequencies change rather slowly in a typical case.

Relative and absolute frequencies in the corpus are connected by the ratio \( n_t = f_t/N_t \), where \( n_t \) and \( f_t \) are series of absolute and relative frequencies, respectively, and \( N_t \) is the corpus size (in words) in different years. Assuming that the true relative frequencies in the book speech change rather smoothly in most cases, one can obtain an estimate of the relative frequency by applying this or that smoothing filter to the series of empirical frequencies \( f_t \). In this paper, we use simple moving average filters. If the Poisson law is assumed, the estimate \( \hat{f} \) is taken in the following form:

\[
\hat{f}_t = \frac{\sum_{\tau=-M}^{M} n_{t+\tau}}{\left(\sum_{\tau=-M}^{M} N_{t+\tau}\right)}
\]  

Having obtained this value, we can estimate the expected absolute frequency \( \hat{n}_i = \hat{f}_t N_t \) and the deviation of the empirical value from the expected one. Then, we determine the conditional mean frequency deviation range as a function of frequency estimation. We have values \( \xi_{i,t} \) and \( \hat{n}_{i,t} \) (where \( i \) is the number of the word in the sample) for all words in the sample and all years. Having selected some reference value of the frequency \( n \), we select all values \( \xi_{i,t} \), for which \( \hat{n}_{i,t} \) differs from \( n \) by no more than a given factor:

\[
n/(1+\varepsilon) \leq \hat{n}_{i,t} \leq n(1+\varepsilon)
\]  

We estimate the range of the sample values selected in accordance with (3) using this or that method. The value \( \varepsilon \) in this expression should be taken proportional to the root of the frequency \( \varepsilon = \alpha \sqrt{n} \), so that the estimations of the range are not overestimated. The last step of the processing is to smooth the obtained dependence of the range on the mean frequency value, in case the statistics is not enough to get a smooth curve.

Probabilistic distribution of frequency fluctuations was tested on the data of the Russian subcorpus of Google Books Ngram. We analysed frequencies of only vocabulary 1-grams, namely the words consisting of letters of the Russian alphabet and, probably, one apostrophe. All 1-grams which differ only by capitalization are considered to be one word form. Frequencies of such 1-grams were summed. POS tags of Google Books Ngram was not used. As stated in [5], the POS markup of the Russian subcorpus contains many errors, so the frequencies were used without taking into account parts of speech. The data were studied within the interval 1920-2009. We chose this interval for two reasons. Firstly, the corpus contains a big amount of data in this period of time. The total amount of Russian words in 1607-2009 is 67 billions. At that, 45 billion of words are found in 1920-2009. Secondly, this helps to avoid the effects associated with the reform of Russian spelling in 1918. We
analysed word forms that are systematically used during the considered time interval. We selected 216 thousand of word forms occurring in the corpus at least once every year in the interval 1920-2009.

3. Data Analysis

The results of calculation of conditional range of frequency fluctuations for the sample of Russian words are shown in figure 1, A. To estimate the average range, a moving average with a 3-year window was used (the parameter \( M \) in formula (2) equals 1). The choice of the window length is a result of a compromise since when the length of the window is increased, on the one hand, one can obtain a more reliable estimate of the average range. On the other hand, the obtained estimates can be distorted due to frequency changes over time. Calculations were performed using different window lengths (3, 5, 7 years). Similar results were obtained.

Interquartile range was used to estimate the range of frequencies. This statistics is resistant to strong spikes, which helps to eliminate the impact of sharp jumps in word frequencies associated with certain historical events. For the Poisson distribution, the interquartile range has the asymptotics \( \sim 1.349 \sqrt{\lambda} \). This dependence is shown in the figure by the dashed line. When calculating the conditional range, we took into account cases when the estimate of the mean frequency differed from the selected value by no more than 25% from the expected (if the Poisson distribution hypothesis is true) standard deviation value (i.e., the parameter \( \epsilon \) in formula (3) is equal to \( \epsilon = 0.25 \sqrt{n} \)).

![Figure 1. A) Dependence of the interquartile range of frequency variations on its average value (solid line, the boundaries of the 95% confidence interval for the estimates are also shown). The dashed line shows the dependence expected in the case of Poisson's distribution; B) Local exponent for the dependence of the interquartile range on the mean frequency value](image)

As can be seen in Figure 1, A, as the average frequency increases, the interquartile range grows much faster than it might be expected, assuming that the frequencies are distributed according to the Poisson law. Fitting of the power curve according to the criterion of the minimum relative square error in the interval \( 10^2-10^8 \) gives the power exponent 0.755 instead of the expected value 0.5 (the power curve with the exponent 0.5 is shown in Figure 1, A by the dashed line).

Following [6, 7], let us consider the local power approximation of the range dependence on the mean frequency value. In this case, the dependence is written as

\[
IQR = A(\lambda) \lambda^{-\nu(\lambda)}
\]

(4)

Here, IQR is the interquartile range, \( \lambda \) is the mean frequency value, and \( A(\lambda) \) and \( \nu(\lambda) \) are assumed to be slowly changing functions. As in [7], we calculate the values \( \nu(\lambda) \) by local approximation of the power dependence in the vicinity of this or that point. The results are shown in Figure 1.B. The figure also shows the levels 0.5 and 0.76. It can be seen that the local exponent varies considerably in different frequency ranges. This may indicate a difference between the factors leading to frequency fluctuations of high-frequency and low-frequency words. In the range \( 10^3-10^5 \), the local values are closest to the value 0.76, obtained by approximation over the entire range.
Thus, the average characteristics show a large deviation from the Poisson law. Let us further consider verification of the distribution hypothesis for the fluctuations of frequencies of individual words. The estimates of the parameters \( \hat{n}_{i,j} \) obtained during the above described calculations allow us to find the likelihood function values under the assumption of the Poisson distribution. However, since these estimates and the likelihood function values are calculated using the same data, it is necessary to perform statistical modeling to determine the p-value and accept or reject the hypothesis.

For this purpose, Poisson series with the parameters \( \hat{n}_{i,j} \) were generated for each word. Their length was the same as the length of the analyzed frequency series (90 years). The procedure of estimating Poisson's law parameters \( \hat{n}'_{i,j} \) was repeated for the generated series with the help of the moving average filter. Then, the obtained parameter values were used to calculate the likelihood function values. Repeating this procedure many times allows us to construct an empirical distribution of the likelihood function and calculate the p-value for the tested hypothesis.

It was found that at the significance level of 0.05, the Poisson distribution hypothesis is not rejected for 2.51% of words in the sample. For a lower significance level of 0.01, the hypothesis is not rejected for 5.62% of words and for 0.001 for 12.22% of words. Thus, for the vast majority of words there are significant deviations from Poisson's distribution.

4. Conclusion

In this paper, we analyzed the ratio of the average value and range of the sampled word frequencies using the Google Books Ngram data. To test the Poisson’s hypothesis, statistical modeling was performed. The results described above show that the Poisson distribution is poorly suited for describing the frequency fluctuations of the vast majority of words. The problem is that the range of frequency variations is significantly larger than expected under the assumption of the Poisson law. This excess is especially large for frequently used words.

It's expected that frequencies of more common words are more stable. This natural assumption is also confirmed by the results of this work. Indeed, the growth of the interquartile range according to the law \( \sim n^{-0.76} \) means a decrease in the relative range of fluctuations. Normalizing the interquartile range to the mean value, we obtain a decrease with increasing frequency as \( \sim n^{-0.24} \) for the relative range of frequency fluctuations. However, this decrease is significantly slower compared to the expected one in the assumption of the Poisson distribution by the \( \sim n^{-0.5} \) dependence. That is, the relative range of frequency fluctuations of high-frequency words decreases in comparison with rare words, but decreases to a lesser extent than expected. The factors that could explain this effect are phenomenon associated with fashion for use of certain words or grammatical constructions, authors' idiolect and style, as well as correlation between the use of different words.

Acknowledgements

This research was financially supported by the Russian Government Program of Competitive Growth of Kazan Federal University, state assignment of Ministry of Education and Science, and by RFBR, grant № 17-29-09163.

References

[1] Baayen R H 2001 Word Frequency Distributions (Dordrecht, Netherlands: Springer)
[2] Baayen R H and Lieber R 1996 Comput. Hum. 30(4) 281
[3] Silva J F, Goncalves C and Cunha J C 2016 2016 IEEE Int. Conf. on Big Data (Washington, DC: IEEE) pp 134-141
[4] Michel J-B, Shen Y K, Aiden A P, Veres A, Gray M K, et al. 2011 Science 331(6014) 176
[5] Bochkarev V V, Solovyev V D and Wichmann S 2014 J. R. Soc. Interface 11 20140841
[6] Bochkarev V V, Lerner E Yu and Shevlyakova A V 2014 J. Phys.: Conf. Ser. 490 012009
[7] Belashova I A, Bochkarev V V and Tyurin V A 2017 J. Phys.: Conf. Ser. 936 012074