Physica A 414 (2014) 378–386

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Scale and time dependence of serial correlations in word-length time series of written texts

E. Rodriguez [a], M. Aguilar-Cornejo [a], R. Femat [b], J. Alvarez-Ramirez [a,*]

[a] División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana-Iztapalapa, Apartado Postal 55-534, Iztapalapa D.F. 09340, Mexico

[b] División de Matemáticas Aplicadas, IPICyT, Camino a la Presa de San José 2055, Lomas 4a Secc, San Luis Potosí, S.L.P., 78290, Mexico

HIGHLIGHTS

- Time and scale dependences of correlations of word-length time series were studied.
- The scaling exponent exhibited variations along the written text.
- Scaling exponent shifts coincide with transitions in narration units.

ARTICLE INFO

Article history:
Received 26 May 2014
Received in revised form 29 June 2014
Available online 23 July 2014

Keywords:
Written texts
Quantitative analysis
Word length sequence
DFA

ABSTRACT

This work considered the quantitative analysis of large written texts. To this end, the text was converted into a time series by taking the sequence of word lengths. The detrended fluctuation analysis (DFA) was used for characterizing long-range serial correlations of the time series. To this end, the DFA was implemented within a rolling window framework for estimating the variations of correlations, quantified in terms of the scaling exponent, strength along the text. Also, a filtering derivative was used to compute the dependence of the scaling exponent relative to the scale. The analysis was applied to three famous English-written literary narrations; namely, Alice in Wonderland (by Lewis Carrol), Dracula (by Bram Stoker) and Sense and Sensibility (by Jane Austen). The results showed that high correlations appear for scales of about 50–200 words, suggesting that at these scales the text contains the stronger coherence. The scaling exponent was not constant along the text, showing important variations with apparent cyclical behavior. An interesting coincidence between the scaling exponent variations and changes in narrative units (e.g., chapters) was found. This suggests that the scaling exponent obtained from the DFA is able to detect changes in narration structure as expressed by the usage of words of different lengths.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Human language is a complex system involving units of different length and function, recurrence, redundancy, etc. Efforts for quantifying the complexity of language were triggered by the pioneering work by Zipf [1], who showed that the frequency counts of words in diverse languages follow a power-law distribution. Besides, words which must be reused repeatedly should be short to minimize the effort of language users. By departing from the fact that human language is

* Corresponding author. Tel.: +52 5558044934.
E-mail address: jjar@xanum.uam.mx (J. Alvarez-Ramirez).

http://dx.doi.org/10.1016/j.physa.2014.07.063
0378-4371/© 2014 Elsevier B.V. All rights reserved.

internally organized by specific construction rules (e.g., grammar) and meaning/connotation structures (e.g., semantics), some efforts have been devoted for modeling language structure and evolution. For instance, evolutionary dynamics has been used for describing the cultural evolution of language and the biological evolution of universal grammar [2]. Considered as a graph of word interactions, human language exhibits small-world features and scale-free distributions [3]. Monte Carlo simulations with evolutionary game theory have been used for studying the evolution of words of systems comprising two interacting species [4]. Interestingly, formulations of equilibrium statistical mechanics have been considered for a qualitative description of important characteristics of language, such as the universality of Zipf's law and the vocabulary size of children [5]. It has been suggested that the word frequency distribution is analogous to the Bose–Einstein distribution, which implies that the temperature of texts can be defined [6].

Language is basically a sequence of words following a complex grammatical order. In this way, language can be seen as a time series of ordered symbols hiding long-range memory effects and recurrence. In this way, methods borrowed from statistical mechanics can be applied for extracting information on the intrinsic organization of language time series. The application of R/S analysis [7] and detrended fluctuation analysis (DFA) [8] showed that word-length time series are not random, but contain weak long-range correlations. In addition, it has been shown that Shannon and Kolmogorov entropies of word-length time series are sensitive to language [9]. Variations in language complexity have been addressed by means of multifractal analysis of written texts [10,11].

Summing up, language texts are internally organized systems as reflected by the presence of long-range serial correlations and non-trivial entropic levels [8,9]. It is expected that for long texts (e.g., literary narrative, political speeches, etc.), these statistical features are not constant, but exhibit variations along the underlying time series. In fact, the variations of the long-range memory degree could be the consequence of stylistic and emotional characteristics of individual authors. This work uses the DFA over word-length time series of written texts for studying scale and time dependences of long-range serial correlations. The results indicated that long-range correlations are not constant along the text, but exhibit important variations that can be linked to changes in narrative units, like chapters and themes.

2. Mapping written texts into numerical time series

Three famous novels from the English literature were considered for analysis. The texts were borrowed from the free-access Internet site Project Gutenberg (www.gutenberg.org); namely, Alice in Wonderland (by Lewis Carrol, 26,541 words), Dracula (by Bram Stoker, 57,613 words) and Sense and Sensibility (by Jane Austen, 119,956 words). These novels are different in extension and style, Sense and Sensibility being the opus with a more involved English usage.

The quantification of serial correlations hidden in written texts was based on the analysis of numerical time series. As in previous reports [7–9], this work considered the sequence of word lengths after removing punctuation symbols. In this way, the words "novel" and "detrended" have length five and nine, respectively. The result is a time series with numerical variations from one to about twenty. In general, pronouns and conjunctions are short-length words, while adjectives, nouns and some verbs usually are long-length words. The ordered combination of pronouns, nouns, verbs and adjectives linked by conjunctions leads to sentences. In turn, a sequence of sentences composes a narration organized within paragraphs, themes and chapters. Expressed in terms of the more primitive quantification form (i.e., word length), one question is whether the underlying time series contain serial correlations that reflect the intrinsic organization of the written text. The previous work using the DFA showed that word-length time series are not random at all, but contain some weak serial correlations [8]. In a next section, the DFA will also be used for showing that such serial correlations depend on the scale and time. For the sake of completeness in presentation, a brief description of the DFA will be given below.

3. Detrended fluctuation analysis

Since the pioneering work by Peng et al. [12], the detrended fluctuation analysis (DFA) is nowadays a widely used method for detecting long-range correlations in time series. Applications include a wide variety of fields, from biology, physics, and image processing to geophysics and meteorology. A brief description of the DFA is as follows. Given a time series, x_k, $k = 1, \ldots, N$, it is integrated

$$Y_k = \sum_{j=1}^{k} \left(x_j - \langle x_k \rangle\right), \quad k = 1, \ldots, N \tag{1}$$

where $\langle x_k \rangle = \frac{1}{N} \sum_{j=1}^{N} x_j$ is the time-series mean. After dividing Y_k into $N_s = [N/s]$ not-overlapping segments of equal length s, a piecewise linear trend $Y_{s,k}$ is estimated within each segment and the detrended series is calculated as $\tilde{Y}_k = Y_k - Y_{s,k}$. The fluctuation function is computed as

$$F(s) = \left(\frac{1}{sN_s} \sum_{j=1}^{sN_s} \tilde{Y}_k^2 \right)^{1/2}. \tag{2}$$

Download English Version:

https://daneshyari.com/en/article/975274

Download Persian Version:

https://daneshyari.com/article/975274

Daneshyari.com