The Relaxed Hilberg Conjecture: A Review and New Experimental Support

Łukasz Dębowski

Institute of Computer Science
Polish Academy of Sciences
ldebowsk@ipipan.waw.pl

Abstract

The relaxed Hilberg conjecture states that the mutual information between two adjacent blocks of text in natural language grows as a power of the block length. The present paper reviews recent results concerning this conjecture. First, the relaxed Hilberg conjecture occurs when the texts repeatedly describe a random reality and Herdan's law for facts repeatedly described in the texts is obeyed. Second, the relaxed Hilberg conjecture implies Herdan's law for set phrases, which can be associated with the better known Herdan law for words. Third, the relaxed Hilberg conjecture is positively tested, using the Lempel-Ziv universal code, on a selection of texts in English, German, and French. Hence the relaxed Hilberg conjecture seems to be a likely and important hypothesis concerning natural language.

Keywords: Hilberg's conjecture, mutual information, Herdan's law, strongly nonergodic processes, grammar-based compression, Lempel-Ziv code

1 Introduction

It is a widely accepted view that texts produced by humans strongly diverge from both pure randomness and pure determinism (Zipf, 1965, p. 187). In the quantitative research of natural language one cannot, however, confine to such a statement but rather one has to investigate mathematical measures of nonrandomness and nondeterminism of texts (cf. e.g. Rao, 2010). Some quantitative measure of predictability of text is conditional entropy. An interesting observation concerning this quantity has been made by Hilberg (1990). Namely, he noticed that the conditional entropy of a letter given n previous letters decays as $n^{-1+\beta}$, where $\beta \approx 0.5$. Hilberg made this observation using the table of conditional entropy estimates in the famous article by Shannon (1951). Although Shannon's data points extended only to $n = 100$ letters of context, Hilberg supposed that a similar relationship holds also for much larger n, being a length of a random text or even longer.

Hilberg's paper (1990) was published in German in a local telecommunications journal and passed unnoticed by linguists. The first interest it received came from physicists seeking to understand properties of so called complex systems (Ebeling and Nicolis, 1991, 1992; Ebeling and Pöschel, 1994; Bialek et al., 2001b,a; Crutchfield and Feldman, 2003). Their main contribution was showing that Hilberg's conjecture implies that so called mutual information between two adjacent text blocks of length n is proportional to n^{β}. This is a weaker and more realistic statement since it does

not imply that human language production is deterministic in the limit.[1] We shall call this statement a relaxed Hilberg conjecture. It should be noted that even the relaxed Hilberg conjecture is incompatible with the conjecture of constant conditional entropy, recently proposed by cognitive scientists and critically examined by Ferrer-i-Cancho et al. (2013).

In this paper we would like to review the research in the relaxed Hilberg conjecture pushed forward in the last decade by Dębowski. He sought to put this hypothesis in the linguistic context while at the same time trying to develop a sound formal theory. Except for the initial paper (Dębowski, 2006), published in this journal, the research was mostly presented in engineering and mathematical periodicals, because of its technical involvement. However, these later results may be also highly interesting for linguists and thus worth popularizing in this venue. A similar review for physicists and researchers working in complex systems was published by Dębowski (2011a).

We hope that this paper may raise interest of general theoretical linguists and cognitive scientists. Seen from a larger perspective, Hilberg's conjecture inspires important and interesting questions about the amount of general and specific knowledge conveyed by texts, emergence of linguistic structure and patterns from a hypothetical stochastic process of speech or thinking, and interplay of randomness and determinism in human cognition as an attempt to model complexity of the world constrained by the organization of the brain and society. Culture, linguistics, pretty deep math, and who knows, maybe some physics, seem finally intertwined by means of a few results.

The main achievement of Dębowski was linking the relaxed Hilberg conjecture with two analogues of Herdan's law on the level of set phrase frequency and of text semantics. The original Herdan law is an observation that the number of distinct words observed in a text of length n is proportional to n^{β} (Kuraszkiewicz and Łukaszewicz, 1951; Guiraud, 1954; Herdan, 1964; Heaps, 1978). By an analogy, we will say that Herdan's law holds for another kind of objects if the number of the respective object types in a text of length n is proportional to n^{β}. In the following we shall speak of two kinds of Herdan's law as mathematical statements rather than empirical facts.

The first kind of Herdan's law we will discuss concerns nonterminal symbols of admissibly minimal grammar-based compressions of texts and will be called shortly Herdan's law for nonterminal symbols. A few words of clarification are needed here. Roughly speaking, admissibly minimal grammar-based codes are compression algorithms that represent a text as the smallest context-free grammar that generates the text as its sole production (Kieffer and Yang, 2000; Dębowski, 2011b). By the results of Charikar et al. (2005), we may suppose that these algorithms are computationally intractable. However, nonterminal symbols of certain grammar-based codes which are similar in principle to admissibly minimal grammar-based codes often correspond to words or set phrases, like *United Kingdom*, when these codes are applied to texts in natural language (Wolff, 1980; de Marcken, 1996; Nevill-Manning, 1996; Kit and Wilks, 1999). Thus we may expect that Herdan's law for nonterminal symbols is a certain approximation of Herdan's law for words. Investigating mathematical properties of admissibly minimal grammar-based codes, Dębowski (2006, 2011b) showed that Herdan's law for nonterminal symbols is a consequence of the relaxed Hilberg conjecture. Namely, if an arbitrary stationary stochastic process satisfies the relaxed Hilberg conjecture then texts generated by this process satisfy Herdan's law for nonterminal symbols. Precisely, a heuristic proof of this theorem was given by Dębowski (2006), whereas a

1 Vanishing entropy rate is equivalent to determinism e.g. by Lemma 4 of Dębowski (2009).

mathematically sound demonstration was provided by Dębowski (2011b).

Subsequently, Dębowski was interested for which stochastic processes the relaxed Hilberg conjecture is satisfied. He supposed that Hilberg's conjecture may stem from the fact that texts convey extremely large amounts of knowledge in a repetitive way, in particular by consistently referring to an external world. To model this phenomenon, Dębowski (2009) defined a class of strongly nonergodic stochastic processes that formalizes the concept of texts which repetitively describe a random reality. The defining feature of these processes is the existence of an infinite number of binary random variables, called facts, which can be learned from any sufficiently long text, i.e., a finite section of the process. Dębowski (2011b) showed that if texts generated by a strongly nonergodic process satisfy Herdan's law for those facts then the process satisfies the relaxed Hilberg conjecture. To prove that such random texts exist, Dębowski (2011b, 2012b) constructed a class of toy processes, called Santa Fe processes. Although highly idealized, Santa Fe processes can be given a natural linguistic interpretation.

Dębowski's results may be viewed as an explanation of Hilberg's conjecture and the distribution of words in texts created by humans. They show that the relaxed Hilberg conjecture and various kinds of Herdan's law have a very natural place in natural language. This conclusion can be supported experimentally. Dębowski (2013a) tested the relaxed Hilberg conjecture directly, using the Lempel-Ziv code (Ziv and Lempel, 1977) on a sample of ten texts in English. He obtained the exponent β", an upper bound of β, close to 0.94. In this paper we will show that the exponent β for a selection of 21 texts in German and French is also close to 0.94. Thus we may formulate a hypothesis that Hilberg's conjecture holds for other languages and a similar value of the exponent may be observed.

In the remaining sections, we will address particular points of Dębowski's research in more detail. In Section 2, we will discuss Hilberg's conjecture, the original and its refinements. In Section 3, we will exhibit some probabilistic models of texts that obey Hilberg's conjecture. Section 4 concerns links between Hilberg's conjecture and grammar-based compression. In Section 5 we present some old and new empirical data. Section 6 offers the conclusion.

2 Hilberg's conjecture and its refinements

Let us begin with a brief introduction to information theory, cf. Cover and Thomas (2006). The basic notion of information theory is the entropy of a random variable on a probability space. For a random text $X_1^n = (X_1, X_2, \ldots, X_n)$, where random variables X_i assume values of consecutive characters, the entropy is defined as

$$H(X_1^n) = -\sum_{x_1^n} P(X_1^n = x_1^n) \log P(X_1^n = x_1^n). \tag{1}$$

where P is the probability measure and $x_1^n = (x_1, x_2, \ldots, x_n)$ is a value that the random variable X_1^n assumes. This concept of entropy can be linked to text compression, or coding. For a uniquely decodable code, denoted C, the expectation of its length $|C(X_1^n)|$ cannot be smaller than the entropy, i.e.,

$$\sum_{x_1^n} P(X_1^n = x_1^n) |C(x_1^n)| \geq H(X_1^n). \tag{2}$$

There exists, however, a uniquely decodable code *C* with lengths

$$|C(x_1^n)| = [-\log P(X_1^n = x_1^n)], \tag{3}$$

where [*y*] is the smallest integer greater or equal *y*. Code (3) is called the Shannon-Fano code and satisfies

$$\sum\nolimits_{x_1^n} P\,(X_1^n = x_1^n)|C(x_1^n)| \leq H(X_1^n) + 1. \tag{4}$$

Whereas entropy pertains to a random variable, the length of the Shannon-Fano code (3) could be considered the information content of an individual (not random) text $x_1^n = (x_{1,}x_{2,}\ldots, x_n)$, such as a particular text in natural language. Alas, we cannot evaluate the length of the Shannon-Fano code if the proper probability distribution cannot be effectively identified and computed. In his seminal paper on defining the concept of information, Kolmogorov (1965) remarked that this may be well the case of texts in natural language. Subsequently, he proposed to define the information content of an individual text x_1^n as the length of the shortest program for a universal computer that makes the computer write x_1^n on its output.[2] This quantity is called Kolmogorov complexity $K(x_1^n)$. For any computable code *C* there exists a constant *c* such that

$$K(x_1^n) \leq |C(x_1^n)| + c. \tag{5}$$

Since Kolmogorov complexity is itself the length of a computable code, we obtain

$$\sum\nolimits_{x_1^n} P\,(X_1^n = x_1^n)K(x_1^n) \geq H(X_1^n) \tag{6}$$

for a random variable X_1^n on a particular probability space. In case of a computable probability distribution the Shannon-Fano code is also computable so, from (4) and (5), we obtain

$$\sum\nolimits_{x_1^n} P(X_1^n = x_1^n)K(x_1^n) \leq H(X_1^n) + c + 1. \tag{7}$$

In view of (6) and (7), the expectation of Kolmogorov complexity for computable distributions is close to entropy. For noncomputable distributions, however, the difference between Kolmogorov complexity and entropy can be arbitrarily large.

Whereas a disadvantage of entropy is the necessity of finding the right probability distribution, a disadvantage of Kolmogorov complexity is that it is not computable itself. Nonetheless there exists a middle path to measuring information content of individual texts, which is universal coding. A universal code is a uniquely decodable computable code *C* which for any stationary stochastic process $(X_i)_{i=-\infty}^{\infty}$ (i.e. an infinite sequence of random variables whose probability distribution is invariant with respect to shifting) asymptotically achieves the optimal compression rate

$$\lim_{n\to\infty} \frac{1}{n} \sum\nolimits_{x_1^n} P\,(X_1^n = x_1^n)|C(x_1^n)| = h, \tag{8}$$

---

2 A universal computer is a mathematical model of a computer that has infinite working memory and is capable of computing any computable function. A Turing machine is a classical model of a universal computer (Li and Vitányi, 2008).

where *h* is the entropy rate, defined as

$$h = \lim_{n\to\infty} \frac{1}{n} H(X_1^n). \tag{8}$$

Some examples of universal codes are the Lempel-Ziv code (Ziv and Lempel, 1977) or grammar-based codes (Kieffer and Yang, 2000; Dębowski, 2011b). Using these codes, we can effectively measure the information content of an individual text. In particular, these codes are used for natural language data compression.

Now we can introduce Hilberg's conjecture. Originally it deals with conditional entropy

$$H(X_n/X_1^{n-1}) = -\sum_{x_1^n} P\,(X_1^n = x_1^n) \log P(X_n = x_n/X_1^{n-1} = x_1^{n-1}). \tag{10}$$

Shannon (1951) attempted to estimate this quantity for printed English using a guessing method. A few decades later, Hilberg (1990) replotted these estimates in the doubly logarithmic scale and observed an approximate power-law relationship

$$H(X_n/X_1^{n-1}) \propto n^{-1+\beta}, \tag{11}$$

where $\beta \approx 0.5$ and $n \leq 100$ characters. When extrapolated to arbitrary n, this relationship implies

$$H(X_1^n) = \sum_{m=1}^{n} H\,(X_m/X_1^{m-1}) \propto \int m^{-1+\beta}\, dm \propto n^{\beta}. \tag{12}$$

Hence we obtain a power law for the joint entropy per token

$$\frac{H(X_1^n)}{n} \propto n^{-1+\beta}. \tag{13}$$

Relationship (13) is the original Hilberg conjecture.

Having derived (13), Hilberg conjectured that the entropy rate (9) amounts to zero. This proposition implies asymptotic determinism of human utterances (i.e. the next character of a text is a function of the infinite past (Dębowski, 2009, Lemma 4)) and theoretical possibility of compression of texts which goes far beyond present state of the art. Whereas one cannot exclude this case a priori, it may be safer to assume that

$$\frac{H(X_1^n)}{n} \approx An^{-1+\beta} + h, \tag{14}$$

where constant *h* is strictly positive. This proposition may be called a relaxed Hilberg conjecture for random texts.

Under the assumption of stationarity, the relaxed Hilberg conjecture can be simpler expressed using mutual information. The Shannon mutual information between two random blocks of length *n* is defined as

$$I_H(X_1^n; X_{n+1}^{2n}) = H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n}). \tag{15}$$

Assuming that $H\left(X_{i+1}^{i+n}\right)$ does not depend on *i* (a form of stationarity!), for relationship (14) we have

$$I_H(X_1^n; X_{n+1}^{2n}) \approx 2An^{\beta} + 2hn - A(2n)^{\beta} - 2hn = Bn^{\beta}, \tag{16}$$

where $B = \left(2 - 2^{\beta}\right)A$. Indeed, relationship (16) is equivalent to (14). To prove it, let us observe that (16) implies

$$H(X_1^n) = \sum_{k=0}^{\infty} \frac{I_H\left(X_1^{2^k n}; X_{2^k n+1}^{2^{k+1} n}\right)}{2^{k+1}} + \lim_{n\to\infty} \frac{H(2^{k+1} n)}{2^{k+1}} \tag{17}$$

$$\approx \left(\sum_{k=0}^{\infty} \frac{2^{\beta k}}{2^{k+1}}\right) B n^{\beta} + hn = A n^{\beta} + hn. \tag{18}$$

Hence, assuming stationarity, the relaxed Hilberg conjecture for random texts states equivalently that mutual information between two adjacent text blocks grows as a power of the block length.

Relationships (14) and (16) are, practically speaking, impossible to test experimentally since the exact probability distribution for texts in natural language is unknown. As a first improvement, it seems more proper to speak of Kolmogorov complexity $K(x_1^n)$ of an individual text x_1^n rather than of the entropy $H(X_1^n)$ of a random variable X_1^n. Thus another plausible modification of Hilberg's conjecture reads

$$\frac{K(x_1^n)}{n} \approx A' n^{-1+\beta'} + h'. \tag{19}$$

Some problem with hypothesis (19) is, however, that it falsely assumes that $K(x_1^n)$ grows uniformly with x_1^n, whereas e.g. for a sequence of independent identically distributed variables X_1^n there arise random fluctuations of Kolmogorov complexity $K(X_1^n)$ of order $\sqrt{n \log\log n}$.[3] We may consider (19) only if $\beta \gg 1/2$. Fortunately, these rather large fluctuations cancel if we consider the algorithmic mutual information between two individual texts blocks, defined as

$$I_K(x_1^n; x_{n+1}^{2n}) = K(x_1^n) + K(x_{n+1}^{2n}) - K(x_1^{2n}). \tag{20}$$

Thus, the term 'the relaxed Hilberg conjecture for individual texts' will be rather used for proposition

$$I_K(x_1^n; x_{n+1}^{2n}) \approx B' n^{\beta'}. \tag{21}$$

Hypothesis (21) is still impossible to test empirically since Kolmogorov complexity is incomputable. Thus for the sake of effective testing, we could consider a universal code *C*, measure its length $|C(x_1^n)|$ for a text x_1^n, and investigate the relationships

$$\frac{|C(x_1^n)|}{n} \approx A'' n^{-1+\beta''} + h'' \tag{22}$$

and

$$I_C(x_1^n; x_{n+1}^{2n}) \approx B'' n^{\beta''}, \tag{23}$$

where $I_C(x_1^n; x_{n+1}^{2n})$ is the code-based mutual information

$$I_C(x_1^n; x_{n+1}^{2n}) = |C(x_1^n)| + |C(x_{n+1}^{2n})| - |C(x_1^{2n})|. \tag{24}$$

---

3 In that case $K(x_1^n)$ is close to $-\sum_{i=1}^{n} \log P(X_i)$, which is a sum of independent identically distributed variables and the fluctuations thereof are described by the law of iterated logarithm (Billingsley, 1979, Section 9).

Proposition (23) will be called a relaxed Hilberg conjecture for code C. As we will show in Section 5 both formulae (22) and (23) fit the empirical data very well for $\beta \approx 1$, $h \approx 0$, and $B'' = \left(2 - 2^{\beta''}\right)A''$. Miraculously, the fluctuations of $|C(x_1^n)|$ are relatively small and dominated by the term $A''n^{\beta''}$.

In certain cases, the three flavors of the relaxed Hilberg conjecture can be related to one another. If relationships (19) and (22) are satisfied for random texts drawn from a stationary process then we obtain $h = h' = h''$ and $\beta \leq \beta' \leq \beta''$ by relationships (2), (5), and (8). Alas, there exist nonstationary stochastic processes for which $h < h' < h''$ and exponents β, β', and β'' cannot be interrelated. Even in the case of natural language, where we have problems with identifying the probability distribution, we need not have a priori $h' = h''$ and $\beta' \leq \beta''$. Nonetheless, the relaxed Hilberg conjecture (23) can be investigated for natural language on its own interest using efficiently computable universal codes.

3 Texts repetitively describing a random reality

One can ask the question why Hilberg's conjecture might be satisfied. In fact, origins of the relaxed Hilberg conjecture may be traced in the narrative coherence of texts. By the narrative coherence we understand the following observation. Before we read the title of a book, we hardly have any preconception what it may be about. But if we begin reading a text in English, we expect that the remaining part of the text is also in English. If we commence reading a manuscript on mathematics, we expect formulas to permeate the manuscript. If we start reading a novel about a heroine called Alice, we expect that Alice remains the heroine until the end of the novel. The number of such narrative constraints in texts is a priori unlimited. It is important to observe that these constraints, although they persist within a text, are mostly random. Thus we may say that texts in natural language repetitively describe a random reality. As we will see, if this random reality can be learned from texts sufficiently fast then the relaxed Hilberg conjecture is satisfied. The intuition is that similar portions of the random reality can be learned independently from two adjacent blocks of text and the mutual information between the blocks cannot be smaller than the amount of the inferred knowledge.

Dębowski (2011b, 2012b) proposed the following mathematical model. Let $(X_i)_{i=-\infty}^{\infty}$ be a sequence of random variables that assume values from a countable set Y, called the alphabet. This sequence will model an infinitely long text, the totality of natural language production, where X_i are consecutive text units. We can imagine that the values of X_i are characters if the alphabet is finite, or words or sentences if the alphabet is infinite. Moreover, let $(Z_k)_{k=1}^{\infty}$be a sequence of independent random variables that assume value 0 or 1 with equal probability. We can imagine that the values of Z_k are logical values (1 = true and 0 = false) of certain systematically enumerated independent propositions which concern the reality described in the text in a repetitive way. In some interpretation, propositions Z_k's resume the knowledge of the described world in the most concise form. Dębowski called these propositions briefly facts. He assumed that the facts are a priori unknown to the reader of the text but can be asymptotically learned from any sufficiently long section of the infinite text. Thus he obtained this definition:

Definition 1 (Dębowski, 2011b)**.** *A stochastic process* $(X_i)_{i=-\infty}^{\infty}$ *is called strongly nonergodic if there exist independent binary variables* $(Z_k)_{k=1}^{\infty}$ *with* $P(Z_k = 0) = P(Z_k = 1) = \frac{1}{2}$ *and functions*

$s_k: Y^* \to \{0,1\}$,[4] *k = 1, 2, 3, ..., such that for all t and k,*

$$\lim_{n\to\infty} P\left(s_k(X_{t+1}^{t+n}) = Z_k\right) = 1. \quad \textbf{(25)}$$

Functions s_k are motivated by the idea that there is a definite method of interpreting finite texts in natural language to infer facts about the random world. This method is simply the human language competence. Definition 1 assumes that facts, or in other words, the knowledge, mentioned in texts can be learned by text readers ultimately, regardless of their starting point. The actual linguistic reality is a bit more complicated. Facts that are mentioned repeatedly can be divided into two classes: (i) facts about the unchangeable objective world (like mathematical or physical constants), which can be discovered and reported independently by successive generations of text creators, and (ii) facts about historical heritage (like fiction, culture, language, or geography), which are subject to distributed creation, accumulation, and lossy transmission from text creators to text readers. Whereas the first class of facts falls under the scope of Definition 1, the second class is not captured because these facts may be forgotten and cannot be learned from the text in the distant future.

An important characteristic of a strongly nonergodic process is the number of facts that can be predicted from a given finite text with a sufficiently high probability. The sets of those facts may be defined as

$$U_\delta(n) := \{k \in N: P(s_k(X_1^n) = Z_k) \geq \delta\}, \quad \textbf{(26)}$$

where δ is a fixed constant close to 1, whereas $|U_\delta(n)|$ will denote the number of elements in $U_\delta(n)$. By an analogy to the word distribution, we will say that Herdan's law is obeyed for facts if the number of facts that can be predicted from the text grows as a power of the text length, i.e,

$$|U_\delta(n)| \propto n^\beta. \quad \textbf{(27)}$$

The subsequent theorem states that if we have Herdan's law for facts then the relaxed Hilberg conjecture must hold with an exponent greater or equal as in the Herdan law for facts.

Theorem 1 (Dębowski, 2011b)**.** *Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary strongly nonergodic process over a finite alphabet Y and define sets (26), where functions s_k satisfy (25). Suppose that inequality*

$$lim\ inf_{n\to\infty} \frac{|U_\delta(n)|}{n^\beta} > 0 \quad \textbf{(28)}$$

holds for some $\beta \in (0, 1)$ and $\delta \in (1/2, 1)$.[5] *Then*

$$lim\ sup_{n\to\infty} \frac{I_H(X_1^n; X_{n+1}^{2n})}{n^\beta} > 0. \quad \textbf{(29)}$$

The world of stochastic processes is very rich and maybe there dwells a process that models human language production sufficiently well. If this process satisfies Herdan's law for facts then, by

---

4 Symbol *Y** denotes the set of all finite strings obtained by concatenating symbols from *Y*.

5 The lower limit of a sequence is defined as $lim\ inf_{n\to\infty} a_n = lim_{n\to\infty}\, i\, nf_{m\geq n} a_m$, where $inf_{m\geq n} a_m$ is the largest number *r* such that $r \leq a_m$ for all $m \geq n$. Analogously, we define the the upper limit of a sequence as $lim\ sup_{n\to\infty} a_n = lim_{n\to\infty}\, s\, up_{m\geq n} a_m$, where $sup_{m\geq n} a_m$ is the smallest number *r* such that $r \geq a_m$ for all $m \geq n$. Th upper and the lower limits exist for any sequence but may be different. We have $lim\ inf_{n\to\infty} a_n \leq lim\ sup_{n\to\infty} a_n$ in general. The upper and the lower limits are equal if and only if $lim_{n\to\infty}\, a_n$. Then $lim_{n\to\infty}\, a_n$ is their common value.

Theorem 1, it also satisfies the relaxed Hilberg conjecture (29). However, an honest mathematician should show that Theorem 1 is not void, that is, there exists at least one process which satisfies Herdan's law for facts. Exhibiting such a simplistic instance in this paper may be also helpful to imagine what may happen in natural language. Dębowski (2011b) introduced the following example, which he called a Santa Fe process.

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

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

where $(Z_k)_{k=1}^{\infty}$ *and* $(K_i)_{i=-\infty}^{\infty}$ *are probabilistically independent,* $(Z_k)_{k=1}^{\infty}$ *are as in Definition 1, whereas* $(K_i)_{i=-\infty}^{\infty}$ *is a sequence of independent variables that satisfy the Zipf-Mandelbrot law*

$$P(K_i = k) \propto k^{-1/\beta}, \tag{31}$$

where $\beta \in (0, 1)$*. Number* β *is the only parameter of the process.*[6]

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 ($z = z'$).

Dębowski (2011b) showed that Santa Fe process satisfies Herdan's law for facts (27). Later, he demonstrated that the relaxed Hilberg conjecture (29) can be strengthened for this process in the following form.

Theorem 2 (Dębowski, 2012b)**.** *Let* $\beta \in (0, 1)$*. The Santa Fe process obeys*

$$\lim_{n\to\infty} \frac{I_H(X_1^n; X_{n+1}^{2n})}{n^{\beta}} > 0. \tag{32}$$

Let us note that the random object$(Z_k)_{k=1}^{\infty}$ described by the Santa Fe process (30) does not evolve in time. As we have said, the actual linguistic reality is a bit more complicated. Besides describing real or fictitious worlds that do not change in time, texts in natural language describe worlds that evolve with a varied speed. To encompass the latter property of texts in natural language, we may generalize the Santa Fe process as follows.

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

$$X_i = \left(K_i, Z_{i,K_i}\right), \tag{33}$$

where processes $(K_i)_{i=-\infty}^{\infty}$ *and* $(Z_{ik})_{i=-\infty}^{\infty}$*, where* $k \in N$*, are independent and distributed as follows. First, variables* K_i *are distributed according to formula (31), as before. Second, each process* $(Z_{ik})_{i=-\infty}^{\infty}$ *is a Markov chain with marginal distribution*

---

6 The Santa Fe can be easily simulated as follows. First we sample $K_i = k$ from distribution (31). Then if Z_k has been previously sampled, we output the previously sampled value $Z_k = z$, otherwise we sample $Z_k = z$ from distribution $P(Z_k = 0) = P(Z_k = 1) = \frac{1}{2}$.

$$P(Z_{ik} = 0) = P(Z_{ik} = 1) = \frac{1}{2} \tag{34}$$

and cross-over probabilities

$$P\left(Z_{ik} = 0/Z_{i-1,k} = 1\right) = P\left(Z_{ik} = 1/Z_{i-1,k} = 0\right) = p_{k.} \tag{35}$$

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

The object $(Z_{ik})_{i=-\infty}^{\infty}$ 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. Dębowski (2012b) proved that the generalized Santa Fe process is ergodic for $p_k \neq 0$, hence it is not strongly nonergodic. However, also in the case of $p_k \neq 0$, we obtain Hilberg's conjecture if the cross-over probabilities decay fast enough.

Theorem 3 (Dębowski, 2012b)**.** *Suppose that $lim_{n\to\infty} p_k/P\,(K_i = k) = 0$. Then the generalized Santa Fe process obeys (32).*

The moral of this section is that some processes satisfying the relaxed Hilberg conjecture can be constructed quite easily and a natural linguistic interpretation can be provided for those. Of course, not all processes with a power-law growth of mutual information can be interpreted linguistically. Dębowski (2013b) presented also a few hidden Markov processes satisfying the relaxed Hilberg conjecture but these are of purely mathematical interest.

4 Links with grammar-based compression

Another important result concerning the relaxed Hilberg conjecture involves grammar-based compression. Originally, grammar-based compression has been proposed as an approximate method for detecting word boundaries in texts lacking spaces between words, such as the output of automatic speech recognition (Jelinek, 1997). It has been observed that strings of characters that are repeated within the text sufficiently many times often correspond to whole words or set phrases like United Kingdom (Wolff, 1980; de Marcken, 1996; Nevill-Manning, 1996; Kit and Wilks, 1999). Such strings can be detected automatically using grammar-based codes (Kieffer and Yang, 2000).

Let us present the necessary mathematical concepts. Grammar-based codes 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 string. We will shortly say that an admissible grammar generates this string. In an admissible grammar, there is exactly one rule per nonterminal symbol and the nonterminals can be ordered so that the symbols are rewritten onto strings of strictly succeeding symbols (Kieffer and Yang, 2000; Charikar et al., 2005). Hence, such a grammar is given by its set of production rules

$$G = \begin{Bmatrix} A_1 \to \alpha_1 \\ A_2 \to \alpha_2 \\ \cdots \\ A_k \to \alpha_k \end{Bmatrix}, \tag{36}$$

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

A function Γ such that $\Gamma(w)$ is an admissible grammar that generates string w is called a grammar transform (Kieffer and Yang, 2000). Many such transforms have been proposed, see Kieffer and Yang (2000). For example, the longest matching grammar transform for the tongue twister

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

returns the admissible grammar

$$\begin{Bmatrix} A_1 \to IA_2 youA_2 we\_allA_3\_for\_iceA_4! \\ A_2 \to A_3,\_ \\ A_3 \to \_sA_4 \\ A_4 \to cream \end{Bmatrix}. \quad (37)$$

In the compressions of longer texts, nonterminal symbols often correspond to words or set phrases, like A_4 in (37), especially if it is additionally required that the secondary nonterminals were defined as strings of only terminal symbols (Kit and Wilks, 1999).

Grammar transforms can be compared according to the so called grammar length. A few distinct definitions for this concept have been proposed. Kieffer and Yang (2000) defined the length of grammar (36) as the total length of the rules' right hand sides,

$$|G| := \sum_{i=1}^{k} |\alpha_i|, \quad (38)$$

where $|\alpha_i|$ is the length of α_i. Subsequently, Charikar et al. (2005) investigated properties of grammar transforms which minimize this kind of length. In contrast, Dębowski (2011b) constructed admissibly minimal grammar transforms that minimize a slightly different length function. The exact definition of admissibly minimal transforms is too technical to present it right here. However, we may say that admissibly minimal transforms resemble the grammar transforms considered by de Marcken (1996) and Kit and Wilks (1999) in the computational linguistic task of detecting word boundaries .

Denote the number of distinct nonterminal symbols in grammar (36) as

$$V(G) := k. \quad (39)$$

By an analogy to the word distribution, we will say that Herdan's law is obeyed for a grammar transform Γ if the number of distinct nonterminal symbols returned by the grammar transform grows as a power of the text length, i.e.,

$$V\big(\Gamma(x_1^n)\big) \propto n^{\beta}. \quad (40)$$

We suppose that Herdan's law for admissibly minimal grammar transforms can be related to Herdan's law for words or set phrases. These two statements are equivalent if the number of distinct nonterminal symbols is proportional to the number of distinct words or set phrases in the encoded text. The latter claim seems likely in view of experiments by de Marcken (1996) and Kit and Wilks (1999). It is a good question, however, whether this claim can be efficiently empirically tested. Certainly, the number of nonterminals returned by the compression schemes proposed by de

Marcken (1996) and Kit and Wilks (1999) can be efficiently computed. Nevertheless, it is known that grammar transforms which minimize length (38) globally are not computable in polynomial time (Charikar et al., 2005). The same may be true for the exact admissibly minimal grammar transforms. This may mean that the number of distinct nonterminal symbols in an admissibly minimal grammar-based compression of a text cannot be efficiently computed.

Although probably hard to compute, admissibly minimal grammar transforms have a theoretical advantage. Dębowski (2011b) showed that if the relaxed Hilberg conjecture is satisfied then Herdan's law for any admissibly minimal grammar transform is approximately obeyed. In the formal statement of this result, the maximal length of a (possibly overlapping) repeated substring in text $w \in Y^*$ will be denoted as

$$L(w) := \max\{|s| : w = x_1 s y_1 = x_2 s y_2 \wedge x_1 \neq x_2\}, \tag{41}$$

where $s, x_i, y_i \in Y^*$. According to the reasoning presented by Dębowski (2011b) we have:

Theorem 4. *Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary process over a finite alphabet Y with entropy rate $h > 0$. Assume that inequality*

$$lim\ inf_{n\to\infty} \frac{I_H(X_1^n; X_{n+1}^{2n})}{n^\beta} > 0 \tag{42}$$

holds for some $\beta \in (0, 1)$. Let Γ be an admissibly minimal grammar transform. Then we have

$$lim\ sup_{n\to\infty} \frac{\sum_{x_1^n} P\,(X_1^n = x_1^n) V\big(\Gamma(x_1^n)\big)\big(1 + L(x_1^n)\big)}{n^\beta} > 0. \tag{43}$$

Theorem 4 states that if the relaxed Hilberg conjecture is satisfied then we have Herdan's law for any admissibly minimal grammar transform provided the length of the longest repeated string in the text does not grow too fast. The latter condition can be asserted in a few interesting cases. For example, for an encoding of the Santa Fe process (30) into a finite alphabet, the maximal length of repeat is proportional to the logarithm of the text length,

$$L(X_1^n) \propto \log n \tag{44}$$

(Dębowski, 2010, 2011b). In contrast, for a sample of texts in English, Dębowski (2012a) checked experimentally that the maximal length of repeat is proportional to a power of the logarithm of the text length,

$$L(x_1^n) \propto (\log n)^\alpha \tag{44}$$

where $\alpha < 4$. Hence we may say that in the relevant cases the relaxed Hilberg conjecture implies Herdan's law for any admissibly minimal grammar transform.

5 Empirical data

The original Hilberg conjecture (13) is hard to test since human guessing estimates of conditional entropy are extremely costly to obtain and prone to large errors. These costs and errors get extreme if we wish to test Hilberg conjecture (13) for very long contexts imagined by Hilberg, such as the order of a novel length. In contrast, the relaxed Hilberg conjecture for universal codes, (22) and (23),

can be tested effectively using texts of any conceivable length.

As our preliminary study shows, the obtained exponent β in conjecture (22) is much larger than β ≈ 0.5 supposed by Hilberg in conjecture (13). For the sake of testing Hilberg's conjecture, we have compressed 31 texts written in English, German, and French.[7] The texts have been downloaded from the Project Gutenberg[8] and are listed in Table 2. We have deleted the licenses contained in the text files and we reduced the alphabet to 27 symbols (26 capital letters and a space), stripping all diacritics, numbers, and punctuation marks, as it has been usually done in previous publications concerning the entropy of English (Shannon, 1951; Cover and King, 1978). Subsequently, we have measured the length of the Lempel-Ziv code for prefixes of the text sequence of an exponentially increasing length. The Lempel-Ziv code was provided by our own implementation done for the alphabet of 27 symbols and featuring an infinite buffer, which assures the universality. The dependence of the compression rate on the block length is given in Figures 1, 3, and 5, for each language separately, whereas the data concerning the mutual information are plotted in Figures 2, 4, and 6.

Using the nonlinear least-squares (NLLS) Marquardt-Levenberg algorithm, we have fitted the following simple model for the compression rate. Let $|C(x_1^n)|$ denote the length of the Lempel-Ziv code for the text of length n (in characters). The fitted model has been (22) with free parameters A'' and β'' whereas h'' has been set to 0. This provided a very satisfactory fit, whereas letting h'' vary resulted in nonconverging computation, for an undiscerned reason. In this way we have obtained:

a) for English:

$$\frac{|C(x_1^n)|}{n} \approx 6.22 n^{-1+0.949} [bpc]; \tag{46}$$

b) for German:

$$\frac{|C(x_1^n)|}{n} \approx 6.73 n^{-1+0.942} [bpc]; \tag{47}$$

c) for French:

$$\frac{|C(x_1^n)|}{n} \approx 6.19 n^{-1+0.950} [bpc]. \tag{48}$$

The parameters of the models for different languages are very similar. The data summarizing the quality of the fitted models are given in Table 1. In the table, SE stands for the standard error and RMSE is the root mean square error.

From the compression rate, we can easily compute the mutual information. The fit for parameters A'' and β'' has been so good, that we have decided to use formula (23) with $B'' = \left(2 - 2^{\beta''}\right)A''$, where A'' and β'' are the estimated parameters. In this way, we have obtained:

a) for English:

$$I_C(x_1^n; x_{n+1}^{2n}) \approx 0.432 n^{0.949} [bits]; \tag{49}$$

7 The results concerning English have been previously published in the conference paper Dębowski (2013a).

8 http://www.gutenberg.org/

b) for German:

$$I_C(x_1^n; x_{n+1}^{2n}) \approx 0.539n^{0.942}[bits]; \quad (\mathbf{50})$$

c) for French:

$$I_C(x_1^n; x_{n+1}^{2n}) \approx 0.430n^{0.950}[bits]. \quad (\mathbf{51})$$

All formulae fit very well.

It may be somewhat surprising that models (46), (47), and (48) fit so well although they contain no constant term $h'' > 0$, supposed in conjecture (22). We know, however, from independent studies that the asymptotic entropy rate $h \leq h''$ for English is less than 1.25 bpc (Cover and King, 1978). In contrast, the lowest compression rate that we observe in Figures 1, 3, and 5 is about 3.0 bpc. Thus we cannot exclude the possibility that the compression rate is asymptotically bounded below by the value of 1.25 bpc or somewhat smaller.

6 Conclusion

The relaxed Hilberg conjecture is a statement that the mutual information between two adjacent blocks of a text in natural language grows very fast, namely, as a power of the block length. In the present paper we have reviewed a few important results concerning this conjecture. First, we have reported that the conjecture occurs when texts in natural language repeatedly describe a random reality and Herdan's law pertaining to facts repeatedly described in the texts is obeyed. Second, we have communicated that the relaxed Hilberg conjecture implies Herdan's law for set phrases. Third, we have shown that the conjecture can be positively tested for texts in English, German, and French using the Lempel-Ziv code. The parameters of the fitted models for those languages are very similar, which is not so surprising given our observation that the origins of Hilberg's conjecture lie in text meaning.

All these observations make the relaxed Hilberg conjecture a very probable and weighty hypothesis concerning natural language. The fundamental importance of Hilberg's conjecture can be corroborated by the fact that Herdan's law for set phrases, a corollary of Hilberg's conjecture, can be probably associated with the better known Herdan law for words, which is in turn a consequence of the celebrated Zipf-Mandelbrot law. Just to recall, the Zipf-Mandelbrot law states that the word frequency is an inverse power of the word rank on the frequency list (Zipf, 1965; Mandelbrot, 1954).

We think that Hilberg's conjecture deserves further research and extensive testing for a wide selection of texts and languages, including also non-Indoeuropean languages, which are typologically different from the considered three languages. The exponent in the relaxed Hilberg conjecture can be connected to the text meaning and thus should be invariant with respect to text translation from one language into another. In contrast, we may expect a priori a larger variation of the exponent across texts that are not related to one another. It should be noted, however, that in the present data this variation is hardly visible when we tried to compress texts using the Lempel-Ziv code. Thus Hilberg's conjecture for the Lempel-Ziv code can be potentially a general quantitative linguistic law, as it was named by Dębowski (2006).

It may be also interesting to investigate Hilberg's conjecture for large corpora. Although corpora are very imperfect models of large-scale human communications (since by their construction, they

are randomized and do not contain longer narrations), they exhibit curious second regimes of Zipf's law for very large ranks (Ferrer i Cancho and Solé, 2001; Montemurro and Zanette, 2002; Petersen et al., 2012). It might be insightful to verify whether an analogous second regime arises also for Hilberg's conjecture for large corpora. Possibly, a positive asymptotic lower bound for the compression rate can be effectively observed as well.

We should also be aware that mathematical investigations of Hilberg's conjecture are not completed yet, either. The estimates of entropy given by the Lempel-Ziv code may be strongly biased and some better estimation procedures should be provided by mathematically inclined researchers. Another bundle of questions is whether the entropy rate of texts in natural language is actually zero, what plausible mathematical models of this phenomenon are, and why typical compression algorithms cannot compress the texts as well as suggested by the original Hilberg conjecture. Finally, an anonymous referee noted that considering mutual information between nonadjacent blocks and nonstationary processes can be also an interesting idea. We suppose that scaling of mutual information for non-adjacent blocks can be a property that distinguishes texts in natural language from the Santa Fe processes discussed in this paper. Developing this idea requires further work.

Acknowledgment

The author wishes to thank Jan Mielniczuk, Ramon Ferrer-i-Cancho, Jacek Koronacki, and an anonymous referee for reading the manuscript and discussion.

References

Bialek, W., Nemenman, I., and Tishby, N. (2001a). Complexity through nonextensivity. *Physica A*, 302:89–99.

Bialek, W., Nemenman, I., and Tishby, N. (2001b). Predictability, complexity and learning. *Neural Computation*, 13:2409.

Billingsley, P. (1979). *Probability and Measure*. New York: John Wiley.

Charikar, M., Lehman, E., Lehman, A., Liu, D., Panigrahy, R., Prabhakaran, M., Sahai, A., and Shelat, A. (2005). The smallest grammar problem. *IEEE Transactions on Information Theory*, 51:2554–2576.

Cover, T. M. and King, R. C. (1978). A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory*, 24:413–421.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory, 2nd ed.* New York: John Wiley.

Crutchfield, J. P. and Feldman, D. P. (2003). Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, 15:25–54.

de Marcken, C. G. (1996). *Unsupervised Language Acquisition.* PhD thesis, Massachussetts Institute of Technology.

Dębowski, Ł. (2006). On Hilberg's law and its links with Guiraud's law. *Journal of Quantitative Linguistics*, 13:81–109.

Dębowski, Ł. (2009). A general definition of conditional information and its application to ergodic decomposition. *Statistics and Probability Letters*, 79:1260– 1268.

Dębowski, Ł. (2010). Variable-length coding of two-sided asymptotically mean stationary measures. *Journal of Theoretical Probability*, 23:237–256.

Dębowski, Ł. (2011a). Excess entropy in natural language: present state and perspectives. *Chaos*, 21:037105.

Dębowski, Ł. (2011b). On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Transactions on Information Theory*, 57:4589–4599.

Dębowski, Ł. (2012a). Maximal lengths of repeat in English prose. In Naumann, S., Grzybek, P., Vulanović, R., and Altmann, G., editors, *Synergetic Linguistics. Text and Language as Dynamic System*, pages 23–30. Wien: Praesens Verlag.

Dębowski, Ł. (2012b). Mixing, ergodic, and nonergodic processes with rapidly growing information between blocks. *IEEE Transactions on Information Theory*, 58:3392–3401.

Dębowski, Ł. (2013a). Empirical evidence for Hilberg's conjecture in single- author texts. In Obradović, I., Kelih, E., and Köhler, R., editors, *Methods and Applications of Quantitative Linguistics—Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO)*, pages 143–151.

Dębowski, Ł. (2013b). On hidden Markov processes with infinite excess entropy. *Journal of Theoretical Probability*. DOI: 10.1007/s10959-012-0468-6.

Ebeling, W. and Nicolis, G. (1991). Entropy of symbolic sequences: the role of correlations. *Europhysics Letters*, 14:191–196.

Ebeling, W. and Nicolis, G. (1992). Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons and Fractals*, 2:635–650.

Ebeling, W. and Pöschel, T. (1994). Entropy and long-range correlations in literary English. *Europhysics Letters*, 26:241–246.

Ferrer-i-Cancho, R., Dębowski, Ł., and del Prado Martin, F. M. (2013). Constant conditional entropy and related hypotheses. *Journal of Statistical Mechanics: Theory and Experiment*, page L07001.

Ferrer i Cancho, R. and Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8(3):165–173.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire*. Paris: Presses Universitaires de France.

Heaps, H. S. (1978). *Information Retrieval—Computational and Theoretical Aspects*. New York: Academic Press.

Herdan, G. (1964). *Quantitative Linguistics*. London: Butterworths.

Hilberg, W. (1990). Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44:243–248.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Cambridge, MA: The MIT Press.

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

Kit, C. and Wilks, Y. (1999). Unsupervised learning of word boundary with description length gain. In Osborne, M. and Sang, E. T. K., editors, *Proceedings of the Computational Natural Language Learning ACL Workshop, Bergen*, pages 1–6.

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

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

Li, M. and Vitányi, P. M. B. (2008). *An Introduction to Kolmogorov Complexity and Its Applications, 3rd ed.* New York: Springer.

Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word*, 10:1–27.

Montemurro, M. A. and Zanette, D. H. (2002). New perspectives on Zipf's law in linguistics: from single texts to large corpora. *Glottometrics*, 4:87–99.

Nevill-Manning, C. G. (1996). *Inferring Sequential Structure*. PhD thesis, University of Waikato.

Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., and Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Report*s, 2(943).

Rao, R. (2010). Probabilistic analysis of an ancient undeciphered script. *IEEE Computer*, 43:76–80.

Shannon, C. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30:50–64.

Wolff, J. G. (1980). Language acquisition and the discovery of phrase structure. *Language and Speech*, 23:255–269.

Zipf, G. K. (1965). *The Psycho-Biology of Language: An Introduction to Dynamic Philology, 2nd ed*. Cambridge, MA: The MIT Press.

Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343.

language	A''	SE of A''	β''	SE of β''	RMSE
English	6.22	0.07	0.949	0.001	0.0987
German	6.73	0.10	0.942	0.002	0.1106
French	6.19	0.07	0.950	0.001	0.0874

Table 1. Summary of the fitted models.

English texts:	
Title	**Author**
First Folio/35 Plays	W. Shakespeare
Critical & Historical Essays	T. B. Macaulay
The Complete Memoirs	J. Casanova
Memoirs of Comtesse du Barry	E. Lamothe-Langon
The Descent of Man	C. Darwin
Gulliver's Travels	J. Swift
The Mysterious Island	J. Verne
Mark Twain, a Biography	A. B. Paine
The Journal to Stella	J. Swift
Life of William Carey	G. Smith
German texts:	
Title	**Author**
Die Abenteuer Tom Sawyers	M. Twain
Alice's Abenteuer im Wunderland	L. Carroll
Also Sprach Zarathustra	F. Nietzsche
Buddenbrooks	T. Mann
Faust	J. W. von Goethe
Die Göttliche Komödie	D. Alighieri
Kritik der reinen Vernunft	I. Kant
Der Tod in Venedig	T. Mann
Die Traumdeutung	S. Freud
Die Verwandlung	F. Kafka
French texts:	

Title	**Author**
20000 Lieues sous les mers	J. Verne
Candide	Voltaire
Le comte de Monte-Cristo, Tome I	A. Dumas
Discours de la méthode	R. Descartes
L'homme qui rit	V. Hugo
Madame Bovary	G. Flaubert
Les misérables, Tome I	V. Hugo
Oeuvres complètes	F. Villon
Le Rouge et le Noir	Stendhal
Les trois mousquetaires	A. Dumas
Voyage au centre de la terre	J. Verne

Table 2: The selection of compressed texts.

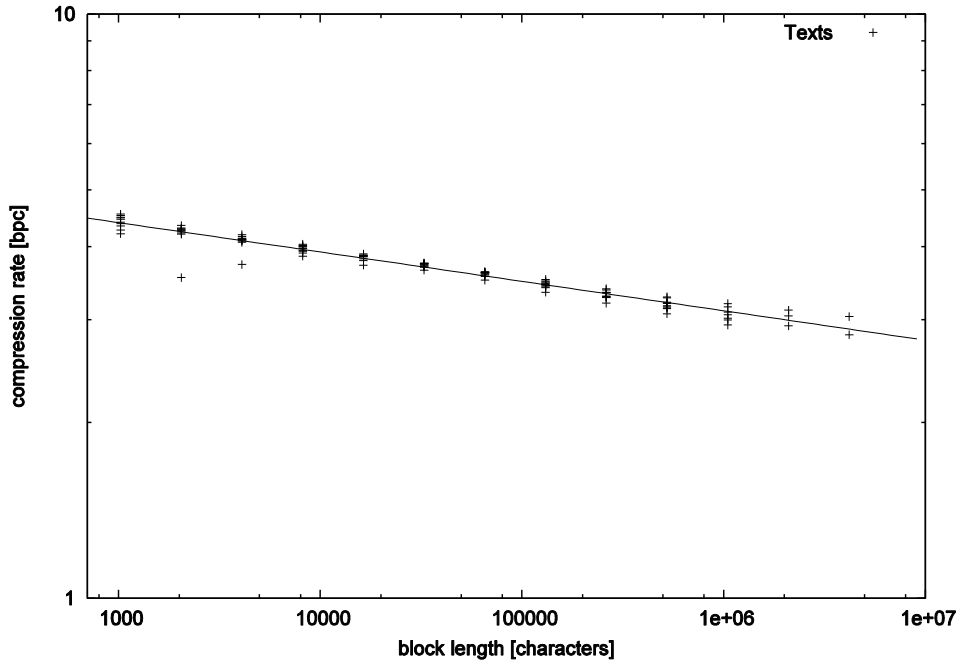

Figure 1: Compression rate vs. block length for the English texts. The solid line corresponds to regression (46).

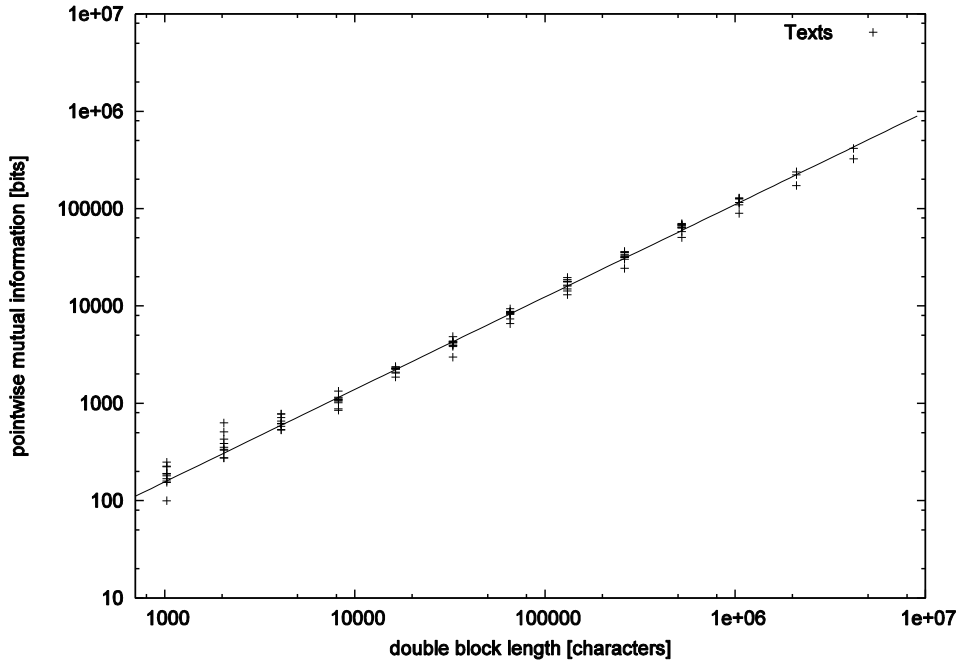

Figure 2: Mutual information vs. double block length for the English texts. The solid line corresponds to regression (49).

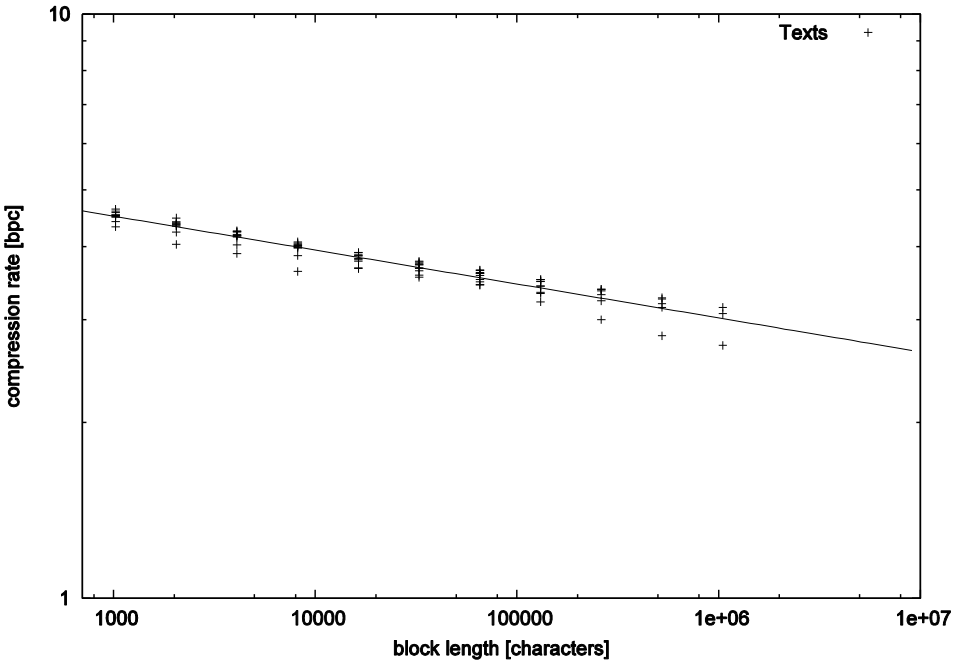

Figure 3: Compression rate vs. block length for the German texts. The solid line corresponds to regression (47).

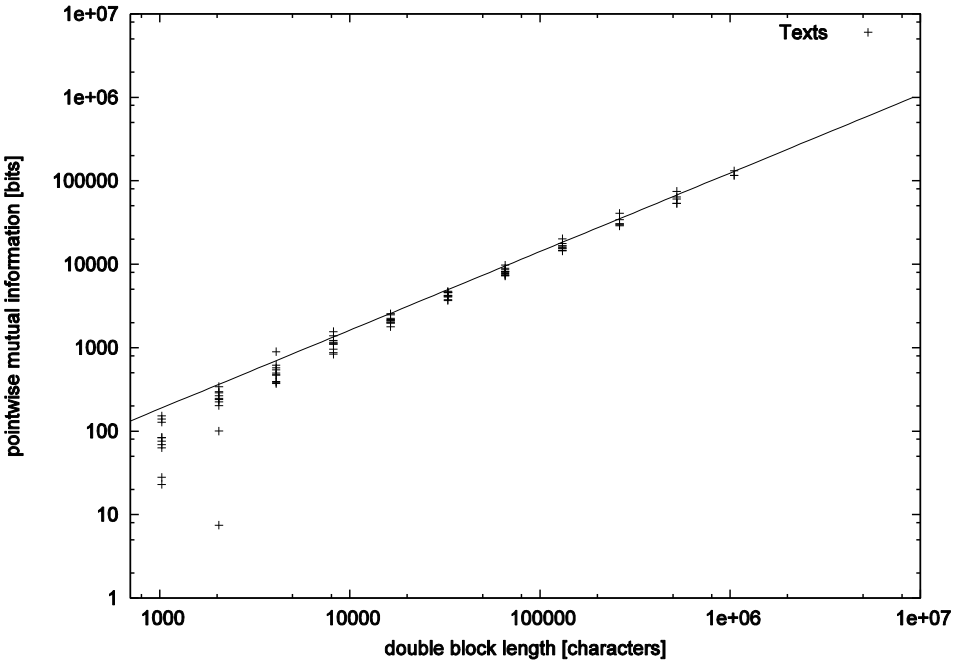

Figure 4: Mutual information vs. double block length for the German texts. The solid line corresponds to regression (50).

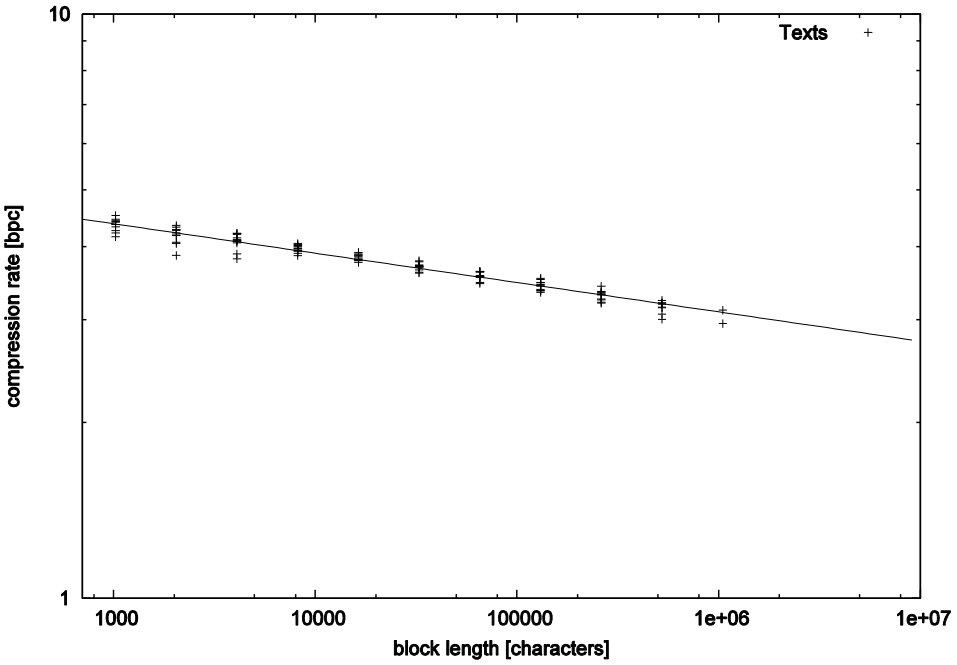

Figure 5: Compression rate vs. block length for the French texts. The solid line corresponds to regression (48).

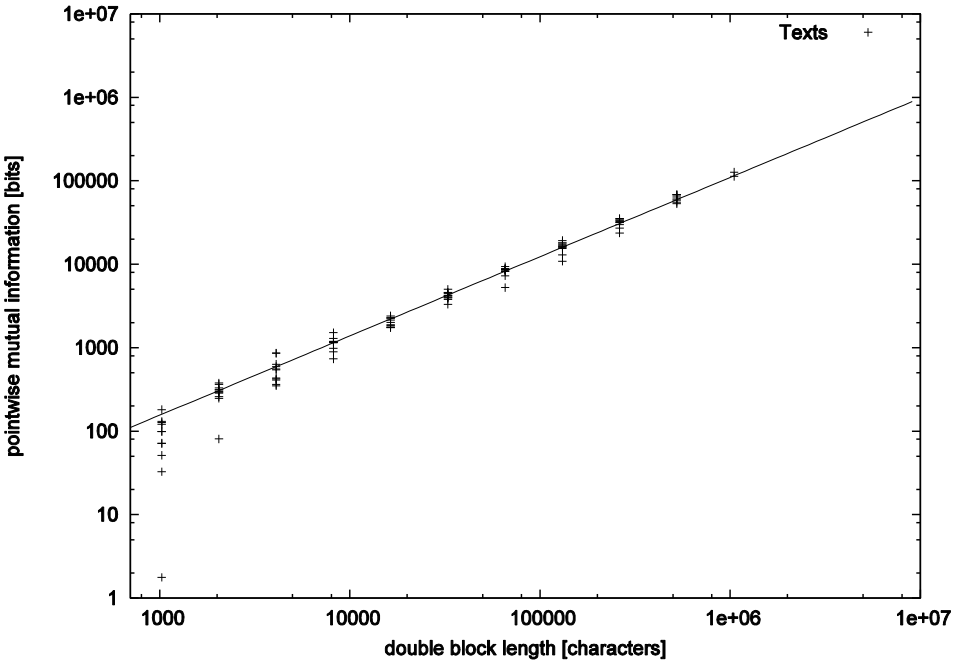

Figure 6: Mutual information vs. double block length for the French texts. The solid line corresponds to regression (51).