applied sciences

MDPI

Article

Written Documents Analyzed as Nature-Inspired Processes: Persistence, Anti-Persistence, and Random Walks—We Remember, as Along Came Writing—T. Holopainen

Omar López-Ortega [1,*,†], Obed Pérez-Cortés [1,†], Heydy Castillejos-Fernández [1,†], Félix-Agustín Castro-Espinoza [1,†] and Miguel González-Mendoza [2,†]

[1] Instituto de Ciencias Básicas e Ingenierías, Área Académica de Computación y Electrónica, Universidad Autónoma del Estado de Hidalgo; Hidalgo 42186, Mexico; obed_perez@uaeh.edu.mx (O.P.-C.); heydy_castillejos@uaeh.edu.mx (H.C.-F.); fcastro@uaeh.edu.mx (F.-A.C.-E.)

[2] Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias, Cd López Mateos 52926, Mexico; mgonza@tec.mx

* Correspondence: lopezo@uaeh.edu.mx; Tel.: +52-771-7172000

† These authors contributed equally to this work.

Received: 18 July 2020; Accepted: 2 September 2020; Published: 12 September 2020

Featured Application: Our proposal to characterize texts by computing the Hurst parameter can be inserted into the e-research paradigm. Internet-based tools and procedures for conducting research and publishing are transforming well-established scientific practices. Submission platforms exist that enable innovative approaches for community interaction among researchers, such as PubMed, Rubriq, ScienceOpen, Libre, GWrit, or PubPeer, to name just a few. Even though the assessment of scientific publications via peer review can be automated through standardized parameters that a scientific writing must fulfil, a quantitative evaluation eliminates any potential bias. Hence, the functionality of such platforms can be improved by incorporating a numeric assessment of the linguistic structure of the manuscript via the Hurst parameter, if its range of values associated to published articles is known.

Abstract: Written communication is pivotal for societies to develop. However, lexicon and depth of information vary greatly among texts according to their purpose. Scientific texts, diffusion of science reports, general and area-specific news are all written differently. Thus, we explore the characterization of different text categories through a nature-inspired feature known as the Hurst parameter. We contend that the Hurst exponent is useful to unveil the rhetorical structure within written documents. We collected and processed texts in five categories: scientific articles, diffusion of science reports, business news, entertainment news, and random texts. Each category contains 350 documents. We found that the median for scientific texts has the highest value of the Hurst parameter (0.575), followed by business news (0.54); the median for randomly-generated texts is 0.48, which lies in the region associated with random walks. The median value for diffusion texts is 0.49, and for entertainment texts is 0.53. However, these two categories present high dispersion. We conclude that the Hurst parameter is a measure that quantifies the structure of communication in the selected categories of texts. Application of our finding in the field of e-research is discussed.

Keywords: hurst parameter; persistence and anti-persistence phenomena; random walk; text analysis

1. Introduction

Written communication is pivotal for societies to develop because they allow people to describe, discuss, modify, categorize, combine and assess the information being presented. In [1] it is explained that interlocutors use language with the goal of communicating information, but they also aim to minimize energetic cost [2]. Molly and Frank [1] also indicate the importance to consider the communicative function of language as a counter-pressure against compression in language, which only accounts for minimizing energy.

Since writing is a form of communication, texts reflect the interaction between *language compression* and *informativity*, according to their context. This implies that the writer aims at minimizing production cost, while the reader seeks to minimize comprehension cost.

Hence, writers aspire to maximize the information within a text, while readers aim for acquiring as much knowledge as possible. For a writer, processing is minimized when a form is clear to express. For the reader, processing is minimized when a form is minimally ambiguous and verbose.

This trade-off between communicating information and minimize energy varies according to the intended purpose. Research theses or journal articles, diffusion of science reports, news about economics, politics and even entertainment, differ in syntax, lexicon and length of phrases, to name only a few linguistic variables. As stated in [1], natural languages are richly structured into lexical and phrasal units of varying length.

To exemplify this premise, we provide two different texts that were produced to report the confirmation of the existance of gravitational waves. The next extract is borrowed from Reuters [3].

> Scientists for the first time have detected gravitational waves, ripples in space and time hypothesized by Albert Einstein a century ago, in a landmark discovery announced on Thursday that opens a new window for studying the cosmos. The waves were unleashed by the collision of the black holes, one of them 29 times the mass of the sun and the other 36 times the solar mass, located 1.3 billion light years from Earth, the researchers said. The scientific milestone was achieved using a pair of giant laser detectors in the United States, located in Louisiana and Washington state, capping a decades-long quest to find these waves. They detected remarkably small vibrations from the gravitational waves as they passed through the Earth. The scientists converted the wave signal into audio waves and listened to the sounds of the black holes merging.

The next piece of text is the abstract of the scientific article confirming the existence of gravitational waves [4].

> On 14 September 2015 at 9.50:45 UTC the two detectors of the Laser Interferometer Gravitational-Wave Observatory simultaneously observed a transient gravitational-wave signal. The signal weeps upwards in frequency from 35 to 250 Hz with a peak gravitational-wave strain of 1.0×10^{-21}. It matches the waveform predicted by the general relativity for the inspiral and merger of a pair of black holes and the ringdown of the resulting single black hole. The signal was observed with a marched-filter signal-to-noise ratio of 24 and false alarm rate estimated to be less than 1 event per 203,300 years, equivalent to a significance greater that 5.1δ...These observations demonstrate the existence of binary stellar-mass black hole systems. This is the first direct detection of gravitational waves and the first observation of a binary hole merger.

The variations of lexical and phrasal units are evident in both texts. For instance, let us compare the expressions that describe the device used to measure the gravitational waves. In the Reuters account it is read *"a pair of giant laser detectors"*, while the scientific text makes clear that *"two detectors of the Laser Interferometer Gravitational-Wave"* were employed.

It has been stated that people unfamiliar with scientific language will find it comparable to a foreign language in its opacity and difficulty [5]. For this reason, scientific articles are transformed wit the presumption to make them easier to comprehend.

The set of changes that texts produced to disseminate new scientific knowledge are analyzed in [6]. They employ two heuristics to examine such transformations in order to understand the processes involved in scientific knowledge dissemination to a variety of audiences. The authors conclude that when scientific knowledge is re-contextualized to be disseminated to different audiences, it is not rephrased or simplified to make it more accessible. Rather, it also undergoes transformational processes that involve issues of social power, authority and access that require new analytical tools to unveil those changes. The resultant variations in lexicon and phrasal structure are the basis that provoke the emergence of text categories, in which the balance between informativity and language compression acquire different interpretations.

1.1. Automated Text Classification

The field of text classification automation aims at being as effective as humans when classifying texts in knowledge fields. An effective automatic clustering over a set of documents contributes to an faster and less expensive classification system.

In [7] it is reported a study aimed at verifying whether an automated clustering process could create the correct clusters for two text corpuses: a scientific corpus having five knowledge fields (Pharmacy, Physical Education, Linguistics, Geography, and History) and a newspaper corpus having five knowledge fields (Human Sciences, Biological Sciences, Social Sciences, Religion and Thought, Exact Sciences). Therefore, the authors had two corpuses already classified by humans and they wanted to measure the effectiveness of the clustering process.

Literary and scientific texts were compared using statistical and network properties [8]. The authors show that Polish texts are described by the Zipf law with the scaling exponent smaller than the one for the English language. They also indicate that scientific texts are typically characterized by the rank-frequency plots with relatively short range of power-law behavior as compared to the literary texts. For the majority of the literary texts, the corresponding networks revealed the scale-free structure, while this is not a pattern in scientific texts.

A study of cognitive structures of scientific texts within the framework of the activity paradigm is presented in [9]. The authors demonstrated how to model the intelligent structure of a scientific text, including an assessment of the textual representation of categories that reflect mental operations.

1.2. e-Research

It is assumed that academic publishers add value to scholarly communications by coordinating reviews and enhancing texts during publication, but some authors challenge this notion [10]. Through a comparative study the authors assess publishers value where pre-print papers from two distinct science, technology, and medicine corpora and their final published counterparts are contrasted. They contend the following: (1) If the publishers' argument is valid, then texts of pre-print papers should vary measurably from their corresponding final published version, and (2) by applying standard similarity measures, such differences must be detected and quantified. The referred analysis revealed that text contents of the scientific papers generally changed very little from their pre-print to final published versions.

Vertiginous advances in technology are transforming well-established scientific practices, particularly, the development of Internet-based tools and procedures for conducting research and publishing. All of these topics belong to the field known as *e-Research* [11–15]. Also, internet-based submission platforms exist that enalbe innovative approaches for community interaction among researchers, such as PubMed, Rubriq, ScienceOpen, Libre, GWrit, or PubPeer, to name just a few [16].

Undoubtedly, in the e-Research context the assessment of scientific publications via peer review, which is one of the most relevant scientific practices [17–19], must incorporate some kind of automatic text

classification. Moreover, e-Research is a model where scientific texts are disseminated efficiently without compromising the reliability of the communication (i.e. the intended informativity of scientific articles). Even though standardized parameters that a scientific writing must fulfil are outlined, our proposal to discriminate texts by computing the Hurst parameter is a contribution to the e-Research paradigm.

Since texts are the result of a natural phenomena intended to optimize the transmission of information and the energetic cost, we want to quantify, through the Hurst parameter, the resultant flow of linguistic structures that are formed in different categories of texts. This characterization, when applied to innovative e-Research platforms, will allow to discriminate whether a manuscript can be placed in the category of scientific texts or not. We are fully aware that human assessing can not be discarded, but our approach is useful to provide more information regarding the linguistic structure of the manuscript.

1.3. Analyzing Written Texts as Natural Phenomena

Our main research question comes from the assumption that written documents reflect trains of thoughts, expressed as linguistic patterns. It has been found that 84 percent of abstracts in scientific articles have, at least, one sentence repeated within the body of the paper [20]. The authors unveil a strong relation between the rhetorical structure of articles and the occurrence of phrases in abstracts. That is to say, a particular train of thought is expressed with quasi-identical phrases all over the analyzed documents.

Hence, we investigate the *persistence of rhetorical structures* within texts. Particularly, we want to discover which categories of the texts in our analysis form linguistic patterns through sentence repetition. On the other hand, we aim at discovering which categories do not suggest this type of written structure and organization.

In other words, we want to calculate the *self-similarity* of texts.

To quantify such phenomena in texts we recur to a nature-inspired metric called the Hurst parameter (H). It owns its name to Harold Edwin Hurst (1880–1978). Originally, Hurst studied the hydrology patterns of the Nile river [21] and proposed the rescaled-range analysis (R/S) of time series. His method is an empirical formulation of long term dependence. It is also an indicator of persistence of the Nile river state. It was argued that periods of droughts and floods are not random, and the changes between these two states are not sudden.

A brief explanation of Hurst method is given next.

Let $x(t)$ be a time series with N points. First, $x(t)$ is partitioned into k non-overlapping sub-series x_k, each of them having size $T = \lfloor(\frac{N}{k})\rfloor$.

Then, the mean values of each segment x_k is computed:

$$\hat{x_k} = \frac{1}{T}\sum_{j=1}^{T} x_k(j), \tag{1}$$

where

$$k = 1, 2, \cdots, \lfloor(\frac{N}{T})\rfloor.$$

A new sub-series $\phi_k(i)$ is constructed:

$$\phi_k(i) = \sum_{j=1}^{i}(x_k(j) - \hat{x_k}), \tag{2}$$

$$i = 1, 2, \cdots, T.$$

The *Ranges* R_k are computed for each sub-series ϕ_k:

$$R_k = max[\phi(i) - min(\phi(i))]; 1 \leq i \leq T. \tag{3}$$

Also, the standard deviations S_k of each non-overlapping sub-series x_k is calculated:

$$S_k = \sqrt{\frac{1}{T}\sum_{j=1}^{T}(x_j - \hat{x_k})}. \tag{4}$$

A statistics called *mean rescaled range* $(\widehat{\frac{R}{S}})_T$ for the corresponding T length is calculated:

$$(\widehat{\frac{R}{S}})_T = \frac{1}{k}\sum_{j=1}^{k}(\frac{R_j}{S_j}). \tag{5}$$

These computations are carried for all sub-series size T.

The *mean rescaled range* follows a power law in relation to the T size:

$$(\widehat{\frac{R}{S}})_T \alpha T^H, \tag{6}$$

where H is the Hurst scaling exponent.

H is related to the fractal dimension of the original time series $x(t)$ by $D = 2 - H$. The Hurst exponent $H \in (0, 1)$, and measures the correlation of neighbour subseries in time series $x(t)$. In [22] it is stated that:

- If $H > 0.5$, then the time series represents a persistent process where the trend of previous steps are likely to be kept i.e., the time series is *self-similar*. A persistent time series possesses long-term memory [23].
- If $H < 0.5$, then the times series represents an anti-persistent process with oscillations. An anti-persistent time series will exhibit higher noise and more volatility.
- If $H = 0.5$, then the times series represent a process with no dependencies, or a random walk. It corresponds to a stochastic process defined by white noise.

According to [24] the Hurst exponent evaluates a system's assimilation capacity, that is to say, the extent to which a system resists change when external stimuli is applied. In [25] Hurst proposed that when a system displays persistence, its response largely depends on its prior behavioural history rather than an instantaneous stimulus. On the other hand, when the system is anti-persistent, its accumulated history is not relevant.

It is important to remark that the procedure to calculate the Hurst parameter carries a normalization of the time series. Thus, time series $x_1(t)$ having N points can be compared with time series $x_2(t)$ having M points, if $N \neq M$.

A numerical approach to calculate the fractal dimension of a time series is by counting the number of circles of a given fixed diameter that are needed to cover the entire time series [23]. That number is related to the diameter of the circle according to Equation (7).

$$Cd^D = 1, \tag{7}$$

where, C = number of circles, d = diameter, and D = fractal dimension.

Equation (7) can be transformed to find the fractal dimension given by Equation (8).

$$D = \frac{\log(N)}{\log(1/d)}. \tag{8}$$

The Hurst Exponent can be obtained from Equation (8) since it is directly related to the fractal dimension. The fractal analysis has been applied in earth science [26], psychology [27], medicine [28] and [29], materials [30], economics [31], environmental [32], to name but a few. Another approach to determine the Hurst exponent is through the rescaled variance statistic [33].

Consequently, we pose the following questions: Do texts mirror persistent, anti-persistent or random phenomena? Do texts in one category are consistently catalogued in one of these three nature-inspired processes?

If common sentences appear as part of the rhetorical structure of texts, then the Hurst parameter must indicate self-similarity. This premise indicates that texts where ideas are being explained recurrently configure a persistent process.

Conversely, if texts do not contain repeated linguistic structures, then they must be deemed as anti-persistent processes and therefore not self-similar.

Finally, the set of random texts must compute values of the Hurst parameter that reflect a random walk.

1.4. Article Organization

After having introduced the research premise and the theoretical background, we proceed to describe the methodology we followed to gather and characterized the corpus (Section 2). Results are presented in Section 3, followed by a discussion in Section 4. Finally, conclusions are outlined in Section 5.

2. Materials and Methods

- The corpus was gathered and prepared by the authors.
- The corpus contains:
  - 350 scientific texts.
  - 350 texts in the category of diffusion of science.
  - 350 randomly generated texts.
  - 350 newspaper articles about business.
  - 350 entertainment texts.
- Title, abstract, keywords, introduction and conclusions were the sections analyzed in scientific articles. We left out tables, equations and discipline-specific symbols.
- Scientific, diffusion, business and entertainment articles are all published texts, gathered randomly from the period between 2014 and 2018.
- The scientific articles were chosen according to a list given by InCites Journal Citation Report under the next parameters: *select categories*: (1) computer science and its different branches; (2) electronics along its different branches.
- The sources of newspaper science articles are the New York Times, BBC news, CNN-tech, ScienceDaily, under the tags *technology, science*.
- Business and entertainment articles were taken from MSN entertainment, Mirror, US weekly, BBC news.
- All texts are written in English.
- The mathematical routines to read and process the texts are coded in Python.
- To quantify the Hurst parameter, texts are mapped into time series $x(t)$ by assigning the corresponding ASCII code to each character appearing in the text. Figure 1 illustrates this mapping.
- The *husrtExp* function *hurst_re* authored by Christopher Scholzel was chosen to compute the Hurst parameter.

3. Experimental Results

As stated in previous lines, the Hurst parameter is applied to determine whether it is suitable to capture *linguistic structures* within a text.

The first result we provide is the existence of linguistic patterns in the scientific article regarding the confirmation of gravitational waves [4]. By inspecting such article it can be noticed that the following linguistic structures appear in the title and abstract:

- Observation of gravitational waves from a binary black hole merger

- Gravitational-wave observatory
- observed
- gravitational-wave
- gravitational-wave
- merger
- black-holes
- black hole
- black hole
- black hole
- gravitational waves
- observations
- binary
- black hole
- gravitational waves
- binary black hole merger

Interestingly, such structures also appear along the body on the text, in different time instants. Figure 1 is the resultant time series of the scientific description regarding the confirmation of gravitational waves [4]. Its resultant Hurst exponent value is 0.677, which places the article in the category of persistent phenomena. It also indicates that the text is self-similar.

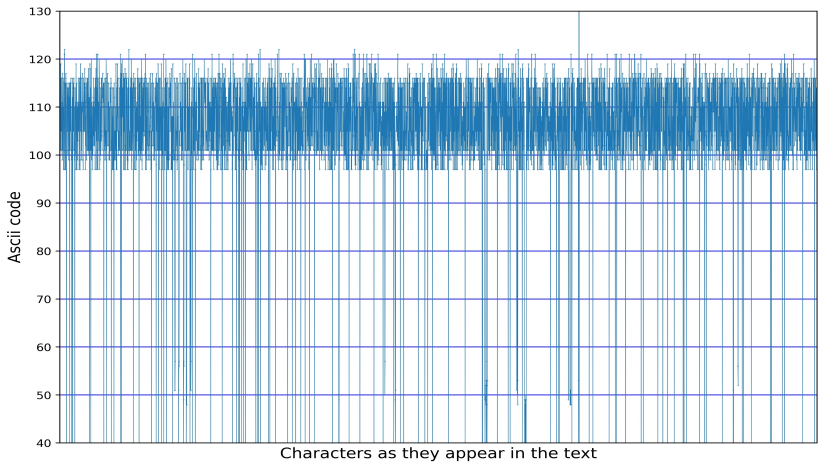

Figure 1. Resultant time series of the text that appear in [4] after it is mapped into ASCII code.

The values of the Hurst parameter for the entire corpus are presented through box-plots and the cumulative probability function (**cpf**). According to [34], box plots provide information about the variability or dispersion of data. They are a standardized way of displaying the distribution of data based on:

1. first quartile (Q1), corresponding to the 25th percentile.
2. third quartile (Q3), corresponding to the 75th percentile.
3. inter-quartile range (IQR): 25th to 75th percentile. That is to say, 50% of the data is contained in this range.
4. median.
5. minimum: Q1 − 1.5IQR.
6. maximum: Q3 + 1.5IQR.
7. Values below *minimum* and above *maximum* are outliers.

Figure 2 shows the values for Hurst parameter.
From the data presented in Figure 2, the median values of the Hurst exponent are:

- 0.54 for business news.

- 0.49 for diffusion of science.
- 0.53 for entertainment mews.
- 0.48 for random texts.
- 0.57 for scientific texts.

It can also be observed that fifty percent of the Hurst parameter values for each category are placed in the following inter-quartile ranges (IQR's):

- $(0.52, 0.57]$ for business news.
- $[0.47, 0.52]$ for diffusion of science.
- $[0.48, 0.57]$ for entertainment news.
- $[0.48, 0.5]$ for random texts.
- $[0.53, 0.6)$ for scientific texts.

Data within IQR represents $(+/-)0.6745\sigma$, where σ is one standard deviation. According to their distribution, data in scientific articles and randomly-generated texts are symmetrical. Randomly-generated texts present the least variability. Also, it is worth noticing that 75% of the Hurst parameter values for scientific texts are equal or greater than 0.53.

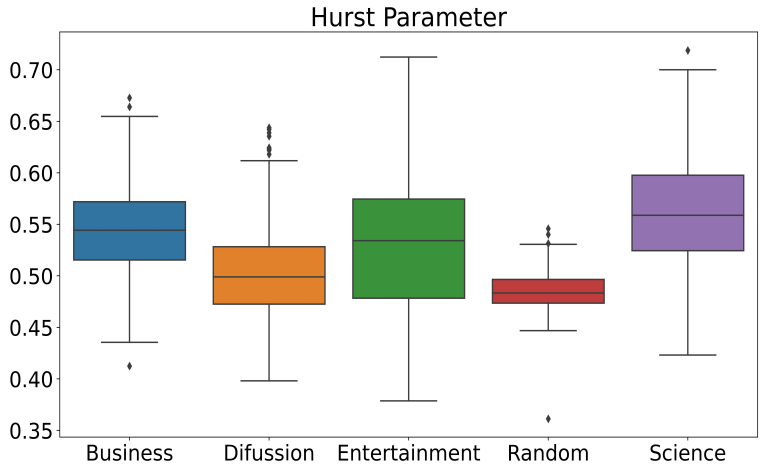

Figure 2. Box plots comparison of the Hurst parameter. From left to right: business, diffusion of science, entertainment, random and scientific texts.

The *cumulative probability functions* that fit the Hurst parameter values are depicted in Figure 3.

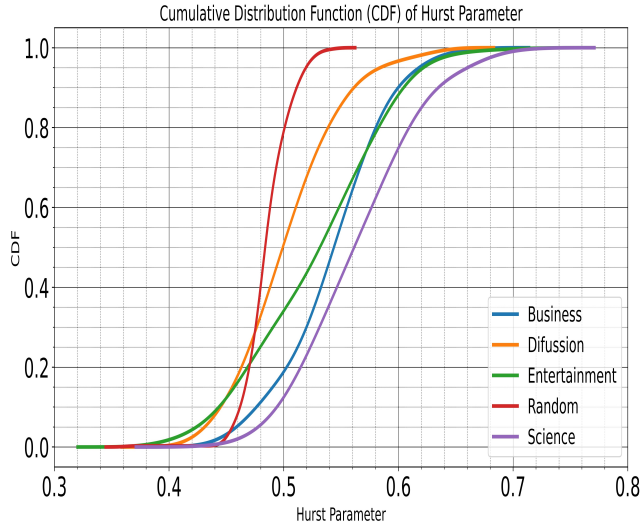

Figure 3. Cumulative Probability Functions of the Hurst parameter values.

4. Discussion

The results of the Hurst parameter, as seen in Figure 2, illustrate that more than 75% of the Hurst parameter values for scientific texts are mainly agglomerated in the interval (0.5, 0.7], suggesting that scientific texts represent persistent phenomena and possess self-similarity. This implies the existence of an intrinsic order that resembles natural phenomena. A *persistent text* has memory i.e., a pattern occurs at time x, then at a time $n(x + \delta)$, as a result of the original event at time x. It means that events in the future are highly correlated to events in the past. This can be explained by the repeatability of sentences as part of the rhetorical structure of scientific texts. Thus, in scientific texts main ideas are presented, and then explained with more and more details. However, every time such main ideas are put forward, they are actually expressed with the same set of concepts, forming a pattern that repeats along the time axis.

Interestingly, business articles also classify as persistent phenomena, but this suggestion is not as strong as it is for scientific articles. The median of the Hurst parameter for this category is 0.54, while the median of the Hurst parameter for scientific texts is 0.57. It can be stated that 75% of the business texts analyzed represent persistent processes, but the upper value in this category is smaller than the upper value in scientific texts.

The median value of the Hurst parameter for randomly-generated texts is 0.48. Interestingly, fifty percent of the random texts are compressed in a very narrow interval. This fact signifies that random texts represent random walk processes. In this analysis a document categorized as random walk is one filled with words having no relation among them. As a matter of fact, randomly-generated texts are a clear example of a Brownian motion.

Figure 3 indicated that, for texts whose Hurst parameter lies in the interval [0.45, 0.54], the probability to be a random text is the highest. The second most probable type of text would be entertainment.

The median for diffusion of science documents lies in 0.49. It means that half the sample is persistent and the other half is anti-persistent. The median values of the Hurst parameter for entertainment texts is 0.53, with a non-symmetrical distribution. Thus, diffusion of science and entertainment news reflect both, anti-persistence and persistence processes. Anti-persistent texts possess no memory, that is to say, the written discourse is elaborated with sentences that do not form linguistic patterns along the time axis. Anti-persistence processes exhibiting higher noise and more volatility than persistence processes.

In summary:

- Scientific texts represent persistent processes, where the trend of previous steps are likely to be kept i.e., the text is *self-similar*, and it possesses long-term memory.
- Business articles are self-similar. However, the values of the Hurst parameter are lower than those of scientific texts.
- Diffusion of science reports and entertainment news represent both, anti-persistent and persistence processes.
- Random texts represent processes with no dependencies, or a random walk, corresponding to stochastic processes defined by white noise.

4.1. What about This Article?

The present article computes as follows:

1. The Hurst parameter is 0.5825. This value lies within the upper 75% of the calculated values. This present paper thus reflects a persistent process where linguistic patterns are formed along the body of the text, making it self-similar.

4.2. Comparison with Related Work

In this section we dissect earlier reports on quantitative analysis of scientific texts, underlying similarities and differences.

The language complexity of scientific texts is calculated in [35]. Language complexity is divided in syntactic complexity and lexical complexity. The formulas used to measure these two types of complexity are based on the statistical properties of the texts, such as the number of words, number of sentences, number of words per sentence, length of nouns, verbs, clauses, to name but a few indicators. In the report, the authors compare scientific texts written in English by what they presumed are native speakers versus those written by what they also pre-supposed are non-native speakers. Interestingly, though, is that differences in lexical complexity are marginal when both sets of scientific articles are compared. We think that this similarity complements our quantitative results regarding the narrow inter-quartile range we present for the Hurst parameter obtained for scientific articles.

Experiments using machine learning to identify the discourse structure in scientific texts are given in [36]. The authors pre-defined 11 Core Scientific Concepts: Hypothesis, Motivation, Goal, Object, Background, Method, Experiment, Model, Observation, Result and Conclusion Then, machine learning classifiers (support vector machines and conditional random fields) were trained on a corpus of 265 full articles in biochemistry and chemistry to recognize such Core Scientific Concepts within texts. One of the steps to train classifiers consisted in dividing the documents into 10 unequal segments. To train the classifiers, they invented 16 different variables to categorize scientific texts. We think their segmentation is arbitrary. By using the Hurst parameter, it is the algorithm that discovers how many sub-series shall be used and how to normalized them. Also, with our approach the Hurst parameter is the only variable needed to discover whether there are linguistic patterns is the texts.

Clustering was applied to analyze scientific texts and newspaper reports [7]. They studied 45 newspaper articles and 36 scientific articles by using three clustering algorithms: Expectation Maximization (EM), SKM and sIB. In both cases the sIB algorithm was the best performer. Their study was aimed, however, to identify what clustering algorithm could be used in a text classification task, rather than drawing conclusions about the characteristic of the texts themselves.

A quantitative study of scientific articles that rests on the Hallidayan notion of *theme* is presented in [37]. The authors claim that understanding of what scientists select as themes offer a vision at how scientific texts are structured in terms of what writers select as the points of departure of their message. They also think that tracing such starting points through the text reveals its structure, since it is from those starting points that the message in the text is developed. The authors employ a quantitative measure known as *thematic density index*. They analyzed 30 articles from the scientific journal called *Cell*. Their data suggest that articles begin with a simple thematic development in the introduction section before settling into an *anchored development* in the rest of the text. Even though the corpus they employ is roughly a tenth of our corpus, and that the discipline of their corpus is biology, they found that writers tend to retain one or a small number of topical themes beyond the introduction section. We concur with such conclusions. However, they felt obliged to analyze each section of the articles and we only used one metric to conclude that scientific texts are self-similar.

In another study ([38]) 500 abstracts of scientific research articles published in 50 journals across five science disciplines (Earth, Formal, Life, Physical and Social Sciences) were analyzed by quantifying how many abstracts adhere to the *Introduction, Purpose, Method, Results and Conclusion* structure. The authors claim that the structure of the abstract should align with the manner the article is structured in order to prompt readers to *predict the content of the upcoming text*. Hence. their premise is similar to ours: there is a self-similarity imbued in scientific articles. Their result indicate that A 4-step structure was adopted in 23.4% of the abstracts and a 3-step structure was employed in 41.2% of the abstracts. Thus, it can be concluded that 64.6% of the abstracts they analyzed contain 3 of the 5 "moves" that were predicted, even though the authors expected that all the abstracts would comply with the 5-step structure they proposed. In light of our analysis, we can argue that self-similarity is that, a process that is similar yet not identical to itself.

Through the study of cause-effect it is found that 65% of the words contained in abstracts and conclusions are part of the set of key-words [39]. The authors indicate that the cause-effect link connects the key concepts to render the core of the text. The referred study uses 39 scientific papers in computer

science. Therefore, if 65% of the keywords are found in abstracts and conclusions, we are witnessing an indirect form to detect self-similarity in scientific texts, which is the conclusion we reached by using the Hurst parameter.

5. Conclusions

The aim of this work is to characterize five categories of texts. We employ a nature-inspired metric called the Hurst parameter. We found that this metric is appropriate to separate the categories of texts used in this study.

The quantitative results reported in this article. obtained by experimental evaluation, lead us to conclude that scientific texts display a process with memory, that is to say, linguistic patterns are found along the text. Through the study of the long range dependence, we found that scientific texts are thus part of phenomenon usually present in persistent natural processes. On the other hand, random texts represent Brownian motion. Diffusion of science and entertainment news do not exhibit strong relations between linguistic patterns.

Future work includes the characterization of scientific papers that were not accepted for publication, to contrast those that were carefully prepared by the authors and accepted by peer reviewers.

To advance the field of e-Research, we want to complement the analysis presented in this article with the cognitive depth of the documents. Cognitive depth aims at measuring to what extent cognition is present in texts. Regarding scientific texts, researchers have found that their readability has been decreasing over time [40]. Thus, a third metric under which characterize texts is the legibility.

Finally, it is necessary to experiment with classifiers and deep learning to enhance the automatic classification of texts. We also want to tackle the analysis of style changes that can be observed in scientific writings during different periods, similar to what has been proposed via stylometry-based approach for detecting writing style changes in literary texts [41].

Author Contributions: Conceptualization, O.L.-O. and O.P.-C.; methodology, H.C.-F.; software, O.P.-C. and F.-A.C.-E.; validation, O.L.-O. and H.C.-F.; formal analysis, O.L.-O. and O.P.-C.; investigation, F.-A.C.-E. and M.G.-M.; writing–original draft preparation, O.L.-O.; writing–review and editing, O.P.-C., H.C.-F., F.-A.C.-E. and M.G.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

H	Hurst Parameter
D	Fractal Dimension
R/S	Rescaled-range analysis

References

1. Lewis, M.L.; Frank, M.C. Linguistic structure emerges through the interaction of memory constratints and communicative pressures. *Behav. Brain Sci.* **2019**, *39*, 38–39.
2. Zipf, G.K. *Human Behaviour and the Principle of Least Effort*; Addison-Wesley: Reading, MA, USA, 1949.
3. Dunham, W.; Malone, S. Einstein's Gravitational Waves Detected in Landmark Discovery. Available online: Reuters.com (accessed on 17 May 2017).
4. Abbott, B.P. Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.* **2016**, *116*, 1–18. [CrossRef] [PubMed]
5. Yager, R.E. The importance of terminology in teaching K-12 science. *J. Res. Sci. Teach.* **1983**, *20*, 577–588. [CrossRef]
6. Gimenez, J.; Baldwin, M.; Breen, P.; Guitierrez, J.; Roque, E. Reproduced, reinterpreted, lost: Trajectories of scientific knowledge across contexts. *Text Talk* **2020**, *40*, 293–324. [CrossRef]

7. Alfonso, A.R.; Duque, C.G. Automated text clustering of newspaper and scientific texts in Brazilian Portuguese: analysis and comparison of methods. *J. Inf. Syst. Technol. Manag.* **2014**, *11*, 415–435.
8. Grabska-Gradzinska, I.; Klig, A.; Kwapien, J.; Drozdz, S. Complex network analysis of literary and scientific texts. *Int. J. Mod. Phys. Comput. Phys.* **2012**, *23*, 1250051–1250060. [CrossRef]
9. Osipov, G.S.; Devyatkin, D.A.; Kusnetzova, Y.M.; Shvets, A.V. The Possibilities for Intelligent Analysis of Scientific Texts by Construction of their Cognitive Models. *Sci. Tech. Inf. Process.* **2019**, *46*, 337–344. [CrossRef]
10. Klein, M.; Broadwell, P.; Farb, S.E.; Grappone, T. Comparing published scientific journal articles to their pre-print versions. *Int. J. Digit. Libr.* **2018**, *4*, 335–350.
11. Balas, E.A. International Collaboration and Competition. In *Innovative Research in Life Sciences: Pathways to Scientific Impact, Public Health Improvement, and Economic Progress*; John Wiley & Sons, Inc.: London, UK, 2019; Chapter 22, pp. 365–380.
12. Sanchez, A.; Carro, B. Internet Services: From Broadband to Ultrabroadband. In *Digital Services in the 21st Century: A Strategic and Business Perspective*; John Wiley & Sons, Inc.: London, UK, 2017; Chapter 2, pp. 9–30.
13. Rees, D.; Laramee, R. A Survey of Information Visualization Books. *Comput. Graph.* **2019**, *38*, 610–646. [CrossRef]
14. Großer, B.; Baumol, U. Virtual teamwork in the context of technological and cultural transformation. *Int. J. Inf. Syst. Proy. Manag.* **2017**, *5*, 21–35. [CrossRef]
15. Sanog, P.; Zhang, C.; Xu, Y.; Xue, L.; Wang, K.; Zhang, C. Asymetrical Interaction in competitive Internet Techonology Diffusion: Implications for the Competition Between Local and Multinacional Online Vendors. In *Global Diffusion and Adoption of Technologies for Knowledge and Information Sharing*; Information Science Reference: Hershey, PA, USA ,2013; Chapter 10, pp. 221–240.
16. PubPeer. About Pubpeer. Available online: *https://pubpeer.com/static/about* (accessed on 3 August 2020).
17. Ward, P.; Graber, K.C.; van der Mars, H. Writing Quality Peer Reviews of Research Manuscripts. *J. Teach. Phys. Educ.* **2015**, *34*, 700–715. [CrossRef]
18. Kulczycki, E.; Rozkosz, E.A. Does an expert-based evaluation allow us to go beyond the Impact Factor? Experiences from building a ranking of national journals in Poland. *Scientometrics* **2017**, *1*, 417–442. [CrossRef] [PubMed]
19. Serenko, A.; Dohan, M. Comparing the expert survey and citation impact journal ranking methods: Example from the field of Artificial Intelligence. *J. Inf.* **2011**, *5*, 629–648. [CrossRef]
20. Atanossova, I.; Bertin, M.; Lariviere, V. On the composition of scientific abstracts. *J. Doc.* **2016**, *72*, 636–647. [CrossRef]
21. Hurst, H.E. Methods of using long-term storage in reservoirs. *ICE Proc.* **1956**, *15*, 519–543. [CrossRef]
22. Molino-Minero, E.; Garcia-Nocetti, F.; Benitez-Perez, H. Application of time-scale local Hurst exponent to time series. *Digit. Signal Process.* **2015**, *37*, 92–99. [CrossRef]
23. Kale, M.; Butar Butar, F. Fractal analysis of time series and distribution properties of Hurst exponent. *J. Math. Sci. Math. Educ.* **2011**, *5*, 8–19.
24. Moreno, C.J.G. Using the Hurst exponent as a monitor and predictor of BWR reactor inestabilities. *Ann. Nucl. Energy* **2010**, *37*, 432–442.
25. Hurst, H.E. A suggested statistical model of some time series which occur in nature. *Nature* **1957**, *180*, 494. [CrossRef]
26. Jiang, C.; Lu, Z.; Zhou, J.; Memon, M.S. Evaluation of fractal dimension of soft terrain surface. *J. Terramechanics* **2017**, *70*, 27–34. [CrossRef]
27. Abboushi, B.; Elzeyadi, I.; Taylor, R.; Sereno, M. Fractals in architecture: The visual interest, preference, and mood response to projected fractal light patterns in interior spaces. *J. Environ. Psychol.* **2019**, *61*, 57–70. [CrossRef]
28. Popovic, N.; Radunovic, M.; Badnjar, J.; Popovic, T. Fractal dimension and lacunarity analysis of retinal microvascular morphology in hypertension and diabetes. *Microvasc. Res.* **2018**, *118*, 36–43. [CrossRef] [PubMed]
29. Ashkenazy, Y. The use of generalized information dimension in measuring fractal dimension of time series. *Phys. Stat. Mech. Its Appl.* **1999**, *271*, 427–447. [CrossRef]
30. Zhokh, A.; Trypolskyi, A.; Strizhak, P. Relationship between the anomalous diffusion and the fractal dimension of the environment. *Chem. Phys.* **2018**, *503*, 71–76. [CrossRef]

31. Batht, S.J.; Dedania, H.V.; Shah, V.R. Fractal dimensional analysis in financial time series. *Int. J. Financ. Manag.* **2015**, *5*, 46–52.
32. Hollingsworth, A. Weather forecasting: Storm hunting with fractals. *Nature* **1986**, *319*, 11–12. [CrossRef]
33. Cajueiro, D.O.; Tabak, B.M. The rescaled variance statistic and the determination of the Hurst exponent. *Math. Comput. Simul.* **2005**, *70*, 172–179. [CrossRef]
34. Galarnyk, M. Understanding Box-Plots. Available online: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 (accessed on 3 August 2020).
35. Lu, C.; Bu, Y.; Wnag, J.; Torvik, V.; Schanaars, M.; Zhang, C. Examining scientific writing styles from the perspective of linguistic complexity. *J. Assoc. Inf. Sci. Technol.* **2018**, *70*, 462–475. [CrossRef]
36. Liakata, M.; Saha, S.; Dobnik, S.; Batchelor, C.; Rebholz-Schuhmann, D. Automatic recognition of conceptualization zones in scientific articles and two life sciences applications. *Bioinformatics* **2012**, *28*, 991–1000. [CrossRef]
37. Leong, A.P.; Toh, A.L.L.; Chin, S.F. Examining Structure in Scientific Research Articles: A Study of Thematic Progression and Thematic Density. *Writ. Commun.* **2018**, *35*, 286–614. [CrossRef]
38. Ngai, S.B.C.; Singh, R.G.; Koon, A.C. A discourse analysis of the macro-structure, metadiscoursal and microdiscoursal features in the abstracts of research articles across multiple science disciplines. *PLoS ONE* **2018**, *13*, e0205417. [CrossRef] [PubMed]
39. Cao, M.; Sun, X.; Zhuge, H. The contribution of cause-effect link to representing the core of scientific paper-The role of Semantic Link Network. *PLoS ONE* **2018**, *13*, e0199303. [CrossRef] [PubMed]
40. Plaven-Sigray, P.; Matheson, G.J.; Schiffler, B.C.; Thompson, W.H. The readability of scientific texts is decreasing over time. *e-Life* **2017**, *6*, 1–14. [CrossRef]
41. Gómez-Adorno, H.M.; Ríos, G.; Posadas-Durán, J.P.; Sidorov, G.; Sierra, G. Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. *Comput. Sist.* **2018**, *22*, 1–18. [CrossRef]