Entropy **2015**, *17*, 5903-5919; doi:10.3390/e17085903

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Maximal Repetitions in Written Texts: Finite Energy Hypothesis *vs.* Strong Hilberg Conjecture

Łukasz Dębowski

Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland; E-Mail: ldebowsk@ipipan.waw.pl; Tel.: +48-22-3800-553; Fax: +48-22-3800-510

Academic Editor: J. A. Tenreiro Machado

Received: 22 May 2015 / Accepted: 19 August 2015 / Published: 21 August 2015

Abstract: The article discusses two mutually-incompatible hypotheses about the stochastic mechanism of the generation of texts in natural language, which could be related to entropy. The first hypothesis, the finite energy hypothesis, assumes that texts are generated by a process with exponentially-decaying probabilities. This hypothesis implies a logarithmic upper bound for maximal repetition, as a function of the text length. The second hypothesis, the strong Hilberg conjecture, assumes that the topological entropy grows as a power law. This hypothesis leads to a hyperlogarithmic lower bound for maximal repetition. By a study of 35 written texts in German, English and French, it is found that the hyperlogarithmic growth of maximal repetition holds for natural language. In this way, the finite energy hypothesis is rejected, and the strong Hilberg conjecture is partly corroborated.

Keywords: finite energy processes; Hilberg's conjecture; entropy rate; maximal repetition; natural language

1. Introduction

Modeling texts in natural language by a stochastic process is a frequent approach in computational linguistics. Classes of stochastic processes, such as hidden Markov processes, constitute state-of-the-art models in natural language processing tasks, such as speech recognition [1] or parts-of-speech tagging [2]. Whereas these kinds of statistical models have the advantages of being easily trained and adapted to the modeled data, they do not capture long-distance dependence in texts, which are connected to coherence of narration, transmission of knowledge or intentionality of language communication.

Therefore, state-of-the-art models in natural language processing overestimate the actual amount of randomness in texts.

In contrast, in this article, we would like to present some novel theoretical and experimental results concerning fundamental properties of texts in natural language as modeled by a stochastic process. These results suggest that the amount of randomness in natural language texts may be much smaller than generally supposed. Briefly speaking, there are two disjoint classes of stochastic processes: finite energy processes (*cf.* [3]), with a strictly positive entropy rate, and Hilberg processes (*cf.* [4]), with a vanishing entropy rate. Experimentally-measuring maximal repetition in some empirical data, we may decide whether the data have been generated by a finite energy processes or by a Hilberg process. Following this idea, in a preliminary study by Dębowski [5], it was discovered that texts in English are not typical of a finite energy process. In the present article, we would like, first, to extend this study by providing empirical data also for French and German and, second, to show that the natural language data may be typical of Hilberg processes.

The further organization of the article is as follows. In Section 2, we present the relevant mathematical models. Section 3 concerns the empirical investigations of natural language. Section 4 contains the conclusion. Finally, in Appendices A and B, we prove two theorems stated in Section 2.

2. Mathematical Models

Let $(\Omega, \mathcal{J}, P)$ be a probability space, *i.e.*, Ω is the event space, $\mathcal{J} \subset 2^{\Omega}$ is the set of events on which the probability measure is defined and P is the probability measure [6]. A discrete (nonstationary) stochastic process $(X_i)_{i\in\mathbb{Z}}$ is a sequence of random variables $X_i : \Omega \to \mathbb{Y}$ on probability space $(\Omega, \mathcal{J}, P)$, where the indices i vary across the set of integers $\mathbb{Z}$ and the alphabet $\mathbb{Y}$ is a finite or countably infinite set of values of X_i. Let $X_k^l = (X_k, X_{k+1}, ..., X_l)$ be blocks of consecutive random variables. A discrete stationary stochastic process $(X_i)_{i\in\mathbb{Z}}$ is such a discrete stochastic process that the distribution of blocks X_{t+1}^{t+n} does not depend on the index t.

There are various classifications of particular stochastic processes. The first class we want to discuss are finite energy processes. The defining feature of a finite energy process is that the conditional probability of any block drawn from the process decreases exponentially with the block length. By $|u|$, we denote the length of a fixed string $u \in \mathbb{Y}^+ = \bigcup_{n=1}^{\infty} \mathbb{Y}^n$, that is the number of symbols in it. Formally, a discrete process $(X_i)_{i\in\mathbb{Z}}$ is called a finite energy process if conditional probabilities of blocks satisfy:

$$P\left(X_{t+|w|+1}^{t+|wu|} = u \middle| X_{t+1}^{t+|w|} = w\right) \leq Kc^{|u|} \tag{1}$$

for all indices $t \in \mathbb{Z}$, all strings u, w and certain constants $0 < c < 1$ and $K > 0$ [3]. (Finite energy processes have nothing to do with the concept of energy in physics. This is just some name coined by a mathematician.)

An important example of finite energy processes is uniformly-dithered processes; *cf.* a formally unproven remark by Shields [3]. Let us recall that i.i.d. processes, called also sequences of independent identically distributed random variables or unigram models in computational linguistics, are processes for which the probability of a block is the product of probabilities of the individual random variables, $P(X_k^l = x_k^l) = \prod_{i=k}^{l} P(X_i = x_i)$. Now, let $(\mathbb{Y}, *)$ be a group. A group $(\mathbb{Y}, *)$ is a pair of a set $\mathbb{Y}$ and a binary operation $* : \mathbb{Y} \times \mathbb{Y} \to \mathbb{Y}$, which is associative, that is $(a * b) * c = a * (b * c)$ has an identity element $e \in \mathbb{Y}$, that is $e * a = a * e = e$, and for each $a \in \mathbb{Y}$, there exists the inverse element a^{-1}, that is $a^{-1} * a = a * a^{-1} = e$. A discrete stochastic process $(X_i)_{i\in\mathbb{Z}}$ is called uniformly dithered if it satisfies:

$$X_i = W_i * Z_i, \tag{2}$$

where $(W_i)_{i\in\mathbb{Z}}$ is an arbitrary discrete process and $(Z_i)_{i\in\mathbb{Z}}$ is an independent i.i.d. process with $\max_{a\in\mathbb{Y}} P(Z_i = a) < 1$. We have this result:

Theorem 1. *Any uniformly-dithered process is a finite energy process.*

Uniformly-dithered processes are processes that are contaminated by random noise in quite a general way. Dębowski [7] supposed that texts in natural language may be contaminated in this way, and so, the appropriate model for the generation of texts in natural language is finite energy. As we will see, this hypothesis is false. Given some data typical of a stochastic process, we can check whether the process is finite energy and, in particular, show for texts in natural language that this is not the case.

The pivotal statistics for our consideration is so-called maximal repetition. For a string of symbols (a text) $w \in \mathbb{Y}^+$, we define the maximal repetition as:

$$L(w) := \max\left\{|s| : w = x_1 s y_1 = x_2 s y_2 \quad \text{and } x_1 \neq x_2\right\}, \tag{3}$$

where s, x_i and y_i range over all admissible substring partitions of text w [8]. In the above definition, strings s, x_i and y_i may be empty, and overlapping repeats are admitted on purpose. For example, $L(w) = 10$ for text w being “Then burst forth the unending argument between the believers and the unbelievers in the societies of the wise and the scientific journals.” because string “believers_” contains 10 characters (nine letters and a space) and appears in the text twice, whereas there is no longer a repeat. The properties of maximal repetition have been studied by both computer scientists [8–11] and probabilists [3,7,12–14]. An efficient algorithm was found for computing the maximal repetition in linear time, so we can compute maximal repetition efficiently for relatively long texts [10].

The object of our interest is how fast the maximal repetition $L(w)$ grows with the length of text $|w|$. In particular, for a finite energy process, the maximal repetition cannot grow faster than logarithmically, *i.e.*, proportionally to the logarithm of the text length.

Theorem 2 (Shields [3])**.** *For a finite energy process $(X_i)_{i\in\mathbb{Z}}$, there exists a constant $C > 0$, such that maximal repetition satisfies:*

$$L(X_1^m) \leq C \log m \tag{4}$$

for sufficiently large text lengths m with probability one.

Law (4) has been studied in mathematics for some time. It was first proven for independent identically distributed (i.i.d.) processes [12,13]. Later, Shields [3] formulated Law (4) under the assumption that the process is stationary finite energy over a finite alphabet. In fact, his proof does not make use of the stationarity or finiteness of the alphabet, so these two conditions may be omitted as above. However, when the process fails to be finite energy, Law (4) need not be satisfied. Shields [14] demonstrated that there exist stationary stochastic processes for which maximal repetition $L(X_1^n)$ grows faster than an arbitrary sublinear function of the text length n.

Now, we may return to the question of whether texts in natural language are typical of a finite energy process. A short empirical study conducted by Dębowski [5] has shown that in printed texts in English, the maximal repetition grows faster than for a finite energy process. The observed growth is hyperlogarithmic, *i.e.*, faster than the logarithm of the text length raised to a certain power $\alpha > 1$. That is, the maximal repetition for texts in English satisfies:

$$L(X_1^m) \geq A\,(\log m)^\alpha. \tag{5}$$

The further question arises then what kind of a stochastic process can be responsible for the generation of texts in natural language. As we are going to show, this question can be related to a hypothesis by Hilberg [4] concerning natural language. This hypothesis involves the entropy of the hypothetical text generation process.

The relevant mathematical background is as follows. The Rényi entropy of order γ of a random variable X is defined as:

$$H_\gamma(X) = \frac{1}{1-\gamma} \log \sum_x P(X=x)^\gamma \tag{6}$$

for $\gamma \in (0,1) \cup (1,\infty)$ [15]. The limiting cases, the topological entropy $H_0(X)$ and the Shannon entropy $H_1(X)$, can be equivalently defined as:

$$H_0(X) = \lim_{\gamma\to 0} H_\gamma(X) = \log \operatorname{card}\{x : P(X=x) > 0\} \tag{7}$$

and:

$$H_1(X) = \lim_{\gamma\to 1} H_\gamma(X) = -\sum_x P(X=x)\log P(X=x). \tag{8}$$

It can be shown that the so-extended Rényi entropy is a decreasing function of γ. In particular, $H_0(X) \geq H_1(X)$. Subsequently, for a discrete stationary stochastic process $(X_i)_{i\in\mathbb{Z}}$, we define block entropy $H_\gamma(n) = H_\gamma(X_{i+1}^{i+n})$ and entropy rate:

$$h_\gamma = \lim_{n\to\infty} \frac{H_\gamma(n)}{n}, \tag{9}$$

if the limit exists. In particular, Limit (9) exists for $\gamma = 0$ [16] and $\gamma = 1$ [17,18].

We will say that a discrete stationary stochastic process $(X_i)_{i\in\mathbb{Z}}$ is a strong Hilberg process of order γ and exponent β if:

$$H_\gamma(n) \in \left[B_1 n^\beta, B_2 n^\beta\right] \tag{10}$$

for some $0 < \beta < 1$ and $B_1, B_2 > 0$. In contrast, we will say that a discrete stationary stochastic process $(X_i)_{i\in\mathbb{Z}}$ is a relaxed Hilberg process of order γ and exponent β if:

$$H_\gamma(n) \in \left[B_1 n^\beta + h_\gamma n, B_2 n^\beta + h_\gamma n\right] \tag{11}$$

for some $h_\gamma \geq 0$, $0 < \beta < 1$ and $B_1, B_2 > 0$. That is, the entropy rate h_γ equals zero for a strong Hilberg process, whereas it can be greater than zero for a relaxed Hilberg process. In particular, if $h_1 = 0$, then the process is asymptotically deterministic, *i.e.*, there exist functions $f_i : \mathbb{Y}^\mathbb{N} \to \mathbb{Y}$, such that $X_i = f_i(X_{i-1}, X_{i-2}, X_{i-3}, \ldots)$ holds with probability one for all $i \in \mathbb{Z}$ (Lemma 4 in [19]).

Let us come back to natural language modeling. Reinterpreting the estimates of entropy for printed English by Shannon [20], Hilberg [4] supposed that texts in natural language can be generated by a strong Hilberg process of order one and exponent $\beta \approx 1/2$. However, any strong Hilberg process of order one has entropy rate $h_1 = 0$, so this condition would imply asymptotic determinism of human utterances. Therefore, in some later works [7,21–24], it was expected that texts in natural language are rather generated by a relaxed Hilberg process of order one with a strictly positive entropy rate $h_1 > 0$.

Now, we want to suggest that a much stronger hypothesis might be true for natural language. In the following, we will observe that for a strong Hilberg process of order zero, the maximal repetition cannot grow slower than hyperlogarithmically.

Theorem 3. *For a strong Hilberg process $(X_i)_{i\in\mathbb{Z}}$ of order zero and exponent β, there exist constants $A, M > 0$, such that the maximal repetition satisfies Condition (5) with probability one for $\alpha = \beta^{-1}$ and $m \geq M$.*

Theorem 3 does not preclude the possibility that a relaxed Hilberg processes (or some other processes) may also satisfy Condition (5). Whereas finding such processes is an interesting open problem, at this moment, let us remark that the empirical observation of the hyperlogarithmic growth of maximal repetition (5) might be explained by a hypothesis that texts in natural language are generated by a strong Hilberg process of order zero. This hypothesis will be called the strong Hilberg conjecture. Since $H_0(X) \geq H_1(X)$, the strong Hilberg conjecture implies that natural language production is also asymptotically deterministic.

Immediately, we would like also to remark that the strong Hilberg conjecture does not imply that texts in natural language are easy to predict or to compress. If we do not know the exact probability distribution of the process, all we can do is universal coding or universal prediction, done for example via the Lempel–Ziv code [25]. It is known that, for a stationary process, the length of the Lempel–Ziv code $|C(X_1^n)|$ divided by the block length n is a consistent estimator of the Shannon entropy rate h_1. However, the convergence rate of $|C(X_1^n)|/n$ to h_1 is very slow, since by Theorem 3 of Dębowski [26], we have:

$$|C(X_1^n)| \geq \frac{n}{L(X_1^n)+1} \log \frac{n}{L(X_1^n)+1}, \tag{12}$$

where $L(X_1^n)$ is the maximal repetition. Therefore, if the maximal repetition grows slower than a power law and the process is a strong Hilberg process, then the length of the Lempel–Ziv code $|C(X_1^n)|$ is orders of magnitude larger than block entropy $H(n)$! Consequently, we cannot estimate block entropy $H(n)$ by the length of the Lempel–Ziv code $|C(X_1^n)|$.

3. Empirical Data

In this section, we check empirically how fast the maximal repetition grows for texts in natural language. Resuming the previous section, if we observe that the maximal repetition in a sample of texts grows faster than the logarithm of the text length, we may infer that the generating process is not finite energy, whereas the strong Hilberg conjecture becomes more likely if we observe hyperlogarithmic growth (5). As we will see, this happens to be the case of natural language.

Let us observe that it is possible to investigate the concerned probabilistic hypotheses on two levels. In the first case, we assume that random variables X_i are consecutive characters of the text, whereas, in the second case, we assume that random variables X_i are consecutive words of the text. Respectively, we have to compute the repetitions in texts as repeated strings of characters or as repeated strings of words. We will test both levels. Moreover, we will show that the maximal repetitions in texts in natural language scale quite differently than for i.i.d. processes, that is the unigram models of the text.

For the experiment, we have downloaded 14 texts in English, 10 texts in German and 11 texts in French from the Project Gutenberg (http://www.gutenberg.org/). To make the finite energy hypothesis more plausible, we have removed legal notices, tables of contents and strings of repeated spaces from the considered text, since they contain long repeats, which dominate repeats in the proper text; *cf*. Dębowski [5]. However, we have not excluded prefaces and afterwords, which sometimes also contain long quotations, but otherwise seem to be non-singular parts of the text. Moreover, using the statistics for the English texts, we have generated four-character-based unigram texts and four-word-based unigram texts (the texts have been generated via sampling with replacement rather than as random permutations). All so prepared texts are presented in Tables 1–4.

Table 1. The selection of texts in German.

Text	**Maximal Repeated Substring**
Mark Twain, Die Abenteuer Tom Sawyers	“selig sind, die da arm sind im Geiste, denn” (9 words, close)
Lewis Carroll, Alice’s Abenteuer im Wunderland	“Edwin und Morcar, Grafen von Mercia und” (7 words)
Friedrich Nietzsche, Also sprach Zarathustra	“stiftete mehr Leid, als die Thorheiten der Mitleidigen? Wehe allen Liebenden, die nicht noch eine Höhe haben, welche über ihrem Mitleiden ist! Also sprach der Teufel einst zu mir: «auch Gott hat seine Hölle: das ist seine Liebe zu den Menschen.» Und jüngst hörte ich ihn diess Wort sagen: «Gott ist todt; an seinem Mitleiden mit den Menschen ist Gott gestorben.»” (61 words)
Thomas Mann, Buddenbrooks	“Mit raschen Schritten, die Arme ausgebreitet und den Kopf zur Seite geneigt, in der Haltung eines Mannes, welcher sagen will: Hier bin ich! Töte mich, wenn du willst!” (28 words)
Goethe, Faust	“Kühn ist das Mühen, Herrlich der Lohn! Und die” (9 words)
Dante Alighieri, Die Göttliche Komödie	“Da kehrt er sich zu mir” (6 words)
Immanuel Kant, Kritik der reinen Vernunft	“als solche, selbst ein von ihnen unterschiedenes Beharrliches, worauf in Beziehung der Wechsel derselben, mithin mein Dasein in der Zeit, darin sie wechseln, bestimmt werden” (25 words, in quotes)
Thomas Mann, Der Tod in Venedig	“diesem Augenblick dachte er an” (6 words)
Sigmund Freud, Die Traumdeutung	“ich muß auch auf einen anderen im sprachlichen Ausdruck enthaltenen Zusammenhang hinweisen. In unseren Landen existiert eine unfeine Bezeichnung für den masturbatorischen Akt: sich einen ausreißen oder sich einen” (29 words, in footnote)
Franz Kafka, Die Verwandlung	“daß sein Körper zu breit war, um” (7 words)

In the proper experiment, for each text, we have considered initial blocks of the text of exponentially-growing length, and the maximal repetition was computed for each block. We report the results in terms of figures and examples of maximal repeats. First, also in Tables 1–4, we have presented the maximal repeated substring in each whole text. For each maximal repeated substring, we give its length and sometimes short comments about its location: "close" means that the repeat was observed within a few paragraphs; "in quotes" means that the repeat was located within quotes one or more times. Further analysis of these repeats is left to more linguistically-oriented researchers.

Table 2. The selection of texts in French.

Text	**Maximal Repeated Substring**
Voltaire, Candide ou l'optimisme	"voyez, tome XXI, le chapitre XXXI du Précis du Siècle de Louis XV. B." (14 words, in footnote)
Alexandre Dumas, Le comte de Monte-Cristo, Tome I	"le procureur du roi est prévenu, par un ami du trône et de la religion, que le nommé Edmond Dantès, second du navire le Pharaon, arrivé ce matin de Smyrne, après avoir touché à Naples et à Porto-Ferrajo, a été chargé, par Murat, d'une lettre pour l'usurpateur, et, par" (49 words, in quotes)
Victor Hugo, L'homme qui rit	"trois hommes d'équipage, le patron ayant été enlevé par un coup de mer, il ne reste que" (17 words, in quotes)
Gustave Flaubert, Madame Bovary	"et madame Tuvache, la femme du maire," (7 words)
Victor Hugo, Les miserables, Tome I	"livres Pour la société de charité maternelle" (7 words, close)
Descartes, Oeuvres. Tome Premier	"que toutes les choses que nous concevons fort clairement et fort distinctement sont toutes" (14 words, many times in paraphrases)
François Villon, Oeuvres completes	"mes lubres sentemens, Esguisez comme une pelote, M'ouvrist plus que tous les Commens D'Averroys sur" (15 words, in quotes in footnote in preface)
Stendhal, Le Rouge et le Noir	"Which now shows all the beauty of the sun And by and by a cloud takes all away!" (18 words, in quotes)
Alexandre Dumas, Les trois mousquetaires	"murmura Mme Bonacieux. «Silence!» dit d'Artagnan en lui" (9 words, close)
Jules Verne, Vingt mille lieues sous les mers	"à la partie supérieure de la coque du «Nautilus», et" (10 words)
Jules Verne, Voyage au centre de la terre	"D0 E6 B3 C5 BC D0 B4 B3 A2 BC BC C5 EF «Arne" (14 words, in quotes)

Table 3. The selection of texts in English.

Text	**Maximal Repeated Substring**
Jacques Casanova de Seingalt, Complete Memoirs	"but not deaf. I am come from the Rhone to bathe you. The hour of Oromasis has begun.»" (18 words, in quotes)
Thomas Babington Macaulay, Critical and Historical Essays, Volume II	"therefore there must be attached to this agency, as that without which none of our responsibilities can be met, a religion. And this religion must be that of the conscience of the" (32 words, in quotes, close)
Charles Darwin, The Descent of Man and Selection in Relation to Sex	"Variability of body and mind in man-Inheritance-Causes of variability-Laws of variation the same in man as in the lower animals–Direct action of the conditions of life-Effects of the increased use and disuse of parts-Arrested development-Reversion-Correlated variation-Rate of increase-Checks to increase-Natural selection-Man the most dominant animal in the world-Importance of his corporeal structure-The causes which have led to his becoming erect-Consequent changes of structure-Decrease in size of the canine teeth-Increased size and altered shape of the skull-Nakedness-Absence of a tail-Defenceless condition of man." (86 words, in the table of contents, undeleted by omission)
Jules Verne, Eight Hundred Leagues on the Amazon	"After catching a glimpse of the hamlet of Tahua-Miri, mounted on its piles as on stilts, as a protection against inundation from the floods, which often sweep up" (28 words, close, probably by mistake)
William Shakespeare, First Folio/35 Plays	"And so am I for Phebe Phe. And I for Ganimed Orl. And I for Rosalind Ros. And I for no woman Sil. It is to be all made of" (30 words, close)
Jules Verne, Five Weeks in a Balloon	"forty-four thousand eight hundred and forty-seven cubic feet of" (9 words, close)
Jonathan Swift, Gulliver's Travels	"of meat and drink sufficient for the support of 1724" (10 words, in quotes, close)
Jonathan Swift, The Journal to Stella	"chocolate is a present, madam, for Stella. Don't read this, you little rogue, with your little eyes; but give it to Dingley, pray now; and I will write as plain as the" (32 words, in quotes in preface)
George Smith, The Life of William Carey, Shoemaker & Missionary	"I would not go, that I was determined to stay and see the murder, and that I should certainly bear witness of it at the tribunal of" (27 words, in quotes)
Albert Bigelow Paine, Mark Twain. A Biography	"going to kill the church thus with bad smells I will have nothing to do with this work of" (19 words, in quotes, close)
Etienne Leon Lamothe-Langon, Memoirs of the Comtesse du Barry	"M. de Maupeou, the duc de la Vrilliere, and the" (10 words)
Jules Verne, The Mysterious Island	"we will try to get out of the scrape" (9 words, in the same sentence)
Willa Cather, One of Ours	"big type on the front page of the" (8 words, close)
Jules Verne, Twenty Thousand Leagues under the Sea	"variety of sites and landscapes along these sandbanks and" (9 words, close)

Table 4. The selection of unigram texts.

Text	Maximal Repeated Substring
Character-based unigram Text 1	" u ti t r "
Character-based unigram Text 2	"e t tloeu "
Character-based unigram Text 3	"o d t eie"
Character-based unigram Text 4	"s ei er e"
Word-based unigram Text 1	"of that A for"
Word-based unigram Text 2	"was the of in"
Word-based unigram Text 3	"in of of the"
Word-based unigram Text 4	"of of of of a"

From the tables, we can learn that there appear repeats in texts in natural language whose lengths exceed a dozen of words. More quantitative data are presented in Figures 1–4, where we study the dependence between the block length and the maximal repetition within the block for both natural language and unigram models. Figures 1 and 2 concern variables X_i being characters, whereas Figures 3 and 4 concern variables X_i being words. It can be seen that there is much variation in the data, but let us try to fit some functional dependence to the data points.

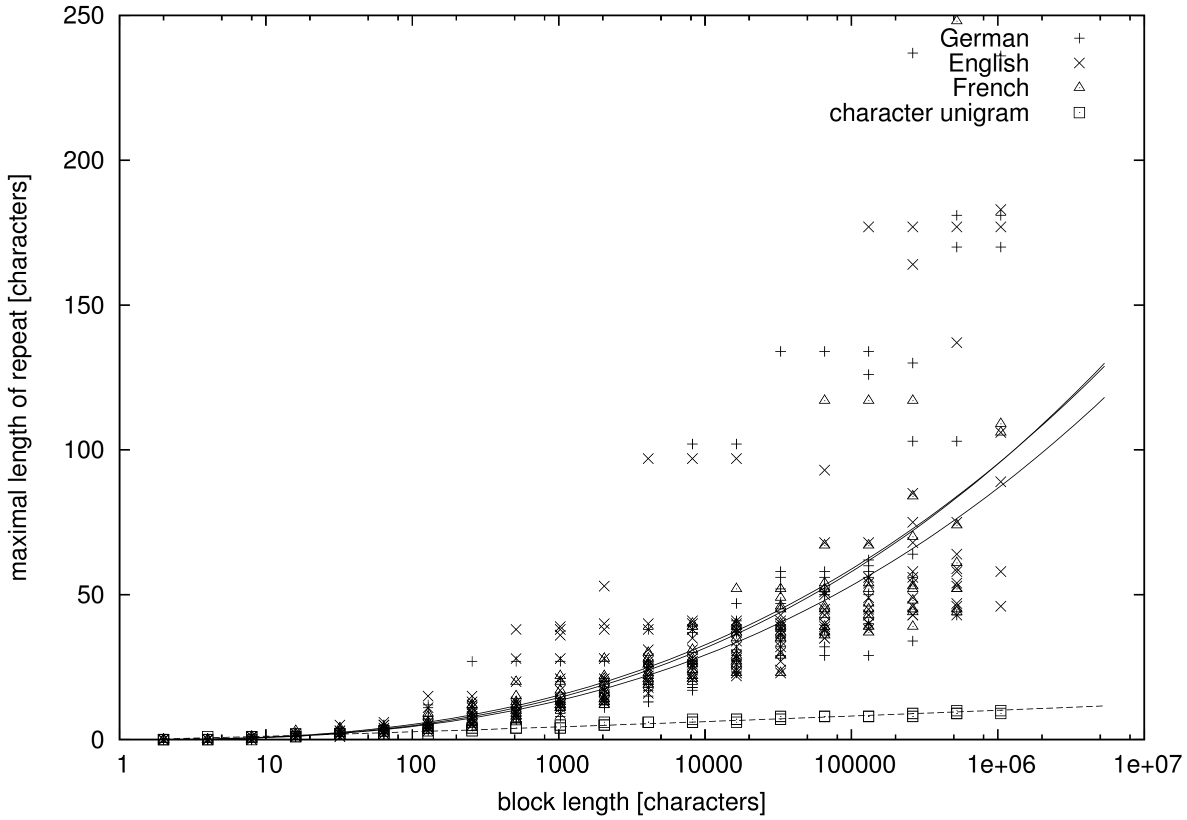

Figure 1. Character-based maximal repetition on the logarithmic-linear scale. The lines are the regression lines.

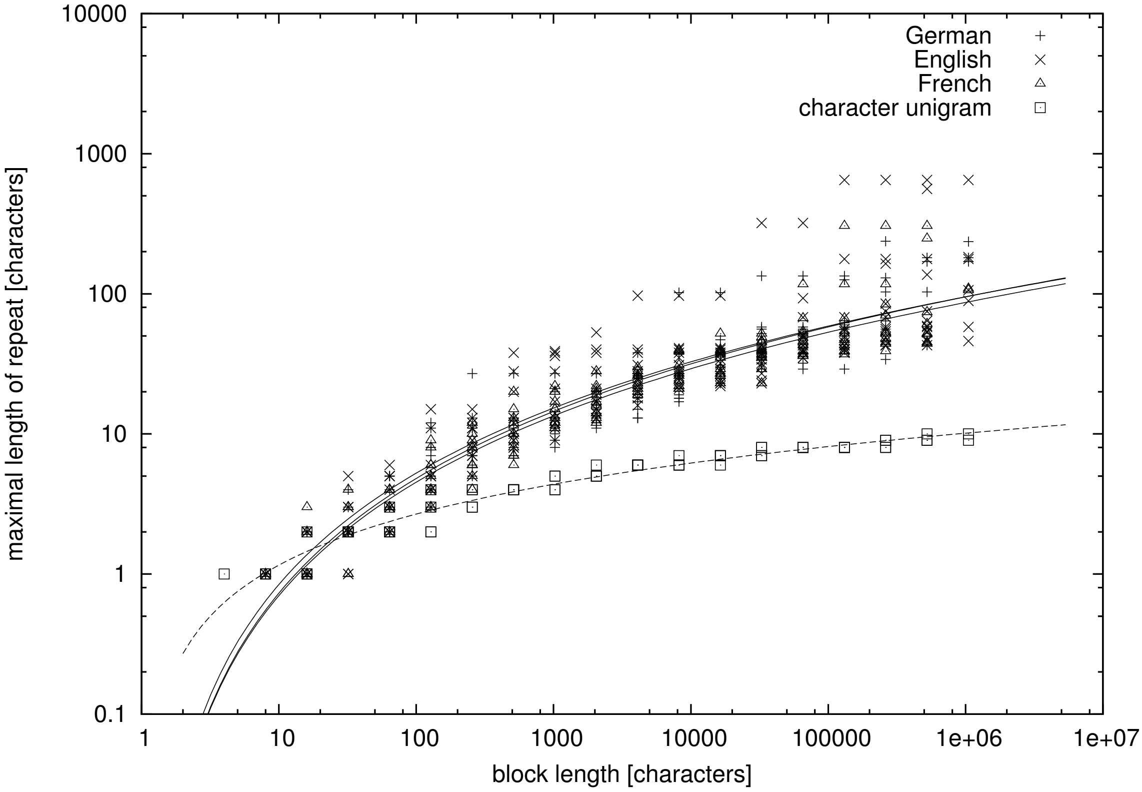

Figure 2. Character-based maximal repetition on the doubly-logarithmic scale. The lines are the regression lines.

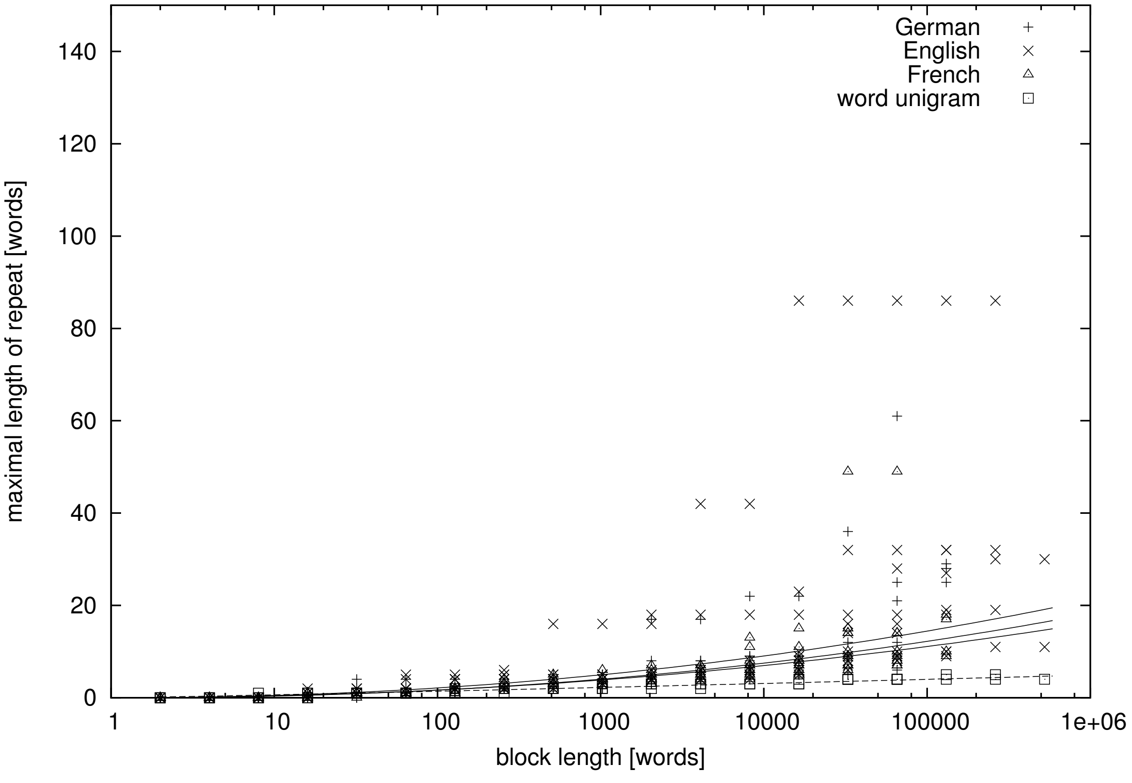

Figure 3. Word-based maximal repetition on the logarithmic linear scale. The lines are the regression lines.

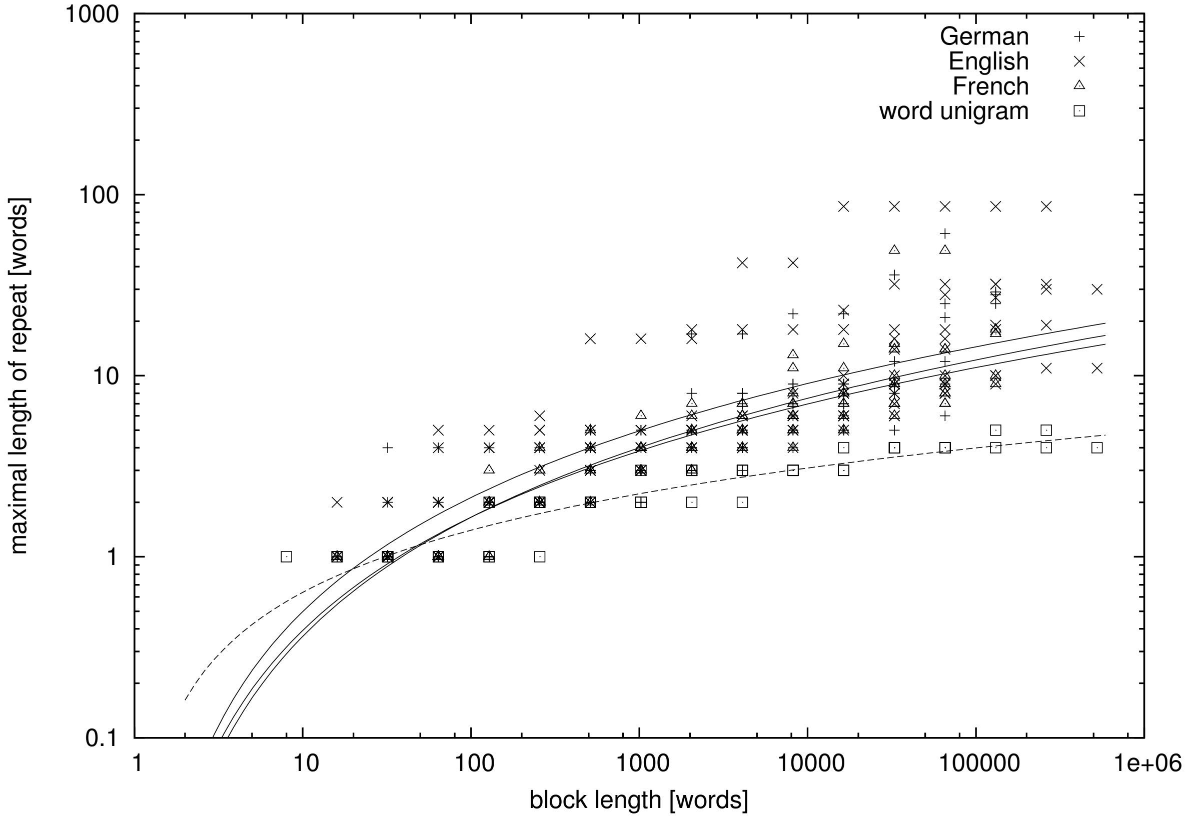

Figure 4. Word-based maximal repetition on the doubly-logarithmic scale. The lines are the regression lines.

In Figures 1 and 3, the abscissa is on the logarithmic scale, and the cloud of points forms a convex (∪) shape. Thus, on average, the maximal repetition grows faster than logarithmically. On the other hand, the clouds of points are concave (∩) in Figures 2 and 4, where both the abscissa and the ordinate are on the logarithmic scale. Hence, on average, the maximal repetition grows slower than a power law. Therefore, let us try to fit the hyperlogarithmic function:

$$L(X_1^m) \approx A(\log m)^\alpha, \tag{13}$$

where A and α are free parameters. If we obtain α significantly larger than one, the investigated process cannot be finite energy. Since the distributions of $L(X_1^m)$ for fixed n are skewed towards large values, to make the finite energy hypothesis more plausible, we fit parameters A and α using the least squares method in the doubly-logarithmic scale. In this way, we obtain smaller values of parameters A and α than when applying the least squares method in the linear scale.

It can be seen in the figures that the maximal repetition behaves quite comparably for texts in English, German and French, whereas it is strikingly different for the unigram texts, for which the mean and variance of $L(X_1^m)$ are much lower. The fitted parameters of Model (13) and their standard errors are presented in Table 5. They were obtained using the standard fitting procedure of the Gnuplot program.

The unigram texts are in theory a finite energy process, so parameter α should not exceed one in that case. In Table 5, we see however that $\alpha \approx 1.21$ on the level of characters and $\alpha \approx 1.14$ on the level of words, where the second digit is significant. This observation needs a definite explanation, which we cannot provide at the moment. The reviewers of this paper suggested that our simple fitting procedure may overestimate parameter α and underestimate its error, analogically as in using the least squares method to fit power law distributions [27]. However, here, we cannot use the improved method of fitting power law distributions by Clauset *et al.* [27], since we are not fitting a probability distribution. As an alternative explanation, let us note that our observation of $\alpha > 1$ for the unigram texts might be due to generating these texts using an imperfect pseudorandom number generator, which does not satisfy the finite energy property. In fact, to generate the unigram texts, we have used the pseudorandom number generator built in the Perl programming language, which need not be the best choice.

Table 5. The fitted parameters of Model (13). The values after the sign ± are the standard errors.

Level of Description	**Class of Texts**	A	α
characters	German	0.076 ± 0.011	2.71 ± 0.07
characters	English	0.093 ± 0.012	2.64 ± 0.06
characters	French	0.074 ± 0.009	2.69 ± 0.06
characters	unigram	0.42 ± 0.03	1.21 ± 0.03
words	German	0.059 ± 0.014	2.18 ± 0.13
words	English	0.086 ± 0.019	2.09 ± 0.11
words	French	0.069 ± 0.010	2.08 ± 0.08
words	unigram	0.24 ± 0.03	1.14 ± 0.06

Having made this remark, let us note that it is obvious by looking at the plots that texts in natural language are of a different class than the unigram texts. The value of α is twice larger, $\alpha \approx 2.6$ on the level of characters and $\alpha \approx 2.1$ on the level of words, for texts in natural language. Hence, we may conclude that texts in natural language, as far as we can trust the estimated parameters and extrapolate the data, are not generated by a finite energy process on the level of characters or on the level of words. Repeats as long as 20 words are expected if they appear in samples of one million word tokens, which is close to the observed maximal length of a text. That the repeats can be so long may seem surprising, but the empirical data, such as the examples in Tables 1–4, confirm this claim. This behavior is strikingly different than for the unigram texts.

4. Conclusions

In this article, we have studied two hypotheses about a possible probabilistic mechanism of generating texts in natural language. According to the first hypothesis, texts are in a certain sense uniformly contaminated by random noise. As we have shown, this conjecture implies that the stochastic process of generating texts is finite energy. According to the second hypothesis, the strong Hilberg conjecture, the number of different admissible texts of a given length is severely restricted, namely the logarithm of it grows as a power law. In other words, the strong Hilberg conjecture assumes some mechanism of intense and very selective replication of texts. According to the presented mathematical results, the finite energy hypothesis and the strong Hilberg conjecture are mutually incompatible and can be tested by investigating the growth rate of the maximal repetition: the maximal length of a repeated substring in a text.

The empirical study performed in this article confirms that in written texts, the maximal repetition grows hyperlogarithmically, *i.e.*, the maximal repetition grows as a power of the logarithm of the text length. As far as we can trust our fitting procedure and extrapolate the empirical data, this falsifies the finite energy hypothesis and partly corroborates the strong Hilberg conjecture. Rejection of the finite energy hypothesis implies that texts written in natural language cannot be formed by transmission through a specific noisy channel without any correction after the transmission. Thus, some mechanism of reducing randomness must be at work during the composition of texts written in natural language.

Although the strong Hilberg conjecture implies the asymptotic determinism of human utterances, it should not be discarded on a purely rational basis. For the further evaluation of the strong Hilberg conjecture, it might be useful to exhibit some abstract examples of strong Hilberg processes. So far, we have constructed some processes, so-called Santa Fe processes and some hidden Markov processes, which are relaxed Hilberg processes with entropy rate $h_1 > 0$ [7,28,29]. There are also known some stationary processes, such as the Thue–Morse process, for which $H_1(n) \approx B \log n$ and $h_1 = 0$ [30,31]. We are now on the way to construct a strong Hilberg process, but this is a topic for a separate paper.

The present study, although extending the previous study by Dębowski [5], covers still a narrow selection of Indo-European languages, for which we obtain very similar results. In future research, it may be interesting to extend the scope to some non-Indo-European languages, especially Chinese or Japanese, which use a very different script, featuring a very large alphabet. However, we suppose that the results may be comparable. The reason is that the strong Hilberg conjecture appears to be not a hypothesis about a particular ethnic language, but rather a hypothesis about fundamental limitations of human memory and attention. On this ground, let us claim that we may have discovered a new language universal. This claim should be thoroughly verified.

In future research, it may be also illuminating to contrast our observation for natural language with genetics, as suggested by Shields [3]. DNA sequences are certain sequences that have been generated by the process of biological evolution, which, like human communication, may be conceptualized as a complicated stochastic process combining randomness and computational mechanisms that decrease the observed disorder. According to Chandrasekaran and Betrán [32], imperfect copying of very long sequences is the main mechanism of creating new genes. Hence, we suppose that repeated strings in DNA are much longer than in texts in natural language. Although we have not found any article concerning the maximal repetition as a function of DNA sequence length, there are publications that concern tools for computing repetitions in DNA [33], topological entropy of DNA [34] and a power law for frequencies of repeated strings in DNA, which resembles Zipf's law for natural language [35].

Finally, let us recall the idea of memetic cultural evolution, an analogue of biological evolution [36]. The principle of that theory is that the human mind tends to select and imitate previously-encountered strings of symbols or ideas, according to their utility, regardless of whether being produced on their own or provided by the environment. It may be obvious that this process occurs heavily in music at the level of melodies or during language acquisition at the level of single words or grammatical constructions; *cf.* Bloom *et al.* [37]. However, the empirical data we provide show that the self-imitation process operates on many length scales of text composition and leads to surprisingly long literal repeats in spite of a conscious intention of not repeating long phrases too often. Moreover, according to the strong Hilberg conjecture, the number of different admissible texts of a given length is severely restricted, that is there must exist some mechanism of intense and very selective replication of texts. Thus, we suppose that the strong Hilberg conjecture may describe a certain idealized equilibrium state in the cultural/biological evolution. This idea may deserve further research. Let us note that a completely different approach to mathematics of cultural/biological evolution has been proposed by Chaitin [38]. In the research of the strong Hilberg conjecture, it might be fruitful to use his ideas.

Acknowledgments

The author wishes to thank Jan Mielniczuk, Jacek Koronacki, Ramon Ferrer-i-Cancho, Kumiko Tanaka-Ishii and the anonymous reviewers for comments and remarks.

Conflicts of Interest

The author declares no conflict of interest.

Appendix

A. Proof of Theorem 1

Consider process $X_i = W_i * Z_i$. First, let $K = 1$ and $c = \max_{a\in\mathbb{Y}} P(Z_i = a)$. Then:

$$P(Z_{t+|w|+1}^{t+|wu|} = u) \le Kc^{|u|}. \tag{A1}$$

For strings $w = w_1...w_n$ and $z = z_1...z_n$, where $w_i, z_i \in \mathbb{Y}$, let us write $w * z = (w_1 * z_1)...(w_n * z_n)$ and $z^{-1} = z_1^{-1}...z_n^{-1}$. Using the fact that $(W_i)_{i\in\mathbb{Z}}$ and $(Z_i)_{i\in\mathbb{Z}}$ are independent, we can further write:

$$\begin{aligned}
&P(X_{t+|w|+1}^{t+|wu|} = u|X_{t+1}^{t+|w|} = w) \\
&= \sum_{z\in\mathbb{Y}^{|u|}} P(W_{t+|w|+1}^{t+|wu|} = u * z^{-1}, Z_{t+|w|+1}^{t+|wu|} = z|X_{t+1}^{t+|w|} = w) &\text{(A2)}\\
&= \sum_{z\in\mathbb{Y}^{|u|}} P(W_{t+|w|+1}^{t+|wu|} = u * z^{-1}|X_{t+1}^{t+|w|} = w)P(Z_{t+|w|+1}^{t+|wu|} = z) &\text{(A3)}\\
&\le \sum_{z\in\mathbb{Y}^{|u|}} P(W_{t+|w|+1}^{t+|wu|} = u * z^{-1}|X_{t+1}^{t+|w|} = w)Kc^{|u|} \le Kc^{|u|} &\text{(A4)}
\end{aligned}$$

since:

$$\begin{aligned}
&\sum_{z\in\mathbb{Y}^{|u|}} P(W_{t+|w|+1}^{t+|wu|} = u * z^{-1}|X_{t+1}^{t+|w|} = w) \\
&= \sum_{v\in\mathbb{Y}^{|u|}} P(W_{t+|w|+1}^{t+|wu|} = v|X_{t+1}^{t+|w|} = w) = 1. &\text{(A5)}
\end{aligned}$$

Hence, process $(X_i)_{i\in\mathbb{Z}}$ is a finite energy process.

B. Proof of Theorem 3

We begin with the following auxiliary result:

Lemma 1. *For a discrete stationary process* $(X_i)_{i\in\mathbb{Z}}$*, where* $H_0(n) < \log(m - n + 1)$*, inequality* $L(X_1^m) \geq n$ *holds with probability one.*

Proof. String X_1^m contains $m - n + 1$ substrings of length n (on overlapping positions). Among them there can be at most $\exp(H_0(n))$ different substrings if X_1^m has a positive probability. Since $\exp(H_0(n)) < m - n + 1$, there must be some repeat of length n. Hence, $L(X_1^m) \geq n$. □

Now, let us suppose that $H_0(n) \leq B_2 n^\beta$. Then, using Lemma 1, we obtain $L(X_1^m) \geq n$ for $B_2 n^\beta < \log(m - n + 1)$. Now, for a constant $C > B_2$, let N be the smallest number, such that $\exp(B_2 N^\beta) + N - 1 < \exp(C N^\beta)$. Then, for $n \geq N$, condition $B_2 n^\beta < \log(m - n + 1)$ follows from condition $C n^\beta = \log m$ or, equivalently $n = A(\log m)^\alpha$, where $\alpha = 1/\beta$ and $A = C^{-\alpha}$. Hence, for $m \geq M = \exp(C N^\beta)$, we obtain:

$$L(X_1^m) \geq A\,(\log m)^\alpha \tag{B1}$$

with probability one.

References

1. Jelinek, F. *Statistical Methods for Speech Recognition*; The MIT Press: Cambridge, MA, USA, 1997.
2. Jurafsky, D.; Martin, J.H. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*; Prentice Hall: Englewood Cliffs, NJ, USA, 2000.
3. Shields, P.C. String matching bounds via coding. *Ann. Probab.* **1997**, *25*, 329–336.
4. Hilberg, W. Der bekannte Grenzwert der redundanzfreien Information in Texten—Eine Fehlinterpretation der Shannonschen Experimente? *Frequenz* **1990**, *44*, 243–248.
5. Dębowski, Ł. Maximal Lengths of Repeat in English Prose. In *Synergetic Linguistics. Text and Language as Dynamic System*; Naumann, S., Grzybek, P., Vulanović, R., Altmann, G., Eds.; Praesens Verlag: Vienna, Austria, 2012; pp. 23–30.
6. Billingsley, P. *Probability and Measure*; Wiley: New York, NY, USA, 1979.
7. Dębowski, Ł. On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts. *IEEE Trans. Inf. Theory* **2011**, *57*, 4589–4599.
8. De Luca, A. On the combinatorics of finite words. *Theor. Comput. Sci.* **1999**, *218*, 13–39.
9. Kolpakov, R.; Kucherov, G. Finding Maximal Repetitions in a Word in Linear Time. In Proceedings of the 40th Annual Symposium on Foundations of Computer Science, New York, NY, USA, 17–19 October 1999; pp. 596–604.
10. Kolpakov, R.; Kucherov, G. On Maximal Repetitions in Words. *J. Discret. Algorithms* **1999**, *1*, 159–186.
11. Crochemore, M.; Ilie, L. Maximal repetitions in strings. *J. Comput. Syst. Sci.* **2008**, *74*, 796–807.

12. Erdős, P.; Rényi, A. On a new law of large numbers. *J. D'Analyse Math.* **1970**, *22*, 103–111.
13. Arratia, R.; Waterman, M.S. The Erdös-Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.* **1989**, *17*, 1152–1169.
14. Shields, P.C. String matching: The ergodic case. *Ann. Probab.* **1992**, *20*, 1199–1203.
15. Rényi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
16. Allouche, J.P.; Shallit, J. *Automatic Sequences. Theory, Applications, Generalizations*; Cambridge University Press: Cambridge, UK, 2003.
17. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: New York, NY, USA, 1991.
18. Yeung, R.W. *First Course in Information Theory*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2002.
19. Dębowski, Ł. A general definition of conditional information and its application to ergodic decomposition. *Stat. Probab. Lett.* **2009**, *79*, 1260–1268.
20. Shannon, C. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64.
21. Ebeling, W.; Nicolis, G. Entropy of Symbolic Sequences: the Role of Correlations. *Europhys. Lett.* **1991**, *14*, 191–196.
22. Ebeling, W.; Pöschel, T. Entropy and long-range correlations in literary English. *Europhys. Lett.* **1994**, *26*, 241–246.
23. Bialek, W.; Nemenman, I.; Tishby, N. Complexity through nonextensivity. *Phys. A* **2001**, *302*, 89–99.
24. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos* **2003**, *15*, 25–54.
25. Ziv, J.; Lempel, A. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory* **1977**, *23*, 337–343.
26. Dębowski, Ł. Hilberg's Conjecture—A Challenge for Machine Learning. *Schedae Inf.* **2014**, *23*, 33–44.
27. Clauset, A.; Shalizi, C.R.; Newman, M.E.J. Power-law distributions in empirical data. *SIAM Rev.* **2009**, *51*, 661–703.
28. Dębowski, Ł. Mixing, Ergodic, and Nonergodic Processes with Rapidly Growing Information between Blocks. *IEEE Trans. Inf. Theory* **2012**, *58*, 3392–3401.
29. Dębowski, Ł. On Hidden Markov Processes with Infinite Excess Entropy. *J. Theor. Probab.* **2014**, *27*, 539–551.
30. Berthé, V. Conditional entropy of some automatic sequences. *J. Phys. A* **1994**, *27*, 7993–8006.
31. Gramss, T. Entropy of the symbolic sequence for critical circle maps. *Phys. Rev. E* **1994**, *50*, 2616–2620.
32. Chandrasekaran, C.; Betrán, E. Origins of new genes and pseudogenes. *Nat. Educ.* **2008**, *1*, 181.
33. Kurtz, S.; Schleiermacher, C. REPuter: Fast computation of maximal repeats in complete genomes. *Bioinformatics* **1999**, *15*, 426–427.
34. Koslicki, D. Topological entropy of DNA sequences. *Bioinformatics* **2011**, *27*, 1061–1067.

35. Wang, J.D.; Liu, H.C.; Tsai, J.J.P.; Ng, K.L. Scaling Behavior of Maximal Repeat Distributions in Genomic Sequences. *Int. J. Cogn. Inform. Nat. Intell.* **2008**, *2*, 31–42.
36. Dawkins, R. *The Selfish Gene*; Oxford University Press: Oxford, UK, 1976.
37. Bloom, L.; Hood, L.; Lightbown, P. Imitation in language development: If, when, and why. *Cogn. Psychol.* **1974**, *6*, 380–420.
38. Chaitin, G. *Proving Darwin: Making Biology Mathematical*; Random House: New York, NY, USA, 2013.