Distinguishing between Natural and Random Texts: a Statistical Measure Linked to Word Clustering

O. S. Kushnir, A. I. Kashuba and V. V. Yaremkiv

Optoelectronics and Information Technologies Department,
Ivan Franko National University of Lviv, 107 Tarnavsky Street, 79017 Lviv, Ukraine

o.s.kushnir@lnu.edu.ua, andriy.kashuba@lnu.edu.ua

Recognition of natural, i.e. semantically filled messages, as opposed to random or randomized (semantically empty) symbolic sequences, represents an interesting problem. It is sometimes believed that the problem can be solved using the approaches of statistical linguistics (see, e.g., [1–3]).

In this work we try to solve the problem mentioned above, using a suitable circumstance that semantically filled texts reveal a property termed as 'burstiness', 'intermittence' or 'clusterization' [4, 5]. Given the temporal positions t_i ($i = 1, 2, ..., F$, with F being the absolute frequency) of different tokens of some word type w in a text, one can calculate the inter-occurrence, or waiting, times τ_i for this word as $\tau_i = t_{i+1} - t_i - 1$. Although there are more complex and reliable measures of temporal clustering of the word w, one can restrict the analysis to the simplest clustering coefficient $R = \Delta\tau / \bar{\tau}$ [4], where $\bar{\tau}$ denotes the mean waiting time and $\Delta\tau$ the corresponding standard deviation. Due to a stochastic (Poisson-like) nature of the time occurrences t_i, the probability $p(\tau_i)$ is given by a negative exponential distribution, so that we have $R \approx 1$ (or the absence of any 'interactions' of different tokens for a given word type) for a large majority of word types. The exception is so-called keywords, for which $R > 1$ or even $R >> 1$ (a case of clusterization of tokens), and rare and untypical words with $R < 1$ ('repulsion' of tokens).

The main idea of this work is to characterize a text by the mean $\overline{R}$ and the standard deviation $\Delta\overline{R}$ found by averaging over all the word types in this text. They represent cumulative measures of the clusterization effect in the text and, therefore, a peculiar metrics for the presence of keywords and semantics in it. By definition, the word tokens in a random text do not 'interact' with each other, so that we have $R = 1$ for all of the word tokens, except for those rare types which suffer finite-size effects. Then the equality $\overline{R} \approx 1$ immediately follows, whereas $\Delta\overline{R}$ should remain small enough. On the contrary, the $\overline{R}$ parameter for any semantics-bearing text can deviate notably from unity, while $\Delta\overline{R}$ can become relatively large.

To test the above hypothesis, we have analyzed a number of natural and random texts, using an original program written in Python. The natural (fiction) texts have been taken from http://www.gutenberg.org/. Among random texts, we have studied the known Miller's monkey texts with different alphabet sizes M, the natural texts randomized (locally or globally) on the linguistic level of words (over 10^9

COLINS'2018, Volume II: Workshop. Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013
http://colins.in.ua, online

randomization cycles), the texts composed according to a preferential-attachment model by Simon, and the Chomsky texts.

Some measures have been taken in order to reduce the finite-size effects. For this aim, we have neglected the statistics of token occurrences for the word types with the absolute frequencies less than some threshold F_{th} (e.g., $F < 20$) and/or for the types for which the relative frequencies f ($f = F/L$, with L being the total text length in the units of words) are less than a corresponding threshold f_{th} (e.g., $f < 10^{-4}$). Besides of these frequency filters, the statistical significance of rare word types has been reduced by calculating the $\overline{R}$ and $\Delta\overline{R}$ parameters, using weighting coefficients proportional to the absolute frequencies of the words. To further evaluate the negative effect of poor statistics typical for the low-frequency word types, we have compared the data for $\overline{R}$ and $\Delta\overline{R}$ with the results obtained using an improved version of clustering coefficient, which has been introduced in the work [6]. Table 1 exemplifies the data obtained with the only filter with $F_{\text{th}} = 20$ and no weighting. Even under such unfavorable statistical conditions, there is a clear distinction of natural and random texts. The $\overline{R}$'s for the latter texts are closer to unity and $\Delta\overline{R}$ notably less than those obtained for the natural texts. In particular, it is instructive to compare $\overline{R}$ and $\Delta\overline{R}$ for the initial natural text and its randomized counterpart.

Table 1. Statistical characteristics of some natural and random texts (see the text).

#	Text	Text length L, 10^5	$\overline{R}$	$\Delta\overline{R}$
1	Natural	1.9	1.33	0.61
2	Randomized natural (10^9 cycles)	1.9	0.97	0.14
3	Monkey ($M = 1$)	10.0	0.96	0.05
4	Simon	1.7	1.02	0.19

Note that the algorithm for producing Simon texts is such that the initial word types from a 'bag of words' are repeated many times in the very beginning of the text and so can, in principle, reveal a spurious clustering in this region, thus increasing the $\overline{R}$ parameter in an uncontrollable manner. However, contrary to these intuitive expectations, it has turned out that the Simon texts reveal the same (nearly unit) $\overline{R}$ coefficients, like the other semantically empty texts.

References

1. A. Cohen, R. N. Mantegna and S. Havlin. *Fractals* **5**, 95 (1997).
2. R. Ferrer i Cancho and R. V. Sole. *Proc. Roy. Soc. Lond. B* **268**, 2261 (2001).
3. D. Kimura and T. Tanaka-Ishii. *J. Natural Language Processing*. **18**, 119 (2011).
4. M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz and A. M. Somoza. *Europhys. Lett.* **57**, 759 (2002).
5. Barabási A.-L. Nature **435**, 207 (2005).
6. P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado and J. L. Oliver. *Phys. Rev. E* **79**, 035102(R) (2009).