Modelling of growth of syntactic relations network in English and Russian

Vladimir V Bochkarev, Anna V Shevlyakova, Eduard Yu Lerner
Kazan Federal University, Kremlevskaya 18, Kazan 420008, Russia
E-mail: vbochkarev@mail.ru

Abstract. Creation of the Google Books Ngram corpus opened up new opportunities for studying language evolution. This corpus consists of a large amount of digitized books written in 8 languages and contains information on frequencies of words, word combinations and syntactic relations of the last 500 years. In this paper, we present data on changes in the key characteristics of syntactic relations in English and Russian and propose a model which allows us to explain the observed changes. We used Google Books Ngram data (1800-2008) and performed modelling of network growth. Then, we compared the characteristics of the obtained model networks with the characteristics of the network of syntactic relations of the English and Russian languages. It was shown that selection of two parameters of the model allows us to obtain a very good correspondence between the changes in the clustering coefficient, the assortativity coefficient, and word distribution by the number of relations in the model network and in real networks of syntactic relations.

1. Introduction
In 2005, a popular article “Syntax for free” was published by Ricard Solé in the journal “Nature” [1]. The article considers networks representing bipartite graphs of words and their possible meanings. At that, the laws of mathematical linguistics (Zipf's and Heap’s laws) and the small distance between words in the context of synonymy are treated as a “free supplement” to the known laws of scale-free networks (the power distribution of the number of edges emanating from the vertices, the law of small worlds, the large cluster coefficient in such networks etc.). Standard models based on the principle of preferential attachments are regarded as models of such networks. Nowadays such attitude to the scale-free models of the development of syntactic networks seems irrelevant.

The fact is that "two-degree" character of Zipf's was noticed in the early 2000s (see, in particular, [2]). To solve this problem, Dorogovtsev S. and Mendes  J. proposed a model of preferential attachment [3], in which the number of edges incident to the vertex attached to the network can be an increasing function of time (the number of the vertex), and the distribution of the degrees of vertices is described by an analytic function with a two-degree asymptotics. To reconcile with the data of a particular novel, this model was refined and expanded in [4]. It considered a network with vertices (word forms from the text of the chapters of the novel) and all kinds of bigrams were considered as edges. It was noticed that the number of edges grows faster than it is provided for by the "classical" models of the preferred attachment. In a series of works by Polish scientists Kulig A, Drozdz S, Kwapien J, Oswiecimka P. (see [5]), a wide corpus of literary works was analysed (specific English, Polish, Russian and German classics). They draw a conclusion that the length of the average shortest
path between word forms does not increase with the increase in the volume of the text. This fact is not consistent with the most cited models based on the principle of preferential attachment and is poorly described by the Dorogovtsev-Mendes model. Finally, in numerous articles by various authors in the book [6], the attitude towards Ricardo Sole's predictions is very cautious, in fact, it is concluded that there are no grounds for his optimism in the relation of syntax as a "free supplement" to the known network models.

Our estimates of the parameters of the networks of syntactic relations between English and Russian are also difficult to explain from the point of view of existing models. In this paper, we propose a new model of network growth (close to the copying model [7]) and compare it with empirical data.

2. Data analysis
Creation of the Google Books Ngram corpus opened up new opportunities for studying language evolution [8]. This diachronic corpus consists of a large amount of digitized books written in 8 languages and contains information on frequencies of words, word combinations and syntactic relations of the last 500 years.

To study the structure and dynamics of the network of syntactic relations, we used the frequency databases for 100000 most frequent (on average for the whole period) English and Russian words. The analysis was limited to this number of frequent words in order to avoid the negative impact of misspelled words contained in the corpus. According to the data in [8], the corpus can contain about 30% of misspelled words, but random checks show that their number can be significantly larger than 30%. First, in order to provide a sufficient volume of statistics, we performed averaging of frequencies over 25-year intervals. In addition, because of the need for sufficient statistics, the analysis of the English language was limited to the period 1800-2008 and of the Russian language – to the period 1925-2008 (this choice allowed us to avoid difficulties associated with the spelling reform in 1918). Unlike the approach adopted in Google Books Ngram Viewer, we take into account only vocabulary 1-grams (consisting of letters of the corresponding alphabet and, possibly, one apostrophe) and regard the word forms as identical if they consist of the same letters but the capitalization is different. After that, we calculated the degrees of “incoming” and “outgoing” relations for each word (the "vertex" of the network), i.e. the number of different connections of the form W + X and X + W for the word W). Certain conclusions can be drawn already on the basis of the type of distribution of the degrees of the network vertices (many models lead to a power distribution). However, a more detailed analysis can be carried out using the assortativity coefficient (the correlation coefficient between the number of neighbouring vertices), the clustering coefficient (shows on average how strong are the relations between the neighbours of the vertex) [9], as well as the average length of the minimum path between the vertices and the diameter of the network. These values were calculated. At that, the average minimum path between the vertices was calculated by the Monte Carlo method because of the large complexity of direct calculation.

The results of the calculations are shown in Tables 1, 2.

Table 1. Parameters of the network of syntactic relations for the English language

| Period  | Clustering coefficient | Assortativity coefficient | Average minimal path | Network diameter |
|---------|------------------------|---------------------------|----------------------|-----------------|
| 1800-1825 | 0.7680                 | -0.2415                  | 2.1305               | 5               |
| 1825-1850 | 0.7695                 | -0.2543                  | 2.0634               | 5               |
| 1850-1875 | 0.7667                 | -0.2622                  | 2.0512               | 4               |
| 1875-1900 | 0.7594                 | -0.2668                  | 2.0390               | 4               |
| 1900-1925 | 0.7517                 | -0.2714                  | 2.0147               | 4               |
| 1925-1950 | 0.7391                 | -0.2711                  | 2.0068               | 4               |
| 1950-1975 | 0.7203                 | -0.2737                  | 1.9948               | 4               |
Table 2. Parameters of the network of syntactic relations for the Russian language

| Period   | Clustering coefficient | Assortativity coefficient | Average minimal path | Network diameter |
|----------|------------------------|---------------------------|----------------------|-----------------|
| 1925-1950 | 0.4242                 | -0.1287                   | 1.9991               | 4               |
| 1950-1975 | 0.4056                 | -0.1261                   | 1.9976               | 3               |
| 1975-2000 | 0.4027                 | -0.1254                   | 1.9963               | 3               |
| 2000-2008 | 0.4118                 | -0.1243                   | 2.0012               | 3               |

Small value of the first two parameters indicates that any proposed model of network development should be a "small world" model. A sufficiently large value of the clustering coefficient is also typical of real networks (this property, by the way, is not fulfilled in the Albert-Barabashi model). It should be noted that all characteristics vary insignificantly with time and, thus, depend little on the size of the network (in the classical models of the preferred attachment, there is a logarithmic increase in the average length of the shortest path).

3. Model description and Simulation Results

A new model of network growth was proposed and tested based on the analysis of the obtained data. Constructing the model, we took into account the prototype theory established by E. Rosch [10]. She suggested that thinking as a whole is based on prototypes and structures of the basic level. Prototype is an abstract image, which is more central than other members of a certain category and takes a salient position in its formation.

The addition of new words in the process of network growth is modelled in three stages:

- random selection of one or more words as a prototype;
- relations of the word-prototype are inherited by a new word with some probabilities;
- a certain number of relations, which the word-prototype doesn’t have, are added at random.

Each of these stages can be realised differently. We considered the simplest variant.

- At the first stage, the word-prototype is selected with equal probability from the existing words;
- probabilities of inheritance of any of the relations of the word-prototype are equal;
- average number of randomly added relations is proportional to the number of relations of the word-prototype.

Thus, two parameters must be specified in this implementation of the model. The key parameter, the probability of inheritance of relations from the word-prototype α, can be estimated directly from the empirical data. To perform this analysis, 15679 words were selected from 100000 most frequent English words, which first appeared in the corpus in 1800 and later. We tried to find an alleged prototype for each of the selected words. A prototype word is understood as a word closest to the given word in terms of frequency distribution of its relations, which appeared in the corpus before the analysed word. Each word can be characterized by a vector of frequencies and using this or that metric, we can select one or several words closest to the given one. In this work, we used the Kullback–Leibler divergence and Jacquard distance. After that, we calculated what proportion of relations of the word-prototype the analysed word enters. The sample average value of this parameter was 75.6%. The second parameter β is the coefficient of proportionality between the number of relations of the prototype-word and the number of randomly added relations. This parameter is selected according to the simulation results.
In [7], the copying model was proposed. This model is also described by two parameters – α and β. However, the first parameter (depending on t) in this paper characterizes the Bernoulli probability of the birth of a new vertex at time t. The prototype relations are copied with a fixed probability of 1-β, while the edge is directed to a random vertex with a probability β. That is, the sum of the copying probabilities and the formation of a relation with a random vertex in standard copying models is equal to 1. Due to this fact, the number of the graph edges satisfies the asymptotic power law. In the proposed model, there is no such relation between the copying probabilities and the formation of a relation with a random vertex (α and β). The objective is to reconcile everything not with the conditions of the theorem on the asymptotic power law of the number of the graph edges but with real Google Books Ngram data.

We performed modelling of growth by setting words, which had appeared in the Google Books Ngram English corpus at least once by 1800 (84321 words), and their syntactic relations in the initial state of the network. The growth was simulated until the total number of relations in the network reached 170 million (the number of relations recorded in the Google Books Ngram corpus by 2008). The clustering coefficient and coefficient of assortativity were calculated for the obtained networks. The growth modelling was repeated many times using different values of the parameter β. In particular, if β = 0.2, we obtain the value of the coefficient of assortativity in the final state of the network, which is equal to -0.275 and corresponds well to the empirical values. The final value of the clustering coefficient is 0.677 under these conditions, which is close to the empirical value.

4. Conclusion
In this paper, we study the structure and dynamics of the network of syntactic relations in English and Russian using the Google Books Ngram corpus. The clustering coefficient for the networks of syntactic relations of both languages is quite large. It is 0.70-0.77 for English (in 1800-2008) and 0.42-0.44 for Russian (in 1925-2008). The assortativity coefficient is negative in both cases, and varies from -0.27 to -0.24 for English and from -0.129 to -0.124 for Russian. A new model of network growth is proposed and its predictions are compared with the empirical data. The results of the network growth modelling show that selection of model parameters allows us to obtain a good correspondence between the changes in the cluster coefficient and the assortativity coefficient in the model network and in the real network of syntactic relations.

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] Solé R V 2005 Nature 434 289
[2] Ferrer R and Solé R V 2001 Santa Fe Institute working papers 01-03-016
[3] Dorogovtsev S N and Mendes J F F 2001 Proceedings of the Royal Society of London. Series B: Biological Sciences 268 (1485) 2603-2606
[4] Masucci A P and Rodgers G J 2006 Physical Review E 74 (2) 026102
[5] Kulig A, Drozdz S, Kwapien J and Owsiencimka P 2015 Phys Rev E 91 032810
[6] Job B, Blanchard P, Lücking A, Mehler A and Banisch S eds. 2016 Towards a Theoretical Framework for Analyzing Complex Linguistic Networks (Berlin: Springer)
[7] Kleinberg J M, Kumar R, Raghavan P, Rajagopalan S and Tomkins A 1999 Lecture Notes in Computer Science 1627 1-17
[8] Michel J-B, et al. 2010 Science 331, 176–182
[9] Newman M E J 2003 SIAM Review 45 (2) 167
[10] Rosch E H 1973 Cognitive Psychology 4, 328-350
[11] Padó S and Lapata M 2007 Computational Linguistics. 33 (2) 161