A Language as a Self-Organized Critical System

Vasilii A. Gromov and Anastasia M. Migrina

School of Applied Mathematics, Oles Honchar Dnipropetrovsk National University, Gagarina Av. 72, Dnipropetrovsk 49010, Ukraine

Correspondence should be addressed to Vasilii A. Gromov; stroller@rambler.ru

Received 2 May 2017; Revised 3 September 2017; Accepted 31 October 2017; Published 19 November 2017

Copyright © 2017 Vasilii A. Gromov and Anastasia M. Migrina. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A natural language (represented by texts generated by native speakers) is considered as a complex system, and the type thereof to which natural languages belong is ascertained. Namely, the authors hypothesize that a language is a self-organized critical system and that the texts of a language are “avalanches” flowing down its word cooccurrence graph. The respective statistical characteristics for distributions of the number of words in the texts of English and Russian languages are calculated; the samples were constructed on the basis of corpora of literary texts and of a set of social media messages (as a substitution to the oral speech). The analysis found that the number of words in the texts obeys power-law distribution.

1. Introduction

Since natural languages gradually came to be regarded as complex systems, a means to study linguistic processes changed from descriptive approaches to formal analysis aiming to construct mathematical model for the operation and the development of language—this change is both a reason and an effect of “the linguistic turn” (the term is due to [1]).

In the frameworks of this approach, it is possible to pursue two main avenues of inquiry: The first one is aimed at constructing a theory of natural languages grammars [2]. The second one is associated with analysis of language statistical characteristics: Primary emphasis is placed, from seminal studies by Zipf [3], on distributions for separate words for written language [4–8] and for various graphs reflecting language features [9]. The principal result of these studies is a class of distributions describing natural language features [10]; the class comprises power (heavy-tail) distributions with various values of exponents: power laws manifest themselves in word’s frequency in language [3], in syntactic networks [5–7, 11], in frequency of letter sequences in vocabularies [8], and so forth.

Meanwhile, a shift in XX century philosophy (akin to the Copernican revolution in astronomy) suggests considering a language as an integral whole: “a man is just a place where a language speaks itself” [12]. Therefore, since above all else a language is a unified communicative tool to convey the meanings and its text (either a literary work or a tweet) is usually a meaningful, complete message, a text becomes a basic unit for such analysis.

The authors hypothesize that a natural language is a self-organized critical system (SOC) [13, 14] and the texts of a language are “avalanches” (as those are defined by Bak) flowing down the word cooccurrence graph of the respective language; large avalanches correspond to literary works, while smaller ones are associated with messages from social media. It is worth noting that a self-organized critical system conventionally features [13, 14]

1. a space of elements able to be in two states, active and passive, along with a set of rules to describe how a change of state of an element affects states of the other ones,
2. the “avalanches” in the space that are chain reactions of elements’ state changes triggered by changes of other elements,
3. a power-law distribution governing avalanche sizes.

For a language system, a semantic space plays a part of the space at issue, and the rules are reduced to syntactic

Hindawi
Complexity
Volume 2017, Article ID 9212538, 7 pages
https://doi.org/10.1155/2017/9212538
and semantic rules of the respective language. In the present study the space was formalized as a cooccurrence graph—the authors are aware that one-to-one correspondence between semantic space and vocabulary is absent, but they assume that the latter approximates the former somehow. Vertices of the graph correspond to words and an edge is present if and only if the words associated with its incident vertices occur simultaneously in the same text of the sample involved, once or more. As indicated above, an avalanche, in this context, is a text of a language, and the hypothesis that sizes of avalanches obey a power-law distribution forms the subject of the present study. One should mention that real-world (unfolding over time) SOC-systems usually exhibit long periods of slow evolution as opposed to short periods of fast evolution when the system space is changed drastically; similar phenomenon is reported to take place for evolving language systems.

Another point of interest here is the emergence of a gigantic volume of information reflecting, in essence, spoken language (parole as opposed to langue according to Saussurean terminology [15], (second) Orality versus Literacy [16]) that is texts posted by users in social networks of every sort and kind (Facebook, Reddit, and so on); this makes it possible to explore this domain of human communication. In this context, the primary problem is to compare statistical characteristics calculated by means of corpus of literary texts, on the one hand, and by means of a set of texts written by social networks users. If these characteristics appear to be statistically equal, this may give proof to the idea of language unity as a complex system; and if so, written and spoken language are merely different projections of internal dynamics of this complex system (synchronic unity of language). On the other hand, the characteristic at issue calculated for different time periods (one is to be restricted in this case to the analysis of written language) can be compared in order to verify unity of the linguistic system unfolding in time (diachronic unity of language). The present work is focused on the study (and comparison) of statistical characteristics of texts sets for English and Russian languages being considered in their synchronic and diachronic aspects.

The first paper to be cited among recent studies of distributions (mainly power-law distribution) observed for oral and written language is a brilliant review [17]. Here, the object of study is a particular text, and its basic unit is a word or a sentence—distributions of characteristics of these objects form the subjects of the overwhelming majority of papers for this line of investigation [18]. For example, Font-Clos et al. [19] examine dependence text length versus statistical properties of word occurrences; the authors reveal that the distribution obeys the power and investigate its relationship with Zipf’s and Heaps’ (Herdan’s) laws (the former states that the vocabulary grows as a power function of text’s length) [20].

Another object of study that generates power-law distributions is a representation of language semantic and syntactic relations using various discrete structures: semantic nets [17], global syntactic dependency trees [21, 22], cooccurrence graphs [18], and others. Both conventional methods aimed at exploring these structures as complex networks [23] and random walks on these structures result in power-law distributions [17, 21, 22, 24, 25]. A basic unit is a word or a sentence likewise.

The subject of the present paper is language as a whole; texts (semantic “avalanches”) are considered as its basic units. The rest of the paper is organized as follows. The next section outlines methods used to estimate distribution parameters; the third provides results for both English and Russian languages. The fourth section discusses results; finally, the last section presents conclusions.

2. Methods

The choice of the languages is determined, apart from the availability of voluminous corpora of texts (for written and spoken languages) for them, by their qualitative difference in grammar structure: Russian is an inflected language, while in English inflections are rather rare; Russian is characterized by flexible word order in a sentence, whereas word order in English language is strict, and rare exceptions are constrained by stringent rules [26, 27].

We used two different approaches to test statistical hypothesis in question. The first one utilizing the concept of data collapse is considered in greater detail in the monograph by Prusssner [14]; the second one (grounded on the Kolmogorov-Smirnov (KS) criterion) is proposed in the source [28]. Both methods not only evaluate the exponents but also cut off the smallest elements of the sample that usually do not fit a power-law. The first method also cuts off the largest nonfitting elements. It is worth noting that this phenomenon (the sharp distinction between the largest [smallest] elements and all others) seems to be a salient characteristic of real-world data following power-law distributions [13, 14].

The first approach we dwell on briefly is that using the concept of data collapse [14]. It assumes that the distribution generating the data has the following probability density function:

$$ p(x) = ax^{-\tau} g \left( \frac{x}{bL^D} \right) $$

with $x$ which can be a continuous or discrete random variable, metric factors $a$ and $b$, characteristic dimension of a system $L$, scaling function $g$, and scaling exponents $\tau$ and $D$ as well. The distribution follows the modified power-law within the interval bounded by the lower and upper cutoffs $x_{\text{min}}$ and $x_{\text{max}}$. The scaling function $g$ (that distinguishes the distribution from a canonical power-law) fits many real-world systems obeying heavy-tail distributions [14]. The quantity $bL^D$ determines the characteristic upper cutoff.

The raw data is taken to be previously binned, that is, grouped together and averaged over the observations belonging to the same group; we used the exponential binning for its suitability for this kind of data [14].

If the null hypothesis (that the sample under study is generated from the distribution equation (1)) is true, then $\overline{f}(x) x^{-\tau}$ plotted against $x = x/L (L)$, where $x(L) = bL^D$, gives the same function $ag(\xi)$ for various $L$, where $\overline{f}(x)$ is the empirical probability density function, and $\tau$ is the true
value of the power exponent. The phenomenon is given the title of data collapse. Thus, for given data generated from the power-law distribution, \( \tilde{P}(x) x^{-\tau} \) (as a function of \( \xi \)) plotted for various \( L \) is superimposed on each other.

Therefore, the respective goodness-of-fit test for power-law distribution involves the following steps:

1. (1) binning of the raw data

2. (2) plotting \( \tilde{P}(x) x^{-\tau} \) against \( \xi = x/x_c(L) \) for various \( L \) using "apparent exponent" \( \tilde{\tau} \) (rough estimate of \( \tau \)—such a plot comprises a nonhorizontal straight line and a characteristic nonlinear curve, whose extremum is called a landmark (\( x_c \) represents its coordinate)

3. (3) the refinement of the value with the employment of the least squares method applied to the landmarks.

As a result, the plots merge into a single horizontal straight line for the section of the domain of definition for which power-law holds true (between the lower and upper cutoffs).

The second approach [28] is applicable to power-law distribution without scaling function:

\[
p(x) = \begin{cases} 
0, & x \leq x_{\text{min}} \\
ax^{-\tau}, & x_{\text{min}} < x 
\end{cases} \tag{2}
\]

with normalization constant \( a = 1/\zeta(\tau, x_{\text{min}}) \), where

\[
\zeta(\tau, x_{\text{min}}) = \sum_{i=1}^{n} (i + x_{\text{min}})^{-\tau} \tag{3}
\]

is the generalized (Hurwitz) zeta function. The method implies that all practicable values of the lower cutoff \( x_{\text{min}} \) are considered; for each \( x_{\text{min}} \) the estimate of the power exponent (with the maximum likelihood principle in mind) is calculated from

\[
\tilde{\tau} = 1 + n \left[ \sum_{i=1}^{n} \ln \frac{x_i}{x_{\text{min}}} \right]^{-1}; \tag{4}
\]

see [28]; for each \( x_{\text{min}} \) the Kolmogorov-Smirnov statistic \( S = \sup_{x \geq x_{\text{min}}} | P(x) - \tilde{P}(x) | \) (where \( P(x) \) is the cumulative distribution function (CDF) with estimated value of \( \tau \) and \( \tilde{P}(x) \) is the empirical CDF) is calculated. The eventual estimate of \( x_{\text{min}} \) minimizes \( S(x_{\text{min}}) \). For real-world data, the function \( S(x_{\text{min}}) \) possesses, usually, several local minima; it is often reasonable not to choose a global minimum but the local minimum closest to \( x_{\text{min}} \), the lower boundary of the domain of definition, provided a value of the statistic at it does not differ significantly from that at a global minimum.

3. Power-Law Distributions for the English and Russian Languages

To test the null hypothesis in question (that the number of words in the texts obeys a power-law distribution; (1) is used to verify data collapse, while the method based on the KS statistics employs (2), (3)), two samples were generated on the basis of corpora of literary texts for these languages and of a set of Reddit messages (or its Russian counterpart Pikabu). The resulting samples sizes for English language are 9820 (literary works), 5016 (Reddit), and 14836 (joint sample); for Russian language they are 12683 (literary works), 6005 (Pikabu), and 18688 (joint sample). For the method based on the concept of data collapse, the size of vocabulary used to generate texts serves as characteristic dimension of a system \( L \). To obtain samples for various \( L \), one resamples down the initial sample deleting randomly \( \lambda \% \) of words from the complete vocabulary and then from all the texts used. This brings about the generation of new samples corresponding to the characteristic dimension of \( L(1 - \lambda/100) \).

Figure 1 presents a dependence of the number of words in a text in double-logarithmic scale on a rank of the text in the sample; namely, Figures 1(a), 1(b), and 1(c) correspond to the joint sample, to the sample constructed on the basis of literary works, and to poetry works for the English language, respectively; Figures 1(d), 1(e), and 1(f) exhibit the same dependence for the Russian, respectively; a dashed straight line in each subfigure corresponds to power distribution with an exponent estimated using data collapse.

Figure 2 (in the coordinates \( x^{-\tau}\tilde{P}(x), \xi = x/x_c \), \( x_c \) is a landmark coordinate) shows data collapse for the joint samples for English (Figure 2(a)) and Russian (Figure 2(d)) languages. Raw data was binned exponentially with bin sizes \( B_j = [cR^j](R-1) \) (\( c = 1 \) and \( R = 1.4 \)). Red colour (with discs) stands for the complete vocabulary (of size \( L \)), grey colour (with squares) is for the vocabulary of size 0.9L, blue colour (with diamonds) is for 0.8L, black colour (with triangles) is for 0.7L, orange colour (with upturned triangles) is for 0.6, and, finally, purple colour (with circles) is for 0.5L; the curves are dragged apart a little bit in order to make it possible to distinguish as they are superimposed owing to data collapse. Figures 2(b) and 2(c) present the same dependence for samples constructed using corpora of literary texts for English (Figure 2(b)) and Russian (Figure 2(e)) languages. Figures 2(c) and 2(f) demonstrate data collapse for poetry samples for English and Russian languages, respectively.

The results obtained using both approaches are presented in Table 1; the table includes results for samples constructed on the basis of literary works and of messages of social media and for joint sample as well for both languages. Each cell contains estimates for a power exponent and (in parentheses) for a lower cutoff \( x_{\text{min}} \). The above results suggest synchrony of both languages because of a good agreement of estimates for power exponents and lower cutoffs calculated for literary works and for social media messages.

In order to deal with the problem of diachronic unity of a language, authors confined themselves to the samples generated on the basis of literary works created before the 20th century and in the 20th century for both the English and Russian languages (respective samples sizes amount to 7179 and 2641 for the English language, 5758 and 6925 for the Russian language). Table 2 exhibits the respective results.
4. Discussion

Data collapse implies that if the distribution obeys power-law, the transformed distributions possess an interval with horizontal line and coincide inside this interval. For real-world data, the line inside the interval may not be so straight, but coincidence must occur as Figure 2 shows (one should take into account the fact that the curves are artificially dragged apart a little bit in order to make it possible to
Figure 2: Collapsed distributions. (a) English texts, joint sample, (b) English literary works, (c) English poetry, (d) Russian texts, joint sample, (e) Russian literary works, and (f) Russian poetry. Red colour (with discs) stands for the complete vocabulary (of size $L$), grey colour (with squares) is for the vocabulary of size 0.9$L$, blue colour (with diamonds) is for 0.8$L$, black colour (with triangles) is for 0.7$L$, orange colour (with upturned triangles) is for 0.6$L$, and, finally, purple colour (with circles) is for 0.5$L$. The curves are artificially dragged apart a little bit in order to make it possible to distinguish them as they are superimposed owing to data collapse. $x^{-\xi} p(x)$ versus $\xi = x/x_c$; $x_c$ is a landmark coordinate.

Distinguish them as they are superimposed owing to data collapse). Analogously, the results produced by the method that uses KS statistics also count in favour of the hypothesis that distributions are power-law.

We would like to emphasize that we regard this assumption as a plausible hypothesis as before; the results of the previous sections are arguments in its favour, not final results. It seems to us extremely important to ensure global visibility of this hypothesis. We strongly hope that other papers concerning this hypothesis will occur, with broader data sets and, probably, with more rigorous statistical methods. Fairly good agreement (for this class of distributions) of parameters for separate distributions of literary works created before the 20th century and in the 20th century (Table 2) suggests that a language (at least written language) is a single system diachronically.
Table 1: Estimated power-law exponents and lower cutoffs for distributions of the number of words in texts of literary works and social media.

<table>
<thead>
<tr>
<th>Sample type</th>
<th>English Estimates based on data collapse</th>
<th>Russian Estimates based on KS criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_{\min}$ (nw)</td>
<td>$\tau$ (dl)</td>
</tr>
<tr>
<td>Literary works</td>
<td>114</td>
<td>1.30</td>
</tr>
<tr>
<td>Poetry works</td>
<td>125</td>
<td>1.91</td>
</tr>
<tr>
<td>Social media</td>
<td>111</td>
<td>2.65</td>
</tr>
<tr>
<td>Joint sample</td>
<td>109</td>
<td>1.39</td>
</tr>
</tbody>
</table>

nw: the number of words; dl: dimensionless.

Table 2: Estimated power-law exponents and lower cutoffs for distributions of the number of words in texts for literary works for XIXth (and before it) and for XXth century.

<table>
<thead>
<tr>
<th>Time limit</th>
<th>English Estimates based on data collapse</th>
<th>Russian Estimates based on KS criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_{\min}$ (nw)</td>
<td>$\tau$ (dl)</td>
</tr>
<tr>
<td>XIX century and early</td>
<td>118</td>
<td>1.37</td>
</tr>
<tr>
<td>XX century</td>
<td>121</td>
<td>1.19</td>
</tr>
</tbody>
</table>

nw: the number of words; dl: dimensionless.

We would like to dwell on a cooccurrence graph as a semanticspace in greater detail as opposed to rather popular global syntactic dependency tree and similar structures. In the present paper, a text is considered as a basic unit of a language; thus a sentence is a means to break (rather arbitrary) this semantic “avalanche” down. Global syntactic dependency tree is a great tool to explore this avalanche locally, but as far as it generally fails to reveal cross-sentence semantic dependencies, it does not seem to be the best tool to examine the avalanche as a whole. Therefore cooccurrence graph is a natural choice; an edge belongs to this graph if the words (corresponding to its vertices) belong to the same text. Generally, the underlying structure does not seem to be of principal importance for the problem considered.

In our opinion, this allows one to distinguish results of the present work and those for global syntactic dependency trees and cooccurrence graphs with the employment of random walks [24, 25, 29, 30]. We explore real-world semantic “avalanches” generated by a particular language, while the algorithms using random walks produces artificial “avalanches” with the employment of a graph complying with a natural language. In particular, such motion continues (theoretically) perpetually, whereas the avalanches considered in this article are of finite sizes, explicitly defined by the authors of the respective texts. In order to emphasize fundamental distinction between these approaches, one could draw the following analogy: the approach of the present paper and those associated with random walks are related to study of joint distribution of random variables and of product of distributions of these variables. Nevertheless, the results of this study are likely to be useful for switcher-random-walks models [30] to estimate realistically switching time.

We would also like to emphasize the difference between the classic Zipf’s laws and the distributions considered in this paper: Zipf studied the laws governing means to represent information, while we attempt to explore laws governing semantic flows and, moreover, semantic flows of a language as a whole.

5. Conclusions

As a result of the above analysis several conclusions may be reached on linguistic systems of English and Russian languages. A language system (given by texts generated in its frameworks) is a self-organized critical system defined on its word cooccurrence graph. Texts of a language are “avalanches” flowing down this graph; the large avalanches correspond to literary works, while the smaller ones are associated with spoken language. A fairly good agreement of parameters for separate distributions of literary works and of social media offers a clearer view of synchronic unity of each linguistic system; on the other hand, an analogous comparison between distributions for literary works of XIXth (and before) and of XXth centuries suggests diachronic unity for the systems. Poetry distributions appear closest to a canonical power-law and therefore poetry may be treated as a kind of supporting column of a language.
Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors are thankful to Mr. Vladimir Marchenko and to Miss Victoria Ankudinova for the manuscript proofreading and language-editing.

References


