Abstract

This note is a brief introduction to theoretical and experimental results concerning Hilberg’s conjecture, a hypothesis about natural language. The aim of the text is to provide a short guide to the literature.

1 What is Hilberg’s conjecture?

In the early days of information theory, Shannon (1951) published estimates of conditional entropy for printed English. A few decades later, Hilberg (1990) replotted these estimates in doubly logarithmic scale and noticed that the entropy $H(n)$ of $n$ consecutive letters of text grows like

$$H(n) \approx Bn^\beta + hn,$$

(1)

where $\beta$ is close to 0.5 and the entropy rate $h$ equals 0, cf. Cover and Thomas (2006); Crutchfield and Feldman (2003). Although Shannon’s data points extended only to $n \leq 100$ letters, Hilberg supposed that relationship (1) with $h = 0$ holds for any natural language and is true for much larger $n$, being the length of a random text or even longer. This hypothesis will be called the original Hilberg conjecture (Dębowski, 2014b).

Hilberg’s conjecture corroborates and strengthens Zipf’s preformal insight that texts produced by humans diverge from both pure randomness and pure determinism (Zipf, 1965, p. 187), albeit they are in a sense both partly random ($H(n) > 0$) and asymptotically deterministic ($h = 0$) (Dębowski, 2014b). Condition $h = 0$ is incompatible with the hypothesis of constant conditional entropy, recently proposed by cognitive scientists and critically examined by Ferrer-i-Cancho et al. (2013). Moreover, $h = 0$ implies that texts in natural language are asymptotically infinitely compressible. If relationship (1) is true with $h = 0$, we need an explanation why texts cannot be fantastically well compressed by modern text compressors, such as the Lempel-Ziv code (Cover and Thomas, 2006). We will address this issue answering Question 3.

However, if we do not believe in asymptotic determinism of human utterances, we may still consider relationship (1) where $h = 0$ does not hold necessarily. Such relationship will be called the relaxed Hilberg conjecture. The relaxed Hilberg conjecture is equivalent to the statement that the mutual information
between two adjacent blocks of text of length \( n \) is proportional to \( n^\beta \) (Dębowski, 2015a). As we will see in the answer to Question 5, this statement can be related to the celebrated Zipf and Herdan laws in quantitative linguistics. Moreover, the relaxed Hilberg conjecture would be a property that distinguishes natural language from \( k \)-parameter sources, for which the mutual information between two adjacent blocks of length \( n \) is proportional to \( \log n \) (Grünwald, 2007).

2 Do there exist stochastic processes that satisfy Hilberg’s conjecture?

Let us observe that it is not trivial that there exist stochastic processes which satisfy the power-law growth of entropy or mutual information. From this point of view, Hilberg’s conjecture may be treated as an interesting mathematical hypothesis. Whereas we still ignore examples of processes that satisfy relation (1) with \( h = 0 \) (i.e., the original Hilberg’s conjecture), there are a few examples of processes that satisfy relation (1) with \( h > 0 \) (i.e., the relaxed Hilberg’s conjecture). These examples include so-called Santa Fe processes (Dębowski, 2009, 2010, 2011a, 2012a, 2014e) and certain hidden Markov processes (Dębowski, 2014a). In particular, Santa Fe processes admit a certain natural, though partly idealized, linguistic interpretation, which we will describe in the answer to Question 5. We suppose that also some processes that satisfy the original Hilberg’s conjecture can be given an interesting linguistic interpretation, connected with the idea of memetic evolution (Dębowski, 2014b).

3 Why texts cannot be so well compressed as apparently suggested by the original Hilberg conjecture even if this hypothesis is true?

This question is very important. It is known that there exist universal codes, whose compression rates equal asymptotically the entropy rate for any stationary process. If relation (1) holds with \( h = 0 \), the respective stochastic process can be compressed with the compression rate tending to zero asymptotically. Shields (1993) has shown, however, that the rate of convergence of the compression rate for any universal code can be arbitrarily slow. For a few modern text compressors, such as the Lempel-Ziv code (Ziv and Lempel, 1977) or grammar-based codes (Kieffer and Yang, 2000; Dębowski, 2011a), it can be further shown that the compression rate is greater than the inverse of maximal repetition (Dębowski, 2015b), i.e., the maximal length of a repeated substring in the text (de Luca, 1999; Shields, 1992, 1997). Moreover, it seems that there exist stochastic processes that satisfy both some version of Hilberg’s conjecture and hyperlogarithmic growth of maximal repetition (Dębowski, 2014b, 2015b). In that case Hilberg’s conjecture and the relatively slow decay of compression rate
would not be mutually incompatible. This question requires, however, further investigation.

4 What is the empirical support of Hilberg’s conjecture?

There is some empirical evidence in favor of Hilberg’s conjecture, although it is indirect or concerns only small block lengths. First, as mentioned in the answer to Question 1, a power-law growth of block entropy for $n \leq 100$ letters was observed by Hilberg (1990) for the estimates provided by Shannon (1951) using human subjects. Second, some other, purely computational but very heuristic estimates of block entropy also corroborate Hilberg’s conjecture (Ebeling and Nicolis, 1991, 1992; Ebeling and Pöschel, 1994). Third, the growth rate of maximal repetition agrees with a lower bound provided by a stronger version of Hilberg’s conjecture with $h = 0$ (Dębowski, 2012b, 2014b). Fourth, scaling of compression rate for universal codes is also compatible with Hilberg’s conjecture with $h = 0$, providing an upper bound $\beta < 0.9$ for $n \leq 10^9$ (Dębowski, 2013a,b, 2014c). Fifth, some estimators of entropy based on subword complexity also corroborate Hilberg’s conjecture with $h = 0$, providing a lower bound $\beta > 0.6$ for $n \leq 10$ (Dębowski, 2014d).

5 Is Hilberg’s conjecture related to Zipf’s law?

Zipf’s law is a celebrated empirical power-law concerning the rank-frequency distribution of words in texts in natural language (Zipf, 1965). In particular, this law implies Herdan’s law, i.e., the observation that the number of distinct words observed in a text of length $n$ is proportional to $n^\gamma$, where $\gamma$ varies in range between $0.5$ and $1$ (Kuraszkiewicz and Łukaszewicz, 1951; Guiraud, 1954; Herdan, 1964; Heaps, 1978). If Hilberg’s conjecture is true, even in its relaxed version ($h > 0$), then the popular explanation of Zipf’s and Herdan’s laws, formulated for a sequence of independent identically distributed variables (IID processes) by Miller (1957), is insufficient. Nevertheless, the relaxed Hilberg conjecture can be linked with two analogues of Herdan’s law on the level of set phrase frequency and of text semantics. Here, we will say that Herdan’s law holds for some kind of objects if the number of the respective object types in a text of length $n$ is proportional to $n^\gamma$.

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. Admissibly minimal grammar-based codes are certain 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, 2011a). Although we may suppose that these algorithms are computationally intractable (Charikar et al., 2005), nonterminal symbols of certain efficiently computable grammar-based codes which
resemble 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, 2011a, 2015b) showed that Herdan’s law for nonterminal symbols is a consequence of Hilberg’s conjecture.

The phenomenology of Herdan’s law for nonterminal symbols is a bit complicated. First, if an arbitrary stationary stochastic process satisfies the relaxed Hilberg conjecture \( h = 0 \) does not hold necessarily) then texts generated by this process satisfy Herdan’s law for nonterminal symbols with \( \gamma \geq \beta \) (Dębowski, 2011a). Second, if a stationary stochastic process satisfies the original Hilberg conjecture \( h = 0 \) and the hyperlogarithmic growth of maximal repetition holds as well then texts generated by this process satisfy Herdan’s law for nonterminal symbols with \( \gamma \) close to 1 (Dębowski, 2015b). In contrast, let us also note that Herdan’s law for nonterminal symbols holds also for IID processes if we apply certain grammar-based codes which are not admissibly minimal (Dębowski, 2007). The last effect seems however some artifact of suboptimal compression. The codes that approximate admissibly minimal grammar-based codes do not detect much structure in random data (Dębowski, 2007).

As for Herdan’s law pertaining to text semantics, let us note that Hilberg’s conjecture itself 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 (2011a) 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. Some examples of such processes are Santa Fe processes (Dębowski, 2011a, 2012a), mentioned in the answer to Question 2.

6 What is the history of Hilberg’s conjecture?

Hilberg (1990) published his hypothesis in German in a local telecommunications journal, which passed unnoticed by quantitative linguists and information theorists. The first interest it received came from physicists seeking to understand properties of complex systems (Ebeling and Nicolis, 1991, 1992; Ebeling and Pöschel, 1994; Bialek et al., 2001a,b; Crutchfield and Feldman, 2003). Later, some research in the mathematical foundations of Hilberg’s conjecture and its empirical verification was done by Dębowski. Firstly, the researcher was interested whether there are some links between Hilberg’s conjecture and
other laws of quantitative linguistics, such as Zipf’s, Herdan’s, and Menzerath’s laws (Dębowski, 2006, 2007). In parallel, Dębowski (2009) discovered Santa Fe processes. A few years later, Santa Fe processes were connected with the relaxed Hilberg conjecture and Herdan’s law in a rigorous way in Dębowski (2011a). Having established this result, the researcher wrote a few survey articles (Dębowski, 2011b, 2015a,b) and sought for experimental support of Hilberg’s conjecture (Dębowski, 2013a,b, 2014b,c,d). It appeared that Hilberg’s conjecture might be true in its original version ($h = 0$) with an additional constraint of hyperlogarithmic growth of maximal repetition (Dębowski, 2014b). For this reason, Dębowski is interested now in the construction of processes that satisfy the original Hilberg conjecture (1) with entropy rate $h = 0$ and hyperlogarithmic growth of maximal repetition. As stated in the answer to Question 2, this problem may be tightly connected with the idea of memetic evolution (Dębowski, 2014b) and failure of known universal codes to compress the data efficiently (Dębowski, 2015b).

References


