Analysis of Dynamics of the Number of Syntactic Dependencies in Russian and English Using Google Books Ngram

Vladimir Bochkarev[1] Valery Solovyev[2], and Anna Shevlyakova[3]

[1] Kazan Federal University, Kazan, Russia
vbochkarev@mail.ru
[2] Kazan Federal University, Kazan, Russia
maki.solovyev@mail.ru
3 Kazan Federal University, Kazan, Russia
anna_ling@mail.ru

Abstract. The work examines the dynamics of the number of syntactic dependencies and 2-grams in Russian and English using the Google Books Ngram diachronic corpus. We counted the total number of 2-grams and syntactic dependencies detected in Google Books Books Ngram at least once in a given year, as well as stable dependencies, which value of pointwise mutual information is above a given threshold. The effective number of dependencies expressed through the perplexity of 2-gram frequency distributions was also calculated. This value is a characteristic number of frequently used word combinations. It was found that quantitatively unchanged core and rapidly growing periphery can be distinguished among the syntactic dependencies of words. It was possible to obtain an estimate of the growth rate of the effective number of syntactic dependencies in the Russian language. The estimate shows that doubling of the effective number of dependencies occurs approximately every 250 years if the corpus size stays unchanged.

Keywords: Google Books Ngram, syntactic dependencies, computational linguistics, correlation models, linguistic databases.

1 Introduction

Emergence of extra-large text corpora and development of new algorithms and methods of linguistic research opens up broad opportunities for studying dynamic processes occurring in a language, and allows us to trace evolution of language phenomena.

Computer processing of large arrays of language data makes it possible to quantify the dynamics of lexicon and development of intralingual relations, classification and clustering of vocabulary. One of the largest corpora of texts is the Google Books library [1, 2]. It includes more than 8 million of digitized books written in 8 languages and is currently the largest digital text resource. The oldest books included in the corpus were written in the 1500s, and the latest book was published in 2009. The Google

Books Ngram services allow frequency analysis of word usage and visualization of the data.

Performing a quantitative analysis of text corpora, researchers solve various problems concerning language complexity [3], interrelations between language and culture (even the special term "culturomics" was introduced) [1], try to detect regularities of emergence and functioning of linguistic units and evolution of grammar. The article [3], in which the growing number of unique phrases in the English language was studied seems to be the most interesting in the context of our work. The author explains that increase in the number of word combinations is due to increasing complexity of culture. Meanwhile, the size of the Google Books Ngram corpus constantly increases (see Figure 1). The corpus growth, by itself, in accordance with Heaps' law, should lead to growth in the number of unique word combinations. The empirical Heaps' law describes the dependence of the number of unique words in a text on the size (length) of this text and states that the number of these words is connected by a power dependence with the size of the text [4, 5]. Despite the fact that the classical formulation of Heaps' law speaks only about the number of unique words, the same applies to the number of word combinations and syntactic dependencies [6]. Also, a certain disadvantage of Juola's work is that all of the conclusions are based on the analysis of the English corpus only.

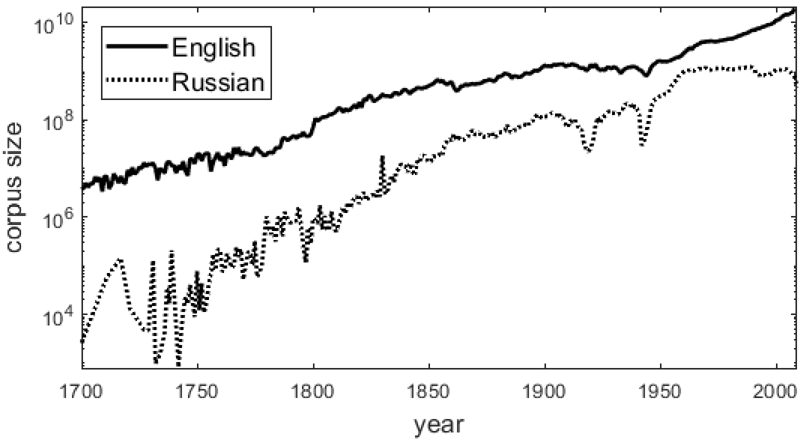

Fig. 1. Size of the common English and Russian sub-corpora included in Google Books Ngram (number of words).

Taking into account the conclusions [3], we set out a goal to analyse the dynamics of the number of syntactic dependencies and 2-grams. A priori, it can be expected that the number of such word relationships increases over time due to two factors: 1) increasing complexity of human culture [7, 8] and emergence of new words providing increase in the number of semantic connections and syntactic dependencies; 2) metaphorization processes, which also increase the number of relationships between words. Also, the number of 2-grams and syntactic dependencies detected in the corpus grows due to increase of the corpus size. The study objective was to identify how the number of 2-grams and syntactic dependencies increases with time, as well as to trace the impact of each of these factors. The Russian and English text corpora, which belong to the diachronic corpus Google Books Ngram, were studied.

2 Data and Methods

The common corpus of the English language and the corpus of the Russian language, which are a part of Google Books Ngram, were analysed.

Raw data are available for download on the project page (https://books.google.com/ngrams/). They contain information on frequency of use of words and n-grams (2-, 3-, 4- and 5-grams) in the books presented in the Google Books electronic library for each year. In our work, we used a base of frequencies of 2 grams, that is, pairs of words which, directly go one after another in the sentence.

A distinctive feature of the version of the 2012-year corpus is the presence of a base of frequencies of syntactic dependencies. Syntactic dependencies are understood as pairwise relationships between words in the same sentence. One of the words is a head, another one is a modifier. Such dependency relations are independent of word order, even though there are often intervening words between the head and the modifier. The data on frequencies of syntactic dependencies available in the Google Books Ngram corpus were also used in this work.

Thus, the term “2-gram” is used in our work when we describe pairs of words, which directly go one after another in the sentence. The term “syntactic dependencies” is used for head-modifier pairwise relationships between two words in a sentence. We study the number of different 2-grams and pairs of words being in a syntactic dependency.

Preliminary data processing was performed before the study. First, we did`t make a distinction between words that differ in case. Accordingly, 2-grams and syntactic dependencies, containing words that differ in case, were considered identical. Secondly, only vocabulary 1-grams were selected. 1-grams are understood to be words composed only of letters of the corresponding alphabet and, possibly, one apostrophe. If not taking into account differences in case, there were 5096 thousand (out of the total number of 8256 thousand) of such 1-grams found in the common English corpus. Accordingly, 4091 thousand 1-grams out of a total number of 5096 thousand 1-grams were selected for the Russian corpus. To normalize and calculate relative frequencies, the number of vocabulary 1-grams was calculated for each year (unlike Google Books Ngram Viewer, where normalization is performed for the total number of 1-grams). Parts of speech are marked in the 2012 version of the database. However, in many cases, parts of speech are determined improperly, which can cause incorrect conclusions based on such data. Therefore, the method introduced in [9] was used. It says that if the number of word forms corresponding to a certain part of speech does not exceed 1% of the total frequency of use of this word form, such word forms should be rejected and not used in further analysis. During the second stage of the survey, 2-grams consisted of the selected 1-grams were analysed.

The analysis was based on the following principles. Many researches attempt to determine the number of word combinations in the language. The easiest way to do it is to count the number of different word combinations in a corpus in a given year. To analyze the number of pairs of words forming dependencies, we counted the total number of 2-grams and syntactic dependencies marked in the Google Books Ngram database at least once in a given year. However, this method has some drawbacks. The first drawback is that a large amount of word pairs located next to each other in a sentence but not forming a dependency is counted. The second drawback is that, according to the authors of the Google Books Ngram project, approximately 30% of unique word forms contained in the database result from misprints. These factors cause an even more significant overestimation of the number of 2-grams and syntactic dependencies. The third drawback is that empirical frequencies of rare words, which are in the majority in the base, highly fluctuate, which also leads to large errors in estimation of the number of 2-grams and syntactic dependencies. Two approaches were used to reduce the impact of these factors. The first one is the following. Not all 2-grams and syntactic dependencies were counted but only frequently used ones, which are in a certain associative connection and are called collocations. Usually collocations are understood as word combinations, where words a located next to each other. However, some researches consider that stable syntactic dependencies can also be called collocations [10].

A value called pointwise mutual information in computational linguistics [11, 12] was used as a measure of associative connection. This value is expressed by the formula:

$$\mathrm{MI} = \log_2 \frac{f_{12}}{f_1 f_2} \tag{1}$$

Here f_{12} is a relative frequency of the word combination, and f_1 and f_2 are relative frequencies of the words, which form the word combination. As can be seen from the formula, the MI value shows to what extent the word combination is found more often in a text or a corpus than in a random text of the same size with an independent choice of words. The selection was carried out according to the value of the MI, which is 0, 3, 6, 9 for a number of threshold values of this quantity. The calculation results for the English and Russian languages are shown in Figure 2.

The second possible solution may be to count the number of word combinations with regard to their informational content. We can used such characteristic of frequency distribution as perplexity [13]. The effective number of syntactic dependencies (2-grams) numerically equal to the perplexity of their frequency distribution was introduced:

$$N_{eff} = 2^h \tag{2}$$

Here h is the entropy of the frequency distribution, calculated by the formula:

$$h = -\sum_i f_i \log_2 f_i \tag{3}$$

where f_i is the frequency of the i-th 2-gram (or syntactic dependency). The introduced value shows the number of frequently used syntactic dependencies (2-grams), taking into account their role in the information exchange. Our approach is close to that used in [3]. However, using perplexity instead of entropy allows us to present the results more vividly, as well as to make comparisons with estimates obtained by other methods. The dynamics of the effective number of syntactic dependencies of both languages is also shown in Figure 2.

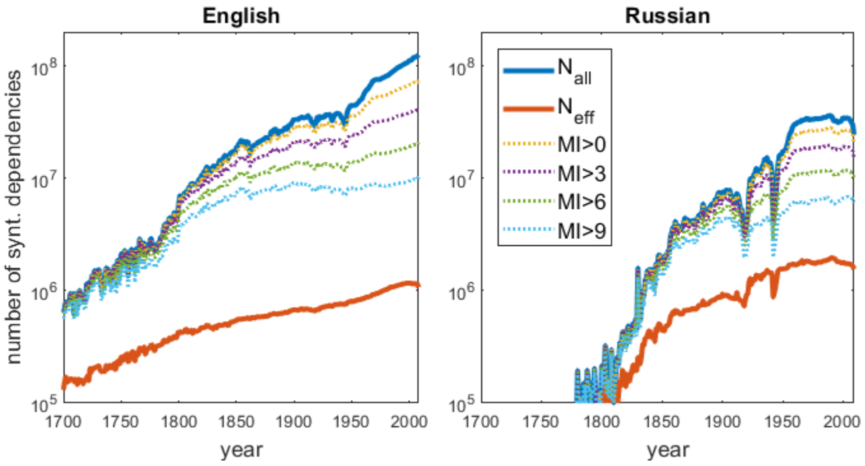

Fig. 2. The number of syntactic dependencies in Russian and English in 1700-2008. The total number of syntactic dependencies, the effective number of syntactic dependencies (perplexity) and the number of syntactic dependencies with MI above the given threshold are shown.

3 Results

As can be seen from Figure 2, the total number of syntactic dependencies in both languages is growing rapidly. At that, the growth rate in different periods changes significantly, the curve responds to various historical events, primarily to wars and revolutions. If we restrict ourselves to stable syntactic dependencies, the curve qualitatively retains its character. However, it shows a slightly lower growth rate. The number of syntactic dependencies with high MI values grows very slowly. All this is true for the number of 2-grams.

Comparing figures 1 and 2, it can be seen that the curves of the total number of syntactic dependencies are similar to the graphs of the corpora size. This observation can be quantified. Table 1 shows the values of the Spearman correlation coefficients between the corpus size and the number of syntactic dependencies (the total number of syntactic dependencies and the number of only stable syntactic dependencies) in English and Russian.

The correlation coefficients will not change in any noticeable way if they are calculated using the limited intervals of 1700-2008 or 1750-2008. Thus, the compared values show a high level of statistical connection, especially for the Russian language.

Table 1. Spearman's correlation coefficient between the corpus size and the number of syntactic dependencies in Russian and English.

	English	**Russian**
Total number of syntactic dependencies	0.890	0.974
Number of syntactic dependencies with MI>0	0.860	0.972

The graph of the effective number of syntactic dependencies has a different character. The curve is much more regular and smooth and responds insignificantly to historical events. The size of the English corpus is substantially larger than the Russian one. It contains approximately 470 billion of words and the Russian corpus includes only 67 billion of words. The English corpus shows no reaction to historical events, and the graph of the effective number of syntactic dependencies can be well described by an exponential dependence (in a logarithmic coordinate system – a linear dependence). The smooth exponential growth of the effective number of syntactic dependencies in the English language is accelerated only after about 1950, which may be a manifestation of globalization processes. It is indisputable that by the end of the 20th century, English becomes the leading world language. Its influence on the processes of international economic, political and cultural integration proceed is great. English has also become the second mother-tongue for many people and develops very fast. The total number of syntactic dependencies in the English language is higher than in Russian, which is a manifestation of Heaps' law.

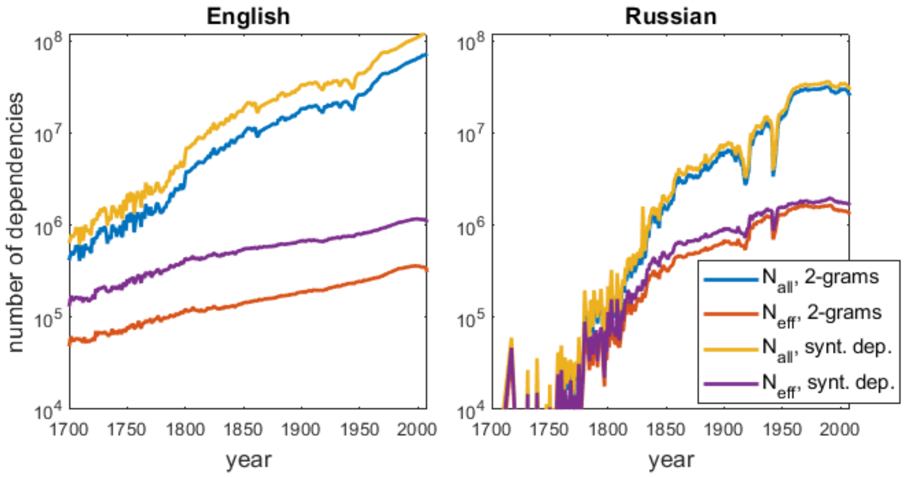

Fig. 3. The number of syntactic dependencies and 2-grams in Russian and English in 1700-2008.

At that, the Russian language has more effective syntactic dependencies than English, which can probably be due to more complicated morphology and word-formation. Thus, applying such indicator as the effective number of syntactic dependencies allows us to perform less subjective comparative analysis of language processes using corpora of various sizes.

Figure 3 shows the number of syntactic dependencies and 2-grams in Russian and English. Both the total and effective number of syntactic dependencies and 2-grams are compared. Attention should be paid to the fact that the ratio of the number of syntactic dependencies to the number of 2-grams is significantly larger in the English language. Probably, this can be due to the fact that the word order in the Russian language is not fixed. As a result, a larger number of 2-grams can be formed. This may also be due to some features of syntactic analysers used for creating a corpus. Otherwise, as can be seen from Figure 3, the number of syntactic dependencies and the number of 2-grams change over time in a similar way.

As it was stated above, increase in the number of syntactic dependencies and 2-grams can be due to growing complexity of culture, increase of a corpus size and metaphorization processes, which cause emergence of new words. Influence of each factor was investigated in the work.

To level the effect of a simple increase in the number of new words, one can count the number of word combinations and syntactic dependencies, which are comprised only of a fixed set of words belonging to the lexicon core. There are various approaches to the problem of determining the lexicon core [14]. To solve the problems mentioned in the article, it seems natural to use the method proposed in [15], according to which we select words recorded in the corpus each year from a certain period. There are approximately 37 thousand of words, which appeared in the common corpus of English each year between 1750 and 2008 (the amount of annual text data was insufficient before that time). Russian words appeared in the corpus every year between 1920 and 2008 were selected. To avoid difficulties associated with the impact of the 1918 spelling reform, the analysis was performed for the stated period. To make the conditions of comparison more equal for both languages, Russian words, which appeared each year at least 10 times, were selected. There were 80 thousand of words, which satisfied the required conditions.

Figure 4 shows the comparison between the change in the effective number (see formulae (2, 3)) of syntactic dependencies of all words and words, which belong to the lexicon core. The number of syntactic dependencies between words from the core grows much slower than that between all words. At that, the number of syntactic dependencies between core words has not grown since 1850. However, a small increase is observed only after 1960. Thus, the growth in the number of syntactic dependencies is largely due to the emergence of new syntactic dependencies for words from the lexicon periphery, as well as syntactic dependencies between words from the lexicon core and periphery.

Let us further consider how the total number of syntactic dependencies and 2-grams varies depending on the number of words in the lexicon. Assuming the validity of Heaps' law for both the number of words and the number of syntactic dependencies, it can be said that there should be power dependence between these quantities. Figure 5 shows the change in the number of syntactic dependencies and 2-grams depending on the number of unique words in English and Russian. Each point on the graph corresponds to the number of words and the number of syntactic dependencies

(2-grams) detected in the corpus in a given year (in the period 1505–2008 for the English language and 1607–2009 for the Russian language).

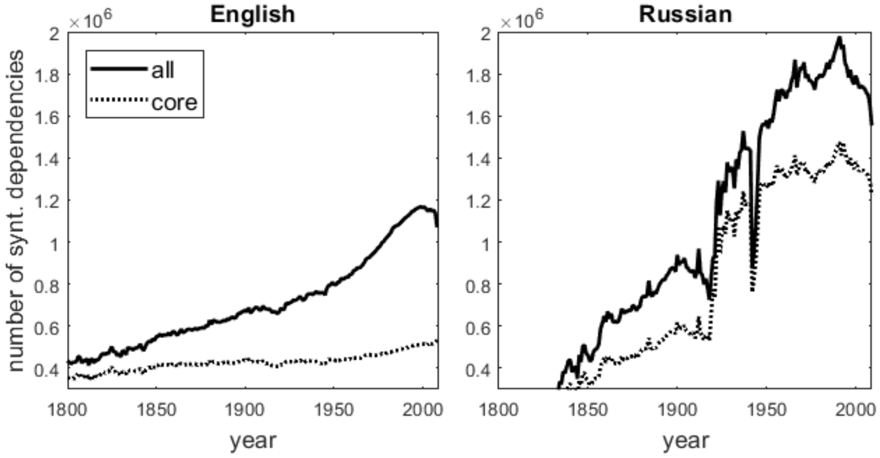

Fig. 4. Effective number of syntactic dependencies in English and Russian (both for the entire lexicon, and for the lexicon core).

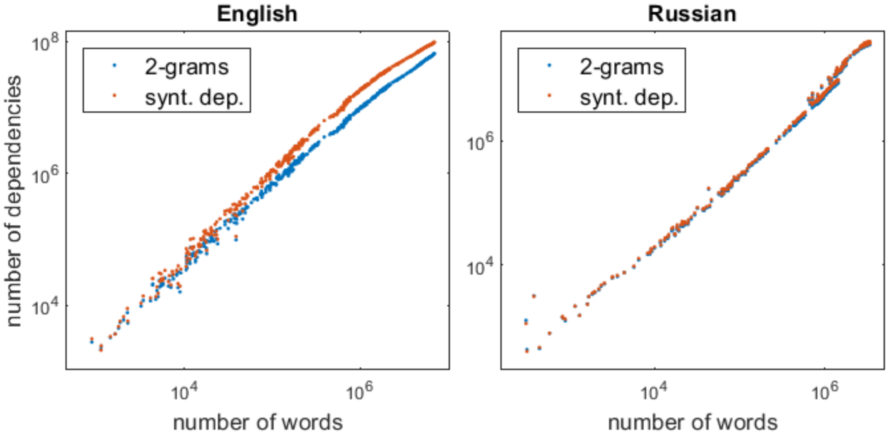

Fig. 5. Dependence of the number of syntactic dependencies and 2-grams detected in the corpus on the number of words in the lexicon.

Dependences shown in Figure 5 are close to a power law, however, differences are also observed. It can be seen that the slope of the graph slightly differs in different areas. These differences may be due to variations of Heaps' exponent with time described in [14]. Performing approximation of the empirical data by a power law on the most important area (for the number of words more than $1.5 \cdot 10^6$), we obtain the value of the power exponent for syntactic dependencies in the English language is equal to 1.174 (for word combinations - 1.169). That is, the number of syntactic dependencies per word grows slowly as the language becomes more complex. However, if we restrict ourselves only to stable syntactic dependencies with MI> 0, the power exponent for the number of syntactic dependencies will be 0.793 (0.815 for word combinations). Thus, the number of stable syntactic dependencies and word combinations per word falls. In both cases, the difference in the values of the power exponent for the number of syntactic dependencies and the number of word combinations is not significant. The difference of the power exponents from 1 is small, however, it can be important, since many growth models of complex networks predict proportionality of the number of network vertices (in our case, vertices are words) and the number of dependencies (in our case, syntactic dependencies) [16].

As for the Russian language, the power exponent for the number of syntactic dependencies is 1.097 (1.11 for the number of phrases) and equals 0.955 for the number of stable syntactic links with MI> 0 (0.96 for the number of stable phrases) under similar conditions. It should be noted that it is more difficult to find a linear segment for the Russian language in Figure 5. Therefore, these results are less reliable. Nevertheless, they are in good agreement with the estimates obtained for the English corpus.

Let us estimate quantitatively the degree of statistical connection between the number of unique words and the number of syntactic dependencies. Table 2 shows the Spearman correlation coefficients between these values for the English and Russian languages.

Table 2. Spearman's correlation coefficient between the number of unique words and the number of syntactic dependencies in Russian and English.

	English	**Russian**
Total number of syntactic dependencies	0.999	0.981
Number of syntactic dependencies with MI>0	0.994	0.983

Comparing with the values given in table 1, it can be seen that the statistical connection between the number of syntactic dependencies and the number of words is even more significant than connection with the corpus size. A more significant increase is observed for the English language. This may be due to the fact that the saturation effect described in [14] (which is more pronounced for a larger English corpus) weakens the dependence of the number of syntactic dependencies and 2-grams on the corpus size.

If the corpus stays unchanged, the number of syntactic dependencies changes in the following way. The number of books represented in the Google Books Ngram Russian sub-corpus varies greatly in different years. The largest amount of books belongs to the period 1960-1991. From 65 to 80 thousand of volumes were published annually in the USSR in this period, and the corpus contains approximately 10 thousand volumes published each year (or 1-1.25 billion words), that is, at least 12% of all published books. Thus, there is a 31-year time period during which the size of the corpus varied within small limits. This provides an opportunity to assess the rate of growth of the number of syntactic dependencies directly, without taking into account the impact associated with the growth of the corpus size.

Figure 6 shows the change in the effective number of syntactic dependencies in the Russian language in the target period. The dotted line shows approximation of the series of the number of syntactic dependencies by exponential dependence using only the data from the period 1960-1990. The exponent rate was $2.74 \cdot 10^{-3}$, which corresponds to a doubling of the effective number of syntactic dependencies within 253 years.

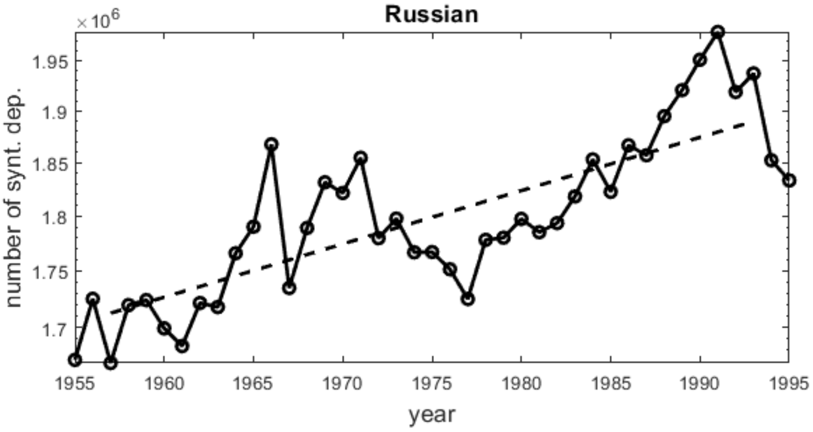

Fig. 6. Change in the effective number of syntactic dependencies in the Russian language in 1955-1995.

There is no period when the English language corpus size changes insignificantly. Nevertheless, if we approximate the curve of the effective number of syntactic dependencies in English in the same interval 1960-1991, the value of the exponent will be $9.36 \cdot 10^{-3}$, which corresponds to doubling of the number of syntactic dependencies within 74 years. If we take the 1850-1950 data (see Figure 4), the exponent will be estimated as $3.49 \cdot 10^{-3}$, which corresponds to doubling of the number of syntactic dependencies within 199 years. The latter value is close enough to the above estimate obtained for the Russian language.

4 Conclusion

The number of 2-grams and syntactic dependencies detected in the Google Books Ngram corpus grows extremely rapidly. It increased by a factor of 160 for the common corpus of English and by a factor of 66 for the Russian corpus over the period 1800-2000. It is obvious that most of this growth is associated not with increase of language complexity, but with an extensive increase of the corpus size. To study the factors causing language complexity, it is more convenient to use not the total number of syntactic dependencies and 2-grams, but the number of stable syntactic dependencies and 2-grams (with MI above a given threshold) or their effective number (calculated as perplexity of frequency distribution). The latter characteristic demonstrates much smoother and regular change compared to the total number of the studied word relationships. The curve of the effective number of syntactic dependencies and 2-grams practically does not respond to historical events and, when calculated using the entire English vocabulary, it shows growth, according to the law close to exponential. However, the effective number of syntactic dependencies and 2-grams detected in the corpus each year over a fairly long time interval (1750–2008 for English and 1920–2008 for Russian) changes very slowly. This can indicate that quantitatively unchanged core and rapidly growing periphery can be distinguished among the syntactic dependencies of words.

It was found that the effects associated with the emergence of new words dominate among the factors influencing the growth in the number of syntactic dependencies and 2-grams. The dependence of the total number of syntactic dependencies and 2-grams on the number of unique words is close to a power law. It is clear that the power law should be considered only as some approximation of the empirical data. However, it should be noted that the power dependence in this case corresponds better to the empirical data than to the dependence of the number of syntactic dependencies and 2-grams on the corpus size (which is expected in accordance with Heaps' law). The same is true for the number of stable dependencies (with MI> 0). At that, the power exponents are slightly greater than 1 (1.1-1.17) for the total number of syntactic dependencies and 2-grams and less than 1 (0.79-0.96) for the number of only stable syntactic dependencies and 2-grams for the studied languages. These facts should be taken into account when building models of growth of a network of syntactic dependencies in natural languages.

It was possible to obtain an estimate of the growth rate of the effective number of syntactic dependencies in the Russian language. If the corpus size stays unchanged, doubling of the effective number of syntactic dependencies should occur in 250 years. The effective number of syntactic dependencies in the English language is characterized by similar growth rates over a long period of time. However, their number increases approximately after 1950. This can be due to the fact that English is a global language.

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, grant agreement № 34.5517.2017/6.7 and by RFBR, grant № 17-29-09163.

References

1. Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., et al.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176-182 (2011).
2. Lin, Y., Michel, J.-B., Aiden, E.L., Orwant, J., Brockman, W., Petrov, S.: Syntactic Annotations for the Google Books Ngram Corpus. In: Li, H., Lin, C.-Y., Osborne, M., Lee, G.G., Park, J.C. (eds.) 50th Annual Meeting of the Association for Computational Linguistics 2012, Proceedings of the Conference, vol. 2, 238-242. Association for Computational Linguistics, Jeju Island, Korea (2012).

3. Juola, P.: Using the Google N-Gram corpus to measure cultural complexity. Literary and Linguistic Computing 28(4), 668–675 (2013).
4. Gerlach, M., Altmann, E.G.: Stochastic Model for the Vocabulary Growth in Natural Languages. Physical Review X 10(3), 021006 (2013).
5. Bochkarev, V.V., Lerner, E.Yu., Shevlyakova, A.V.: Deviations in the Zipf and Heaps laws in natural languages. Journal of Physics: Conference Series 490(1), 012009 (2014).
6. Williams, J. R., Lessard, P.R., Desu, S., Clark, E.M., Bagrow, J.P., Danforth, C.M., Dodds, P.: Zipf's law holds for phrases, not words. Scientific Reports 5, 12209 (2015).
7. Chick, G.: Cultural complexity: The concept and its measurement. Cross-Cultural Research 31, 275–307 (1997).
8. Arbib, M.A.: How the Brain Got Language: The Mirror System Hypothesis. Oxford University Press, Oxford (2012).
9. Bochkarev, V.V., Solovyev, V.D., Wichmann, S.: Universals versus historical contingencies in lexical evolution. J. R. Soc. Interface 11, 20140841 (2014).
10. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., et. al.: Syntactic dependency-based N-grams as classification features. LNAI 7630, 1-11 (2012).
11. Fano, R.M.: Transmission of Information: A Statistical Theory of Communications. M.I.T. Press, Cambridge Mass. (1961).
12. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990).
13. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Lai, J.C., Mercer, R.L.: An Estimate of an Upper Bound for the Entropy of English. Journal of Computational Linguistics 18(1), 31-40 (1992).
14. Solovyev, V.D., Bochkarev, V.V., Shevlyakova, A.V.: Dynamics of core of language vocabulary. CEUR Workshop Proceedings 1886, 122-129 (2016).
15. Buntinx, V., Bornet, C., Kaplan, F.: Studying Linguistic Changes over 200 Years of Newspapers through Resilient Words Analysis. Frontiers in Digital Humanities 4, 1-10 (2017).
16. Durrett, R.: Random Graph Dynamics. Cambridge University Press, Cambridge (2007).