A comparative study of power law scaling in large word-length sequences

J Quezada-González¹, D Aguilar-Velázquez¹, T Contreras-Uribe¹, B Obregón-Quintana², L Guzmán-Vargas¹,³

¹ Unidad Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Instituto Politécnico Nacional, Av. IPN No. 2580, L. Ticomán, México, D.F., 07340, México
² Facultad de Ciencias, Universidad Nacional Autónoma de México, Circuito Exterior s/n, Ciudad Universitaria, Coyoacán, México, D.F., 04510, México

E-mail: lguzmanv@ipn.mx

Abstract. A study of the correlation of lengths of words in large literary texts is presented. We use the statistical tools based on Allan factor and fractal dimension for estimating the fractal indices associated with the presence of correlations in the original sequences. We found that there is a scaling behavior at large scales for Allan factor with an exponent value, which reveals positive correlations. By means of the fractal dimension, we confirmed the presence of memory in the sequences of lengths of words. In addition, when considering a random rearrangement of the lengths of words in the original text, the fractal exponents are consistent with an uncorrelated process.

1. Introduction

The texts have been instrumental throughout the history of mankind to explore and document the progress in the development of human knowledge, from hieroglyphs, scrolls and printed books until the appearance of e-books. Human communication is ruled by grammar and syntax that allow to give meaning to the sentences which are comprised by several words [1, 2]. In 1851, the work of August of Morgan began the study of the frequency in the language. Morgan found that the average length of words (letters per word) was an indicator of individual style. For details of this work and state of the art, see [3]. In the decade of the 40s of the last century, George Kingsley Zipf, established that the frequency of the words shows a power law behavior with the rank, i.e., if \( f(r) \) is the frequency of a word and \( r \) the rank of the word, then \( f(r) \sim 1/r^\alpha \), with \( \alpha \approx 1 \) [4]. These type of studies are widely known and have been applied to other areas, particularly with regard to language are: content of information [5], polarity and information [6], recurrence times [7], correlations [8-10], allometries [11], and many others [2, 5, 12-19].

The work of Piantadosi et al., noted that the length of a word is also important, because the length has a nonlinear relationship with the frequency and information, that is, if the duration or length of a word is longer, it contains more information [5]. Garcia et al. established a relationship between the frequency of a word and its positive or negative content, being more frequent positive words, however, the negative ones contain more information [6]. In other studies, the relation between words has been considered taking into account their sound, if two words are pronounced similarly, then there is a relationship between them [16-17]. Moreover, words have a life time or time of use in everyday language

³ To whom any correspondence should be addressed.
where the frequency is not the only factor in determining survival (birth, survival and death) as it was shown in the work of Petersen et al. [11]. On the other hand, diverse works have been done on specific texts manipulated as time series to identify characteristics of language considering temporal correlations with respect to the length or frequency of words [8, 20-22].

However, these previous works have not widely studied the interdependence between lengths of words and the relationship between ideas and stories developed in large literary texts, therefore, it is important to study the all these properties and not only the frequency in isolation. In this paper we consider the temporal correlation between the length of the words for several large literary texts. First, we consider the sequence of the length of words as a point process and calculate the Allan factor to determine the scaling regime characterized by the correlation exponent. Next, we use the fractal dimension method to estimate the fractality index for the word-length sequences. We find that both methods show the presence of positive correlations, indicating that word-length time series is not a random process. This paper is organized as follows, in the Section II, we explain the Allan and fractal dimension methods and a description of the texts in our study. Next, in Section III, we present the results, and finally the conclusions are given.

2. Methods and data

We use the point process statistics to determine the presence of scaling properties in word-length sequences. A point process can be described as events (spaces between words in our case) located along a time axis. We use the so-called Allan factor (AF), which is defined as the variance of successive counts for a specific window time, divided by twice the average number of events in that count [23]. For a process where this ratio remains constant and close to the unit for different time windows, it is identified as a Poisson processes (with no memory). In contrast, for a process with a scaling behavior between AF and the time window $T$, i.e., $A(T)$ behaves like a power law function of the time window $T$, $A(T) \sim T^\alpha$, the process is known to exhibit positive correlations and the exponent $\alpha$ characterizes the behavior of the strength of the correlations.

We also use the fractal dimension method (FDM) [24, 25] to estimate the value of the dimension of the set represented by the sequence of lengths of words. The FDM basically consists of the following steps: given the time series $x_1, x_2, \ldots, x_N$, new series are constructed $x_{m}^k$, which are defined as $x_m, x_{m+k}, x_{m+2k}, \ldots, x_{m+k[N/k]}$, with $m = 1, 2, \ldots, k$; $[\ ]$ denotes Gauss notation, i.e., the greatest integer, $m$ and $k$ are integers representing the initial time and the time interval, respectively. The length of the curve is defined as:

$$L_m(k) = \frac{1}{k} \left[ \sum_{i=1}^{N-m|N/k|} |x(m + ik) - x(m + (i - 1)k)| \right] \frac{N - 1}{k|N - m|}$$

where $(N - 1)|((N - m)/k)|k^{-1}$, represents a normalization factor. In order to get the total length of the original curve for the time interval $k$, it is considered the length of each sequence $x_{m}^k$. Finally, if $\langle L(k) \rangle \propto k^{-D}$, then the curve is fractal with dimension $D$ [24, 25].

For a Brownian type process resulting from the integration of an uncorrelated process, typically we get $D = 1.5$, whereas for a process with $D < 1.5$, we have positive correlations or persistence and $D > 1.5$ reflects anticorrelations. There is a relationship between the Allan factor exponent $\alpha$ and $D$, given by $\alpha = 3 - 2D$ [24, 25].

Our database consists of 10 electronic e-books in English language (see table 1), obtained from the Project Gutenberg (http://www.gutenberg.com) and Project Gutenberg Australia (http://gutenberg.net.au). For the selection of titles, we considered books published in different times and from different genres, and in particular that were large enough texts (number of words $> 10^5$) in order evaluate the asymptotic scaling behavior.
Table 1. List of books used in the study. $M$ denotes the number of words and $N_U$ the number of different words

| No. | Title and author           | $M$   | $N_U$ |
|-----|----------------------------|-------|-------|
| 1   | Ulysses, J. Joyce          | 272,415 | 29,898 |
| 2   | The Bible, Ed. King James  | 884,964 | 12,806 |
| 3   | On the Origin of Species, C. Darwin | 156,811 | 7,273 |
| 4   | Frankenstein, M. Shelley   | 104,753 | 8,078 |
| 5   | Gulliver’s Travels, J. Swift | 104,797 | 8,188 |
| 6   | Hopscotch, J. Cortazar     | 195,702 | 15,338 |
| 7   | Moby Dick, H. Melville    | 218,704 | 17,150 |
| 8   | Pride and Prejudice, J. Austen | 122,878 | 6,450 |
| 9   | The Koran (Al-Qur'an)      | 210,967 | 11,058 |
| 10  | The Idiot, F. Dostoyevsky  | 247,952 | 10,102 |

3. Results

In our case we are interested in comparing the asymptotic behavior of the scaling exponents for large time scales. First, we apply the Fano factor to word-length sequences. For this purpose, the event of interest is the space between words, and the interevent times is given by the word-lengths. In figure 1 we show the results of analysis of Allan for the 10 books in our study. It is observed that at small scales, $A(T)$ is close to 1, indicating that the process is Poissonian; for intermediate scales, there is a valley with values smaller than 1, reflecting an increase in the regularity. At large scales, a scaling regime can be identified, that is, $A(T) \sim T^\alpha$, with an average exponent $\alpha = 0.57 \pm 0.09$, denoting the presence of correlations. Our results of exponent values indicate that when word-lengths are observed at large scales (windows), they exhibit persistence, i.e., word-lengths separated by large intervals display memory whereas at short intervals the dynamics is uncorrelated. Moreover, when we build a shuffled version of the word-lengths in the original text, the statistics $A(T)$ behaves differently than the original, with values near and below the unit value (homogeneous Poisson process), indicating that the memory observed in the original sequences is destroyed by the shuffled process (see figure 1).

![Figure 1](image-url)  
Figure 1. Plot of $A(T)$ vs. $T$ for the books in our study. We identify a scaling regime for the scales $10^3 < T < 10^6$, with an average exponent close to 0.5. For shuffled data, the values exhibit a flat distribution, indicating that the memory is destroyed.

To further explore the behavior of the power-law scaling at large scales, we proceeded to estimate the value of the local exponent $\alpha^l$ given by the following expression: $\alpha^l = \frac{\delta \log A(T)}{\delta \log (T)}$. For a better accuracy in the estimation of the exponent, we use the centered derivative. The results of these calculations are shown in figure 2. We observe a smooth transition from the regime corresponding to the Poisson process, to an average value close to 0.5, confirming that in the asymptotic region the local exponent is consistent with the presence of positive correlations and long-range memory. We also noticed that there is a greater dispersion at large scales, possibly associated with the heterogeneity of our database, a slightly different trend in the level of correlations is identified for texts like the Bible.
and the Koran (see figure 1), i.e., some texts emphasize the presence of correlations in the long term while others exhibit more oscillations in the statistics.

![Figure 2](image2.png)

**Figure 2.** Behavior of the average local exponent $\alpha_l$ in terms of the time scale $T$. We observe a smooth transition from the Poissonian regime to the correlated one. We show the mean value $\pm 1$ standard deviation.

To complement our study, we use the FDM method for estimating the values of the fractal dimension of the same of word-length sequences. For a better estimation of the fractal exponent $D$, an integration of the original word-length time series is performed prior the application of the FDM.

Figure 3 shows the results of $\langle L(k) \rangle$ vs. $k$ for the ten books in our dataset. After an inspection, it is possible to identify two trends in the slopes that fit the data, that is, for short time intervals, it is observed that the statistics is characterized by the average exponent value of the exponent $D = 1.45 \pm 0.03$, which corresponds to an uncorrelated process, whereas for large scales, the fractal index is $D = 1.25 \pm 0.04$, confirming that the process is persistent with long-term memory. Additionally, the estimation of the local exponent was performed to quantify the asymptotic behavior of this scale invariant. In Figure 4 we present the results of the mean value of the local exponent $D_l$, we also observe a smooth transition from the uncorrelated regime ($D = 1.5$) until it stabilizes around $D \approx 1.3$, indicating a good concordance with the global exponent. It is also worth to mention that the obtained results for the fractal dimension are in general agreement with the values for the invariants obtained by means of the Allan factor, i.e., both methods confirm the presence of long-term memory in the word-length sequences. More specifically, if use the relationship $\alpha = 3 - 2D$, for $D \approx 1.25$ we get $\alpha^* \approx 0.5$ which is in good concordance with the value $\alpha = 0.57$ obtained by means of the Allan Factor.

![Figure 3](image3.png)

**Figure 3.** Plot of $L(k)$ vs. $k$ for the ten books in our dataset. We observe that the regime at large scales is characterized by the average exponent $D \approx 1.25$
Figure 4. Behavior of the average local exponent $D^l$ in terms of the time interval $k$. We observe a smooth transition from the uncorrelated regime to the correlated one. We show the mean value $\pm 1$ standard deviation.

4. Conclusions
We have presented the Allan factor and fractal dimension analyses for word-length sequences obtained from large literary texts. The interest has focused on the estimation of the stability of the local exponent and its agreement with the global exponent that characterizes the correlations in the long term. It has been shown that the exponent describing the scaling in Allan factor, has an average value $\alpha = 0.57$, which reflects the presence of positive correlations, while the trend of the local index displays a smooth transition from the Poisson regime to the correlated one, achieving stability in the very large scales, but exhibiting a large spread. We also found that the fractal dimension value $(D \approx 1.5)$ is consistent with the presence of positive correlations and confirm the results of the Allan statistics. The local behavior of this scaling index at large scales, reflects a stability with a moderate dispersion. This result assigns an advantage to this methodology regarding the estimation of a fractality index for irregular time series.

In summary, we have shown that when we observe a word-length sequence as a point process, the Allan factor reflects the presence of scaling with an index that is consistent with the dimensionality of the original series, being more locally stable the exponent extracted by means of the FDM algorithm.

Acknowledgments
This work was partially supported by COFAA-IPN, EDI-IPN and SNI-Conacyt.

References
[1] Nowak M A, Plotkin J B and Jansen V A A 2000 Nature 404 495-498
[2] Solé R V, Corominas-Murtra B, Valverde S and Steels L 2010 Complexity 15 20-26
[3] Grzybek P 2006 History and Methodology of Word Length Studies: The State of the Art Contributions to the Science of Text and Language (Springer Amsterdam, The Netherlands vol 31) pp 15-90
[4] Zipf G K 1935 The Psycho-Biology of Language: An Introduction to Dynamic Philology (MIT Press, Cambridge, MA, USA)
[5] Piantadosi S T, Tily H and Gibson E 2011 Proc. Natl. Acad. Sci. 108 3526-3529
[6] Garcia D, Garas A and Schweitzer F 2012 EPJ Data Sci. 1:3
[7] Altmann E G, Pierrehumbert J B and Motter A E 2009 PLoS ONE 4 e7678
[8] Kosmidis K, Kalampokis A and Argyrakis P 2006 Physica A 370 808-816
[9] Nuñez A, Lacasa L, Valero E, Gómez J P and Luque B 2012 Int. J. Bifurc. Chaos 22
[10] Lacasa L, Nuñez A, Roldán E, Parrondo J and Luque B 2012 Eur. Phys. J. B 85
[11] Petersen A, Tenenbaum J, Havlin S and Stanley H E 2012 Sci. Rep. 2
[12] Kalimeri M, Constantoudis V, Papadimitriou C, Karamanos K, Diakonos F K and Papageorgiou H 2012 Int. J. Bifurc. Chaos 22
[13] Kalimeri M, Constantoudis V, Papadimitriou C, Karamanos K, Diakonos F K and Papageorgiou
[14] Rêgo H H A, Braunstein L A, D’Agostino G, Stanley H E and Miyazima S 2014 *PLoS ONE* **9** e110213

[15] Michel J B, Shen Y K, Aiden A P, Veres A, Gray M K, Team T G B, Pickett J P, Hoiberg D, Clancy D, Norvig P *et al.* 2011 *Science* **331** 176-182

[16] Arbesman S, Strogatz S H and Vitevitch M S 2010 *Int. J. Bifurc. Chaos* **20** 679-685

[17] Arbesman S, Strogatz S H and Vitevitch M S 2010 *Entropy* **12** 327-337

[18] Chan K Y and Vitevitch M S 2010 *Cognit. Sci.* **34** 685-697

[19] Samoilenko A, Karimi F, Edler D, Kunegis J and Strohmaier M 2016 *EPJ Data Sci.* **5** :9

[20] Guzmán-Vargas L, Obregón-Quintana B, Aguilar-Velázquez D, Hernández-Pérez R and Liebovitch L 2015 *Entropy* **17** 7798-7810

[21] Rodriguez E, Aguilar-Cornejo M, Femat R and Alvarez-Ramirez J 2014 *Physica A* **414** 378-386

[22] Ausloos M 2012 *Phys. Rev. E* **86** 031108

[23] Allan D W 1966 *Proc. IEEE* **54** 221-230

[24] Higuchi T 1990 *Physica D* **46** 254-264

[25] Guzmán-Vargas L and Angulo-Brown F 2003 *Phys. Rev. E* **67** 052901