11-20-2015

Word-Length Correlations and Memory in Large Texts: A Visibility Network Analysis

Lev Guzmán-Vargas
Instituto Politécnico Nacional

Bibiana Obregón-Quintana
Universidad Nacional Autónoma de México

Daniel Aguilar-Velázquez
Instituto Politécnico Nacional

Ricardo Hernández-Pérez
Instituto Politécnico Nacional

Larry S. Liebovitch
CUNY Queens College

Entropy **2015**, *17*, 7798-7810; doi:10.3390/e17117798

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Word-Length Correlations and Memory in Large Texts: A Visibility Network Analysis

Lev Guzmán-Vargas [1,*], Bibiana Obregón-Quintana [2], Daniel Aguilar-Velázquez [1], Ricardo Hernández-Pérez [3] and Larry S. Liebovitch [4,5,6]

[1] Unidad Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Instituto Politécnico Nacional, Av. IPN No. 2580, L. Ticomán, México D.F., 07340, Mexico; E-Mail: zafskumo@hotmail.com
[2] Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, México D.F., 04510, Mexico; E-Mail: bobregon@ciencias.unam.mx
[3] Departamento de Física, Escuela Superior de Física y Matemáticas, Instituto Politécnico Nacional, Edif. No. 9 U.P. Zacatenco, México D.F., 07738, Mexico; E-Mail: ricardohdzpz@gmail.com
[4] Departments of Physics and Psychology, Queens College, City University of New York, 65-30 Kissena Boulevard, SB B322, Flushing, NY 11367, USA; E-Mail: Larry.Liebovitch@qc.cuny.edu
[5] Adjunct Senior Research Scholar, Advanced Consortium on Cooperation, Conflict, and Complexity (AC4), Earth Institute, Columbia University, New York, NY 10027, USA
[6] Physics Program, The Graduate Center, City University of New York, New York, NY 10016, USA

* Author to whom correspondence should be addressed; E-Mail: lguzmanv@ipn.mx; Tel.: +52-55-5729-6000 (ext. 56873).

Academic Editor: J. A. Tenreiro Machado

Received: 27 August 2015 / Accepted: 13 November 2015 / Published: 20 November 2015

Abstract: We study the correlation properties of word lengths in large texts from 30 ebooks in the English language from the Gutenberg Project (www.gutenberg.org) using the natural visibility graph method (NVG). NVG converts a time series into a graph and then analyzes its graph properties. First, the original sequence of words is transformed into a sequence of values containing the length of each word, and then, it is integrated. Next, we apply the NVG to the integrated word-length series and construct the network. We show that the degree distribution of that network follows a power law, $P(k) \sim k^{-\gamma}$, with two regimes, which are characterized by the exponents $\gamma_s \approx 1.7$ (at short degree scales) and $\gamma_l \approx 1.3$ (at large degree scales). This suggests that word lengths are much more strongly correlated at large distances between words than at short distances between words. That finding is also supported by the detrended fluctuation analysis (DFA) and recurrence time distribution. These results provide

new information about the universal characteristics of the structure of written texts beyond that given by word frequencies.

Keywords: words frequency; words recurrence; syllables; texts

1. Introduction

A widely-recognized property of language is Zipf's law, in which the frequency of words exhibits a power law behavior in terms of the rank, that is if $f(r)$ is the frequency of a word and r the rank of that word, then $f(r) \sim 1/r^{\alpha}$, with $\alpha \approx 1$ [1]. As is well known, language is a system composed of grammatical rules applied to a vocabulary or lexicon, where the words represent an essential unit, and the order or sequence is related to the need to transmit information or ideas [2]. Since Zipf discovered this property in the 1940s, several studies have focused on this direction and recently, other properties, such as information content [2], polarities and information [3], recurrence times [4], correlations [5–7], allometries [8,9], the length of words [2,10,11], and many others [12–17].

For instance, Piantadosi *et al.* [2] found that the word length has a non-linear relationship with the frequency and that an efficient communication process is concomitant with the fact that the word length increases with information content. Furthermore, Garcia *et al.* [3] noted that words with positive emotional content are used more often, and they tend to carry less information than negative ones. In the same direction, other studies have noted the important role of word length in meaning, emotional and information content in the organization of human language. Particularly, word length has been systematically studied in quantitative linguistics since 1851, when August de Morgan suggested the use of word lengths as a hallmark of the text style and a possible factor in determining authorship [11]. For a review about this topic, see [18]. Recently, Chen *et al.* [19] reported that the increase of word length is an essential ingredient in the evolution of written Chinese. These recent studies have contributed new approaches to the complex analysis of texts and, at the same time, have opened up new questions about the underlying complexity of language, particularly in written texts. For example, longer words are more likely to be used to express more abstract ideas [3].

An important trait of written texts is the appearance of temporal correlations as ideas or stories are created. However, the direct evaluation of these correlations is not feasible, because words can be used in different manners, which can make a quantitative analysis difficult. In past years, diverse methods have been used to explore the presence of temporal correlations in texts [5,20,21], mainly focused on the length or frequency of words. Very recently, it has been reported that there are some differences between European languages when they are compared in terms of the frequency and correlations of the word lengths [11]. As is recognized, written language is the conformation of grammar properties and semantic connotations with the purpose of expressing ideas or information. It has been reported that the temporal organization of word-length sequences from written texts can be characterized by the presence of slightly positive correlations [5,20] and with local variations of the scaling exponents related to the temporal organization of the texts [21].

However, these studies have not considered the temporal organization over a wide dataset of large literary texts, where the concatenation of ideas or stories is the most important trait. In this work, we study the "temporal" correlations of word lengths in large literary texts by means of the natural visibility graph algorithm (NVG) [22]. Specifically, we consider 30 ebooks in the English language from which we extract the word-length time series. To evaluate the presence of correlations in the word-length sequences, we use the NVG to relate every two words [22]. The NVG method has the advantage of providing potential further insight into the temporal organization of sequences, by exploring the emerging network structures in a quantitative manner. The NVG has been used to explore organizational features in complex time series from different systems ranging from chaotic signals [23], heartbeat variability [9] and economics to seismology [24,25].

Our results show that the resulting degree distribution follows a power law, $P(k) \sim k^{-\gamma}$, which exhibits two different regimes of correlations over the short and long distances between words. These findings are complemented with the application of the detrended fluctuation analysis (DFA) and the calculations of recurrence times.

The paper is organized as follows. In Section 2, we provide a brief description of the NVG and the collection of texts that we studied. The results are described in Section 3. Finally, some concluding remarks are given in Section 4.

2. Methods and Data

2.1. Natural Visibility Graph Algorithm

The NVG algorithm [22] was proposed to transform irregular time series into networks, with the idea of exploring the complexity of the original time series with the help of methodologies from network science. Consider a time series $y_1, y_2, y_3, ...y_n$. For any two data values, (t_a, y_a) and (t_b, y_b), where the time t_i refers to the time of event y_i, we define a link between them if there is no other element, (t_c, y_c), placed in between that intercepts the line connecting both values, that is (t_c, y_c) fulfills,

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}. \quad (1)$$

In this way, we can define a node for every single item of the sequence, so that the resulting network is always connected, undirected and invariant under affine transformations of the series [22]. Lacasa *et al.* [23] have shown that, for example, stochastic time series lead to networks characterized by their degree distributions, $P(k)$, which follow a power-law function, $P(k) \sim k^{-\gamma}$, with γ an exponent that reflects the level of correlations; in particular, $\gamma = 4 - \beta$, where β is the exponent of the power spectra of fractional Brownian motions (fBm) [23]. In addition to the success of the NVG, modified versions of the original method, such as the horizontal visibility algorithm and its directed version [26], have also previously proven to be rather useful to characterize features, such as correlation [26] and reversibility [7], when applied to a number of other time series, such as chaotic ones [26], polluted periodic signals [6], as well as other randomly-correlated ones.

2.2. Data

Our input data consisted of 30 ebooks in the English language downloaded mainly from the websites of the Gutenberg Project (http://www.gutenberg.org) and the Project Gutenberg Australia (http://gutenberg.net.au). There was not a particular strategy to select the titles, other than considering well-known works, as well as some that have been considered polemic or that treated polemic topics; and the selected books were first published in different epochs, therefore giving diversity in time. Table 1 lists the books considered in our study, including the total number of words M and of different words N_U. For a simple comparison, Figure 1 shows the scatter plot of M *vs.* N_U.

Table 1. Books considered in our study. The number of words and the number of different words are denoted by M and N_U, respectively.

#	**Title and Author**	M	N_U
1	The Aeneid, Virgil	112,478	8250
2	Animal Farm, G. Orwell	30,383	3921
3	Around the World in 80 Days, J. Verne	63,759	6822
4	The Bible, King James Ed.	884,964	12,806
5	Madame Bovary, G. Flaubert	117,536	10,298
6	The Catcher in the Rye, J. D. Salinger	77,555	4024
7	The Conquest of Bread, P. Kropotkin	72,016	7473
8	On the Origin of Species, C. Darwin	156,811	7237
9	The Picture of Dorian Gray, O. Wilde	80,407	6819
10	Dracula, B. Stoker	162,316	9291
11	The Diary of Anne Frank	104,753	8078
12	The Great Gatsby, F. S. Fitzgerald	50,102	5820
13	The Grapes of Wrath, J. Steinbeck	187,578	8696
14	Gulliver's Travels, J. Swift	104,797	8188
15	Hopscotch, J. Cortázar	195,702	15,338
16	The Iliad, Homer	157,581	9117
17	Mein Kampf, A. Hitler	273,387	12,697
18	Moby Dick, H. Melville	218,704	17,150
19	Pierre and Jean, G. de Maupassant	46,543	5400
20	Pride and Prejudice, J. Austen	122,878	6450
21	The Koran (Al-Qur'an)	210,967	11,058
22	Siddhartha, H. Hesse	39,773	3550
23	The Idiot, F. Dostoyevsky	247,952	10,102
24	Three Men in a Boat, J. K. Jerome	68,804	6779
25	The Time Machine, H. G. Wells	32,775	4594
26	The Trial, F. Kafka	86,391	4776
27	Ulysses, J. Joyce	272,415	29,898
28	War and Peace, L. Tolstoy	572,627	17,543
29	The War of the Worlds, H. G. Wells	60,896	6758
30	Thus Spake Zarathustra, Nietzsche	92,400	7534

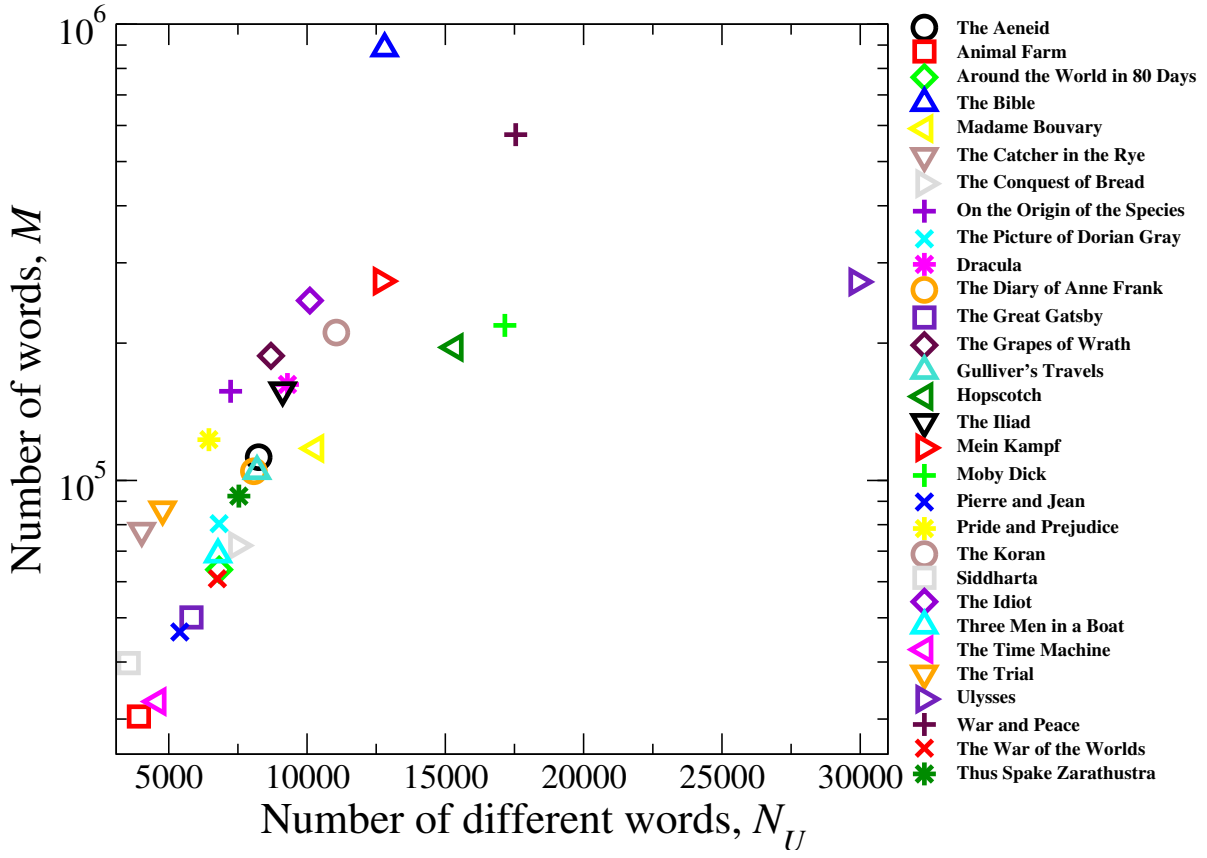

Figure 1. Scatter plot of the number of words M *vs.* the number of different words N_U for the books considered in our study. Notice that the longest book is *The Bible* with $M = 884{,}964$, whereas the shortest one is *Animal Farm* with $M = 30{,}383$. For new words in the text, that is how much innovation is present, the highest is *Ulysses*, while the lowest is *Siddhartha*.

3. Results

We consider the sequences of word- length obtained from the 30 books described above. We emphasize that the word length is given in terms of the number of letters forming a word. First, we notice that a direct application of the NVG to the word length sequences reveals that the NVG is highly inaccurate at capturing temporal correlation/organization structures in signals close to the transition from anti-persistent to persistent behavior [23]. For a more reliable application of the NVG to the word-length data, the sequence $\{l(1), l(2), ..., l(N)\}$ is first integrated to obtain the profile, $L(i) = \Sigma_{j=1}^{j=i}(l(j) - \bar{l})$, where $l(j)$ is the j-th word length, $\bar{l}$ is the mean value and $i = 1, .., N$ (see Figure 2). In this way, the integrated signal is within the fractional Brownian motion regime, and the NVG provides more reliable information of the emerging networks by means of, for example, quantities, like the degree distribution.

The NVG was applied to the integrated word length sequences in the dataset to get the corresponding networks, where nodes are the words (length value), and a link exists if there is a direct visibility between two values in the sequence, as shown in Figure 2.

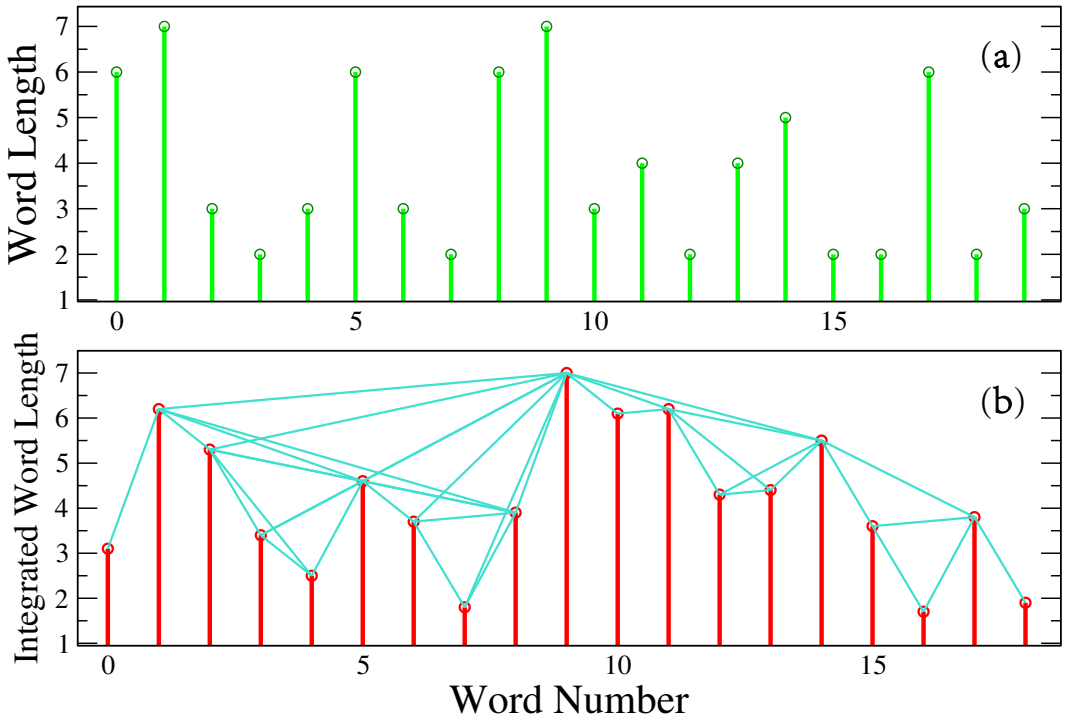

Figure 2. (**a**) Representative word-length sequence. (**b**) Visibility graph method applied to the integrated sequence of word lengths. Two values are connected if there is "visibility" between them.

The number of nodes of the resulting networks correspond to the number of words in the original literary text, while the order (number of edges) depends on the "visibility" restriction. First, we construct the probability degree distribution $P(k)$ from this network in order to characterize the connectivities. In Figure 3a, we depicted the degree distribution of the visibility networks from individual data. We find that a power-law behavior is observed for each book (Figure 3a), that is $P(k) \sim k^{-\gamma}$, with approximately two regimes; for the range $10 \leq k \leq 10^2$, the average exponent is $\gamma_s = 1.72 \pm 0.02$, while for $10^2 < k \leq 10^3$, the value is $\gamma_l = 1.34 \pm 0.07$. A significant difference is observed between γ_s and γ_l when comparing both groups of exponents (p-value $< 10^{-3}$ by student's test). It is important to evaluate the extent to which these distributions from different books correspond to the same distribution. To this end, we used the Kolmogorov–Smirnov (K-S) test to accept or reject the null hypothesis that any pair of distributions (books) has the same distribution. We computed the p-value between the cumulative distributions from all of the pairs of books.

In Figures 3b,c, we present the results obtained from the application of the K-S test to our dataset. We observe that, at the 5% level of significance, in all cases, we accept the null hypothesis that any two books have the same distribution, justifying that we can pool the data (degrees) from all books to get better statistics (see the caption of Figure 3 for a description). The results of pooling the data are shown in Figure 3b, where we find that $P(k)$ is consistent with a power law with two regimes separated by the crossover degree scale located at $k^* \approx 100$; over short scales, the probability of degrees decays following an exponent $\bar{\gamma}_s = 1.70 \pm 0.01$, which is bigger than the one corresponding to the large scales $\bar{\gamma}_l = 1.32 \pm 0.01$, confirming that the connectivities exhibit two different tendencies.

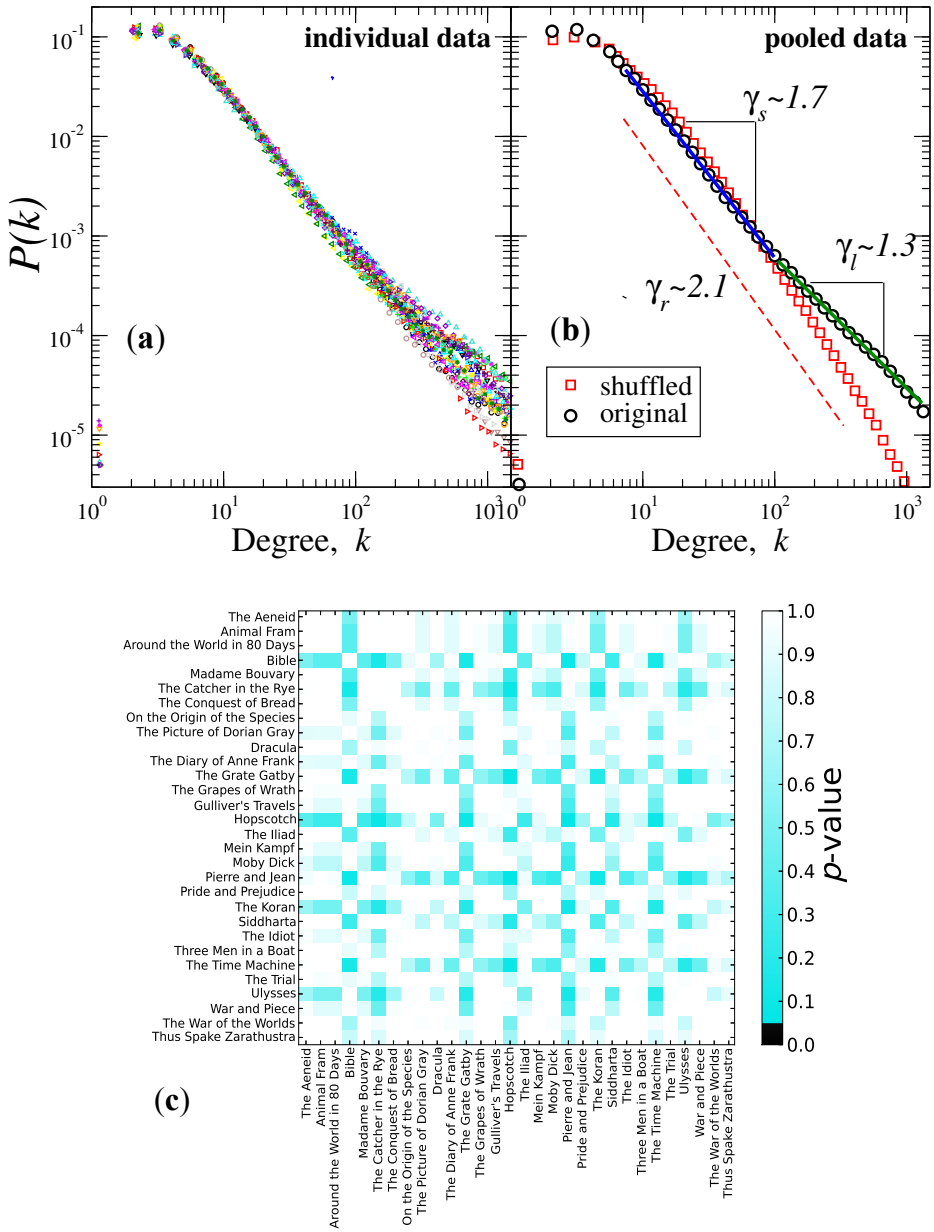

Figure 3. Visibility graph method (natural visibility graph (NVG)) analysis. Degree distributions of the degree probability $P(k)$ *versus* the degree k of the number of visible connections from the graph of the integrated word-length sequences. (**a**) Log-log plot of $P(k)$ *vs.* k of individual data. We observe that each book follows an overall function of the form $P(k) \sim k^{-\gamma}$, but with two apparent regimes having different exponents. (**b**) As in (a), but for pooled original and shuffled data. The original data follow a power-law function with two regimes separated by the degree scale value $k^* \approx 10^2$; over short degree values, the scaling exponent is $\gamma_s \approx 1.7$, while for large scales, $\gamma_l \approx 1.3$. We also show the distribution from shuffled data, and we observe that they follow a power law with exponent value $\gamma_r \approx 2.1$. (**c**) The matrix of *p*-values from the application of the Kolmogorov–Smirnov (K-S) test to all pairs of individual distributions shown in (a). We observe that at the 5% level of significance, for all of the books in our dataset, we cannot reject the null hypothesis that any pair of books has the same distribution. Therefore, we can pool the degree data from all of the texts and improve the statistics.

Lacasa *et al.* [23] have reported that the exponent γ, which characterizes the connectivities in fractional Brownian motion (fBm), is related to the exponent of power spectrum β through the relationship $\gamma = 4 - \beta$. Since our work here is based on the NVG analysis of the integrated word-length series, the corresponding relationship for the increments (*i.e.*, the original word-length values) would be $\gamma = 2 - \beta^{'}$, with $\beta^{'} = \beta - 2$ [27]. Given our values of $\gamma_s \approx 1.7$ and $\gamma_l \approx 1.3$, this means that the corresponding exponents of the power spectrum are given by $\beta_s^{'} \approx 0.3$ and $\beta_l^{'} \approx 0.7$, in agreement with previous results obtained for small datasets [5,20,21].

For a comparison, we repeated our procedure, but for the case of randomized versions of the word-length sequences, *i.e.*, the initial word-length series are shuffled in order to destroy correlations, and then, the NVG is applied to construct the degree distribution of those integrated series. The results are shown in Figure 3b for the case of pooled data. The randomized data lead to a degree distribution, which follows a power law $P_{random}(k) \sim k^{-\gamma_r}$, with $\gamma_r = 2.1$. According to the relation $\gamma = 4 - \beta$, the shuffled data lead to the value $\beta^{'} = \beta - 2 = 0.1$, in good agreement with the expected exponent for uncorrelated sequences.

We also compared these results from the NVG with the detrended fluctuation analysis (DFA) [28]. The DFA is a reliable method to detect long-range correlations in time series. In the DFA, the original time series is integrated; the resulting series is divided into boxes of size n, and for each box, a straight line is fitted to the points. Next, the root-mean-square fluctuation $F(n)$ is computed of the detrended sequence within each box. If a scaling function of the form $F(n) \sim n^{\alpha}$ is present, then the correlation exponent α characterizes the original signal. It is known that $\alpha = 0.5$ corresponds to white noise (non-correlated signal) and that $\alpha = 1$ corresponds to a long-range correlated process. In many cases, the scaling behavior in the fluctuation is not expressed through a single exponent, and two or more of them are necessary to characterize the signal [29,30]. For these cases and in order to get a good estimation of the α-values and the crossover point, we consider the following procedure: given the statistics of $F(n)$, a sliding pointer along the scale n is considered to perform linear regression fits to the values on the left and to the elements on the right. At each position of the pointer, we calculate the errors in the fits (e_l and e_r), monitor the total error defined by $e_t = e_l + e_r$ and find the position of the minimum of e_t. We then define two stable exponents (α_s and α_l) as the power law fit to the left and right, respectively, of e_t. For our data, e_t reaches its minimum value, and the position of the crossover point is within the interval $10 \leq n \leq 10^3$.

We use the DFA method to verify the presence of long-range correlations in the word-length sequences. We notice that in this case, the DFA method is applied to the original world-length series. As shown in Figure 4, the scaling behavior is characterized by two regimes; over short scales ($n < n_{\times}$), the average exponent is $\alpha_s = 0.52 \pm 0.04$, whereas for large scales ($n > n_{\times}$), the value is $\alpha_l = 0.71 \pm 0.07$. We observe that for short scales, the average exponent is close to 0.5, indicating an uncorrelated behavior, while over large scales, the average exponent is larger than 0.5, revealing a persistent behavior with long-range correlations. A significant difference is observed between α_s and α_l for the whole dataset (p-value $< 10^{-3}$ by Student's test). For a comparison, we shuffled the original word-length sequences in order to destroy correlations and repeated our procedure. The results are also depicted in Figure 4.

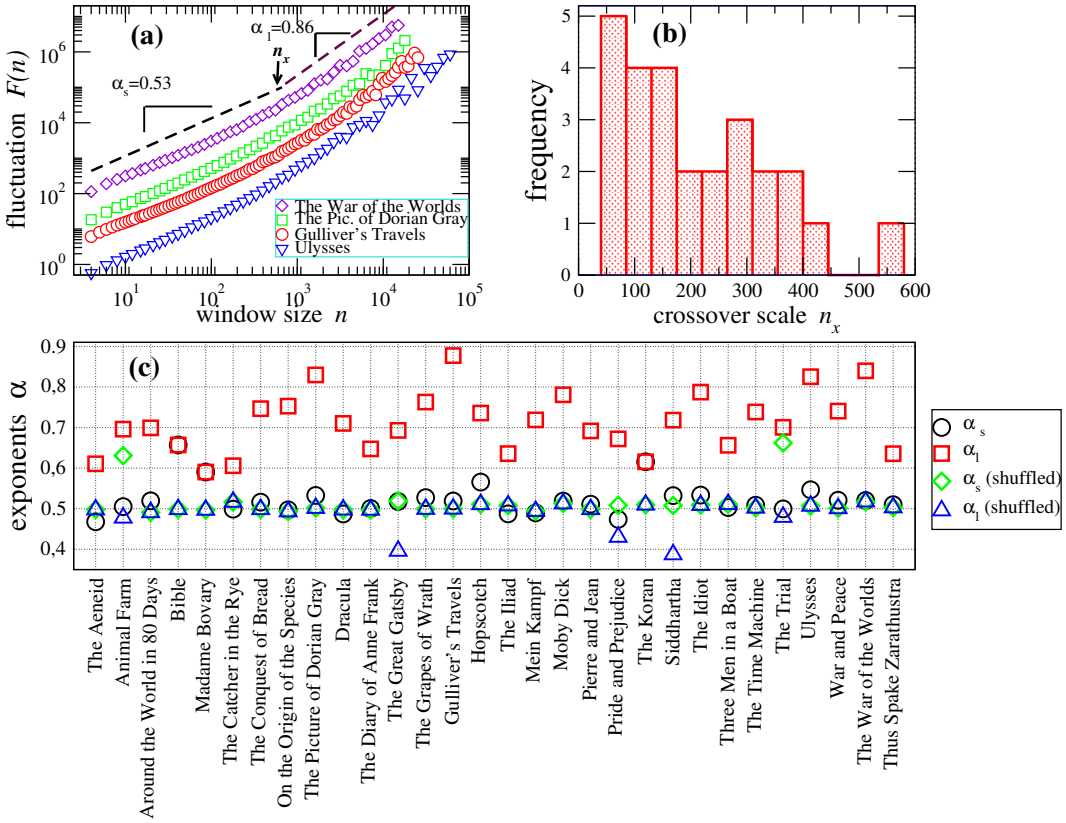

Figure 4. Detrended fluctuation analysis (DFA). Correlation exponents α_s and α_l of the detrended fluctuation analysis (DFA) for the books in the dataset. (**a**) Plots of $F(n)$ *vs.* n for four representative cases of crossover scaling in word-length sequences from the books *The Picture of Dorian Grey*, *Gulliver's Travels*, *Ulysses* and *The War of Worlds*. Two regimes in the exponent values are identified, and they are separated by the crossover scale $n_\times$, as is indicated. (**b**) Histogram of crossover scale position $n_\times$ for the DFA data. Most of the cases lie between the range of a 50–400 window size. (**c**) Correlation exponents α for the 31 books in our study. Here, the exponents' values were determined by means of two linear fittings for which the error is minimum (see the text for details). For most cases, the scaling exponents ($\alpha_s \approx 0.5$) from the small scales (squares) are close to that for the randomized data from the small and large scales (diamonds, triangles), indicating that there are no significant correlations in word lengths at small distances between the words. This contrasts with the behavior of most of the data at large scales (squares), where ($\alpha_l > 0.5$) indicates the presence of positive long-range correlations. However, for three books, *The Bible*, *Madame Bovary* and *The Koran*, the exponent values, α_s and α_l, at both small and large scales, are the same.

For all books, both scaling exponents collapse to the value of 0.5, confirming that the randomization procedure has destroyed the correlations. We notice that the correlation exponents from large scales are not uniform, and some deviations with respect to the mean are observed for specific texts, such as *Gulliver's Travels*, *The Picture of Dorian Grey*, *Ulysses* and *The War of the Worlds*, which exhibit an exponent value above 0.8. It is also worth noticing that the obtained values of DFA exponents are in qualitative concordance with the values observed under the visibility method, according to the relationships $\gamma = 2 - \beta'$ and $\alpha = (\beta' + 1)/2$ [28,31]; the exponents α and γ are related through

$\gamma = 3 - 2\alpha$. Given our values of $\alpha_s \approx 0.5$ ($\alpha_l \approx 0.7$), the corresponding exponents of the power spectrum and degree distribution are given by $\beta'_s \approx 0.0$ ($\beta'_l \approx 0.4$) and $\gamma_s \approx 1.9$ ($\gamma_l \approx 1.5$), respectively. We recall that the results from the DFA indicate that for small scales, there is very weak correlation, while for large scales, there are positive correlations, consistent with our results from the NVG.

Next, in order to test the presence of memory effects in word-length sequences, we consider the return times of word-length values equal or bigger than a given threshold [32]. The return time is given by the number of word lengths until the word length in question appears again. The threshold lengths $\ell_0 = 2, 3, 4, 5$ are considered. For each value of ℓ_0, we construct the cumulative probability distribution of the return times of all of the books in our study. Figure 5 shows four representative cases of the calculations. We observe that the return times for all thresholds can be approximately described by a stretched exponential distribution of the form $H(\tau) \sim e^{-a\tau^b}$, with a and b two fitting parameters, which reveal the information of the behavior of the distribution. This distribution is more skewed than a single exponential distribution and less skewed than a power law distribution. As $b \to 1$, it approaches a single exponential distribution, and as $b \to 0$, it approaches a power-law distribution. The fits to the individual data lead to the average values $\bar{b}_2 = 0.1.1 \pm 0.14$, $\bar{b}_3 = 0.93 \pm 0.06$, $\bar{b}_4 = 0.86 \pm 0.05$ and $\bar{b}_5 = 0.84 \pm 0.06$, where the subindex indicates the threshold value.

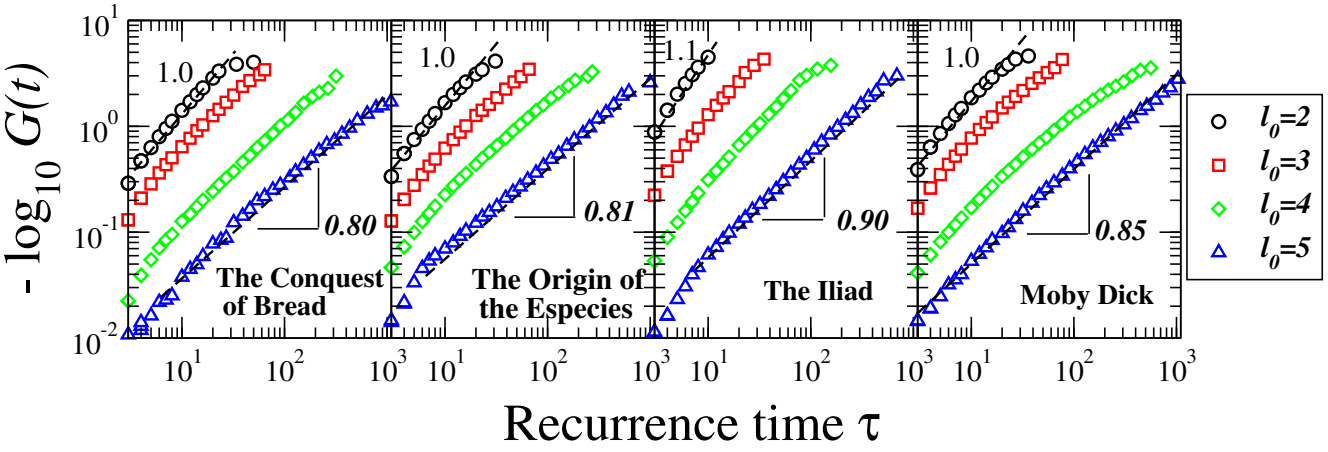

Figure 5. Recurrence time analysis. Plot of $-\log_{10} G(\tau)$ *versus*. recurrence time τ for four representative books in our study. We show the cases of threshold values $\ell_0 = 2$ (circles), $\ell_0 = 3$ (squares), $\ell_0 = 4$ (diamonds) and $\ell_0 = 5$ (triangles). The exponent $b = 1$ here represents the single exponential case, while $b < 1$ is a stretched exponential distribution. Consistent with the NVG and DFA analysis in Figures 3 and 4, the exponent b is closer to one, indicating no memory effects in word length for small values and $b < 1$ indicating the presence of memory in return intervals of large word-length values.

As shown in Figure 5, for $\ell_0 = 2$, the distributions are close to the exponential limit, indicating that the mechanism of selecting a word with a length equal to or above the threshold is time independent and can be explained as a simple Poisson process. As the value of the threshold increases, the value of b decreases, revealing that the burstiness of larger words tends to increase. This burstiness in larger words is also consistent with our results that the structure of the correlations changes at larger scales as these larger words appear at lower frequency and, therefore, over larger scales in the text.

4. Concluding Remarks

The frequency of word distributions has been well known for some time [1], but much less is known about the correlations b etween words and word lengths. Here, we have studied the correlation properties of word-length sequences from large literary texts. Our results, based on the visibility graph method (NVG), reveal that the degree distribution of the integrated word-length visibility networks can be described by a power-law function with approximately two regimes, while the corresponding distribution from shuffled data is characterized by the exponent value $\gamma = 2.1$, as expected for uncorrelated data [23]. Specifically, the correlation behavior is different at short and large word length degrees. For word-length degrees between 10 and 10^2, $\gamma_s = 1.7$, while for degrees between 10^2 and 10^3, $\gamma_l = 1.3$. These results are corroborated by our results from the detrended fluctuation analysis (DFA) where we found that for small scales, the sequences possess $\alpha_s \approx 0.5$, indicating no correlation, while for large scales, the word-length sequences have positive correlations, as expressed by the exponent $\alpha_l \approx 0.7$. Furthermore, the recurrence time distributions also exhibit a deviation with respect to pure exponential behavior as the threshold parameter increases, that is for small values ℓ_0, the low word lengths dominate the dynamics with no memory, while for the large ℓ_0, the high word lengths are distributed with recurrence times, which exhibit memory ($b < 1$). Thus, all three methods find that the lack of correlation in word lengths at small scales is replaced by a positive correlation in word lengths over large separations between words. The large-scale linguistic structures in these books are different, in an important way, compared to the small-scale structures. We hope that these new findings will serve as the basis for a linguistic or semantic interpretation of what they tell us about language or, more specifically, about written language.

Acknowledgments

This work was partially supported by COFAA-IPN , EDI-IPN and SNI-Conacyt, México. Lev Guzmán-Vargas thanks Conacyt, México (Grant 246568), and the Physics Department of Queens College for the hospitality. Support for this project was provided by a PSC-CUNY Award, jointly funded by The Professional Staff Congress and The City University of New York. This work was carried out during a sabbatical leave Lev Guzmán-Vargas.

Author Contributions

Lev Guzmán-Vargas and Larry S. Liebovitch designed the research. Lev Guzmán-Vargas, Daniel Aguilar-Velázquez, Bibiana Obregón-Quintana and Ricardo Hernández-Pérez, performed the research and analyzed the data. Lev Guzmán-Vargas and Larry S. Liebovitch wrote the paper. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Zipf, G.K. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*; M.I.T. Press: Cambridge, MA, USA, 1935.
2. Piantadosi, S.T.; Tily, H.; Gibson, E. Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3526–3529.
3. Garcia, D.; Garas, A.; Schweitzer, F. Positive words carry less information than negative words. *EPJ Data Sci.* **2012**, *1*, doi:10.1140/epjds3.
4. Altmann, E.G.; Pierrehumbert, J.B.; Motter, A.E. Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. *PLoS ONE* **2009**, *4*, e7678.
5. Kosmidis, K.; Kalampokis, A.; Argyrakis, P. Language time series analysis. *Physica A* **2006**, *370*, 808–816.
6. Nuñez, A.; Lacasa, L.; Valero, E.; Gómez, J.P.; Luque, B. Detecting series periodicity with horizontal visibility graphs. *Int. J. Bifurc. Chaos* **2012**, *22*, doi:10.1142/S021812741250160X.
7. Lacasa, L.; Nuñez, A.; Roldán, E.; Parrondo, J.; Luque, B. Time series irreversibility: A visibility graph approach. *Eur. Phys. J. B* **2012**, *85*, doi:10.1140/epjb/e2012-20809-8.
8. Petersen, A.; Tenenbaum, J.; Havlin, S.; Stanley, H.E. Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death. *Sci. Rep.* **2012**, *2*, doi:10.1038/srep00313.
9. Qian, M.C.; Jiang, Z.Q.; Zhou, W.X. Universal and nonuniversal allometric scaling behaviors in the visibility graphs of world stock market indices. *J. Phys. A Math. Theor.* **2010**, *43*, 335002.
10. Kalimeri, M.; Constantoudis, V.; Papadimitriou, C.; Karamanos, K.; Diakonos, F.K.; Papageorgiou, H. Entropy analysis of word-length series of natural language texts: Effects of text language and genre. *Int. J. Bifurc. Chaos* **2012**, *22*, doi:10.1142/S0218127412502239.
11. Kalimeri, M.; Constantoudis, V.; Papadimitriou, C.; Karamanos, K.; Diakonos, F.K.; Papageorgiou, H. Word-length Entropies and Correlations of Natural Language Written Texts. *J. Quant. Linguist.* **2015**, *22*, 101–118.
12. Rêgo, H.H.A.; Braunstein, L.A.; D'Agostino, G.; Stanley, H.E.; Miyazima, S. When a Text Is Translated Does the Complexity of Its Vocabulary Change? Translations and Target Readerships. *PLoS ONE* **2014**, *9*, e110213.
13. Solé, R.V.; Corominas-Murtra, B.; Valverde, S.; Steels, L. Language networks: Their structure, function, and evolution. *Complexity* **2010**, *15*, 20–26.
14. 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*. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* **2011**, *331*, 176–182.
15. Arbesman, S.; Strogatz, S.H.; Vitevitch, M.S. The structure of phonological networks across multiple languages. *Int. J. Bifurc. Chaos* **2010**, *20*, 679–685.
16. Arbesman, S.; Strogatz, S.H.; Vitevitch, M.S. Comparative Analysis of Networks of Phonologically Similar Words in English and Spanish. *Entropy* **2010**, *12*, 327–337.
17. Chan, K.Y.; Vitevitch, M.S. Network Structure Influences Speech Production. *Cognit. Sci.* **2010**, *34*, 685–697.

18. Grzybek, P. History and Methodology of Word Length Studies: The State of the Art. In *Contributions to the Science of Text and Language*; Springer: Amsterdam, The Netherlands, 2006; Volume 31, pp. 15–90.
19. Chen, H.; Liang, J.; Liu, H. How Does Word Length Evolve in Written Chinese? *PLoS ONE* **2015**, *10*, e0138567.
20. Ausloos, M. Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series. *Phys. Rev. E* **2012**, *86*, 031108.
21. Rodriguez, E.; Aguilar-Cornejo, M.; Femat, R.; Alvarez-Ramirez, J. Scale and time dependence of serial correlations in word-length time series of written texts. *Physica A* **2014**, *414*, 378–386.
22. Lacasa, L.; Luque, B.; Ballesteros, F.; Luque, J.; Nuño, J.C. From time series to complex networks: The visibility graph. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4972–4975.
23. Lacasa, L.; Luque, B.; Luque, J.; Nuño, J.C. The visibility graph: A new method for estimating the Hurst exponent of fractional Brownian motion. *Europhys. Lett.* **2009**, *86*, 30001.
24. Aguilar-San Juan, B.; Guzmán-Vargas, L. Earthquake magnitude time series: Scaling behavior of visibility networks. *Eur. Phys. J. B* **2013**, *86*, doi:10.1140/epjb/e2013-40762-2.
25. Telesca, L.; Lovallo, M. Analysis of seismic sequences by using the method of visibility graph. *Europhys. Lett.* **2012**, *97*, 50002.
26. Luque, B.; Lacasa, L.; Ballesteros, F.; Luque, J. Horizontal visibility graphs: Exact results for random time series. *Phys. Rev. E* **2009**, *80*, 046103.
27. Malamud, B.D.; Turcotte, D.L. Self-affine time series: Measures of weak and strong persistence. *J. Stat. Plan. Inference* **1999**, *80*, 173–196.
28. Peng, C.K.; Buldyrev, S.V.; Havlin, S.; Simons, M.; Stanley, H.E.; Goldberger, A.L. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **1994**, *49*, 1685–1689.
29. Guzmán-Vargas, L.; Munoz-Diosdado, A.; Angulo-Brown, F. Influence of the loss of time-constants repertoire in pathologic heartbeat dynamics. *Physica A* **2005**, *348*, 304–316.
30. Munoz-Diosdado, A.; Guzmán-Vargas, L.; Ramírez-Rojas, A.; Del Rio-Correa, J.; Angulo-Brown, F. Some cases of crossover behavior in heart interbeat and electroseismic time series. *Fractals* **2005**, *13*, 253–263.
31. Barabási, A.; Stanley, H. *Fractal Concepts in Surface Growth*; Cambridge University Press: Cambridge, MA, USA, 1995.
32. Eichner, J.F.; Kantelhardt, J.W.; Bunde, A.; Havlin, S. Statistics of return intervals in long-term correlated records. *Phys. Rev. E* **2007**, *75*, 011128.