There Are Fewer Facts Than Words: Communication With A Growing Complexity

Łukasz Dębowski
Institute of Computer Science
Polish Academy of Sciences
01-248 Warszawa, Poland
ldebowsk@ipipan.waw.pl

Abstract

We present an impossibility result, called a theorem about facts and words, which pertains to a general communication system. The theorem states that the number of distinct words used in a finite text is roughly greater than the number of independent elementary persistent facts described in the same text. In particular, this theorem can be related to Zipf's law, power-law scaling of mutual information, and power-law-tailed learning curves. The assumptions of the theorem are: a finite alphabet, linear sequence of symbols, complexity that does not decrease in time, entropy rate that can be estimated, and finiteness of the inverse complexity rate.

1 Introduction

In several recent large-scale computational experiments in statistical language modeling, there were observed power-law tails of learning curves [Takahira et al., 2016, Hestness et al., 2017, Hahn and Futrell, 2019, Braverman et al., 2020, Kaplan et al., 2020, Henighan et al., 2020, Hernandez et al., 2021, Tanaka-Ishii, 2021]. Namely, the difference between the cross entropy rate of the statistical language model and the entropy rate of natural language decays as a power law with respect to the amount of training data. Equivalently, this is tantamount to a power-law growth of mutual information between increasing blocks of text—the first observation thereof attributed to Hilberg [1990], see also [Crutchfield and Feldman, 2003]. This power-law growth occurs for languages typologically as diverse as English, French, Russian, Chinese, Korean, and Japanese. Moreover, we observe a universal language-independent value of the power-law exponent: the mutual information between two blocks of length n is proportional to $n^{0.8}$ [Takahira et al., 2016, Tanaka-Ishii, 2021].

In this paper, we would like to advertise some mathematical theory of this phenomenon, covering its potential causes and effects. We have been developing this theory for several years. Our results were resumed in the recently published book [Dębowski, 2021a] and the subsequent article [Dębowski, 2021b]. This paper supplies an updated brief overview for a venue of machine learning. The novel thing is a simple generalization to non-stationary processes with a growing complexity.

The basic theory of power-law-tailed learning curves can be simply stated as furnishing the proof of a general statement of form:

The number of distinct words used in a finite text is roughly greater than the number of independent elementary persistent facts described in this text.

We call this sort of a statement a theorem about facts and words. In fact, the theorems about facts and words come into a few distinct flavors and can be proved relatively easily provided a certain attention is paid to the formal understanding of the concepts of a fact and of a word.

Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

The theorems about facts and words can be regarded as an impossibility result that pertains to a general communication system. Simply speaking, one cannot communicate about a certain amount of independent facts in a repetitive fashion without effectively using at least as many distinct words. This result seems paradoxical since we might think that combining words we may express many more independent facts. Moreover, the above result links information theory and discrete stochastic processes with linguistics, semantics, and cognition. In principle, it applies to any kind of a communication system consisting of a finite number of discrete signs stringed into longer messages. Besides natural language, some obvious examples are computer programs, DNA, and music.

In particular, the statements and the proofs of the theorems about facts and words combine:

- Zipf's and Herdan-Heaps' laws for word frequency distributions [Zipf, 1935, Mandelbrot, 1954, Guiraud, 1954, Herdan, 1964, Heaps, 1978],
- universal coding based on grammars [de Marcken, 1996, Kieffer and Yang, 2000, Charikar et al., 2005] and on normalized maximum likelihood [Shtarkov, 1987, Ryabko, 1988, 2008],
- consistent (hidden) Markov order estimators [Merhav et al., 1989, Ziv and Merhav, 1992],
- the concept of infinite excess entropy [Hilberg, 1990, Ebeling and Nicolis, 1991, Ebeling and Pöschel, 1994, Bialek et al., 2001, Crutchfield and Feldman, 2003],
- the ergodic theorem and the ergodic decomposition [Birkhoff, 1932, Rokhlin, 1962, Gray and Davisson, 1974], and
- Kolmogorov complexity and algorithmic randomness [Kolmogorov, 1965, Martin-Löf, 1966, Li and Vitányi, 2008].

As we can see, there are many interacting mathematical concepts. There are also many open problems in the surrounding theory. In the following, we present some particular version of a theorem about facts and words, which pertains to algorithmic randomness [Dębowski, 2021a] and consistent Markov order estimation [Dębowski, 2021b]. An informal discussion is relegated to Appendix A.

2 Preliminaries

Consider a string $x_j^k := (x_j, x_{j+1}, ..., x_k)$ over a countable alphabet. Its prefix-free Kolmogorov complexity is denoted $K(x_j^k)$ [Li and Vitányi, 2008]. The algorithmic mutual information between strings u and v is $J(u, v) := K(u) + K(v) - K(u, v)$. The expectation of random variable X is denoted $\mathbf{E}\, X$. The Shannon entropy of X is $H(X) := \mathbf{E}\,[-\log P(X)]$ [Cover and Thomas, 2006].

As in [Dębowski, 2021a, Definition 8.1], we will use the Hilberg exponent of a sequence, defined as

$$\operatorname*{hilb}_{n\to\infty} S(n) := \left[\limsup_{n\to\infty} \frac{\log S(n)}{\log n}\right]_+, \quad r_+ := r\,\mathbf{1}\{r \ge 0\}, \tag{1}$$

so that $\operatorname{hilb}_{n\to\infty} n^\beta = \beta$ for $\beta \ge 0$. We recall that if limit $s = \lim_{n\to\infty} S(n)/n$ exists then

$$\operatorname*{hilb}_{n\to\infty} [S(n) - sn] \le \operatorname*{hilb}_{n\to\infty} [2S(n) - S(2n)]\,. \tag{2}$$

For a discrete one-sided stochastic process $(X_i)_{i\in\mathbb{N}}$, we consider conditions:

(A) The complexity rate $h = \lim\limits_{n\to\infty} \mathbf{E}\, K(X_1^n)/n$ exists and $\operatorname*{hilb}\limits_{n\to\infty} [hn - H(X_1^n)] = 0$.

(B) The complexity does not decrease in time: $\mathbf{E}\, K(X_1^n) \le \mathbf{E}\, K(X_{n+1}^{2n})$.

(C) The inverse complexity rate is finite, $H := \limsup\limits_{n\to\infty} \mathbf{E}\, \dfrac{n}{K(X_1^n)} < \infty$ (thus $h > 0$ for (A)).

(D) The alphabet is finite: $X_i : \Omega \to \{a_1, a_2, ..., a_D\}$, where $D \in \mathbb{N}$.

In particular, conditions (A) and (B) are satisfied by any stationary process that satisfies (D).

For conditions (A) and (B), we also obtain that the power-law rate of redundancy is dominated by the power-law rate of mutual information,

$$\operatorname*{hilb}_{n\to\infty} [\mathbf{E}\, K(X_1^n) - hn] \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^n). \tag{3}$$

Condition $\operatorname*{hilb}\limits_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^n) > 0$ is called the Hilberg condition, after Hilberg [1990].

3 Facts

The concept of independent elementary persistent facts can be most easily understood on the example of a certain stationary ergodic process over a countably infinite alphabet called a Santa Fe process [Dębowski, 2011]. Let $(K_i)_{i\in\mathbb{N}}$ be an IID process in natural numbers with the Zipfian marginal distribution

$$P(K_i = k) = \frac{k^{-\alpha}}{\zeta(\alpha)}, \tag{4}$$

where $k \in \mathbb{N}$, $\alpha > 1$ is a fixed parameter, and $\zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}$ is the Riemann zeta function. Distribution (4) is a formal model of Zipf's law from quantitative linguistics [Zipf, 1935, Mandelbrot, 1954]. Moreover, let $(z_k)_{k\in\mathbb{N}}$ be an algorithmically random sequence, i.e., a sequence of particular (= fixed) bits (= coin flips) such that the Kolmogorov complexity of any string z_1^k is the highest possible, $K(z_1^k) \ge k - c$ for a certain constant $c < \infty$ and all lengths $k \in \mathbb{N}$ [Li and Vitányi, 2008]. Then the Santa Fe process $(X_i)_{i\in\mathbb{N}}$ is a sequence of pairs

$$X_i = (K_i, z_{K_i}). \tag{5}$$

In the following, bits z_k will be called facts.

The Santa Fe process can be understood as a model of an infinite text that consists of random statements of form „the k-th fact equals z_k". Importantly, these statements are non-contradictory, namely, if statements X_i and X_j describe the same fact ($K_i = K_j$) then they assert the same value of this fact ($z_{K_i} = z_{K_j}$). Moreover, we observe that facts z_k are in some sense independent (the Kolmogorov complexity of their concatenation is the highest possible), elementary (they assume only two distinct values), and persistent (described faithfully at any time instant i).

We will say that a finite text x_1^n describes exactly first l facts of a fixed sequence $(z_k)_{k\in\mathbb{N}}$ by means of a computable function g if $l = U_g(x_1^n; z_1^\infty) - 1$, where

$$U_g(x_1^n; z_1^\infty) := \min\left\{k \in \mathbb{N} : g(k, x_1^n) \ne z_k\right\}. \tag{6}$$

For a Santa Fe process we can easily construct a function g such that $U_g(X_1^n; z_1^\infty) = \min\left\{k \in \mathbb{N} : k \notin K_1^n\right\}$ by reading off the values of z_k for all $k \in K_1^n$. Hence it can be proved that the expected number of initial facts described by a random text X_1^n grows as a power law

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n; z_1^\infty) = 1/\alpha \in (0, 1). \tag{7}$$

Power laws (4) and (7) are related to each other as Zipf's law is related to Herdan-Heaps' law [Guiraud, 1954, Herdan, 1964, Heaps, 1978]. Processes $(X_i)_{i\in\mathbb{N}}$ such that $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n; z_1^\infty) > 0$ for a certain algorithmically random sequence $(z_k)_{k\in\mathbb{N}}$ and a certain computable function g are called perigraphic. All perigraphic processes have incomputable probability distributions. We notice that Santa Fe processes are perigraphic.

These results can be linked to the redundancy rate. As shown in [Dębowski, 2021a, Eqs. (8.117)–(8.119)], for any discrete stochastic process $(X_i)_{i\in\mathbb{N}}$, any algorithmically random sequence $(z_k)_{k\in\mathbb{N}}$, and any computable function g, we obtain inequality

$$\mathbf{E}\, U_g(X_1^n; z_1^\infty) - 6\log \mathbf{E}\, U_g(X_1^n; z_1^\infty) - c_g \le \sup_{k\in\mathbb{N}} \mathbf{E}\, J(X_1^n; z_1^k) \le \mathbf{E}\, K(X_1^n) - H(X_1^n). \tag{8}$$

where constant $c_g < \infty$ depends on function g. Hence, the power-law rate of the number of facts is dominated by the power-law rate of redundancy,

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n; z_1^\infty) \le \operatorname*{hilb}_{n\to\infty} \left[\mathbf{E}\, K(X_1^n) - hn\right] \tag{9}$$

if condition (A) holds. The above inequality can be chained with inequality (3).

4 Words

In the remaining move, we will relate the algorithmic mutual information to the number of words used in a given text. This can be done in several ways. One way, pursued by Dębowski [2011], is to apply the minimal grammar-based coding [Kieffer and Yang, 2000, Charikar et al., 2005], which applies a mathematical concept of a word that resembles words in a linguistic sense [de Marcken, 1996] but this method is a bit lengthy to describe formally. Therefore, here, we will apply a different approach which is based on consistent Markov order estimation [Merhav et al., 1989, Ziv and Merhav, 1992]—that approach was pursued by Dębowski [2021b].

There is a function called subword complexity that counts how many distinct substrings of a given length there are in a given string, namely,

$$V(k|x_1^n) := \#\left\{x_{i+1}^{i+k} : 0 \le i \le n-k\right\}. \tag{10}$$

Function $V(k|x_1^n)$ will be a proxy for the number of distinct words in text x_1^n. The only problem is to choose a motivated length k of the substrings. In principle, this k may depend on string x_1^n.

Quite a natural choice of k is the estimator of the Markov order of the process defined as

$$M(x_1^n) := \min\left\{k \ge 0 : -\log L_k(x_1^n) \le K(x_1^n)\right\}, \tag{11}$$

where $K(x_1^n)$ is the Kolmogorov complexity and $L_k(x_1^n)$ is the maximum likelihood of order k,

$$L_k(x_1^n) := \max_Q \prod_{i=k+1}^{n} Q(x_i|x_{i-k}^{i-1}), \qquad Q(x_i|x_{i-k}^{i-1}) \ge 0, \qquad \sum_{x_i} Q(x_i|x_{i-k}^{i-1}) = 1. \tag{12}$$

Function $M(x_1^n)$ is a strongly consistent and asymptotically unbiased estimator of the Markov order. Namely, for any stationary ergodic process $(X_i)_{i\in\mathbb{N}}$ over a finite alphabet we have

$$\lim_{n\to\infty} M(X_1^n) = M \text{ almost surely}, \qquad \lim_{n\to\infty} \mathbf{E}\, M(X_1^n) = M. \tag{13}$$

where we denote the Markov order of the process as

$$M := \inf\left\{k \ge 0 : P(X_{k+1}^n|X_1^k) = \prod_{i=k+1}^{n} P(X_i|X_{i-k}^{i-1}) \text{ for all } n > k\right\}, \quad \inf\emptyset := \infty. \tag{14}$$

Let us write succinctly the number of substrings of the supposedly optimal length as

$$V(x_1^n) := V(M(x_1^n)|x_1^n). \tag{15}$$

Then, using the universal code by Ryabko [1988], we can prove inequality

$$J(x_1^n; x_{n+1}^{2n}) \le 2\left[DV(x_1^{2n}) + \frac{4n\log D}{K(x_1^{2n})} + c_1\right](\log n + c_2), \tag{16}$$

where D is the cardinality of the alphabet and $c_i < \infty$ are certain small constants [Dębowski, 2020, Theorems 11–12]. Applying Hilberg exponents and expectations, we obtain that the power-law rate of mutual information is dominated by the power-law rate of the number of words,

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^n) \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, V(X_1^n), \tag{17}$$

if conditions (C) and (D) are satisfied. The above inequality can be chained with inequalities (3) and (9). The asymptotic power law $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, V(X_1^n) > 0$ resembles Herdan-Heaps' law for words in the linguistic sense [Guiraud, 1954, Herdan, 1964, Heaps, 1978].

5 Conclusion

Chaining inequalities (3), (9), and (17) under conditions (A)–(D), we obtain the sandwich bound

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n; z_1^\infty) \le \operatorname*{hilb}_{n\to\infty} [\mathbf{E}\, K(X_1^n) - hn] \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^n) \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, V(X_1^n), \tag{18}$$

which yields a formal statement of a certain theorem about facts and words. In particular, we can infer from inequalities (18) that no perigraphic process can be a Markov process, i.e., a process with a finite Markov order, $M < \infty$. These two classes of processes are disjoint.

Of course, the above inequalities rest on a repeated use of Kolmogorov complexity. Therefore they are ineffective in some sense. We note that there are other statements of theorems about facts and words that apply effective notions. However, they are more complicated to formulate. See also Appendix A in this paper for some informal discussion of our formal assumptions.

The theorems about facts and words, as an impossibility result, raise questions about their applicability to empirical data as well as questions about further examples of perigraphic processes that are more complex or more realistic than Santa Fe processes. Some of these questions were addressed or stated as open problems in book [Dębowski, 2021a] and article [Dębowski, 2021b].

References

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

G. D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America*, 17:656–660, 1932.

M. Braverman, X. Chen, S. M. Kakade, K. Narasimhan, C. Zhang, and Y. Zhang. Calibration, entropy rates, and memory in language models. In *2020 International Conference on Machine Learning (ICML)*, 2020.

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

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

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

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

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

Ł. Dębowski. On a class of Markov order estimators based on PPM and other universal codes. `https://arxiv.org/abs/2003.04754`, 2020.

Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. New York: Wiley & Sons, 2021a.

Ł. Dębowski. A refutation of finite-state language models through Zipf's law for factual knowledge. *Entropy*, 23:1148, 2021b.

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

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

R. M. Gray and L. D. Davisson. Source coding theorems without the ergodic assumption. *IEEE Transactions on Information Theory*, 20:502–516, 1974.

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

M. Hahn and R. Futrell. Estimating predictive rate-distortion curves via neural variational inference. *Entropy*, 21:640, 2019.

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

T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. `https://arxiv.org/abs/2010.14701`, 2020.

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

D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish. Scaling laws for transfer. `https://arxiv.org/abs/2102.01293`, 2021.

J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. `https://arxiv.org/abs/1712.00409`, 2017.

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

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. `https://arxiv.org/abs/2001.08361`, 2020.

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

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

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

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

P. Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

N. Merhav, M. Gutman, and J. Ziv. On the estimation of the order of a Markov chain and universal data compression. *IEEE Transactions on Information Theory*, 35(5):1014–1019, 1989.

G. A. Miller. Some effects of intermittent silence. *American Journal of Psychology*, 70:311–314, 1957.

V. A. Rokhlin. On the fundamental ideas of measure theory. *American Mathematical Society Translations, series 1*, 10:1–54, 1962.

B. Ryabko. Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series. In *2008 IEEE Information Theory Workshop, Porto*, pages 271–275. Institute of Electrical and Electronics Engineers, 2008.

B. Y. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24(2):87–96, 1988.

Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(2):3–17, 1987.

R. Takahira, K. Tanaka-Ishii, and Ł. Dębowski. Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364, 2016.

K. Tanaka-Ishii. *Statistical Universals of Language: Mathematical Chance vs. Human Choice*. New York: Springer, 2021.

G. K. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston: Houghton Mifflin, 1935.

J. Ziv and N. Merhav. Estimating the number of states of a finite-state source. *IEEE Transactions on Information Theory*, 38(1):61–65, 1992.

A An informal discussion of the adopted mathematical model

A certain problem about the presented theorems about facts and words is that they appear quite abstract. This abstraction is an advantage from a mathematical point of view since it allows to apply the same result to very different systems—for example to discuss the internal complexity of communication in the field of mathematics. However, any abstraction also begs for some concrete examples of applications that would explain the intuitions standing behind the adopted formal models. As a mathematician, we prefered to present the hard results in the main matter of this paper, whereas we relegate the intuitions to the present appendix—written upon the request of the reviewers.

First of all, one should be aware that theorems about facts and words are an asymptotic result—for the aesthetic virtue of applying the power law rate exponents, called succinctly Hilberg exponents. The discussed inequalities for Hilberg exponents stem, however, from two non-asymptotic inequalities (8) and (16). Thus, if we could somehow make an informed (and necessarily fallible) guess about the Kolmogorov complexity of a particular text in natural language or another communication system then we might expect a sort of a theorem about facts and words also for finite amounts of empirical data. Assumptions such as perfect stationarity or operations like the expectation were applied to get rid of some relatively small deviations from aesthetically appealing general trends.

Another important remark, theorems about facts and words are just an impossibility result: In a certain precise sense, there are always roughly fewer facts described than words used. For Santa Fe processes, these two quantities are of a similar order. For other processes or communication systems, the number of facts can be significantly smaller than the number of words. If it were so for natural language, it would be extremely interesting. In such a case, language communication would be much less complex than suggested by the mere power-law growth of the lexicon. This hypothesis is so counterintuitive that it asks for a further consideration.

From this point of view, it is advisable to bridge the adopted formal notions of a fact and a word with a more usual understanding of these concepts. Let us begin with the concept of a fact. As noted by one reviewer, the representational content of a fact is irrelevant for our theorem. This makes the theorem both powerful from a mathematical point of view and quite difficult to digest by empirical researchers. However, asking what the facts are in their essence is not the most fortunate question to answer. It is more appropriate to ask what the distinctive features of facts are in our model, namely, how they behave.

Taking Santa Fe processes as a working mathematical model of facts, there are three basic properties of our facts. First, they are assumed to be binary variables, like bits, coin flips, or spins. Second, they are independent, either probabilistically or algorithmically, namely, their Kolmogorov complexity is the highest possible. Third, they are persistent, recurrent, or eternal, in the sense that the same fact is repeatedly described infinitely many times by the information source that generates the text.

The question whether such eternal and algorithmically random facts can be actually described at a power law rate by the totality of human culture is intellectually challenging but important. It concerns not only the approximate or real randomness of cultural conventions but also their origins in the physical sources of noise. The question cannot be honestly answered unless we know exactly how much randomness there is in the physical world. Fortunately, in the spirit of algorithmic information theory, we may ultimately falsify some class of wrong answers—given enough computation time.

Moreover, it may be helpful to imagine our sequence of facts as a sort of an unknown real parameter of an information source. When we observe or learn this source, we learn particular binary digits or facts of this parameter at a rate that is specific for the information source. For Santa Fe processes, this rate is given by the power law (7). For Bernoulli processes or finite-state sources, the same number of described facts grows only logarithmically, so the respective Hilberg exponent is zero. By contrast, for natural language, we cannot be sure how large it is. However—by the theorems about facts and words—we already know that the number of independent facts is upper bounded by the estimates of mutual information or by the number of distinct words, which grow roughly as $n^{0.8}$, where n is the length of the text [Takahira et al., 2016, Tanaka-Ishii, 2021].

In this way, we proceed smoothly to discuss the concept of a word. From the point of view of formal linguistics, the only rigorous definition of a word that makes sense is by enumeration of the lexicon. This can be successful only if the lexicon is finite or given by a finite formal grammar. Nevertheless, this definition cannot fully accommodate creativity and openness of the lexicon and of other language or culture conventions. Thus in order to count words, we need a certain operational definition of what we want to count in general.

One such way is to count strings of letters that appear empirically and are delimited by separators such as spaces or pauses. However, when we count such strings in a stream of probabilistically independent letters then we obtain a spurious power law [Mandelbrot, 1954, Miller, 1957]. We have a general intuition that a memoryless mechanism cannot be responsible for the power-law distribution of words in natural language. Thus we had better made words operational in a different way.

For example, we may use phrases defined by universal coding. Since there are many different universal codes of varying encoding rates for finite data, it is not clear which one we should use. Quite an interesting idea seems to apply grammar-based codes. These codes represent a text as a specific context-free grammar that produces the text as the only production. If we heuristically minimize the length of such a grammar then we obtain non-terminals that usually span across whole words [de Marcken, 1996]. That is, such minimal non-terminals can be considered a proxy for orthographic words. There is also a theorem about facts and words that applies to the number of minimal non-terminals [Dębowski, 2011]. The problem is that this theorem applies the global minimization of the grammar length, which is likely intractable [Charikar et al., 2005].

For this reason, we searched for yet another approach. In the main matter of this paper, we presented an operational proxy for words as overlapping strings of the constant length equal to the Markov order estimate computed for the considered text. This approach is also quite intuitive. Its advantage is that it does not lead to a spurious discovery of an unbounded vocabulary in Markovian sources. However, the Markov order estimate is usually smaller than the average length of a word for moderately sized texts. Thus the number of Markovian substrings is only an imperfect proxy for the number of words.

Thus there remains an open problem of finding an operational definition of a word that could be applied to any sort of symbolic data and would possess the following properties:

1. The number of distinct operational words should be theoretically lower bounded by the block mutual information—and hence by the number of facts. (Thus, a rich semantics implies a rich vocabulary.)
2. The number of distinct operational words should be theoretically upper bounded for a finite Markov order of the process. (Thus, a meager semantics implies a meager vocabulary.)
3. Parsing of the input text into operational words should be efficiently computable in a time close to linear in the length of the text.
4. The operational words should be similar in shape and in number to the orthographic words for natural language data.

We hope that such a satisfactory operational definition of a word exists.

Concluding this appendix, there seem to be a few somewhat different formal statements that fall under the umbrella of theorems about facts and words. In our work, we tried to identify a few of them but the topic has not been exhausted, in our opinion.