Schedae Informaticae Vol. 23 (2014): 33–44
doi: 10.4467/20838476SI.14.003.3020

Hilberg's Conjecture – a Challenge for Machine Learning

Łukasz Dębowski
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
e-mail: *ldebowsk@ipipan.waw.pl*

Abstract. We review three mathematical developments linked with Hilberg's conjecture – a hypothesis about the power-law growth of entropy of texts in natural language, which sets up a challenge for machine learning. First, considerations concerning maximal repetition indicate that universal codes such as the Lempel-Ziv code may fail to efficiently compress sources that satisfy Hilberg's conjecture. Second, Hilberg's conjecture implies the empirically observed power-law growth of vocabulary in texts. Third, Hilberg's conjecture can be explained by a hypothesis that texts describe consistently an infinite random object.

Keywords: statistical language modeling, Hilberg's conjecture, maximal repetition, grammar-based codes, Santa Fe processes.

1. Introduction

May generation of texts in natural language be described by a probabilistic model? Researchers from several domains worked on this problem: linguists, mathematicians, engineers, and physicists. The theoretical difficulties inspired a few important concepts in applied mathematics, such as Markov chains [32], entropy [35, 36], fractals [30, 31], and algorithmic complexity [26]. A common intuition is that texts are a result of a process that is neither purely deterministic nor purely random [41, p. 187], which makes them so difficult to model using established mathematical frameworks. A few statistical laws of language have been, however, observed, such as the Zipf-Mandelbrot law [31, 41], Herdan's law [18, 19, 20, 28], Menzerath's law [33], and a few other [25]. This suggests that some well-defined stochastic process may describe the generation of texts in natural language, although we have tremendous problems in identifying or approximating this process efficiently.

The question of probabilistic language modeling is no longer of only theoretical importance. Statistical language modeling is highly relevant for practical applications, for instance, for speech recognition. The problem of speech recognition is to provide a system for automated transcription of human speech (acoustic wave) into written text. This task has been partially solved by providing two imperfect models: $P(A|W)$ – probability of speech A corresponding to a given text W, and $P(W)$ – probability of text W. Consequently, to obtain the most probable text corresponding to a given speech signal, we use Bayes theorem

$$\max_W P(W|A) = \max_W P(A|W)P(W). \tag{1}$$

The quality of a speech recognition system depends on both models $P(A|W)$ and $P(W)$. State-of-the-art modeling of $P(W)$ consists in approximating the hypothetical stochastic process by Markov chains but we know this is not optimal [22].

Why is modeling $P(W)$ so difficult? Can some insight into the practical task of statistical language modeling be provided by more theoretically-oriented research? We may suppose that randomness of texts must be strongly constrained by the existence or the search for meaning but the meaning itself may manifest as a form of apparent randomness. Here we may invoke the idea of halting probability, the number Ω, which is apparently random but stores an infinite amount of mathematical knowledge [29, Section 3.6.2]. Something highly puzzling must be going about information in natural language but we should not lose hope that we may get some insight into the phenomena. We have both to look at empirical data and to build idealized mathematical models.

In this article we would like to show that there is a hypothetical property of natural language, called Hilberg's conjecture, which can model some of the observed regularities and difficulties. For a discrete random variable X on a probability space $(\Omega, \mathcal{J}, P)$ denote the random variable $P(X)$, which takes value $P(X) = P(X = x)$ for $X = x$. Then the Shannon entropy of X is defined as

$$H(X) = \mathbf{E}\left[\, -\log P(X)\,\right], \tag{2}$$

where $\mathbf{E}X$ is the expectation of X and log denotes the binary logarithm. For a stochastic process $(X_i)_{i\in\mathbb{Z}}$ we denote blocks of variables X_i as $X_k^l = (X_i)_{i=k}^l$ and, if the process is (strictly) stationary, we define block entropy $H(n) = H(X_1^n)$. The core of Hilberg's conjecture [21] is that texts in natural language can be modeled by a stationary process, where the variables take values in characters and the block entropy grows according to a power law,

$$H(n) \approx Bn^\beta, \tag{3}$$

where $\beta \approx 0.5$ and B is some positive constant. According to this hypothesis, texts in natural language situate somewhere between determinism ($\beta = 0$) and randomness ($\beta = 1$), just as stated in [41, p. 187]. In a certain way, they are both random and deterministic.

There are three important theoretical developments about Hilberg's conjecture, which we will describe in the following sections in more detail. First, considerations concerning maximal repetition in texts indicate that universal codes such as the Lempel-Ziv code may fail to efficiently compress sources that satisfy Hilberg's conjecture. Hence Hilberg's conjecture constitutes some challenge for machine learning. Second, Hilberg's conjecture implies Herdan's law, i.e., the empirically observed power-law growth of vocabulary in texts [5]. Third, Hilberg's conjecture can be explained by a hypothesis that texts describe consistently an infinite random object (a random world) [5].

There is some empirical evidence in favor of Hilberg's conjecture, although it is indirect or concerns only small block lengths. First, relationship (3) with $n \leq 100$ was observed in [21] for the estimates of conditional entropy provided by [36] using human subjects. Second, some other estimates of block entropy also corroborate (3) [15, 16, 14, 10]. Third, growth rate of maximal repetition agrees with a lower bound provided by a stronger version of Hilberg's conjecture [6, 11]. Fourth, scaling of compression rate for universal codes is also compatible with relationship (3) [8, 9, 12].

Probably, there may be fundamental problems with verifying relationship (3) for large n and the validity of Hilberg's conjecture must be partly decided using rational (i.e., nonempirical) arguments. Hence we think that the greatest present problem concerning Hilberg's conjecture is a construction of examples of stochastic processes that satisfy this relationship. For the relaxed Hilberg conjecture,

$$H(n) \approx Bn^{\beta} + hn, \tag{4}$$

where the entropy rate is $h > 0$, a few such constructions were given in [7, 13]. Some of these constructions, called Santa Fe processes [7], are motivated linguistically. But we are not aware of any example of a process that would satisfy relationship (3), i.e., (4) with entropy rate $h = 0$. In particular, we ignore whether construction of such processes can be motivated linguistically.

The further organization of the paper is as follows. In Section 2., we discuss links of Hilberg's conjecture with maximal repetition. In Section 3., we link Hilberg's conjecture with power-law growth of vocabulary. In Section 4., we connect Hilberg's conjecture with describing an infinite random object. We conclude in Section 5.

2. Maximal repetition

Scaling of maximal repetition provides some basic insight into Hilberg's conjecture. Let $|w|$ denote the length of text w. Maximal repetition $L(w)$ is the maximal length of a repeated substring in text w [2, 37, 39], formally defined as

$$L(w) := \max \{|s| : w = x_1 s y_1 = x_2 s y_2 \text{ and } x_1 \neq x_2\}, \tag{5}$$

where s, x_i, and y_i range over all substrings of text w. For a given text w, $L(w)$ can be computed in time $O(|w|)$ [27].

For typical stationary processes, the maximal repetition grows logarithmically with the length of the text.

Theorem 1 ([39]) *Suppose that $(X_i)_{i\in\mathbb{Z}}$ is a finite energy process, i.e.,*

$$-\log P(X_1^n|X_{-k+1}^0) = \Omega(n) \tag{6}$$

holds with probability 1 *for all $k \geq 0$. Then*

$$L(X_1^n) = O\left(\log n\right) \tag{7}$$

holds with probability 1*.*

Passing an arbitrary stationary process through a symmetric memoryless noisy channel yields a finite energy process [11, 39]. Hence there are many such processes. However, in [6, 11], maximal repetition was computed for a collection of texts in natural language and the empirical data satisfy a strictly larger lower bound

$$L(X_1^n) = \Omega\left((\log n)^\alpha\right) \tag{8}$$

with $\alpha \approx 2.7$.

Relationship (8) can be linked to some modification of Hilberg's conjecture. We define topological entropy of random variable X as

$$H_{\text{top}}(X) = \log \operatorname{card}\left\{x : P(X = x) > 0\right\}. \tag{9}$$

It can be easily shown that $H_{\text{top}}(X) \geq H(X)$. For a stationary process $(X_i)_{i\in\mathbb{Z}}$, we put $H_{\text{top}}(n) = H_{\text{top}}(X_1^n)$. Obviously $H_{\text{top}}(n) \geq H(n)$.

Theorem 2 ([11]) *Suppose that a stationary process $(X_i)_{i\in\mathbb{Z}}$ satisfies relationship*

$$H_{top}(n) = O(n^\beta), \tag{10}$$

where $\beta \in (0, 1]$. Then (8) holds for $\alpha = 1/\beta$ with probability 1*.*

We suppose that links between Hilberg's conjecture and maximal repetition are stronger. The natural language data support also an upper for maximal repetition

$$\mathbf{E}L(X_1^n) = O\left((\log n)^\alpha\right), \tag{11}$$

cf. [6, 11]. Hence if generation of texts in natural language may be modeled by a process that satisfies the original Hilberg conjecture (3), this definition is not void.

Definition 1 *A stationary process $(X_i)_{i\in\mathbb{Z}}$ is called a regular Hilberg process if*

$$H(n) = \Theta(n^\beta), \tag{12}$$
$$\mathbf{E}L(X_1^n) = \Theta\left((\log n)^\alpha\right) \tag{13}$$

for certain $\beta \in (0, 1)$ and $\alpha \geq 1/\beta$.

Now we will show that regular Hilberg processes, if they exist, cannot be compressed efficiently using the Lempel-Ziv universal code, cf., [38]. This result stems from the following observation (a new result of this paper):

Theorem 3 *The length $|C(w)|$ of the Lempel-Ziv code* [42] *for text w satisfies*

$$|C(w)| \geq \frac{|w|}{L(w)+1} \log \frac{|w|}{L(w)+1}. \tag{14}$$

Proof. The length of the Lempel-Ziv code for text w is greater than $A \log A$, where A is the number of phrases in the Lempel-Ziv parsing of the text. Each phrase in the Lempel-Ziv parsing is a concatenation of a repeated substring and a single bit, thus its length cannot be greater than $L(w)+1$. Moreover, the sum of lengths of the phrases equals $|w|$. Hence $A(L(w)+1) \geq |w|$, from which the claim follows. □

Let us recall that a code C is called universal if $\mathbf{E}\,|C(X_1^n)| - H(n) = o(n)$ for any stationary process. In contrast, here we will say that a code C compresses a process $(X_i)_{i\in\mathbb{Z}}$ efficiently if

$$\mathbf{E}\,|C(X_1^n)| = O(H(n)). \tag{15}$$

This condition is nontrivial for processes with zero entropy rate, such as regular Hilberg processes. For these processes, by (14) and $\mathbf{E}\left[\,Y^{-1}\,\right] \geq \left[\,\mathbf{E}Y\,\right]^{-1}$, the length of the Lempel-Ziv code is lower bounded by

$$\mathbf{E}\,|C(X_1^n)| = \Omega\left(\frac{n}{(\log n)^\alpha} \log \frac{n}{(\log n)^\alpha}\right) \tag{16}$$

and condition (15) cannot be satisfied. In fact, growth rate close to (16) has been observed for texts in natural language. For the Lempel-Ziv code, the empirical data support relationship

$$|C(X_1^n)| = \Theta(n^\gamma), \tag{17}$$

where $\gamma \approx 0.9$ [9]. In the considered range of text lengths, the observed relationship (17) does not preclude the theoretical bound (16).

In plain words, the above result shows that regular Hilberg processes pose some challenge for machine learning. Namely, these sources could be very much compressed if we knew their exact distribution and we used the appropriate Shannon-Fano code. But if we do not know their distribution and we have to learn it using a universal compression scheme, such as the Lempel-Ziv code, the resulted compression length is orders of magnitude larger than the entropy. The problem lies in that universal codes such as the Lempel-Ziv code cannot capture the order in the process that is hidden beyond the maximal repetition. In the next section we will show that this deficiency applies also to some improvement of the Lempel-Ziv code, called grammar-based codes, and leads to emergence of some hierarchical structure.

3. Grammar-based codes

Grammar-based codes [5, 23] compress texts by transforming them first into special grammars, called admissible grammars, and then encoding the grammars back into texts according to a fixed simple method. An admissible grammar is a context-free grammar that generates a singleton language, i.e., a language that contains only one text. We will say that an admissible grammar generates this text. In an admissible grammar, there is exactly one rule per nonterminal symbol and the nonterminals can be ordered so that the symbols are substituted by strings of strictly succeeding symbols [1, 23]. Hence, such a grammar is given by its set of production rules

$$G = \left\{ \begin{array}{l} A_1 \to \alpha_1, \\ A_2 \to \alpha_2, \\ ..., \\ A_k \to \alpha_k \end{array} \right\}, \tag{18}$$

where A_1 is the start nonterminal symbol, other A_i are secondary nonterminal symbols, and the right-hand sides of rules satisfy $\alpha_i \in (\{A_{i+1}, A_{i+2}, ..., A_k\} \cup \mathbb{X})^*$, where $\mathbb{X}$ is the set of terminal symbols.

A function that for a given text returns some admissible grammar that generates this text is called a grammar transform [23]. Many such transforms have been proposed, see [23]. For example, the longest matching grammar transform for the tongue twister

I scream, you scream, we all scream for icecream!

returns the admissible grammar

$$\left\{ \begin{array}{l} A_1 \to \mathbf{I}A_2\mathbf{you}A_2\mathbf{we\_all}A_3\mathbf{\_for\_ice}A_4\mathbf{!} \\ A_2 \to A_3\mathbf{,\_} \\ A_3 \to \mathbf{\_s}A_4 \\ A_4 \to \mathbf{cream} \end{array} \right\}. \tag{19}$$

In the compressions of longer texts in natural language, nonterminal symbols often correspond to words or set phrases, like A_4 in (19) [3, 34, 40], especially if it is additionally required that the secondary nonterminals were defined as strings of only terminal symbols [24].

Certain grammar transforms can be turned into universal codes if we apply a certain encoding of an arbitrary grammar into a string [5, 23]. In this way we obtain for instance admissibly minimal grammar-based codes [5]. The exact definition of admissibly minimal grammar-based codes is too technical to reproduce it in this paper. However, we may say that admissibly minimal grammar-based codes resemble the grammar transforms considered in [3, 24, 34, 40] for detecting word boundaries in texts. Hence, as we will show, there is some link between Hilberg's conjecture and so called Herdan's law.

Herdan's law is an empirical observation which states that the number of distinct words in a text of length n is proportional to n^γ, where $\gamma \approx 0.5$ [18, 19, 20, 28]. We may suppose that the number of distinct words in a text X_1^n can be approximated by the number of distinct nonterminals $V(X_1^n)$ in an admissibly minimal grammar-based code $C(X_1^n)$ for X_1^n. Frankly speaking, this claim may be practically impossible to verify since the computational complexity of the exact admissibly minimal grammar-based codes is probably NP-hard, cf., [1]. Nevertheless, there are intimate connections between statement $V(X_1^n) = \Omega(n^\gamma)$ and Hilberg's conjecture, which stem from the remarkable inequality

$$|C(u)| + |C(v)| - |C(uv)| \le BV(uv)(1 + L(uv)) \tag{20}$$

where $|C(u)|$ is the length of the admissibly minimal grammar-based code for text u, B is some positive constant, and $L(uv)$ is the maximal repetition in text uv as in the previous section [5].

Expression on the left hand side of (20) is an estimate of mutual information

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{21}$$

If Hilberg's conjecture is satisfied in the original form (3) or the relaxed form (4), then mutual information $I(X_1^n; X_{n+1}^{2n})$ is proportional to n^β. Consequently we may apply the following result:

Theorem 4 ([5]) *Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary process over a finite alphabet with a strictly positive entropy rate. Assume that $I(X_1^n; X_{n+1}^{2n}) = \Omega(n^\beta)$ holds for some $\beta \in (0,1)$. Then the number of distinct nonterminals $V(X_1^n)$ in an admissibly minimal grammar-based code $C(X_1^n)$ satisfies*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\,[\,V(X_1^n)(1 + L(X_1^n))\,]}{n^\beta} > 0. \tag{22}$$

Theorem 4 states that if the relaxed Hilberg conjecture is satisfied then we have Herdan's law provided the maximal repetition does not grow too fast. For regular Hilberg processes this bound could be much strengthened (a new result of this paper):

Theorem 5 *Let $V(X_1^n)$ be the number of distinct nonterminals in an admissibly minimal grammar-based code $C(X_1^n)$. If for a stationary process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet the code satisfies $\mathbf{E}\,|C(X_1^n)| = o(n)$ then*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\left[\,V(X_1^{2n})(1 + L(X_1^{2n}))\,\right]}{n\left[\frac{1}{\mathbf{E}L(X_1^n)} - \frac{1}{\mathbf{E}L(X_1^{2n})}\right]} > 0. \tag{23}$$

Remark: In [5], it was proved that admissibly minimal grammar-based codes are almost universal, i.e., they satisfy $\mathbf{E}\,|C(X_1^n)| - H(n) = o(n)$ if $H(n) = \Theta(n)$. We suppose that this result remains valid for sources with $H(n) = o(n)$. This would guarantee $\mathbf{E}\,|C(X_1^n)| = o(n)$ for regular Hilberg processes, and then we would have

$$\limsup_{n\to\infty} \frac{\mathbf{E}\,[\,V(X_1^n)(1 + L(X_1^n))\,]}{n/(\log n)^{\alpha+1}} > 0. \tag{24}$$

Proof. The length of an admissibly minimal grammar-based code for text X_1^n is greater than A – the number of terminal or nonterminal symbols in the right-hand side of the rule for the start symbol. On the other hand, the length of the expansion of a secondary nonterminal symbol cannot be greater than $L(X_1^n)$ since every secondary nonterminal must appear at least twice in the right-hand sides of some rules. Thus, since the expansion of the start symbol is X_1^n, we obtain $AL(X_1^n) \ge n$. Hence

$$\mathbf{E}\,|C(X_1^n)| \ge \mathbf{E}\left[\frac{n}{L(X_1^n)}\right] \ge \frac{n}{\mathbf{E}L(X_1^n)}, \tag{25}$$

where we used inequality $\mathbf{E}Y\mathbf{E}Y^{-1} \ge 1$. Now, the excess-bounding lemma [5] states that for any function $G(n) \ge 0$ which satisfies $G(n) = o(n)$, we have $2G(n) - G(2n)$ for infinitely many n. In view of bound (25) we can apply this lemma to $G(n) = \mathbf{E}\,|C(X_1^n)| - n/\mathbf{E}L(X_1^n)$. Hence, for infinitely many n,

$$\mathbf{E}\left[\,2\,|C(X_1^n)| - \left|C(X_1^{2n})\right|\,\right] \ge 2n\left[\frac{1}{\mathbf{E}L(X_1^n)} - \frac{1}{\mathbf{E}L(X_1^{2n})}\right]. \tag{26}$$

To obtain the claim we chain inequality (26) with inequality (20). □

The above results show that processes that satisfy the relaxed Hilberg conjecture exhibit some rich structure which can be detected using grammar-based coding. As we will see in the next section this structure may be somewhat apparent because there exist very simple processes which satisfy the relaxed Hilberg conjecture.

4. Examples of processes

As we have stated in the introduction, we do not know any example of a process that would satisfy the original Hilberg conjecture (3) but there are a few examples of processes that satisfy the relaxed Hilberg conjecture (4). Here we will present two such constructions called Santa Fe processes. As we have mentioned, they are motivated linguistically to a certain extent.

The first example is as follows.

Definition 2 (Santa Fe process) *Let the process have the form*

$$X_i := (K_i, Z_{K_i}), \tag{27}$$

where $(Z_k)_{k=1}^{\infty}$ and $(K_i)_{i=-\infty}^{\infty}$ are probabilistically independent, $(Z_k)_{k=1}^{\infty}$ is a sequence of independent random binary variables that satisfy

$$P(Z_k = 0) = P(Z_k = 1) = 1/2, \tag{28}$$

whereas $(K_i)_{i=-\infty}^{\infty}$ is a sequence of independent variables that satisfy Zipf's law

$$P(K_i = k) \propto k^{-1/\beta}, \tag{29}$$

where $\beta \in (0,1)$.

The Santa Fe process can be given such an idealized linguistic interpretation. Imagine that $(X_i)_{i=-\infty}^{\infty}$ is a sequence of consecutive statements extracted from an infinitely long text that describes an infinite random object $(Z_k)_{k=1}^{\infty}$ consistently. In this description, each statement $X_i = (k, z)$ reveals both the address k of a random bit of $(Z_k)_{k=1}^{\infty}$ and its value $Z_k = z$. Logical consistency of the description is reflected in this property: If two statements $X_i = (k, z)$ and $X_j = (k', z')$ describe bits of the same address ($k = k'$) then they always assert the same bit value ($z = z'$).

In fact, the Santa Fe process satisfies the relaxed Hilberg conjecture:

Theorem 6 ([7]) *Let $\beta \in (0,1)$. The Santa Fe process obeys $I(X_1^n; X_{n+1}^{2n}) = \Theta(n^\beta)$.*

Let us note that the random object $(Z_k)_{k=1}^\infty$ described by the Santa Fe process (27) does not evolve in time which causes the Santa Fe process to be nonergodic [5]. This condition may be slightly relaxed so as to produce an ergodic process.

Definition 3 (generalized Santa Fe process) *Let the process have the form*

$$X_i = (K_i, Z_{i,K_i}), \tag{30}$$

where processes $(K_i)_{i=-\infty}^\infty$ and $(Z_{ik})_{i=-\infty}^\infty$, where $k \in \mathbb{N}$, are independent and distributed as follows. First, variables K_i are distributed according to formula (29), as before. Second, each process $(Z_{ik})_{i=-\infty}^\infty$ is a Markov chain with marginal distribution

$$P(Z_{ik} = 0) = P(Z_{ik} = 1) = 1/2 \tag{31}$$

and cross-over probabilities

$$P(Z_{ik} = 0 | Z_{i-1,k} = 1) = P(Z_{ik} = 1 | Z_{i-1,k} = 0) = p_k. \tag{32}$$

Numbers p_k are additional parameters of the process, in principle not related to $P(K_i = k)$.

The object $(Z_{ik})_{k=1}^\infty$ described by the generalized Santa Fe process is a function of time i and the probability that the k-th bit flips at a given instant equals p_k. For vanishing cross-over probabilities, $p_k = 0$, the generalized Santa Fe process collapses to the original Santa Fe process. For $p_k \neq 0$, the generalized Santa Fe process is ergodic [7]. However, also for $p_k \neq 0$, we obtain Hilberg's conjecture if the cross-over probabilities decay fast enough.

Theorem 7 ([7]) *Suppose that $\lim_{k\to\infty} p_k / P(K_i = k) = 0$. Then the generalized Santa Fe process obeys $I(X_1^n; X_{n+1}^{2n}) = \Theta(n^\beta)$.*

The careful reader will notice that Santa Fe processes satisfy the relaxed Hilberg conjecture but they do not satisfy the full premise of Theorem 4 because they are not processes over a finite alphabet (the condition of finite energy is, however, satisfied). To overcome this deficiency, in [4], stationary variable-length coding of Santa Fe processes was considered and it was shown that the coded processes satisfy the full premise of Theorem 4 indeed.

5. Conclusion

In this paper we have sketched some issues connected with Hilberg's conjecture – a hypothesis about a power-law growth of entropy of texts in natural language. Hilberg's conjecture formalizes the received intuition that texts are generated by a process which is neither purely random nor purely deterministic. Among other results, we have argued that universal codes such as the Lempel-Ziv code may fail to efficiently compress some sources that satisfy Hilberg's conjecture. Thus Hilberg's conjecture poses some challenge for machine learning. An interesting open question is whether there exist simple examples of processes which satisfy this condition. For some relaxation of that problem we know that the answer is positive.

6. References

[1] Charikar M., Lehman E., Lehman A., Liu D., Panigrahy R., Prabhakaran M., Sahai A., Shelat A., *The smallest grammar problem,* IEEE Transactions on Information Theory 51, 2005, pp. 2554–2576.

[2] de Luca A., *On the combinatorics of finite words,* Theoretical Computer Science 218, 1999, pp. 13–39.

[3] de Marcken C. G. *Unsupervised Language Acquisition*, *PhD thesis,* Massachussetts Institute of Technology, 1996.

[4] Dębowski Ł., *Variable-length coding of two-sided asymptotically mean stationary measures,* Journal of Theoretical Probability 23, 2010, pp. 237–256.

[5] Dębowski Ł., *On the vocabulary of grammar-based codes and the logical consistency of texts,* IEEE Transactions on Information Theory 57, 2011, pp. 4589–4599.

[6] Dębowski Ł., *Maximal lengths of repeat in English prose.* S. Naumann, P. Grzybek, R. Vulanović, G. Altmann (Eds.), Synergetic Linguistics. Text and Language as Dynamic System, Praesens Verlag, Wien 2012, pp. 23–30.

[7] Dębowski Ł., *Mixing, ergodic, and nonergodic processes with rapidly growing information between blocks,* IEEE Transactions on Information Theory 58, 2012, pp. 3392–3401.

[8] Dębowski Ł., *Empirical evidence for Hilberg's conjecture in single-author texts.* In: I. Obradović, E. Kelih, R. Köhler (Eds.), Methods and Applications of Quantitative Linguistics – Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO), Academic Mind, Belgrade, 2013, pp. 143–151.

[9] Dębowski Ł., *A preadapted universal switch distribution for testing Hilberg's conjecture,* http://arxiv.org/abs/1310.8511, 2013.

[10] Dębowski Ł., *Estimation of entropy from subword complexity,* http://www.ipipan.waw.pl/˜ldebowsk/, 2014.

[11] Dębowski Ł., *Maximal repetitions in written texts: Finite energy hypothesis vs. strong Hilberg conjecture,* http://www.ipipan.waw.pl/˜ldebowsk/, 2014.

[12] Dębowski Ł., *A new universal code helps to distinguish natural language from random texts,* http://www.ipipan.waw.pl/˜ldebowsk/, 2014.

[13] Dębowski Ł., *On hidden Markov processes with infinite excess entropy,* Journal of Theoretical Probability 27, 2014, pp. 539–551.

[14] Ebeling W., Pöschel T., *Entropy and long-range correlations in literary English,* Europhysics Letters 26, 1994, pp. 241–246.

[15] Ebeling W., Nicolis G., *Entropy of symbolic sequences: The role of correlations,* Europhysics Letters 14, 1991, pp. 191–196.

[16] Ebeling W., Nicolis G., *Word frequency and entropy of symbolic sequences: A dynamical perspective,* Chaos, Solitons and Fractals 2, 1992, pp. 635–650.

[17] Graham R.L., Knuth D.E., Patashnik O., *Concrete Mathematics. A Foundation for Computer Science*, Addison-Wesley, 1994.

[18] Guiraud P., *Les caractères statistiques du vocabulaire*, Presses Universitaires de France, Paris, France, 1954.

[19] Heaps H.S., *Information Retrieval – Computational and Theoretical Aspects*, Academic Press, New York, USA, 1978.

[20] Herdan G., *Quantitative Linguistics*, Butterworths, London, England, 1964.

[21] Hilberg W., *Der bekannte Grenzwert der redundanzfreien Information in Texten – eine Fehlinterpretation der Shannonschen Experimente?* Frequenz 44, 1990, pp. 243–248.

[22] Jelinek F., *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, Massachusetts, USA, 1997.

[23] Kieffer J.C., Yang E., *Grammar-based codes: A new class of universal lossless source codes,* IEEE Transactions on Information Theory 46, 2000, pp. 737–754.

[24] Kit Ch., Wilks Y., *Unsupervised learning of word boundary with description length gain,* M. Osborne and E.T.K. Sang (Eds.), *Proceedings of the Computational Natural Language Learning ACL Workshop, Bergen*, 1999, pp. 1–6.

[25] Köhler R., Altmann G., Piotrowski R.G. (Eds.), *Quantitative Linguistik. Ein internationales Handbuch / Quantitative Linguistics. An International Handbook*, Walter de Gruyter, Berlin, Germany, 2005.

[26] Kolmogorov A.N., *Three approaches to the quantitative definition of information,* Problems of Information Transmission 1 (1), 1965, pp. 1–7.

[27] Kolpakov R., Kucherov G., *On maximal repetitions in words,* Journal of Discrete Algorithms 1, 1999, pp. 159–186.

[28] Kuraszkiewicz W., Łukaszewicz J., *The number of different words as a function of text length,* Pamiętnik Literacki (In Polish) 42 (1), 1951, pp. 168–182.

[29] Li M., Vitányi P.M.B., *An Introduction to Kolmogorov Complexity and Its Applications,* 2nd ed. Springer, New York, USA, 1997.

[30] Mandelbrot B., *An informational theory of the statistical structure of languages,* W. Jackson (Ed.), Communication Theory, Butterworths, London, England, 1953, pp. 486–502.

[31] Mandelbrot B., *Structure formelle des textes et communication*, Word 10, 1954, pp. 1–27.

[32] Markov A.A., *An example of statistical investigation of the text 'Eugene Onegin' concerning the connection of samples in chains,* Science in Context 19, 2006, pp. 591–600.

[33] Menzerath P., *Über einige phonetische Probleme,* Actes du premier Congres international de linguistes, Sijthoff, Leiden, Holland, 1928.

[34] Nevill-Manning C.G., *Inferring Sequential Structure*, PhD thesis, University of Waikato, 1996.

[35] Shannon C., *A mathematical theory of communication,* Bell System Technical Journal 30, 1948, pp. 379–423, 623–656.

[36] Shannon C., *Prediction and entropy of printed English,* Bell System Technical Journal 30, 1951, pp. 50–64.

[37] Shields P.C., *String matching: The ergodic case,* The Annals of Probability 20, 1992, pp. 1199–1203.

[38] Shields P.C., *Universal redundancy rates don't exist,* IEEE Transactions on Information Theory IT-39, 1993, pp. 520–524.

[39] Shields P.C., *String matching bounds via coding,* The Annals of Probability 25, 1997, pp. 329–336.

[40] Wolff J.G., *Language acquisition and the discovery of phrase structure,* Language and Speech 23, 1980, pp. 255–269.

[41] Zipf G.K., *The Psycho-Biology of Language: An Introduction to Dynamic Philology,* 2nd ed. The MIT Press, Cambridge, Massachusetts, USA, 1965.

[42] Ziv J., Lempel A., *A universal algorithm for sequential data compression,* IEEE Transactions on Information Theory, 23, 1977, pp. 337–343.