Physica A 445 (2016) 328–334

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

# On variation of word frequencies in Russian literary texts

Vladislav Kargin

*Binghamton University, United States*

## HIGHLIGHTS

- We examine a large online library of Russian literary texts.
- The variation in the word frequencies across texts is related to the average word frequency by a non-linear power law.
- The finding is consistent with "burstiness" (increased relative variation) of rare words.
- A latent Dirichlet allocation (LDA) model is estimated.
- The non-linearity result can be explained by asymmetry in the distribution of latent factors.

## ARTICLE INFO

*Article history:*
Received 22 June 2015
Received in revised form 4 October 2015
Available online 19 November 2015

*Keywords:*
Burstiness
Word frequency variation
Latent Dirichlet allocation

## ABSTRACT

We study the variation of word frequencies in Russian literary texts. Our findings indicate that the standard deviation of a word's frequency across texts depends on its average frequency according to a power law with exponent $\frac{1}{2} < \alpha < 1$, which shows that the rarer words have a relatively larger degree of frequency volatility (that is, higher "burstiness").

A latent factor model has been estimated to investigate the structure of the word frequency distribution. The findings suggest that the dependence of a word's frequency volatility on its average frequency can be explained by the asymmetry in the distribution of latent factors.



## 1. Introduction

The study of word frequency variation in different texts arose first in the problem of author attribution [1–3]. Recently, the explosive growth in the computing power and in the text data volume led to many new applications. For example, the text indexing problem asks to associate documents with queries for fast retrieval; the authorship profiling problem requires to describe features of the author (sex, age, religious and political beliefs, etc.) based on texts that the author produced. In addition, the classic authorship attribution problem found new applications in security and forensics (see surveys by Holmes [4], Juola [5], Koppel, Schler, and Argamon [6] and Stamatatos [7]).

For all these applications, the fundamental statistical issue is the distribution of word frequencies[1] in different texts. For example, if a word in a query has its frequency in a document higher than its average frequency, then this document can be regarded as more relevant to the query.

Some properties of the word frequency distribution were noticed a long time ago. For example, Zipf's law [1] describes the distribution of word frequencies in a particular text, and Heaps' law (p. 207 in Ref. [8], p. 75 in Ref. [9]) relates the number of distinct words in a text to its length. Some new research on these laws was done in Refs. [10–12], and [13]. See also surveys

*E-mail address:* vladislav.kargin@gmail.com.

[1] In this paper we use the term "frequency" as usual in statistics, that is, the number of the word occurrences in a document divided by the document's total number of words.

http://dx.doi.org/10.1016/j.physa.2015.11.014
0378-4371/

in Refs. [14,15]. This paper focuses on a different set of properties and investigates the variation of word frequencies across documents.

One has to understand the structure of the word-document frequency matrix for applications in the information retrieval, in order to handle the problems of word synonymity and polysemy. For this purpose, there have been recently developed tools such as LSA ("latent semantic analysis", Deerwester et al. [16]), pLSA ("probabilistic latent semantic analysis", Hofmann [17]), and LDA ("latent Dirichlet allocation", Blei, Ng, and Jordan [18]). The main idea of these methods is the dimension reduction. The variation of word frequencies across texts is assumed to stem mainly from the variation in relatively small amount of factors (or "topics") across texts.

The goal of this study is to establish basic facts about the fluctuations of word frequencies across documents such as the dependence of the fluctuation size on the average word frequency. In order to clarify this dependence, we will apply a latent factor technique, the LDA.

The paper is organized as follows. First, in Section 2 we describe the data. Then, in Section 3 we study how the size of frequency fluctuations across texts depends on the word's average frequency. Next, in Section 4 we apply a latent factor model to analyze the variation of vocabulary across texts in more detail. Finally, Section 5 concludes.

## 2. A preliminary look at the data

We use data from Flibusta, a Russian online library. It covers Russian and translated fiction works from many historical periods and literary genres. The data is freely available either via the torrent network or by an automatic download. To the author's best knowledge, it have not been used previously for linguistic research.

Currently, it has between 200,000 and 300,000 texts by about 85,000 authors, where the author is understood to include translators and sometimes organizations that published a particular text. Our analysis uses only a part of this dataset (around 25,000 books). In particular, we use only books which are available in a text format (more precisely, in the "FB2" book format) and we exclude the documents that are available only as pdf, djvu, doc, and other binary formats.

The library works using the wiki principle and the texts are uploaded by users, therefore the number of texts depends both on how many texts were written by the author and on how many of them were uploaded by users.

To illustrate the content of the library, the two authors with the largest number of texts are the American and Russian science fiction writers Ray Bradbury and Kir Bulychev, with 550 and 540 texts, respectively. Many of the other top authors are authors and translators of books in popular genres such as science fiction, mystery, romance, action, historical fiction, sensational and how-to literature.

If the authors working in the genres associated with popular culture are excluded, then we find many well-known classic authors, most of whom are short story writers. To illustrate, for the first 25 of these authors the number of texts in the online library ranges from 446 for Anton Chekhov to 144 for Franz Kafka.

## 3. Variation of word frequencies across texts

In this section, as a first step we establish that there is significant variation in word frequencies across different texts. Then we connect the size of the variation to the average frequency of the word in a given text. We find a power function dependence between these two quantities.

Let $\xi_{b,w}^{(t)}$ be an indicator variable which equals 1 if the word at place $t$ in book $b$ equals $w$. Then, the frequency of word $w$ in book $b$ can be written as

$$x_{b,w} = \frac{1}{T_b}\sum_{t=1}^{T_b}\xi_{b,w}^{(t)}, \tag{1}$$

where $T_b$ is the length of the book $b$.

Suppose that for a given $w$ the random variables $\xi_{b,w}^{(t)}$ are independent and identically distributed with the expectation parameter $p_w$, which does not depend on $b$. Then $\mathbb{E}x_{b,w} = p_w$, and

$$\mathbb{V}\left(x_{b,w}\right) = \frac{p_w(1-p_w)}{T_b}. \tag{2}$$

To test this hypothesis, we estimate $p_w$ by using the whole sample:

$$\widehat{p}_w := \frac{1}{T}\sum_{b=1}^{B}\sum_{t=1}^{T_b}\xi_{b,w}^{(t)}, \tag{3}$$

where $T$ is the total number of words in the data and $B$ is the number of texts. Then we compute the normalized variance of $x_{b,w}$ across books.

$$V_w = \frac{1}{\widehat{p}_w(1-\widehat{p}_w)}\frac{1}{B}\sum_{b=1}^{B}T_b(x_{b,w}-\widehat{p}_w)^2. \tag{4}$$

This statistic should be compared with 1.

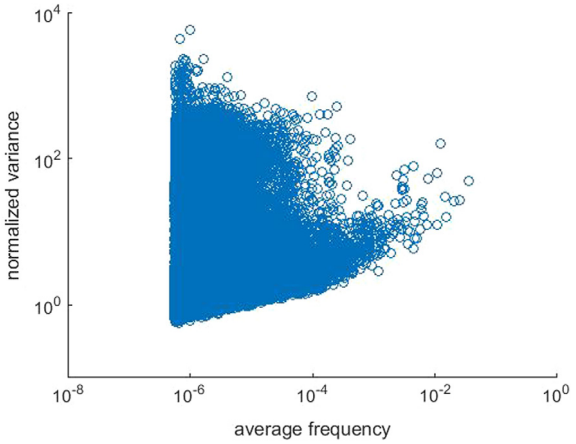


**Fig. 1.** Normalized variance $V_w$ vs. average frequency $\widehat{p}_w$.

The results are shown in Fig. 1. The figure shows the statistics $V_w$ for various words. They suggest that this model is not acceptable and that there is a significant degree of variation in the distribution of $\xi_{b,w}$ across texts.

This observation is not really surprising, since the variation in the word frequencies across texts is at the heart of many linguistic algorithms. However, it appears that the systematic study of this phenomenon's properties is relatively recent [19]. The phenomenon is often called *burstiness*.[2] One interesting observation of Church and Gale is that the words with an unusually high frequency variability are often content words: they have an additional linguistic load.

Now, let $\xi_{b,w}^{(t)}$ be independent random variables, which are identically distributed conditional on $b$ and $w$ and have the expectation parameter $p_{b,w}$. That is, the parameter is allowed to change from text to text and we are interested in learning how it is distributed across texts.

The simplest estimate for $p_{b,w}$ is $x_{b,w} = \frac{1}{T_b}\sum_{t=1}^{T_b}\xi_{b,w}^{(t)}$. It is reliable only if the standard deviation of the estimate is sufficiently small: $p_{b,w} \gg \sqrt{\frac{p_{b,w}(1-p_{b,w})}{T_b}}$, or $p_{b,w} \gg T_b^{-1}$. In our database, the average text length is of the order of $3 \times 10^4$ words and therefore we can expect that $x_{b,w}$ reliably estimates $p_{b,w}$ only if $p_{b,w} \geq 10^{-4}$.

Let us define the average word frequency:

$$\bar{x}_w = \frac{1}{B}\sum_{b=1}^{B} x_{b,w}, \tag{5}$$

and the cross-text variance:

$$\sigma_w^2 = \frac{1}{B}\sum_{b=1}^{B}(x_{b,w} - \bar{x}_w)^2. \tag{6}$$

The pictures in Fig. 2 suggest that in general the variance declines together with the average frequency, so it is natural to ask about the law of this dependence.

(The smooth upper bound for rare words in Fig. 2 corresponds to an extremal situation when a word occurs in just one text $i$. In this case the average (across texts) frequency for this word is $E(x_w) = \frac{1}{B}x_{w,i}$ and the second moment of the frequency is $E(x_w^2) = \frac{1}{B}x_{w,i}^2$, which gives the upper bound $E(x_w^2) = B[E(x_w)]^2$.)

First, from plots in Fig. 3 it appears that neither variance nor standard deviation is linearly related to the average frequency.

However, these plots suggest that the dependence has the form of a power law. A linear regression of $\log(\sigma_w^2)$ on $\log(x_w)$ gives the following estimate:

$$\log(\sigma_w^2) \approx -5.17 + 1.20\log(x_w).$$

(The confidence interval for the regression coefficient is (1.18, 1.21) at the 95% confidence level. For estimation, we used the first 2000 different words that appeared in the data.)

[2] The name "burstiness" comes from the observation that if a rare word has occurred at least once in a document, then it is likely to occur more times in the same document than it is predicted by a Poisson distribution with the word's average frequency. This is easily explained by the variability of the word frequencies across texts, since if it is known that a word occurs in a document then the posterior average of the word frequency in this document is higher than the average frequency in the whole corpus.

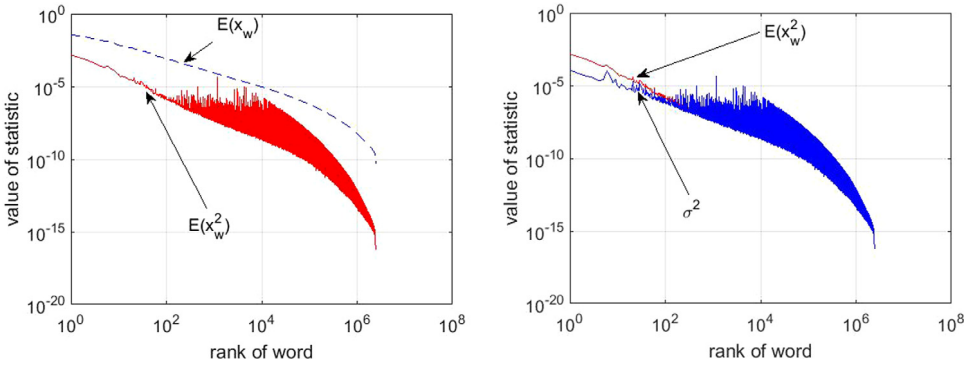


**Fig. 2.** The (estimated) expectation, second moment, and variance of the word frequency distribution. The words are ranked in the declining order by their average frequency.

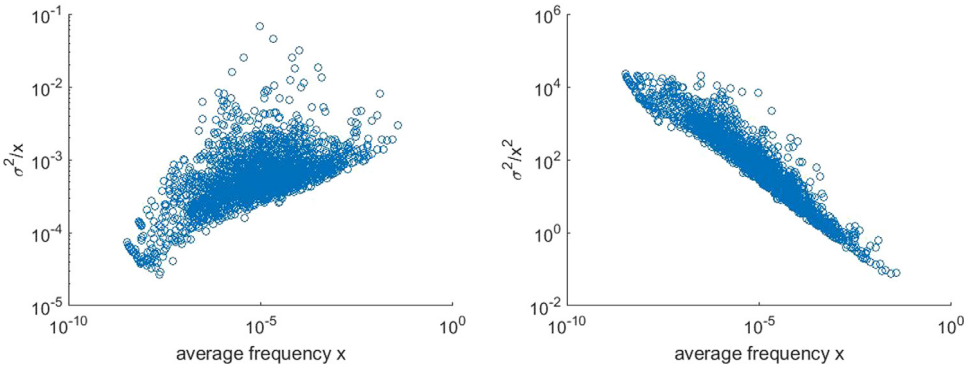


**Fig. 3.** Dependence of word variance on average frequency. (Figure shows a sample of 2000 words.)

Fig. 4 shows the fit of the regression, and Fig. 5 shows the residuals. The results of the regression and the figures suggest that the power law

$$\sigma^2 \sim a\bar{x}^{1.20} \tag{7}$$

gives a good approximation to the data. In terms of the ratio of the standard deviation to the mean, this law can be written as

$$\frac{\sigma}{\bar{x}} \sim a^{1/2}\bar{x}^{-0.4}. \tag{8}$$

In particular, the ratio is larger for rarer words.

In summary, these observations show that there is a power dependence between the variance of document word frequencies and the average frequency. The frequent words have larger variation in frequency across texts. However, the ratio of the standard deviation to frequency declines as the average frequency grows. This dependence follows a power law.

This relation can be seen as a quantification of the burstiness phenomenon. In particular, it shows that burstiness is in general more pronounced for rarer words. This suggests an implication that if volatility of a word's frequency (i.e., its burstiness) is used to evaluate the amount of content associated with the word, then the volatility should be normalized by a function of its frequency.

In the next section, we will try to uncover the structure in the variation of document word frequencies using a probabilistic model.

## 4. LDA model

The LDA ("Latent Dirichlet allocation") model is a particular case of pLSA ("probabilistic latent semantic analysis"), which postulates that the changes in word frequencies across texts can be explained by a relatively small number of factors. This approach is especially convenient for very large collections of data, when we are interested in reducing the complexity of the data, or, in other words, in "reducing the dimensionality" of an observed phenomenon.

According to the pLSA approach, the true word frequencies in a document are modeled as a mixture of a few probability distributions, and these distributions are interpreted as word distributions corresponding to a factor (or a "topic" in the

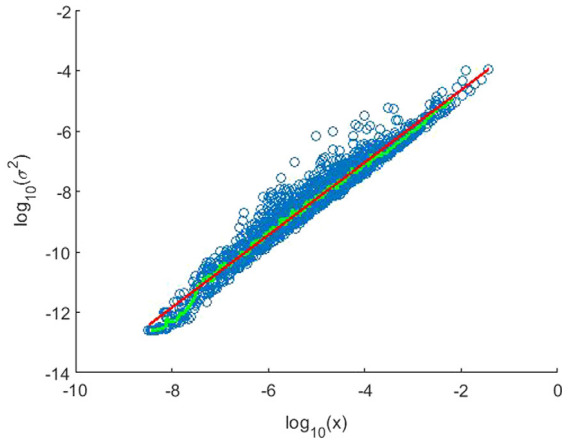


**Fig. 4.** Normalized variance vs. average frequency. The red solid line shows the regression fit. The blue line shows the moving median. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

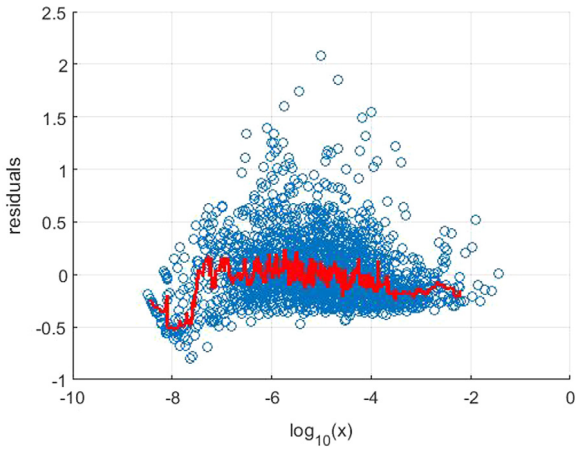


**Fig. 5.** Normalized variance vs. average frequency. The residuals of the regression. The red solid line shows the moving median of the residuals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

terminology of text indexing literature):

$$P(w|b) = \sum_{z=1}^{s} P(w|z)P(z|b). \tag{9}$$

The interpretation is that for each word in a book $b$ we randomly select a topic $z$ and then select the probability of a word $w$ on the basis of this topic. In other words, given topic $z$, the probability of a word $w$ is independent of the book $b$.

The LDA ("latent Dirichlet allocation") is a variant of the pLSA that treats the conditional probabilities $P(z|b)$ as a random variable drawn from a Dirichlet distribution.

In detail, let $\beta_{zw} = P(w|z)$ and $(\theta_b)_z = P(z|b)$. The joint distribution of the mixture $\theta$, and sequences of words $\{w_i\}$ and topics $\{z_i\}$ in a text $b$ is

$$P(\theta, \{w_i\}, \{z_i\}|\alpha, \beta) = P(\theta|\alpha) \sum_{i=1}^{N_b} \theta_{z_i} \beta_{z_i w_i}, \tag{10}$$

where $P(\theta|\alpha)$ is the Dirichlet distribution with parameter $\alpha$.

The main task is to estimate the parameters $\alpha$ and $\beta$ and compute the posterior distribution $P(\theta|\{w_i\})$. This is a non-trivial computational problem. Several approximation algorithms are available. For details, see paper by Blei, Ng, and Jordan [18]. In our experiments we used the code developed in Ref. [20].

The advantage of the LDA model for our purposes is that it can be used to investigate the burstiness phenomenon. (For a related model, the Dirichlet compound multinomial model, the burstiness was investigated in Ref. [21].)

In particular, we will use the LDA model to clarify results found in Section 3. First, note that the probability of word $w$ in a book $b$ equals $(\theta\beta)_{bw} = \sum_{z=1}^{s} \theta_{bz}\beta_{zw}$. Here $\theta_b$ is a realization of a random vector $\theta$ distributed according to the Dirichlet distribution with parameter $\alpha$. The joint moments of the Dirichlet distribution are well-known:

$$E\left(\prod_{z=1}^{s}\theta_i^{k_i}\right) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\Gamma\left(\sum_i(\alpha_i + k_i)\right)} \times \prod_i \frac{\Gamma(\alpha_i + k_i)}{\Gamma(\alpha_i)}, \tag{11}$$

and therefore one can easily compute the moments of the linear combinations of $\theta_i$.

Consider, for simplicity, the case with only two factors and the symmetric Dirichlet distribution. So, let $s = 2$ and $\alpha_1 = \alpha_2 = \alpha$. Then the probability that a particular word in a book is a word $w$ has a distribution with the expectation:

$$\mathbb{E}(p_w) = \frac{1}{2}(\beta_{1w} + \beta_{2w}) \tag{12}$$

and the variance can be computed as

$$\mathbb{V}(p_w) = \frac{1}{4}\frac{1}{2\alpha + 1}(\beta_{1w} - \beta_{2w})^2. \tag{13}$$

If $\xi_w = |\beta_{2w} - \beta_{1w}|/2$, then we could recover the findings in Section 3 provided that $\xi_w \sim (\mathbb{E}p_w)^{\kappa/2}$ with $\kappa = 1.20$. The problem with this interpretation, is that this relation is impossible for small $\mathbb{E}p_w$. Indeed, the positivity of $\beta_{1w}$ and $\beta_{2w}$ implies that $\xi_w \leq \mathbb{E}p_w$ and this contradicts the previous relation for small $\mathbb{E}p_w$. This can also be seen from the fact that $\mathbb{V}(p_w) \leq (\mathbb{E}p_w)^2$ in this model.

This can be rectified by using an asymmetric model. Take for example $s = 2$, $\alpha_1 = 1$ and $\alpha_2 = \alpha$. In this case,

$$\mathbb{E}(p_w) = \frac{1}{1+\alpha}\beta_{1w} + \frac{\alpha}{1+\alpha}\beta_{2w}, \tag{14}$$

$$\mathbb{V}(p_w) = \frac{\alpha}{(2+\alpha)(1+\alpha)^2}(\beta_{1w} - \beta_{2w})^2. \tag{15}$$

Let $\alpha \ll 1$, $\beta_{1w} = \gamma_w\alpha \ll \beta_{2w}$. Then,

$$[\mathbb{E}(p_w)]^2 \sim (\gamma_w + \beta_{2w})^2\alpha^2, \tag{16}$$

and

$$\mathbb{V}(p_w) \sim \frac{\beta_{2w}^2}{2}\alpha. \tag{17}$$

Hence, $\mathbb{V}(p_w) \gg [\mathbb{E}(p_w)]^2$ provided that $\gamma_w$ is not too large relative to $\beta_{2w}$.

Intuitively, the second topic occurs very rarely ($\alpha \ll 1$). However, it is associated with much larger conditional probability to observe the word $w$: $\beta_{2w} \gg \beta_{1w}$. This leads to a relatively large variance of the frequency distribution for the word $w$. In other words, the high burstiness of the word $w$ is due to its being a marker of a rare topic.

Next, we observe that when $\alpha$ is small and fixed, the power relation $\mathbb{V}(p_w) = [\mathbb{E}(p_w)]^\kappa$ is possible but only if $\gamma_w \gg \beta_{2w}$. Since $\gamma_w = \beta_{1w}/\alpha$, it follows that the relationship can occur in a limited range when $\beta_{1w} \ll \beta_{2w} \ll \beta_{1w}/\alpha$. This range is wide only if $\alpha$ is small.

In summary, the power relation observed in Section 3 appears to be due to the asymmetry in the distribution of topics vector $\theta$, and, in particular, it is due to the existence of rare topics that are associated with some specific words ("topic markers").

In order to demonstrate the asymmetry in the distribution of topics in the data, we show the estimates of the parameter $\alpha$ which is the Dirichlet parameter for topics, and the parameter $\beta_z = p(w|z)$ for one of the rare topics $z$.

The left plot in Fig. 6 shows the distribution of $\alpha$, which ranges from 0.04 to 0.45. The right plot shows that a rare topic is indeed associated with marker words. In this example, for the topic with $\alpha = 0.04$, there are three relatively infrequent words with $\beta > 0.03$. They are "всё" ("all"), "ещё" ("yet"), and "её" ("her"). Their average frequencies are $3.8 \times 10^{-4}$, $2 \times 10^{-4}$, and $1.9 \times 10^{-4}$, respectively. The common feature of these words is the presence of the letter "ё". This letter is often substituted by the letter "е" to economize on typography costs, and its presence indicates that either the book is intended for children or it has been published recently with the help of computerized typography.

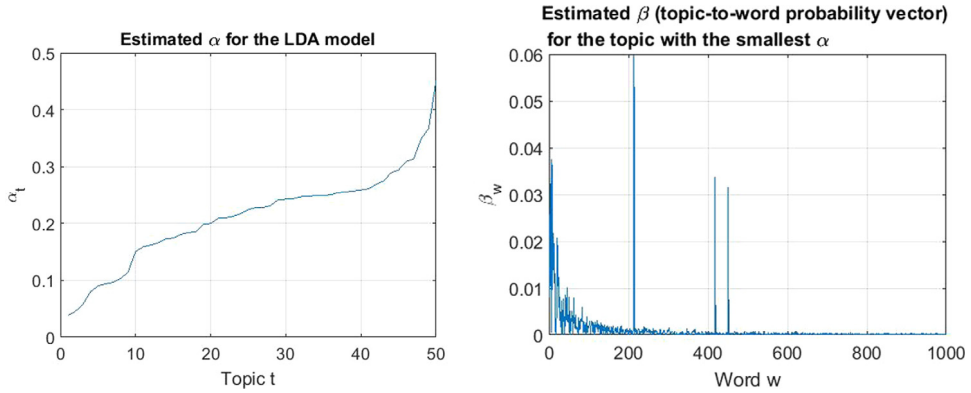


**Fig. 6.** The estimated parameters of the LDA model with 50 topics and 1000 most frequent words.

## 5. Conclusion

In this paper we studied the variation in the vocabulary of Russian literary texts from a large online database.

First, we detected a significant variation in the distribution of word frequencies across texts, and found that the variance of this distribution is in general larger for words with higher frequency. We found that the dependence of the word frequency volatility on its mean has a form of power law with the exponent $1/2 < \alpha < 1$, (in this data $\alpha \approx 0.60$), which quantifies the observation that rarer words have greater degree of "burstiness".

In order to study the variation in word frequencies across texts, we applied the Latent Dirichlet Allocation model. An analysis of the LDA model suggests that the power dependence of the frequency volatility on its mean can be explained by an asymmetry in the prior distribution of topics.

## References

[1] G.K. Zipf, Selected Studies of the Principle of Relative Frequency in Language, Harvard University Press, Cambridge, MA, 1932.
[2] G.U. Yule, The Statistical Study of Literary Vocabulary, Cambridge University Press, 1944.
[3] Frederick Mosteller, David L. Wallace, Inference and Disputed Authorship: The Federalist, Addison-Wesley Publishing Company, Inc., 1964.
[4] D.I. Holmes, The evolution of stylometry in humanities scholarship, Lit. Linguist. Comput. 13 (1998) 111–117.
[5] Patrick Juola, Authorship Attribution, in: Foundations and Trends(r) in Information Retrieval, Now Publishers Inc., 2008.
[6] Moshe Koppel, Jonathan Schler, Shlomo Argamon, Computational methods in authorship attribution, J. Assoc. Inf. Sci. Technol. 60 (2009) 9–26.
[7] Efstathios Stamatatos, A survey of modern authorship attribution methods, J. Amer. Soc. Inf. Sci. Technol. 60 (2009) 538–556.
[8] H.S. Heaps, Information Retrieval: Computational and Theoretical Aspects, Academic Press, New York, 1978.
[9] G. Herdan, Advanced Theory of Language as Choice and Chance, Springer-Verlag, New York, 1966.
[10] Francesc Font-Clos, Gemma Boleda, Alvaro Corral, A scaling law beyond Zipf's law and its relation to Heaps' law, New J. Phys. 15 (2013) 093033.
[11] Martin Gerlach, Eduardo G. Altmann, Stochastic model for the vocabulary growth in natural languages, Phys. Rev. X 3 (2013) 021006.
[12] Martin Gerlach, Eduardo G. Altmann, Scaling laws and fluctuations in the statistics of word frequencies, New J. Phys. 16 (2014) 113010.
[13] Steven T. Piantadosi, Zipf's word frequency law in natural language: a critical review and future directions, Psychon. Bull. & Rev. 21 (2014) 1112–1130.
[14] Damian H. Zanette, Statistical patterns in written language, 2014. Available at: arxiv:1412.3336.
[15] Eduardo G. Altmann, Martin Gerlach, Statistical laws in linguistics, in: Proceedings of the Flow Machines Workshop: Creativity and Universality in Language, Paris, June 18 to 20, 2014, 2015. Available at: arxiv:1502.03296.
[16] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990) 391–407.
[17] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.
[18] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[19] Kenneth W. Church, William A. Gale, Poisson mixtures, Nat. Lang. Eng. 1 (1995) 163–190.
[20] Jacob Verbeek, Latent Dirichlet Allocation/Probabilisic Latent Semantic Analysis, 2006. http://lear.inrialpes.fr/~verbeek/software.php.
[21] Rasmus E. Madsen, David Kauchak, Charles Elkan, Modeling word burstiness using the Dirichlet distribution, in: Proceedings of the 22Nd International Conference on Machine Learning, ICML'05, ACM, New York, NY, USA, ISBN: 1-59593-180-5, 2005, pp. 545–552.