Vertex Degree Distribution for the Graph of Word Co-Occurrences in Russian

Victor Kapustin
Faculty of Philology
Saint-Petersburg State University
Saint-Petersburg, Russia 199178
vak@icape.nw.ru

Anna Jamsen
Faculty of Philology
Saint-Petersburg State University
Saint-Petersburg, Russia 199178
anna_zheleznova@mail.ru

Abstract

Degree distributions for word forms co-occurrences for large Russian text collections are obtained. Two power laws fit the distributions pretty good, thus supporting Dorogovtsev-Mendes model for Russian. Few different Russian text collections were studied, and statistical errors are shown to be negligible. The model exponents for Russian are found to differ from those for English, the difference probably being due to the difference in the collections structure. On the contrary, the estimated size of the supposed kernel lexicon appeared to be almost the same for the both languages, thus supporting the idea of importance of word forms for a perceptual lexicon of a human.

1 Introduction

Few years ago Ferrer and Solé (2001a) draw the attention of researchers to the fact that the lexicon of a big corpus (British National Corpus – BNC –in the case) most probably consists of two major components: a compact kernel lexicon of about 10^3 – 10^4 words, and a cloud of all other words. Ferrer and Solé studied word co-occurrence in BNC in (2001b). Two word forms[1] in BNC were considered as "interacting" when they appeared in the same sentence and the words' distance didn't exceed 2. Ferrer and Solé (2001b) treated also some other notions of word interaction, but the results obtained don't differ qualitatively. The interacting words form a graph, where the vertices are the words themselves, and the edges are the words' co-occurrences. The fact of the collocation considered to be important, not the number of collocations of the same pair of words. Ferrer and Solé (2001b) studied vertices degree distribution and found two power laws for that distribution with a crossover at a degree approximately corresponding to the previously found size of the supposed kernel lexicon of about 10^3 – 10^4 words. In (Solé et al, 2005) word co-occurrence networks were studied for small (about 10^4 lines of text) corpora of English, Basque, and Russian. The authors claim the same two-regime word degree distribution behavior for all the languages.

Dorogovtsev and Mendes (2001, 2003: 151-156) offered an abstract model of language evolution, which provides for two power laws for word degree distribution with almost no fitting, and also explains that the volume of the region of large degrees (the kernel lexicon) is almost independent of the corpus volume. Difference between word (lemma) and word form for an analytic language (e.g. English) seems to be small. Dorogovtsev-Mendes model certainly treats word forms, not lemmas, as vertices in a corpus graph. Is it really true for inflecting languages like Russian? Many researchers consider a word form, not a word (lemma) be a perceptual lexicon unit (Zasorina, 1977; Ventsov and Kassevich, 1998; Verbitskaya et. al., 2003; Ventsov et. al., 2003). So a hypothesis that word forms in a corpus of an inflecting language should exhibit degree distribution similar to that of BNC looks appealing. An attempt to investigate word frequency rank sta-

[1] Strictly speaking, word forms, not words.

TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pages 89–92,
Rochester, April 2007

tistics for Russian was made by Gelbukh and Sidorov (2001), but they studied only Zipf law on too small texts to reveal the kernel lexicon effects. To study the hypothesis one needs a corpus or a collection[2] of texts comparable in volume with the BNC part that was examined in (Ferrer and Solé, 2001b), i.e. about 4.10^7 word occurrences. Certainly, texts that were analyzed in (Solé et al, 2005) were much smaller.

Recently Kapustin and Jamsen (2006) and Kapustin (2006) studied a big (~5.10^7 word occurrences) collection of Russian. The collection exhibited power law behavior similar to that of BNC except that the vertex degree at the crossover point and the average degree were about 4-5 times less than that of BNC. These differences could be assigned either to a collection nature (legislation texts specifics) or to the properties of the (Russian) language itself. We shall reference the collection studied in (Kapustin and Jamsen, 2006; Kapustin, 2006) as "**RuLegal**".

In this paper we present a study of another big collection of Russian texts. We have found that degree distributions (for different big sub-collections) are similar to those of BNC and of RuLegal. While the exponents and the kernel lexicon size are also similar to those of BNC, the average degree for these collections are almost twice less than the average degree of BNC, and the nature of this difference is unclear still.

The rest of the paper has the following structure. **Technology** section briefly describes the collection and the procedures of building of co-occurrence graph and of calculation of exponents of power laws. In **Discussion** section we compare the results obtained with those of Kapustin and Jamsen (2006), Kapustin (2006), and (Ferrer and Solé, 2001b). In **Conclusion** some considerations for future research are discussed.

2 Technology

At present Russian National Corpus is unavailable for bulk statistical research due to copyright considerations. So we bought a CD ("World Literature in Russian") in a bookstore – a collection of fiction translations to Russian. We'll call the collection **WLR**. The size of the whole collection is more than 10^8 word occurrences. The source format of the collection is HTML, but its files contain essentially no formatting, just plain paragraphs. We made three non-overlapping samples from WLR (WLR1–3). The samples were approximately of the same size. Each sample was processed the same way. The idea behind using more than one sample was to estimate statistical errors.

We used open source Russian grapheme analysis module (Sokirko, 2001) to strip HTML and to split the texts into words and sentences. Word co-occurrences were defined as in (Ferrer and Solé, 2001b): two words are "interacting" if and only if they: (a) appear in the same sentence, and (b) the word distance is either 1 (adjacent words) or 2 (one word or a number or a date in-between). A found co-occurred pair of words was tried out against MySQL database of recorded word pairs, and if it wasn't found in the database, it was put there. Then we use a simple SQL query to get a table of count of vertices **p(k)** vs. vertex degree **k**.

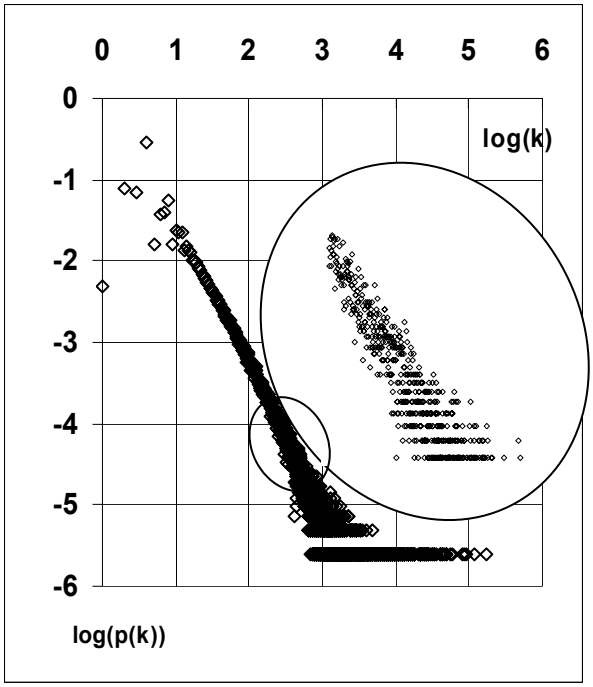

Figure 1. Raw degree distribution for WLR1.

The raw results for one of the samples are shown on Fig. 1. For the two other samples the distributions are similar. All distributions are almost linear (in log-log coordinates, that means that they obey power law), but fitting is impossible due to high fluctuations. As noted by Dorogovtsev and Mendes (2003: 222-223), cumulative distribution

[2] We consider a corpus to be a special type of a text collection, which comprises text samples chosen for language research purposes, while a more general term "collection " refers to a set of full texts brought together for some other purpose.

$P(k) = \Sigma_{K \geq k}\ p(K)$ fluctuates much less, so we calculated the cumulative degree distributions (Fig.2). Cumulative degree distributions for all three WLR samples are very similar.

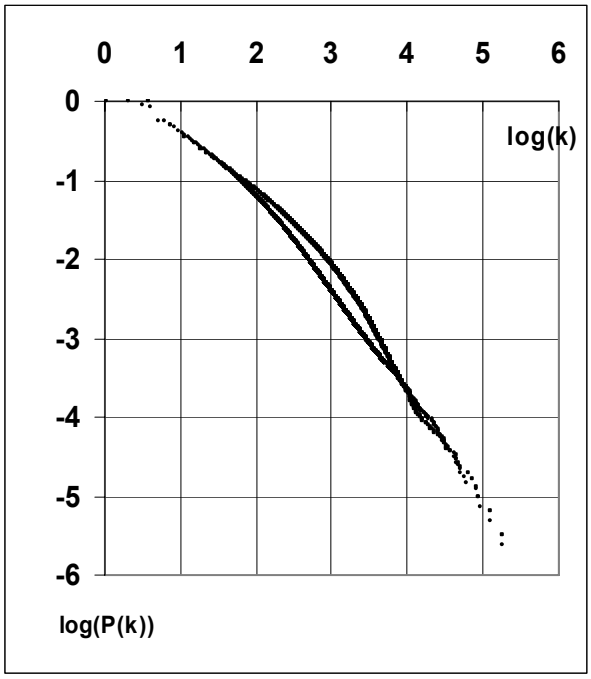

Figure 2. Cumulative degree distributions for WLR1 (lower curve) and RuLegal (upper curve).

3 Discussion

To estimate statistical errors we have normalized the distributions to make them comparable: the degree ranges were reduced to the largest one, then the cumulative degree distribution was sampled with the step of 1/3, as in (Ferrer and Solé, 2001a, Dorogovtsev and Mendes, 2003: 222-223). When we use WLR samples only, the statistical errors are less than 7% in the middle of the curves and reach a margin of 77% in small degrees region. With the inclusion of RuLegal sample, difference between samples becomes larger – up to 13% in the middle of the curves), but are still small enough.

In both cases (with and without RuLegal) we attempted to fit either a single power law (a straight line in log-log coordinates) or two/three power laws with one/two crossover points. Strong changes and large statistical errors of the distributions in the low degree region prevent meaningful usage of these points for fitting. We have made attempts to fit all three approximations for all points, and omitting one or two points with the lowest degrees. To choose between the hypotheses we minimized Schwarz information criterion (Schwarz, 1978):

$$SIC = N * \ln\left(\sum_i (p_i - \hat{p}_i)^2 / N\right) - m * \ln(N)$$

where p_i – cumulative distribution at i-th point;
$\hat{p}_i$ – fitting law at the same point;
N – number of sampling points (13–15, depending on the number of omitted points);
m – number of fitting parameters (2, 4 or 6)

Omitted points	SIC (1/2/3 power laws)	
	WLR1–3	WLR1–3 + RuLegal
0	–44 / –85 / –68	–42 / –93 / –77
1	–46 / **–85** / –65	–43 / **–93** / –73
2	–47 / –80 / –60	–44 / –86 / –67

Table 1. Fitting power laws to averaged degree distributions – Schwarz information criterion

	WLR1–3	WLR1–3 + RuLegal	RuLegal	BNC
γ_1	–0.95	–0.95	–0.95	–0.5
γ_2	–1.44	–1.46	–1.75	–1.7
k_{cross}	670	670	510	2000
V_{kernel}	4.10^3	4.10^3	4.10^3	5.10^3
$k_{average}$	36	31	15	72
Collection size	3.10^7	14.10^7	5.10^7	4.10^7

Table 2. Parameters of the best fit two power laws for the cumulative distributions

Clearly two power laws fit the curves better. The exponents, the crossover degree and estimated size of the kernel lexicon (number of vertices with high degrees above the crossover) for the best fits (two powers, zero/one omitted point) are shown in Table 2. The exponents for the raw distributions are γ_1 and γ_2 minus 1.

Disagreement between English and Russian seems to exist. Probably, the differences are still due to the collections' nature (the difference between different Russian collections is noticeable).

4 Conclusion

We found that ergodic hypothesis for word form degree distribution seems to work for large text collections – differences between the distributions

are small (except for the few smallest degrees). At least, a single big enough sample permits reliable calculation of degree distribution parameters.

Dorogovtsev-Mendes model, which yields two power laws for the degree distribution for the word forms graph, gives pretty good explanation both for an analytic language (English) and for an inflecting one (Russian), though numeric parameters for both languages differ. The estimated sizes of the supposed kernel lexicons for the both languages are almost the same, the fact supports the point that word form is a perceptual lexicon unit.

To make more rigorous statements concerning statistical properties of various languages, we plan to calculate other important characteristics of the co-occurrence graph for Russian: clustering coefficient and average shortest path. Also we hope that legal obstacles to Russian National Corpus usage will have been overcome. Other statistical language graph studies are also interesting; among them are investigation of networks of lemmas, and statistical research of agglutinated languages.

Acknowledgements

The authors are grateful to the anonymous reviewers, the comments of whom were of much help.

The work is supported in part by Russian Foundation for Basic Research, grants 06-06-80434 and 06-06-80251.

References

Sergey N. Dorogovtsev., José F. Mendes, 2001. Language as an evolving word web. *Proceedings of the Royal Society of London B*, 268(1485): 2603-2606

Sergey N. Dorogovtsev., José F. Mendes, 2003. Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford.

Ramon Ferrer and Ricard V. Solé. 2001a. Two regimes in the frequency of words and the origin of complex lexicons. *Journal of Quantitative Linguistics* 8: 165-173.

Ramon Ferrer and Ricard V. Solé. 2001b. The Small-World of Human Language. *Proceedings of the Royal Society of London B*, 268(1485): 2261-2266

Victor Kapustin, 2006. Капустин В.А. Ранговые статистики совместной встречаемости словоформ в большой монотематической коллекции. Труды третьей международной конференции «Корпусная лингвистика», 11–13 октября 2006 г., – СПб.: Изд-во С. Петерб. ун-та, 2006. – С. 135-142 (*Rank Statistics of Word Co-Occurrences in a Big Monothematic Collection*. Proc. 3rd International Conf. "Corpus Linguistics", Oct. 11-13, 2006. Saint-Petersburg State Publishing: 135-142).

Alexander Gelbukh and Grigory Sidorov, 2001. *Zipf and Heaps Laws' Coefficients Depend on Language*.. Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics (February 18–24, 2001, Mexico City), Lecture Notes in Computer Science, Springer-Verlag. (2004): 332-335. (ISSN 0302-9743, ISBN 3-540-41687-0)

Victor Kapustin and Anna Jamsen. 2006. Ранговая статистика встречаемости слов в большой текстовой коллекции. Труды 8ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2006, Суздаль, Россия, 2006. – С. 245–251 (*Rank Statistics of Word Occurrence in a Big Text Collection*. Proc. 8th National Russian Research Conference "Digital libraries: advanced methods and technologies, digital collections", Oct. 17-19, 2006: 245-251).

Alexey Sokirko, 2001. *A short description of Dialing Project*. http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html

Ricard V. Solé, Bernat Corominas, Sergi Valverde and Luc Steels. 2005. Language Networks: their structure, function and evolution. SFI-WP 05-12-042, SFI Working Papers

Gideon Schwarz, 1978. *Estimating the dimension of a model*. Annals of Statistics 6(2): 461-464.

Anatoly V. Ventsov and Vadim B. Kassevich, 1998. Венцов А.В., Касевич В.Б. *Словарь для модели восприятия речи*. Вестник Санкт-Петербургского университета, сер. 2, вып. 3, с. 32-39 (*A Dictionary for a Speech Perception Model*. Vestnik Sankt-Peterbugskogo Universiteta, 2(3): 32-39).

Anatoly V. Ventsov, Vadim B. Kassevich and Elena V. Yagounova, 2003. Венцов А.В., Касевич В.Б., Ягунова Е.В. *Корпус русского языка и восприятие речи*. Научно-техническая информация.– Серия 2, № 6, с.25-32 (*Russian Corpus and Speech Perception*. Research and Technical Information, 2(6): 25-32).

Liudmila A. Verbitskaya, Nikolay N. Kazansky and Vadim B. Kassevich, 2003. Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка. Научно-техническая информация.– Серия 2, № 5, с.2-8 (*On Some Problems of Russian National Corpus Development*. Research and Technical Information, 2(5): 2-8).

Lidia N. Zasorina, ed., 1977. *Частотный словарь русского языка*. Под ред. Л.Н. Засориной. М.: Русск. яз. (*Frequency Dictionary of Russian*. Russian Language, Moscow, 1977).