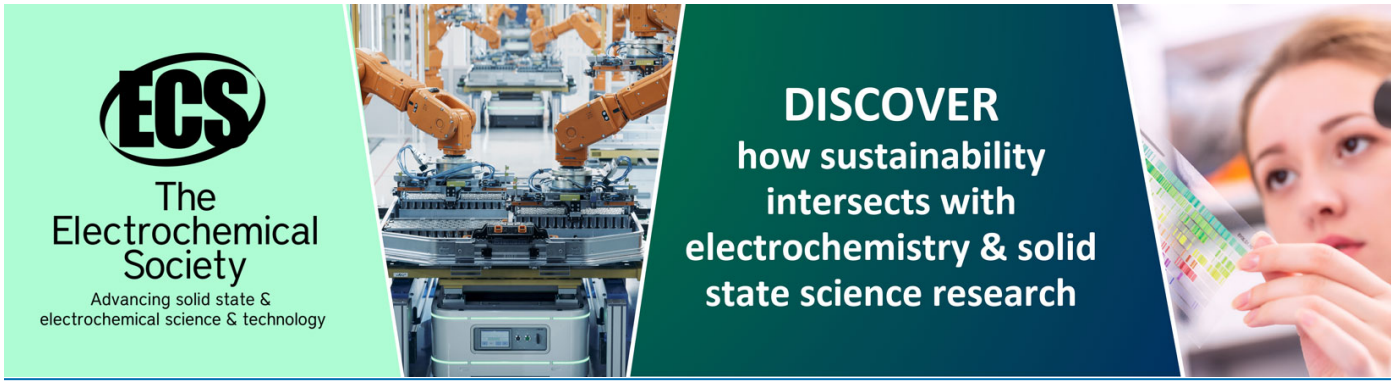
Cluster analysis of word frequency dynamics



Cluster analysis of word frequency dynamics

Yu.S. Maslennikova, V.V. Bochkarev, I.A. Belashova

Kazan Federal University, Kremlevskaya str.18, Kazan 420018, Russia

E-mail: JSMaslennikova@kpfu.ru, vbochkarev@mail.ru

Abstract. This paper describes the analysis and modelling of word usage frequency time series. During one of previous studies, an assumption was put forward that all word usage frequencies have uniform dynamics approaching the shape of a Gaussian function. This assumption can be checked using the frequency dictionaries of the Google Books Ngram database. This database includes 5.2 million books published between 1500 and 2008. The corpus contains over 500 billion words in American English, British English, French, German, Spanish, Russian, Hebrew, and Chinese. We clustered time series of word usage frequencies using a Kohonen neural network. The similarity between input vectors was estimated using several algorithms. As a result of the neural network training procedure, more than ten different forms of time series were found. They describe the dynamics of word usage frequencies from birth to death of individual words. Different groups of word forms were found to have different dynamics of word usage frequency variations.

1. Introduction

Literature is one of the main forms of artistic expression. Its temporal dynamics, both in content and style, allows us to track the historical aspects of the evolution of culture. To date, a lot of research has been conducted on the basis of the quantitative analysis of literary texts and frequency dictionaries. Through such research, the behavior word set of a language is approached as a kind of complex system. Continuing this line of investigation we here explore the stages of birth, evolution and death of individual word-forms. Significant progress in the understanding of the statistics of use of word-forms was achieved by Baayen in numerous works, e.g. [1, 2]. It should be noted, that processing techniques, the metric selected and the databases used play essential roles in the quantitative analysis of the evolution of language, including that of separate word-forms. By way of introduction we will review some of the most important studies in this field.

The authors of [3] conducted their study using a 108-word database from an online discussion group and a 1011 item word collection of digitized books, provided by Google Corporation. The model, presented in the article shows the existence of a strong relation between changes in word dissemination and changes in frequency. Google Books digital library and The Google Books Ngram corpus provide new opportunities for the quantitative analysis of a language evolution [4]. For example, using this database the authors of [5] showed that Heaps' law, which describes the number of distinct words in a document (or set of documents) as a function of the document length, does not always adequately describe the empirical facts: the exponent in the Heaps' law is not constant, it may vary with time, and, additionally, may exhibit oscillations, sometimes of a quasi-periodic character. This may mean that the influx of new words constantly varies. There are also papers devoted to research on narrowly defined lexical categories such as names. In [6], the authors studied the dynamics of the frequency of names that parents choose for their children, showing that names that

quickly come into fashion quickly go out of fashion. This interesting observation contributes to shedding light on the genesis and extinction of different cultural tastes in general.

Ref. [7] presents an analysis of 10^7 words from the English, Spanish, and Hebrew parts of the Google Ngram database. The authors study trends in the rate of birth and death of words. For new words a peak in usage fluctuations peak is usually observed around 40 years after a word has gained wide circulation. Pronounced changes in the dynamics in times of war shows that the dynamics depends on co-evolutionary social, technological and political factors. The authors also showed that trends in birth and death of individual words generally have a similar shape. Ref. [8], however, shows the existence of difference between individual words dynamics. The authors of that paper analyzed the frequency of usage of function words using artificial neural networks. Here we propose an analysis of a broader database of words using an improved approach to clustering based on Kohonen neural networks.

2. Database

Our data are drawn from the 2012 version of the Google Books Ngram corpus of English words [4]. We did some filtering, selecting only 1-grams exclusively containing letters of the Latin alphabet and not more than one apostrophe. For each selected word-form we calculated a series of relative frequencies. While the interactive Google Books Ngram Viewer normalizes frequencies by the total number of 1-gram in the corpus, we normalized frequencies by the number of 1-grams retained after filtering.

Through this selection frequency charts for 6.78 million word forms was obtained. We prepared a reduced sample of the word forms that are among the 100,000 most frequent over the entire period which we focus on. This is the period from 1800to. For the period before 1800 the corpus is relatively small and the small sample size influences variation too much.. Moreover, the British database suffers from the additional problem of changes in fonts in years preceding 1800, something which leads to a large number of errors of optical character recognition.

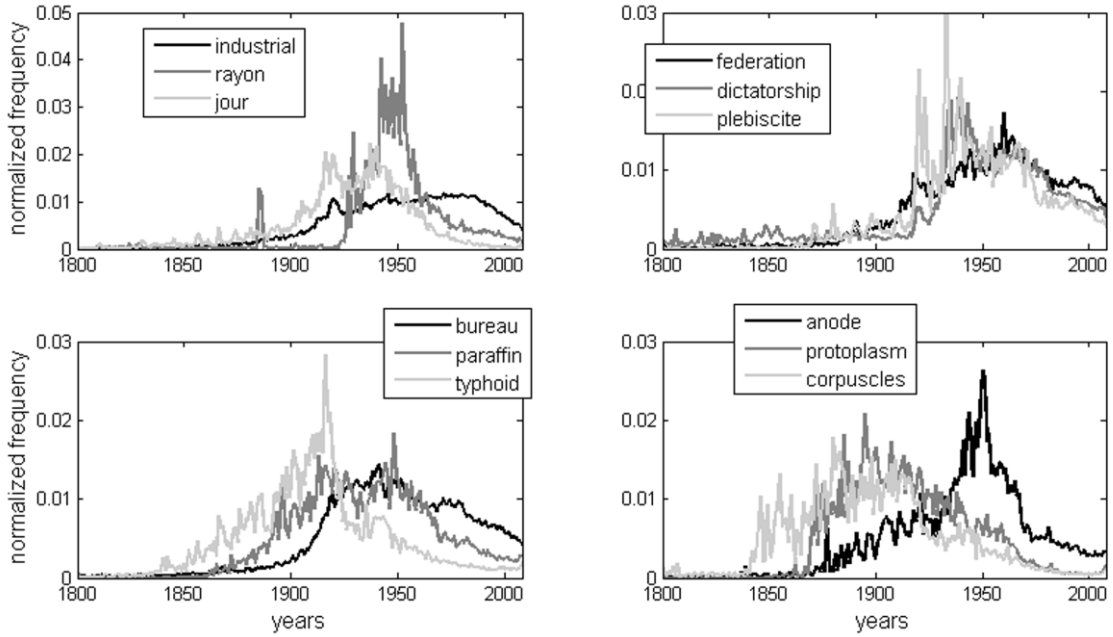


Figure 1. Dynamics of normalized frequencies for the 12 selected words.

There is a great variety in the types of frequency dynamics which can be observed. Even if we could overcome the technical difficulties involved in clustering the entire volume of the vocabulary, would still be difficult to visualize and interpret all types encountered. Therefore, in this paper we restrict the analysis to a narrow class of words which display clear 'bursts' in their frequency graphs,

i.e., words for which we can distinguish a phase of increase, a clear maximum and a phase of decrease in frequency for the period under consideration). Examples of the frequency graphs for such words are shown in Fig.1. This shows the dynamics of the normalized frequencies for the 12 selected words: '*industrial', 'rayon', 'jour', 'federation', 'dictatorship', 'plebiscite', 'bureau', 'paraffin', 'typhoid', 'anode', 'protoplasm', 'corpuscles'* (normalization was performed according to the formula: $f_{norm}(t) = f(t) / \sum_t f(t)$). The selection of words with of the 'bursting' type was as accomplished in the follows way. For each word we calculated the center and the half-width of the frequency dependency according to the following formulas:

$$t_c = \frac{\sum_t t \cdot f(t)}{\sum_t f(t)}, \qquad \sigma_t^2 = \frac{\sum_t (t - t_c)^2 f(t)}{\sum_t f(t)}.$$

We selected word-forms that satisfy two requirements: first, we took only the words with a half-width σ_t of no more than 60 years. Secondly, to be able to observe all phases of the evolution we selected word forms for which $1850 \le t_c \le 1958$. Among the 100,000 most frequent words 4,445 satisfied these conditions whereas 48,969 word forms in the total corpus satisfy the conditions. First a comparison of different measures and clustering methods was performed on the short list (4445 word forms). Then, after determining the optimal approach, we extended clustering to the complete list of 48,969 word forms.

3. Clustering method

We used hierarchical clustering methods as well as clustering using the neural network of the Kohonen layer [9]. Hierarchical methods have the advantage of simple interpretation of the clustering results, and do not require a prior setting of the number of clusters. The advantage of the Kohonen layer is that it requires less computation.

To solve the clustering problem we need to choose the measure of the difference so that it is dependent only on the shape of the frequency dependencies compared and will not change with variation in absolute values and time shift. Achieving this requirement is possible with following method. Choosing some measure of the difference $D(x,y)$, we evaluate its value with different time shifts for a series compared against another one, just as is done in the calculation of the cross-correlation function. Thereafter, the difference between series x and y is calculated:

$$\min_{\tau} D(x(t+\tau), y(t))$$

We tested the following distance metrics:

- Correlation metric (via Pearson's coefficient);
- Correlation metric (via Spearman's coefficient);
- Distance metric based on L_1 norm. In this case, the distance measure is defined as the L_1-norm of the difference between vectors, normalized to unit length (according to $f_{norm}(t) = f(t) / \sum_t f(t)$) of frequencies;
- Kullback–Leibler metric. This measure arises naturally when deciding on a frequency vector's membership to the certain cluster based on the maximum likelihood method (it is necessary to take into account that the sampling frequencies have a Poisson distribution).

We used the following algorithms for computing distances between clusters: unweighted average distance (UPGMA), furthest distance, shortest distance, weighted average distance (WPGMA). Selection of the clustering method was performed by visual analysis. We checked how similar frequency dependencies are combined into one cluster. We also gave preference to methods which provided a more uniform grouping of words.

Let's consider the example of calculating the distance between two different words. Figure 2a shows dynamics of normalized frequency for words 'bolshevist' and 'isolationists'. Let's use the

correlation metric based on Pearson's correlation coefficient. Figure 2b shows how the correlation distance depends on a shift of the first curve relative to the second one. We can see that the zero shift corresponds to the correlation coefficient 0.137 and a sufficiently large distance that is equal to 0.863. The minimum value of the distance that is equal to 0.097 (Pearson's coefficient equals 0.903), achieved with a shift for 22 years. It can be seen that curves are similar using the given shift. Therefore, for further analysis we use the found distance (0.097) as a distance between two considered words.

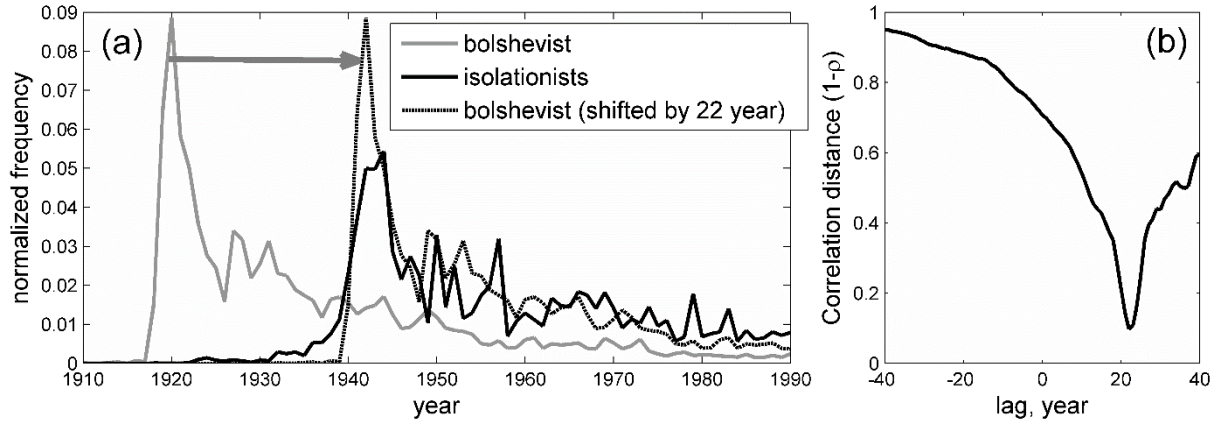


Figure 2. (a) Normalized frequencies for words 'bolshevist' and 'isolationists', frequency plot for the word 'bolshevist' shifted by 22 years is shown by dotted line; (b) Correlation distance at different shift values.

4. Results

Concerning the results of clustering, first of all, we note a wide variety of identified types of dynamics of frequency of word usage. When using any of the measures discussed above, the differences cannot be reduced to a classification into less than 70-100 clusters. Otherwise, with an increase in the threshold of clustering, the algorithm starts combining very different frequency trends. Figure 3 shows the normalized average frequency for 6 clusters containing the largest number of words. The dependencies presented were obtained by clustering the short list of words with a hierarchical method using the L_1-norm. In total, 6 clusters shown include 89.5% of the words. The remaining 10.5% of the words are distributed over 94 clusters of small size, which confirms that there is a great variability of types of frequency dynamics.

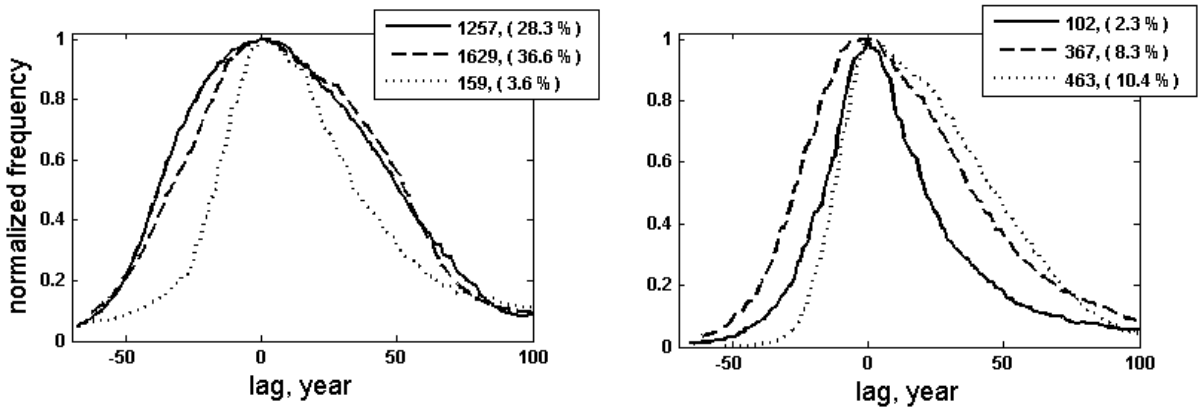


Figure 3. The normalized average frequency for 6 clusters containing the largest number of words.

The asymmetry of most dependencies is noteworthy: the phase of decrease in the frequency is longer than the phase of increase in the frequency. This impression can be confirmed by direct calculations. For each word from the complete list (48,969 word-forms) we evaluated the maximum and the point of intersection of the graph with the half of the maximum. Thus, for each word we estimated the duration of the growth phase of frequency (T_1) and the phase of decline of frequency (T_2). The median value of the ratio T_2 / T_1 was equal to 1.723, with T_2 greater than T_1 in 71% of the cases. Similar asymmetry was shown in frequency dynamics of child names in the paper [6] that we early considered. The duration of frequency increase phase is probably determined by rate of information exchange in society and social interaction. The duration of frequency decrease phase is determined by rate of falling a word into disuse and replacing it by frequently used words. As these two processes are substantially different, there are no a priori reasons to expect that the duration of these two phases will be equal. In fact, it can be seen that the second phase requires much more time.

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

References

[1] Baayen R H and Lieber R 1996/1997 Word frequency distributions and lexical semantics. *Computers and the Humanities* **30(4)** 281-291
[2] Baayen R Word Frequencies. Word Frequency Distributions 2001 *Text, Speech and Language Technology* **18** 1-38
[3] Eduardo G Altmann and Zakary L Whichard Adilson E. Motter Identifying Trends in Word Frequency Dynamics 2013 *J Stat Phys* **151** 277–288.
[4] The Google Books Ngram Corpuses, available at http://books.google.com/ngrams/datasets
[5] Bochkarev Vladimir V, Lerner Eduard Yu, Shevlyakova Anna V Deviations in the Zipf and Heaps laws in natural languages 2014 *Journal of Physics*: *Conference Series* **490**
[6] Berger J et al. How adoption speed affects the abandonment of cultural tastes 2009 *PNAS* **106(20)** 8146-50
[7] Petersen A M, Tenenbaum J, Havlin S, Stanley H E Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death 2012 *Scientific Reports* **2** 313
[8] Maslennikova Yu S, Bochkarev V V, Voloskov D S Modelling of word usage frequency dynamics using artificial neural network 2014 *J. Phys.: Conf. Ser.* **490**
[9] Kohonen T Self-Organizing Maps. Third, extended edition 2001 *Springer*