v7.n10.2026

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

# *LONG-RANGE DEPENDENCE IN LEXICAL RENEWAL SERIES*

# DEPENDÊNCIA DE LONGO ALCANCE EM SÉRIES DE RENOVAÇÃO LEXICAL

# *DEPENDENCIA DE LARGO ALCANCE EN SERIES DE RENOVACIÓN LÉXICA*

Regis Nunes Vargas[1], Raquel Romes Linhares[2]

e7109165
https://doi.org/10.47820/recima21.v7i10.9165

PUBLISHED: 10/2026

***ABSTRACT***

*This work proposes a numerical series to describe the long-range structure of written texts: each word is mapped to one of four frequency classes, and the number of times the class changes from one position to the next is counted in successive blocks along the text. Classical Detrended Fluctuation Analysis (DFA) and its wavelet- and clustering-based variant (DFA-WK) are applied to this series, using a corpus of public-domain literary works. The resulting exponents indicate long-range dependence in most texts, a result confirmed by a control test involving word shuffling, which preserves the frequency distribution while destroying the original order. The exponent proved stable across genres, authors, and eras, yet it did not distinguish between texts with high and low lexical repetition; this suggests it captures a more general structural property of writing rather than serving as a measure of vocabulary richness. Since large language models learn the statistical regularities of text corpora, this long-range structure is a promising candidate property for comparing human-written text with machine-generated text, a possibility not tested here but supported by the observed stability of the exponent.*

***KEYWORDS:*** *Long-range dependence. Detrended fluctuation analysis. Renewal series. Quantitative linguistics. Language models.*

**RESUMO**

Este trabalho propõe uma série numérica para descrever a estrutura de longo alcance de textos escritos: cada palavra é mapeada em uma de quatro classes de frequência e conta-se, em blocos sucessivos ao longo do texto, quantas vezes a classe muda de uma posição para a seguinte. Aplica-se a essa série a Análise de Flutuação Destendenciada (DFA) clássica e a sua variante baseada em wavelets e agrupamento (DFA-WK), usando um corpus de obras literárias em domínio público. Os expoentes resultantes indicam dependência de longo alcance na maior parte dos textos, resultado confirmado por um teste de controle com embaralhamento das palavras, que preserva a distribuição de frequências e destrói a ordem original. O expoente mostrou-se estável entre gêneros, autores e épocas, mas não distinguiu textos com alta e baixa repetição lexical; isso sugere que ele capta uma propriedade estrutural mais geral da escrita, em vez de servir como medida de riqueza vocabular. Como os grandes modelos de linguagem aprendem as regularidades estatísticas de corpora textuais, essa estrutura de longo alcance é uma propriedade candidata promissora para comparar texto escrito por humanos com texto gerado por máquina, possibilidade não testada aqui, mas sustentada pela estabilidade observada do expoente.

[1] PhD in Electrical Engineering from the School of Electrical Engineering at the Federal University of Uberlândia and teacher at the Uberlândia Municipal Department of Education.
[2] PhD in Mathematics from the Federal University of Rio Grande do Sul (UFRGS) and research professor at the Institute of Mathematics and Statistics of the Federal University of Uberlândia.



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

**PALAVRAS-CHAVE:** Dependência de longo alcance. Análise de flutuação destendenciada. Série de renovação. Linguística quantitativa. Modelos de linguagem.

***RESUMEN***

*Este trabajo propone una serie numérica para describir la estructura de largo alcance de los textos escritos: cada palabra se asigna a una de cuatro clases de frecuencia y se cuenta, en bloques sucesivos a lo largo del texto, cuántas veces cambia la clase de una posición a la siguiente. Se aplica a esta serie el Análisis de Fluctuación sin Tendencia (DFA) clásico y su variante basada en wavelets y agrupamiento (DFA-WK), utilizando un corpus de obras literarias de dominio público. Los exponentes resultantes indican dependencia de largo alcance en la mayoría de los textos, resultado confirmado por una prueba de control con barajado de palabras, que preserva la distribución de frecuencias y destruye el orden original. El exponente se mostró estable entre géneros, autores y épocas, pero no distinguió textos con alta y baja repetición léxica; esto sugiere que capta una propiedad estructural más general de la escritura, en lugar de servir como medida de riqueza de vocabulario. Dado que los grandes modelos de lenguaje aprenden las regularidades estadísticas de los corpus textuales, esta estructura de largo alcance es una propiedad candidata prometedora para comparar texto escrito por humanos con texto generado por máquina, posibilidad no evaluada aquí, pero respaldada por la estabilidad observada del exponente.*

***PALABRAS CLAVE:*** *Dependencia de largo alcance. Análisis de fluctuación sin tendencia. Serie de renovación. Lingüística cuantitativa. Modelos de lenguaje.*

## INTRODUCTION

Linhares (2023) proposed modeling DNA sequences as a renewal process: a renewal occurs whenever the nitrogenous base changes between consecutive positions, and the numerical series analyzed is built by counting these renewals in fixed-size blocks. The central advantage of this construction is that it avoids a well-known problem in this area: symbolic sequences are not numerical, and transformations that assign arbitrary values to symbols can alter the very conclusion about the existence of long-range dependence. Linhares's (2023) method sidesteps this problem because it does not depend on any property specific to DNA: any symbolic sequence can, in principle, be described in the same way.

Natural-language texts are another class of symbolic sequences, and it is natural to ask whether they exhibit the same kind of long-range correlated structure documented in DNA. Text, however, has a crucial difference in alphabet size: rather than four symbols, it has thousands of vocabulary types. This matters because the definition of renewal depends on alphabet size: consecutive identical words are rare, and applying the definition literally produces a series that is practically constant, as documented in the subsection on reducing text to a four-class alphabet.

This work proposes and evaluates the transposition of the renewal-frequency series to text documents. The contributions are threefold: it is shown that this transposition requires first reducing the vocabulary to a small alphabet, and a mapping based on frequency quartiles is



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

proposed together with an alternative control scheme; the resulting series is submitted to classical DFA and to DFA-WK on a corpus of 200 literary works; and the result is subjected to a shuffling control, which establishes that the detected dependence stems from word order and reveals a residual positive bias of the estimator, whose magnitude is quantified here.

The general objective is therefore to determine whether the renewal-frequency series, originally formulated for DNA, describes the long-range structure of written text once the vocabulary has been reduced to a small alphabet. Three specific objectives follow from it: to define and justify a reduction of the vocabulary to four frequency classes that keeps the renewal series informative; to estimate the scaling exponent of the resulting series with classical DFA and with DFA-WK over a corpus of 200 literary works; and to establish, by means of a shuffling control, whether the exponents obtained reflect the order of the words or an artifact of the construction itself. The problem that motivates the study is that the renewal construction, as originally stated, simply does not work on text: with thousands of vocabulary types, almost every position is a renewal and the series carries no information. The justification lies in what a positive answer would make available, namely a description of the long-range structure of a text that depends only on the notion of identity between symbols, with no arbitrary numerical coding of the vocabulary.

## 1. THEORETICAL FRAMEWORK

Long-range correlations are recurring features of natural complex systems, manifesting as non-summable autocovariances and unbounded spectral density at the origin (Beran, 1994). Among natural symbolic sequences, DNA sequences are the most extensively studied case, from the initial identification of correlations in nucleotides (Peng et al., 1992) to investigations of their structural origin (Karlin and Brendel, 1993; Peng et al., 1994; Oliver et al., 2004), their scaling properties in non-coding regions (Stanley et al., 1999; Crato et al., 2011), their entropic characterization (Karmeshu and Krishnamachari, 2004; Melnik and Usatenko, 2014), and their analysis via long-memory methods (Lopes and Nunes, 2006; Podobnik et al., 2007).

The property at stake in all of these studies is the rate at which dependence decays with distance. In a process without long memory, the influence of one observation on another vanishes quickly enough for the autocovariances to be summable, and averages taken over long stretches converge at the usual rate. In a process with long-range dependence, that decay is slow enough for the sum to diverge, so that distant parts of the sequence remain statistically linked (Beran, 1994). This is what makes the phenomenon interesting in a symbolic sequence: it says that the structure of the object is not exhausted by the local rules that generate it, whether those rules are



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

the biochemical constraints of a genome or the grammar and subject matter of a text. It is also what makes the phenomenon delicate to measure, since a slowly decaying dependence and a slowly varying trend produce similar signatures in the raw autocorrelation, which is precisely the difficulty that detrended methods were designed to address.

A methodological problem runs through this literature: symbolic sequences are not numerical, and analysis requires a prior transformation. Approaches that assign arbitrary values to symbols produce results that may depend on the chosen assignment, and it remains debated to what extent the type of transformation alters the conclusion about long-range dependence (Linhares, 2023). Linhares (2023) proposed an alternative that circumvents this problem: modeling the sequence as a renewal process (in which a renewal occurs whenever the symbol changes between consecutive positions) and building the series by counting renewals in fixed-size blocks. This construction requires only the notion of identity between symbols, and it connects to the Fractional Poisson Process (Laskin, 2003; Cahoy, Uchaikin, and Woyczynski, 2010). The scaling exponent is estimated via Detrended Fluctuation Analysis (Peng et al., 1994), whose statistical properties were studied by Crato, Linhares, and Lopes (2010), and from which variants such as smoothed DFA (Linhares, 2016) and DFA-WK (Linhares and Vargas, 2025a, 2025b) derive.

Two features of this estimation route deserve to be made explicit, because they determine how the results reported below should be read. The first is the reason for a detrended estimator. Removing a local trend inside each block, rather than subtracting a single global mean, is what allows the fluctuation to be measured without confusing slowly decaying dependence with a slow drift of the level of the series (Peng et al., 1994); for the renewal series studied here, such a drift would arise from nothing more exotic than a long descriptive passage or a stretch of dialogue, which changes the local mixture of function words and content words. The second is the reading of the exponent itself. By construction, the value $\alpha = 0.5$ corresponds to the absence of long memory; values above it indicate persistence, in the sense that blocks with many renewals tend to be followed by blocks with many renewals; and values below it indicate the opposite tendency. The exponent is therefore a summary of one property of the sequence, not a general measure of complexity, and the question of what textual property it tracks has to be settled empirically, which is what the comparison by lexical repetition reported in Section 3 sets out to do.



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

## 2. METHODOLOGY

### 2.1. Corpus

The corpus comprises 200 public-domain literary works in English from Project Gutenberg (Project Gutenberg, 2026): novels, short stories, essays, philosophy, drama, and poetry, by authors from different periods. The texts were obtained through the individual mirrors of the GITenberg project on GitHub (GITenberg, 2026), which distributes each work as a separate Git repository. Texts range from 20,303 to 904,897 tokens (median 82,731). Each document was lowercased, had Project Gutenberg headers and footers removed, and was tokenized into alphabetic sequences. Exact content duplicates were identified by checksum and removed.

Each of these preprocessing decisions serves the construction that follows. Lowercasing prevents the same word from being split into two vocabulary types by sentence position alone, which would shift its frequency and therefore its class. Removing the Project Gutenberg headers and footers matters because that material is editorial boilerplate repeated across the collection, and it would contribute renewal counts unrelated to the work itself. Restricting tokens to alphabetic sequences keeps the vocabulary comparable across works published under different typographic conventions. Removing exact duplicates by checksum avoids counting the same text twice in a corpus assembled from independent repositories, which would give the exponents of the repeated work undue weight in the aggregate statistics reported below.

### 2.2. Reducing text to a four-class alphabet

Applying the renewal definition directly to words (a renewal whenever $w_t \neq w_{t-1}$) does not work. This was tested on one work from the corpus: the renewal rate was 99.9% of positions, with a standard deviation of only 1.04 renewals per block of 1,000 tokens. The series comes out practically constant, and DFA has nothing to estimate. The problem is alphabet size: DNA has four bases, so repeating the same base at the next position is common; a text has thousands of distinct words, and the same word almost never repeats at the next position.

The solution is to map each token to one of four classes, the same way DNA uses four bases. The construction proceeds through three chained series, shown in Figure 1 and explained step by step below.

**Step 1: frequencies by type.** The number of occurrences of each distinct vocabulary type is counted over the entire document. The total frequency of type $w$ is denoted by $freq(w)$.

**Step 2: series X.** The text is traversed in the order it was written, and the frequency of the type present at each position $t$ is recorded:



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

$$X(t) = freq(w_t), \qquad t = 1, \dots, n.$$

Series $X$ therefore has the same length $n$ as the text, and the value of a type is repeated at every position it occupies. In the opening excerpt of *Pride and Prejudice* shown in panel (A) of Figure 1, the word *a*, which occurs 2,012 times in the novel, receives the value 2,012 at each of its two occurrences, while *truth* receives 27 and *universally* receives 3.

**Step 3: quartiles of X.** The three quartiles $Q_1, Q_2, Q_3$ of the distribution of $X$ are computed. One detail matters here: the reference distribution is that of the $n$ values of $X$, weighted by token, not the frequencies of the distinct types. This choice is not a convenience but a necessity: in the corpus used here, 38.7% of distinct words are *hapax legomena* (occurring exactly once) and 53.3% occur at most twice, so quartiles computed over distinct types would give $Q_1 = Q_2 = 1$, coincident cutoffs that make it impossible to define four classes. Weighting by token, the cutoffs come out well defined: in the example work, $Q_1 = 42$, $Q_2 = 360$, and $Q_3 = 1{,}450$ (panel B of Figure 1). Each document has its own cutoffs, computed independently.

**Step 4: series Y.** Each position receives the class of the quartile in which its $X$ value falls, generating series $Y$ in the same order as the text:

$$Y(t) = \begin{cases} C_1, & X(t) \le Q_1, \\ C_2, & Q_1 < X(t) \le Q_2, \\ C_3, & Q_2 < X(t) \le Q_3, \\ C_4, & X(t) > Q_3. \end{cases}$$

Because frequency is a property of the word, the class is too: in a given document, *a* will always be $C_4$, and *truth* always $C_1$. Series $Y$ is the text rewritten in a four-symbol alphabet, the direct equivalent of DNA's four bases A, G, C, T. Note that $X$ and $Y$ are never reordered: sorting happens only inside the quartile computation and is discarded immediately afterward, because it is the order of the words in the text that carries the information to be measured.

The four classes come out balanced in number of tokens (25.2%, 25.1%, 24.9%, and 24.8%), but strongly unequal in number of distinct words: $C_4$ has only 12 types, the most common function words, versus 6,366 types in $C_1$. This is Zipf's law showing up, and it is what makes the renewal series register, at bottom, the alternation between function words and content words.

As a control, the entire procedure was repeated using word length instead of frequency, with quartiles computed the same way.

### 2.3. Renewal series and estimation of the scaling exponent

**Step 5: renewals.** A renewal occurs at position $t$ when $Y(t) \neq Y(t-1)$, that is, when the class changes relative to the previous position. In panel (C) of Figure 1, *truth universally*



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

*acknowledged* stays entirely in $C_1$ and generates no renewal; the transition to *that* ($C_4$) generates one.

**Step 6: series N.** Series $Y$ is divided into $l$ blocks of $t$ positions, and the number of renewals contained in each block $B_j$ is counted:

$$N(j) = \#\{\, t \in B_j : Y(t) \neq Y(t-1) \,\}, \qquad j = 1, \ldots, l.$$

It is this series $N$, of length $l < n$, that is analyzed; it corresponds to Linhares's (2023) definition. Block counting is not a dispensable detail: applying DFA directly to the renewal indicator, position by position, would yield a sequence of zeros and ones close to pure noise, without the smoothing effect that grouping into blocks produces.

The value $l = 0.01\,n$ is used, which gives blocks of $t \approx 100$ tokens across all documents. This value balances two sides: blocks large enough for the renewal count to actually vary, and numerous enough blocks to estimate the exponent with stability. With this, the median was 827 blocks per document, and the mean renewal rate was 0.787. In the example work, $l = 1{,}251$ blocks, with $N(j)$ having mean 79.0 and standard deviation 4.44 (panel D of Figure 1).

Detrended Fluctuation Analysis is applied to the series $\{N(j)\}_{j=1}^{l}$, following Peng et al. (1994) and Linhares's (2023) formulation. First the series is integrated by summing deviations from the mean:

$$P_t = \sum_{j=1}^{t} \big(N(j) - \overline{N}\big).$$

This profile $P_t$ is divided into blocks of $k$ observations; in each block a least-squares line is fitted, $\hat{P}_j^{\,k}$, and this local trend is subtracted: $Z_j^k = P_j - \hat{P}_j^{\,k}$. The mean-squared fluctuation at scale $k$ is

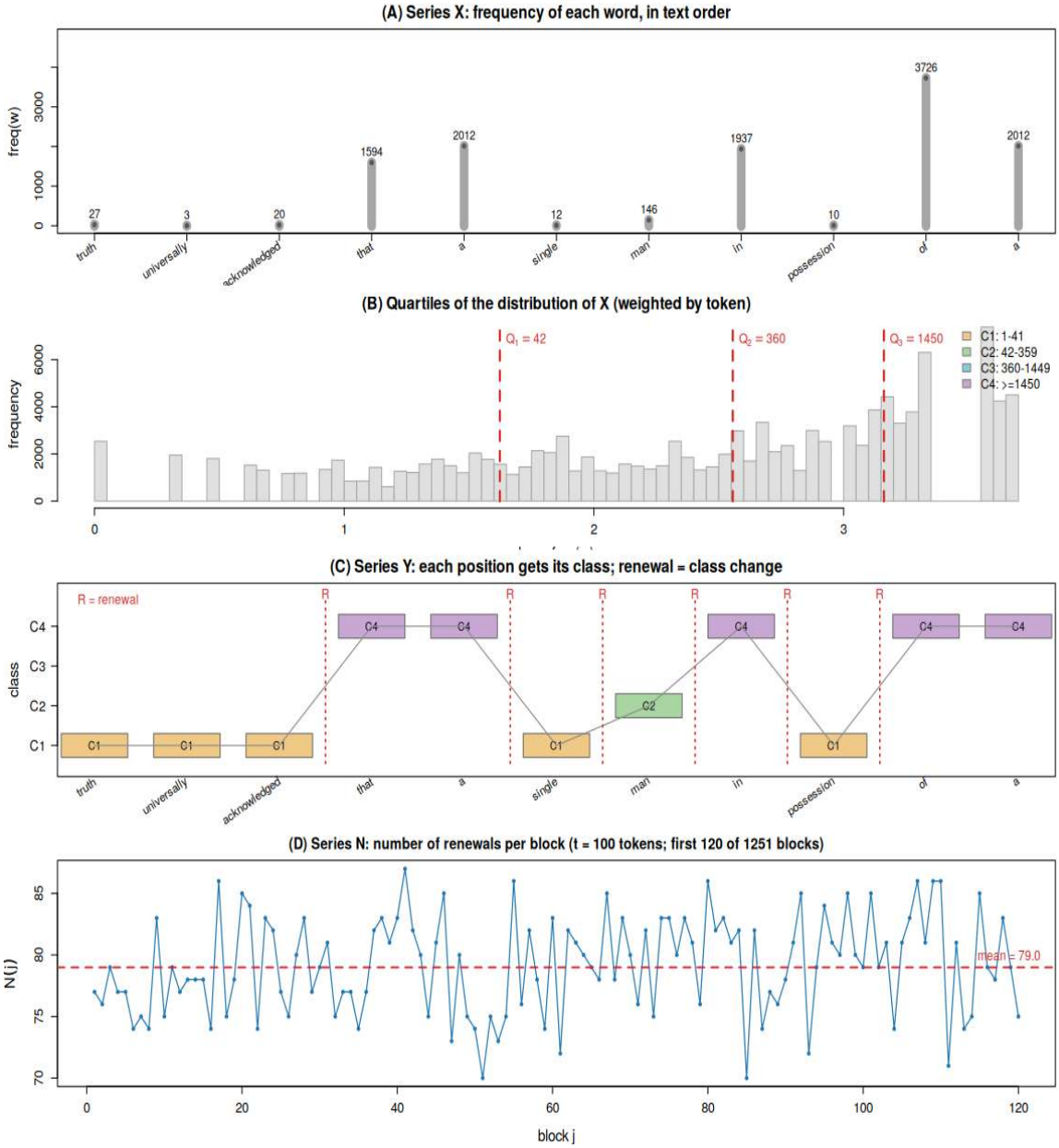
$$F(k) = \sqrt{\frac{1}{\tilde{l}} \sum_{j=1}^{\tilde{l}} (Z_j^k)^2}, \qquad \tilde{l} = k \cdot \left\lfloor \frac{l}{k} \right\rfloor,$$

and the exponent $\alpha$ is the slope of the line of $\ln F(k)$ against $\ln k$, since $F(k) \sim \varphi\, k^{\alpha}$. The choices $g(l) = \lfloor l/10 \rfloor$ and scales $k$ from 4 to $g(l) - 3$ are adopted here; the 95% confidence interval for $\alpha$ comes from the standard error of that slope. By convention, $\alpha = 0.5$ means no long memory, and $\alpha > 0.5$ indicates long-range dependence.



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

**Figure 1.** Construction of the renewal series, illustrated with the opening excerpt of *Pride and Prejudice*. (A) Series $X$: frequency of each word in the document, in text order. (B) Distribution of $X$ (log scale) and the three quartiles defining the four classes. (C) Series $Y$: the class assigned to each position; R marks indicate renewals. (D) Series $N$: number of renewals per block of $t = 100$ tokens, first $120$ of $1{,}251$ blocks.




Source: Prepared by the authors (2026).



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

The DFA-WK (Linhares and Vargas, 2025a) is also applied, which combines the Continuous Wavelet Transform, $K$-means, and local DFA. The CWT

$$W(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} x\,(t)\,\psi^{*}\left(\frac{t-b}{a}\right) dt$$

uses the Morlet wavelet, $\psi(t) = \pi^{-1/4} e^{i\omega_0 t} e^{-t^2/2}$ with $\omega_0 = 6$ (Daubechies, 1992; Percival and Walden, 2000), across five logarithmically spaced scales adapted to the length of the series (from $2$ to $l/20$). From the power spectrum $|W(a,b)|^2$, three features are extracted for each position: local energy $E_t$, variance across scales $\sigma_t^2$, and the ratio $R_t$ between high- and low-frequency components. $K$-means (MacQueen, 1967), with $K = 3$, groups the series into homogeneous regimes based on these features; DFA runs separately on each regime with at least $80$ points, yielding $H_k$, and the final exponent is the weighted average by regime size:

$$H_{\text{global}} = \sum_{k} w_k\, H_k, \qquad w_k = \frac{|C_k|}{\sum_j |C_j|}.$$

### 2.4. Shuffling control

To check whether the exponents really come from word order, and not from some artifact of how the series is built, the entire procedure was repeated on shuffled versions of each document. Shuffling the tokens preserves the frequency distribution, the vocabulary, and the four-class alphabet; only the order changes. If long-range dependence is a real property of the sequence, shuffling should erase it, and the exponents should fall close to $0.5$.

The control is what separates two explanations that would otherwise produce the same evidence. An exponent above $0.5$ may reflect a genuine ordering structure in the text, or it may be produced by the construction itself, since both the reduction to four classes and the counting of renewals in blocks impose structure on the data before any estimation takes place. Shuffling keeps intact everything the second explanation relies on and removes only what the first one requires: the vocabulary, the frequency distribution and the class cutoffs are exactly those of the original document, and the only thing destroyed is the sequence. Whatever exponent survives shuffling is therefore attributable to the construction rather than to the text, which is what makes the shuffled value, and not the theoretical $0.5$, the appropriate baseline for the comparison reported below.



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

## 3. RESULTS AND DISCUSSION

Table 1 summarizes the exponents across the 200 works. Both methods gave similar values, around $0.59$, with low dispersion. In $187$ of the $200$ texts $(93.5\%)$, the 95% confidence interval of $\alpha$ lay entirely above $0.5$, that is, statistically significant long-range dependence.

**Table 1.** Scaling exponents of the renewal-frequency series across 200 literary works.

| Method | Mean | Median | SD | Range |
|---|---|---|---|---|
| Classical DFA ($\alpha$) | 0.5919 | 0.5830 | 0.0572 | 0.490 - 0.885 |
| DFA-WK ($H_{\text{global}}$) | 0.5986 | 0.5917 | 0.0580 | 0.474 - 0.908 |

Source: Prepared by the authors (2026).

The $\alpha$ and $H_{\text{global}}$ are correlated (Pearson $r = 0.638$; Spearman $0.612$), but they are not the same thing, as panel (D) of Figure 2 shows. An interesting detail: within each document, the three regimes found by $K$-means have very similar local exponents (means of $0.580$, $0.581$, and $0.585$; mean amplitude of only $0.063$). This means the lexical renewal series is fairly homogeneous: it does not have the regime shifts that justify DFA-WK in financial or physiological series.

Shuffling the tokens knocked down both exponents (Table 2). The differences are large relative to their dispersion and clearly significant: in $89.5\%$ of texts, the original exponent came out above the shuffled one.

**Table 2.** Token-shuffling control (n = 200 pairs).

| Method | Original | Shuffled | Difference | 95% CI of diff. | p (paired t) |
|---|---|---|---|---|---|
| Classical DFA | 0.5919 | 0.5225 | +0.0694 | [+0.060; +0.079] | $3.9 \times 10^{-34}$ |
| DFA-WK | 0.5986 | 0.5331 | +0.0655 | [+0.056; +0.075] | $1.3 \times 10^{-28}$ |

Source: Prepared by the authors (2026).

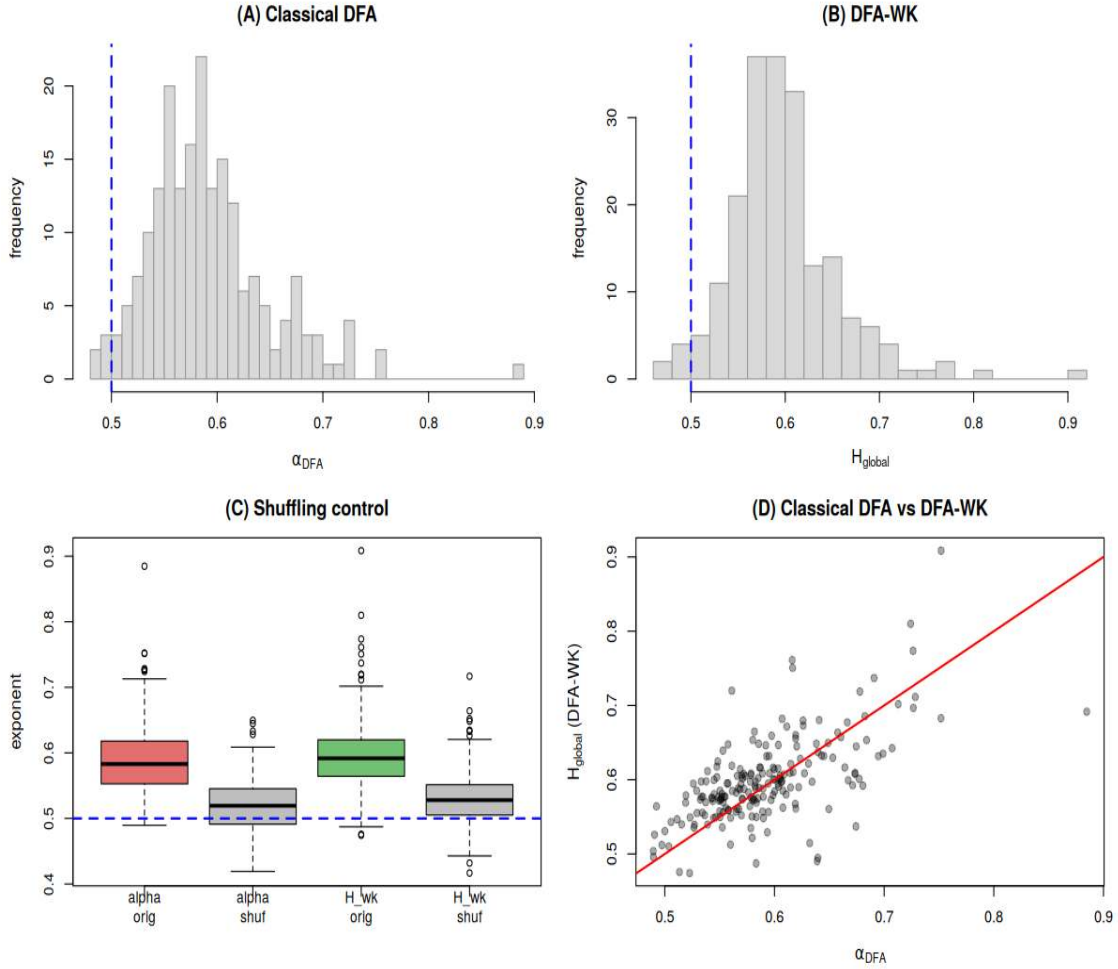
The corpus was split by the median Yule's $K$ (Yule, 1944; $92.63$) into two groups of $100$ texts, high and low lexical repetition. Neither exponent differed between groups: $\alpha = 0.5941$ versus $0.5898$ (difference $+0.0043$; 95% CI $[-0.012;\ +0.020]$; $p = 0.60$) for classical DFA, and $H_{\text{global}} = 0.5975$ versus $0.5996$ ($p = 0.80$) for DFA-WK. Correlations with continuous Yule's $K$ were



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

essentially null ($-0.050$ and $-0.068$), and the word-length control gave the same null result ($p = 0.26$).

**Figure 2.** (A) Distribution of $\alpha$ (classical DFA). (B) Distribution of $H_{\text{global}}$ (DFA-WK). (C) Comparison between original and shuffled series. (D) Relationship between the two estimators. The dashed line marks the value $0.5$.




Source: Prepared by the authors (2026).



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

The renewal-frequency series, carried over from DNA to text, finds long-range dependence consistently: $93.5\%$ of the works analyzed had an exponent significantly above $0.5$. The shuffling control shows that this comes from word order, not from some artifact of how the series is built, a check worth making, since binning into blocks and reducing to four classes could, in principle, create spurious correlation.

That same control reveals a detail worth recording: the shuffled series did not fall exactly to $0.5$, but to $0.522$ (classical DFA) and $0.533$ (DFA-WK), values statistically different from $0.5$ ($p = 6.5 \times 10^{-13}$). There is a residual positive bias of $0.02$ to $0.03$, likely due to series length or to how the series is constructed. It is therefore recommended that the shuffled exponent be used as a practical baseline instead of the theoretical $0.5$, since it is the method's real baseline, not its ideal one.

The size of this bias is small in absolute terms, but it is of the same order as the effects that a study of this kind would typically want to detect, which is what makes it worth reporting rather than rounding away. An exponent of $0.55$ measured against the theoretical baseline would look like evidence of long-range dependence, while measured against the shuffled baseline of $0.522$ it is a much weaker claim. The practical recommendation that follows is that the shuffled series be computed and reported whenever this construction is applied to a new corpus, at the cost of doubling the computation, since the bias depends on series length and on the reduction scheme, and therefore cannot be assumed to carry over unchanged from the values reported here.

The exponents obtained ($0.49$ to $0.91$) fall in a range comparable to what Linhares (2023) reported for DNA from chromosomes 21 and X ($0.47$ to $0.99$), suggesting similar magnitude across the two domains, although a more rigorous comparison would require matching block size and series length, which was not done here.

The null result on lexical repetition has a direct reading: the exponent was too stable to discriminate, with a standard deviation of only $0.057$ across a corpus spanning seventeenth-century verse plays to novels and philosophical treatises. This suggests the exponent does not measure vocabulary richness, but something more general about the structure of written language, common across genres and eras; this stability is itself the study's main finding.

### 3.1. Limitations and future work

Three limitations bound the scope of what has been shown. The first is the corpus: it is English-only and restricted to public-domain works, which places its bulk in a period that ends in the early twentieth century, so neither generalization to other languages nor generalization to



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

contemporary writing is supported by these data. The second is the reduction to four classes. It is necessary, in the sense that the renewal definition yields nothing without it, but it is one choice among several. The word-length control gave the same null result on lexical repetition, which is evidence that the finding does not hinge on the frequency criterion in particular; it is not evidence that every reduction would behave alike, and a reduction by grammatical class, for instance, might well track a different property. The third is the estimation itself: block size and the range of scales were fixed at values that work across documents of very different lengths, and the residual bias documented above shows that these choices leave a footprint on the exponent.

Each limitation suggests its own continuation. Repeating the procedure on corpora in other languages would show whether the stability of the exponent is a property of written language or of English in particular, and the comparison is direct, since the construction requires nothing beyond identity between symbols. Testing alternative reductions would establish how much of the result belongs to the frequency criterion and how much to the mere fact of reducing the alphabet. A sensitivity analysis over block size and scale range would quantify how much of the exponent is method and how much is text. Finally, since large language models learn the statistical regularities of the corpora they are trained on, the long-range structure measured here is a candidate property for comparing human-written text with machine-generated text; that comparison was not attempted in this study, but the stability of the exponent across genres, authors and eras is what makes it worth attempting.

## 4. CONCLUSION

The renewal-frequency series, adapted from DNA to text via a four-class alphabet, detects long-range dependence robustly, as confirmed by the shuffling control, and stably across genres, authors, and eras. It does not, however, serve as a measure of vocabulary richness: it did not distinguish texts with high and low lexical repetition. The method thus proves useful for describing the long-range correlational structure of a text, but that structure appears to be a general property of writing, not a substitute for the usual measures of lexical diversity. The original texts come from Project Gutenberg (Project Gutenberg, 2026), obtained through the mirrors of the GITenberg project on GitHub (GITenberg, 2026).

## REFERENCES

BERAN, J. *Statistics for Long Memory Processes*. New York: Chapman & Hall, 1994.

CAHOY, D. O.; UCHAIKIN, V. V.; WOYCZYNSKI, W. A. Parameter estimation for fractional Poisson processes. *Journal of Statistical Planning and Inference*, v. 140, p. 3106-3120, 2010.



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

CRATO, N.; LINHARES, R. R.; LOPES, S. R. C. Statistical properties of detrended fluctuation analysis. *Journal of Statistical Computation and Simulation*, v. 80, n. 6, p. 625-641, 2010.

CRATO, N.; LINHARES, R. R.; LOPES, S. R. C. $\alpha$-stable laws for noncoding regions in DNA sequences. *Journal of Applied Statistics*, v. 38, n. 2, p. 261-271, 2011.

DAUBECHIES, I. *Ten Lectures on Wavelets*. Philadelphia: SIAM, 1992.

GITENBERG. *GITenberg*: Git repositories mirroring Project Gutenberg texts. [S. l.], 2026. Available at: https://github.com/GITenberg. Accessed: Sep. 2026.

KARLIN, S.; BRENDEL, V. Patchiness and correlations in DNA sequences. *Science*, v. 259, n. 5095, p. 677-680, 1993.

KARMESHU; KRISHNAMACHARI, A. Sequence variability and long-range dependence in DNA: An information theoretic perspective. *Lecture Notes in Computer Science*, v. 3316, p. 1354-1361, 2004.

LASKIN, N. Fractional Poisson process. *Communications in Nonlinear Science and Numerical Simulation*, v. 8, n. 3-4, p. 201-213, 2003.

LINHARES, R. R. Smoothed detrended fluctuation analysis. *Journal of Statistical Computation and Simulation*, v. 86, n. 17, p. 3388-3397, 2016.

LINHARES, R. R. Fractional Poisson process: long-range dependence in DNA sequences. *Model Assisted Statistics and Applications*, v. 18, p. 33-43, 2023.

LINHARES, R. R.; VARGAS, R. N. DFA-WK: an integrated approach to estimate the Hurst exponent using DFA, K-means, and Wavelet Transform. *Cadernos Cajuína*, v. 10, n. 5, e1374, 2025a.

LINHARES, R. R.; VARGAS, R. N. Unsupervised classification of ECG signals using DFA and wavelet entropy. *Observatorio de la Economía Latinoamericana*, v. 23, n. 8, e11007, 2025b.

LOPES, S. R. C.; NUNES, M. A. Long memory analysis in DNA sequences. *Physica A: Statistical Mechanics and its Applications*, v. 361, n. 2, p. 569-588, 2006.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 5., 1967, Berkeley. *Proceedings* [...]. Berkeley: University of California Press, 1967. v. 1, p. 281-297.

MELNIK, S. S.; USATENKO, O. V. Entropy and long-range correlations in DNA sequences. *Computational Biology and Chemistry*, v. 53, p. 26-31, 2014.

OLIVER, J. L.; CARPENA, P.; HACKENBERG, M.; BERNAOLA-GALVÁN, P. IsoFinder: Computational prediction of isochores in genome sequences. *Nucleic Acids Research*, v. 32, p. 287-292, 2004.



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

PENG, C. K.; BULDYREV, S. V.; GOLDBERGER, A. L.; HAVLIN, S.; SCIORTINO, F.; SIMONS, M.; STANLEY, H. E. Long-range correlations in nucleotide sequences. *Nature*, v. 356, p. 168-170, 1992.

PENG, C. K.; BULDYREV, S. V.; HAVLIN, S.; SIMONS, M.; STANLEY, H. E.; GOLDBERGER, A. L. Mosaic organization of DNA nucleotides. *Physical Review E*, v. 49, n. 5, p. 1685-1689, 1994.

PERCIVAL, D. B.; WALDEN, A. T. *Wavelet Methods for Time Series Analysis*. Cambridge: Cambridge University Press, 2000.

PODOBNIK, B.; SHAO, J.; DOKHOLYAN, N. V.; ZLATIC, V.; STANLEY, H. E.; GROSSE, I. Similarity and dissimilarity in correlations of genomic DNA. *Physica A: Statistical Mechanics and its Applications*, v. 373, p. 497-502, 2007.

PROJECT GUTENBERG. *Project Gutenberg*. Salt Lake City: Project Gutenberg Literary Archive Foundation, 2026. Available at: https://www.gutenberg.org. Accessed: Sep. 2026.

STANLEY, H. E.; BULDYREV, S. V.; GOLDBERGER, A. L.; HAVLIN, S.; PENG, C. K.; SIMONS, M. Scaling features of noncoding DNA. *Physica A: Statistical Mechanics and its Applications*, v. 273, n. 1, p. 1-18, 1999.

YULE, G. U. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press, 1944.



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.