entropy

MDPI

Article

A Refutation of Finite-State Language Models through Zipf's Law for Factual Knowledge

Łukasz Dębowski

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

Abstract: We present a hypothetical argument against finite-state processes in statistical language modeling that is based on semantics rather than syntax. In this theoretical model, we suppose that the semantic properties of texts in a natural language could be approximately captured by a recently introduced concept of a perigraphic process. Perigraphic processes are a class of stochastic processes that satisfy a Zipf-law accumulation of a subset of factual knowledge, which is time-independent, compressed, and effectively inferrable from the process. We show that the classes of finite-state processes and of perigraphic processes are disjoint, and we present a new simple example of perigraphic processes over a finite alphabet called Oracle processes. The disjointness result makes use of the Hilberg condition, i.e., the almost sure power-law growth of algorithmic mutual information. Using a strongly consistent estimator of the number of hidden states, we show that finite-state processes do not satisfy the Hilberg condition whereas Oracle processes satisfy the Hilberg condition via the data-processing inequality. We discuss the relevance of these mathematical results for theoretical and computational linguistics.

Keywords: statistical language modeling; Zipf's law; Hilberg's hypothesis; algorithmic mutual information; hidden Markov processes; perigraphic processes

check for updates

Citation: Dębowski, Ł. A Refutation of Finite-State Language Models through Zipf's Law for Factual Knowledge. *Entropy* **2021**, *23*, 1148. https://doi.org/10.3390/e23091148

Academic Editors: Kathleen Currie Hall, Uriel Cohen Priva and Richard Futrell

Received: 16 June 2021
Accepted: 30 August 2021
Published: 1 September 2021

1. Introduction

The goal of this article is to show that finite-state statistical language models can be refuted using a hypothetical argument that is based on semantics rather than syntax. This semantic argument is rooted in recent theoretical research in information theory. Even if some hypotheses thereof do not pertain to natural language, we suppose that our reasoning may still be appealing enough for computational and theoretical linguistics and it points out interesting directions of future research. In the following, first, we sketch the historical context of our research line (Section 1.1) and, next, we describe the particular technical aims of this article (Section 1.2).

1.1. Historical and Conceptual Research Context

In the famous critique of Burrhus Skinner's book [1], Noam Chomsky refuted finite-state models for human language as implausible since they could not express context-free syntax with central embeddings of an unbounded depth [2–5]. In turn, this refutation produced doubt among linguists about whether information theory and statistical language modeling are relevant for language studies and stimulated a fast growth of purely formal linguistics [6]. Probabilities were relegated mostly to natural language engineering, where some completely new ideas were developed and gradually radiated back to linguistics. Focusing specifically on the innovations of language engineering, probabilistic finite-state models were initially applied for speech recognition [7,8] and part-of-speech tagging [9], to be followed by probabilistic context-free grammars for sentence parsing [10–12] and were replaced by long short-term memory (LSTM) neural networks [13], word embeddings [14], and transformers [15], which achieved an apparently human-like quality of text prediction and generation [16–18]. All of this progress is breath-taking, and language theories can not keep up with these technical achievements. We can be concerned with whether there is a fundamental statistical theory of language, for the successes of neural statistical language models suggest that the most accurate description of language is of a probabilistic nature. However, can there be a language theory more concise and more transparent than a neural network with millions or even billions of parameters?

Actually, we should entertain the idea that there is no finite theory of human language more seriously in the obvious and narrow sense that we constantly update the neural network wiring of our brains. What may exist is rather a universal language learning mechanism—though not necessarily exactly one proposed by Chomsky [19]—that is updated with the unbounded influx of stimuli and random drift. In particular, an important phenomenon that may not have caught enough attention in the formally oriented linguistic literature is the interaction between the language theories and the potential unboundedness of factual knowledge conveyed by means of language. It is an assumption of some linguistic theories that the description of the core language system can be sharply delineated from the factual knowledge expressed in texts. However, when we perform statistical language modeling for speech recognition or machine translation, we cannot afford to ignore factual knowledge. Taking factual knowledge into account is essential for a good performance of respective computer applications [20]. Statistical modeling of texts, called deceptively statistical language modeling, requires that we model not only language as a system but also things that are expressed in language, and these seem to come as a large number of rare events [21,22]. Under Zipf's law [23,24], roughly half of the vocabulary of a text are *hapax legomena*, i.e., words that appear only once. This skewness of distribution may also apply to concepts or facts.

In our opinion, the fields of computational and theoretical linguistics lack a corresponding baseline probabilistic model of an unbounded accumulation of factual knowledge; see also Bar-Hillel and Carnap [25] for some fairly old ideas with regard to linking formal semantics and information theory, and compare it with Claude Shannon's disregard for semantics in information theory [26]. Having such an idealized model, we could try to explain and better understand why certain kinds of language theories have to grow unboundedly as we have more and more data and why statistical language models have to be continually trained. We want to argue that such a model can be provided by accumulated developments in information theory and quantitative linguistics. The core idea is to operate with an idealized stochastic model of the distribution of factual knowledge in texts, known as perigraphic processes.

Roughly speaking, perigraphic processes introduced in [27], whereof simple examples are Santa Fe processes [28,29] and whereof some less trivial examples may be random hierarchical association (RHA) processes [30], are stationary stochastic processes for which effectively inferrable mentions of independent elementary facts are distributed according to the Zipfian power laws. By the Zipfian power laws, we collectively understand the Zipf–Mandelbrot law for the word rank-frequency distribution [23,24] and the Herdan–Heaps law for the growth of the number of word types [31–34], where the former implies the latter, cf. [35] and [36] (Section 1.3). To make our model mathematically precise, for each perigraphic process, we assume that elementary facts are bits (binary symbols) of a fixed algorithmically random sequence, i.e., an infinite sequence of bits in which the shortest description is the sequence itself [37,38], and there exists a computable function that allows us to ultimately infer elementary facts from any sufficiently long subsequence of the process [27,29].

In plain words, perigraphic processes define a model of factual knowledge that is infinite, time-independent, compressed losslessly as much as possible, and effectively described in random texts at a power-law rate. Namely, the number of initial bits of factual knowledge that are correctly described in a text of length n equals roughly n^{β^+}, where $\beta^+ \in (0,1)$ is a free parameter, such as in the Herdan–Heaps law for the growth of the number of word types. Each elementary fact, i.e., each bit of factual knowledge, is described infinitely often in the infinite random text generated by a perigraphic source, but the facts located earlier in the sequence of factual knowledge are described more frequently, roughly according to the Zipf–Mandelbrot law. The function that computes facts from finite sections of the infinite text, called the knowledge extractor, can be quite arbitrary, but within our model, we assume that it is computable. Making connections to natural language, we can regard the knowledge extractor as a mathematical model of a sort of language competence.

An important open problem in the theory of perigraphic processes is whether the mentioned power-law exponent β^+ can be consistently estimated. Namely, the open question is whether there exists a computable function of finite texts that returns some estimates converging to β^+ almost surely. If such a function exists, then we could empirically verify whether natural language is a perigraphic process or, rather, to what degree it resembles a perigraphic process. However, regardless of the uncertain success of this research project, we stress that there are some other measurable side effects of perigraphicness. Namely, when an algorithmically random sequence is described repetitively in texts, then the algorithmic mutual information between the previous text and the forthcoming text must grow unboundedly as we increase the text length. Moreover, since we can estimate the algorithmic mutual information to a certain extent, e.g., using universal codes [39], this growth effect should be approximately empirically measurable. Thus, for any stationary streams of data that do not satisfy a power-law growth of computable estimates of algorithmic mutual information, we can effectively tell that they are not perigraphic.

In this way, we proceed to another important topic, namely, Hilberg's hypothesis. The proponent of this hypothesis was the German engineer Wolfgang Hilberg [40], who replotted the famous guessing estimates of conditional entropy for English by Claude Shannon [41] in the doubly logarithmic scale. In the replotted graph, Hilberg's eyes saw a straightish line, meaning a hypothetical power-law growth of block entropy. Hilberg's hypothesis of the power-law growth of entropy has dwelled on the peripheries of mainstream language sciences, where it gradually matured. The idea was first seriously considered by physicists [42–45], who reformulated Hilberg's hypothesis as a power-law growth for block Shannon mutual information—getting rid of the dubious asymptotic determinism of the statistical language model. We took up the topic in 2000, and we devoted to it twenty years of mathematical research resumed in the book in [36]; see also a more empirically oriented monograph by Tanaka-Ishii [46]. In parallel, suggestive upper bounds for the power-law growth of mutual information and partial evidence for infinite excess entropy, i.e., the divergent mutual information between the past and the future [45], were provided by several independent large-scale computational experiments [47–53]. For languages as diverse as English, French, Russian, Chinese, Korean, and Japanese, the upper bounds for mutual information grow universally as roughly $n^{0.8}$ [46,47]. Thus, all of these languages seem equally hard to learn to predict.

The most important achievement of our mathematical theory of Hilberg's hypothesis are so-called theorems about facts and words, cf. [27,29,54] and [36] (Section 8.4), that connect this hypothesis with Zipfian power laws for words and for bits of the compressed factual knowledge called facts. The theorems about facts and words make a quantitative connection between the unbounded accumulation of factual knowledge (measured by the number of distinct inferrable facts) and the unbounded growth of some primitive linguistic theories (measured by the total length of distinct discernible words, cf. [55]). According to the theorems about facts and words,

- The expected number of distinct binary facts that can be learned from a finite text is roughly less than the mutual information between two halves of the text.
- The mutual information between two halves of the text is roughly less than the expected total length of distinct words that can be found in the text.

These statements pertain to texts generated by arbitrary stationary stochastic sources over a finite alphabet. They are purely mathematical theorems but with a linguistic twist. The rough inequalities are understood as precise inequalities of so-called Hilberg exponents. Not only facts but also words are understood as effectively inferrable. Namely, words can be detected in the text via the prediction by a partial matching (PPM) universal code [56–59] or via shortest grammar-based compression [29,60,61], which roughly agrees with the orthographic parsing of texts into words for human languages [55].

Using the theorems about facts and words, perigraphic processes not only satisfy Hilberg's hypothesis but also satisfy Zipfian power laws for words. This result shows not only that Hilberg's hypothesis can be connected on a theoretical level with some abstract semantics but also that the abstract semantic properties of a random text imply double articulation of the text, i.e., discreteness of words, which is the rudiment of structures studied by linguists. In consequence, we suppose that perigraphic processes are a promising class of abstract statistical language models in which linguistically interpretable properties can be investigated deductively and partly motivated empirically.

1.2. Aims and Organization of the Article

Having made such a long historical and conceptual introduction, let us state the particular aim of this article. Continuing our line of research, in this article, we solve open problem no. 4 from the conclusion of the book in [36]. The conjecture was that no finite-state process is perigraphic, even if the finite-state process has uncomputable transition probabilities. We show that this proposition holds indeed, which sheds another beam of light onto the debate between Skinner and Chomsky. In the very beginning of our acquaintance with Hilberg's hypothesis, we realized that it can be also used for refuting finite-state models for human language. The reason for this is that excess entropy, i.e., the mutual information between the infinite past and the infinite future [45], is finite for finite-state models by the data-processing inequality, whereas it is obviously infinite if Hilberg's hypothesis is satisfied. This statement holds straightforwardly in the framework of the Shannon information theory, which assumes that we have a definite statistical language model, i.e., a distribution of a stochastic process.

However, a large part of our later theorizing dealt with technical and conceptual problems around the ergodic decomposition of the statistical language model [62,63]. To make the long story short, it is natural to assume that the subjective probabilities in our minds contain certain priors and, hence, that they are computable but nonergodic. By contrast, the resulting relative frequencies in the unbounded stream of our speech are typical ergodic components of subjective probabilities, and hence, they are ergodic but uncomputable. This distinction results in two complementary versions of an idealized statistical theory of language: one seen from the perspective of a language user and another seen through the lens of a fixed generated text. Whereas from the language user's perspective Shannon information theory seems sufficient, from the perspective of a particular text, we need to apply algorithmic information theory [38,64,65]. Not everything that can proven easily in Shannon information theory can be proven as easily in algorithmic information theory. This is exactly the case with refuting finite-state language models.

Thus, as the main goal of this article, we show that Hilberg's hypothesis and the accumulation of factual knowledge at a power-law rate are incompatible with finite-state models also in the algorithmic framework. This solves open problem no. 4 from the conclusion of the book in [36], i.e., we show that no perigraphic process can be a finite-state process—even if we admit uncomputable transition probabilities. To deal with these issues, we apply techniques inspired by the aforementioned theorems about facts and words—manifested most prominently in the proofs of Theorems 5 and 7. In this way, we provide a complete argument against finite-state models, which is orthogonal to the Chomskyan argument, since it is more related to an idealized model of semantics than to an idealized model of syntax.

As a secondary goal of this article, we also present a simple example of a perigraphic process over a finite alphabet, called Oracle processes. The first constructed examples of a perigraphic process are the Santa Fe processes [28,29], which are even simpler but constitute processes over a countably infinite alphabet. In Reference [66] (see also the book in [36]), we have provided quite a complicated encoding of Santa Fe processes in a finite alphabet. By contrast, Oracle processes constitute a much simpler encoding, of which the construction applies the monkey-typing explanation of Zipf's law by Benoît Mandelbrot and George Miller [24,67].

Ironically, the composition of this article follows a central embedding: We delve gradually into mathematical considerations and eventually emerge from them to come back to linguistic interpretations towards the end. To be concrete, we begin with recalling some established classes and examples of discrete stochastic processes in Section 2. Subsequently, we discuss Hilberg's hypothesis at length in Section 3. In Section 4, we show that no finite-state process satisfies Hilberg's hypothesis. By contrast, in Section 5, we discuss that all perigraphic processes satisfy Hilberg's hypothesis. To exhibit a simpler example of such processes over a finite alphabet, we construct so-called Oracle processes in Section 6. In Section 7, we discuss the relevance of these mathematical results for theoretical and computational linguistics. Section 8 concludes the article. All proofs of theorems are deferred to Appendix A. Salient mentions of important formal concepts are typeset in boldface. Intentionally, we try to write this article in a more popular fashion than an average mathematical paper to reach some audience in language research.

2. Some Classes of Processes

To provide an introduction for readers who are less versed in measure-theoretic probability, we begin with discussing some basic classes and examples of discrete stochastic processes. We may imagine those as a progression of rudimentary statistical language models, i.e., the conditional probability distributions that predict the next letter or the next word given a sequence of previous ones. Since we work with discrete distributions, we can avoid measure theory in the beginning but the reader should be aware that it exists and that it takes care of what is not explicitly explained or even noticed during the first reading, cf. [36] (Chapters 2–4) and [68]—especially in the treatment of stationary and ergodic processes.

Within our framework, stochastic processes are infinite sequences of discrete random variables, indexed by natural numbers ($\mathbb{N}$) or by integers ($\mathbb{Z}$). The linguistic interpretation is that the indices point to specific symbols in an idealized random text or a corpus of texts, which extends toward an infinite future ($\mathbb{N}$) or towards both an infinite future and an infinite past ($\mathbb{Z}$). To specify the probability measure on such infinite sequences of symbols, it is enough to specify all finite-dimensional distributions—or conditional distributions of a single symbol given any sequence of previous symbols. In the following, notation x_j^k denotes string $x_j x_{j+1} ... x_k$ of particular symbols. The same convention applies to blocks of random variables $X_j^k := X_j X_{j+1} ... X_k$, which are, technically speaking, functions from elementary events to strings of particular symbols.

Let us proceed to defining some classes of discrete processes. Process $(X_i)_{i \in \mathbb{N}}$ over a countable alphabet $\mathbb{X}$ is called a **Markov process** when the conditional probability of the next symbol x_i depends only on the directly preceding symbol x_{i-1}, i.e.,

$$P(X_1 = x_1) = \pi(x_1), \tag{1}$$

$$P(X_i = x_i | X_1^{i-1} = x_1^{i-1}) = \sigma(x_i | x_{i-1}) \tag{2}$$

for certain functions $\pi : \mathbb{X} \to [0, 1]$ (initial distribution) and $\sigma : \mathbb{X} \times \mathbb{X} \to [0, 1]$ (transition matrix). A Markov process such that $\sigma(x_i | x_{i-1}) = \pi(x_i)$ is called an **IID (independent identically distributed) process**. Whereas IID processes are central to the theory of mathematical statistics, Markov processes exhibit some rudimentary dependence—exactly only on the directly preceding observation—and were in fact proposed by Andrey Markov [69,70] as some primitive statistical language models.

By contrast, process $(X_i)_{i\in\mathbb{N}}$ over a countable alphabet $\mathbb{X}$ is called a **hidden Markov process** with respect to a Markov process $(Y_i)_{i\in\mathbb{N}}$ over a countable alphabet $\mathbb{Y}$ when the conditional probability of the next symbol x_i depends only on the hidden state y_i, i.e.,

$$P(X_i = x_i | Y_1^i = y_1^i, X_1^{i-1} = x_1^{i-1}) = \varepsilon(x_i|y_i) \tag{3}$$

for a certain function $\varepsilon : \mathbb{X} \times \mathbb{Y} \to [0,1]$ (emission matrix). Elements of $\mathbb{X}$ are called symbols, whereas elements of $\mathbb{Y}$ are called (hidden) states. Hidden Markov processes were the state-of-the-art of statistical language modeling for speech recognition and part-of-speech tagging in the 1990s [7–9,12]. As we can see, the dependence between a symbol x_i and its past is bottlenecked by the hidden state y_i, which in turn is a result of a Markov process. The modeling power of hidden Markov processes depends on what we assume about the hidden states and about their structure. When these hidden states are closer to mental states, we may suppose that the resulting process of emitted symbols is closer to human utterances. In practice, we consider much simpler models. In particular, a **finite-state process** is such a hidden Markov process that the set of states $\mathbb{Y}$ is finite. By contrast, a **unifilar process** $(X_i)_{i\in\mathbb{N}}$ with respect to a Markov process $(Y_i)_{i\in\mathbb{N}}$ is such a hidden Markov process that

$$Y_{i+1} = \tau(Y_i, X_i) \tag{4}$$

for a certain function $\tau : \mathbb{Y} \times \mathbb{X} \to \mathbb{Y}$ (transition table). Unifilar processes are a probabilistic version of deterministic automata in automata theory. To specify a unifilar process, it suffices to fix initial distribution π, emission matrix ε, and transition table τ, since transition matrix σ follows from them.

To be concrete, let us discuss some further examples of stochastic processes. First, ***n*-th order Markov processes**, called also $(n+1)$-gram models, are unifilar processes such that $Y_i = X_{i-n}^{i-1}$. A subclass of these processes with $n = 2$, called trigram models, constitutes particularly effective statistical language models, which were applied in computational linguistics of the 1990s [7–9,12]. Another important examples are **computable processes**, which are processes such that function $w \mapsto P(X_1^{|w|} = w)$ is computable. It can be seen that the class of these processes is the class of hidden Markov processes with countably infinite $\mathbb{X}$ and $\mathbb{Y}$ and with computable functions π, σ, and ε, since it suffices to state that $Y_i = X_1^{i-1}$. The last example shows that hidden Markov processes can model pretty much anything if we do not impose a finite number of hidden states or a particular structure of the transition and emission matrices.

The post-Chomskyan linguistics refuted the class of finite-state processes on the account that they cannot model a context-free syntax with central embeddings of an unbounded depth [2–5]. This raised some doubt in the general utility of stochastic processes for theoretical linguistics. However, the class of discrete stochastic processes is much richer than simply finite-state processes. There are of course probabilistic context-free grammars (PCFGs), a model useful in the parsing of sentences in natural language [10–12,71]. However, PCFGs define probability distributions on finite trees or finite strings rather than infinite sequences. By contrast, here, we are interested in a model of text that can be unboundedly extended with time. Formally, let us define the **hidden Markov order** of process $(X_i)_{i\in\mathbb{N}}$ as the number of states in its minimal hidden Markov presentation,

$$M_{HM} := \inf\{|\mathbb{Y}| : (X_i)_{i\in\mathbb{N}} \text{ is hidden Markov with respect to } (Y_i)_{i\in\mathbb{N}}\} \tag{5}$$

with the convention that the infimum of the empty set is infinite. That is, we have equality $M_{HM} = \infty$ if and only if $(X_i)_{i\in\mathbb{N}}$ is not a finite-state process. Analogously, we can define the **unifilar Markov order** of process $(X_i)_{i\in\mathbb{N}}$ as the number of states in its minimal unifilar presentation,

$$M_U := \inf\{|\mathbb{Y}| : (X_i)_{i\in\mathbb{N}} \text{ is unifilar with respect to } (Y_i)_{i\in\mathbb{N}}\}, \tag{6}$$

cf. [72–74]. We have inequality $M_U \geq M_{HM}$. The minimal unifilar presentation of a process, called the **ϵ-machine**, is unique and given by the equivalence classes of conditional probability of infinite future given infinite past [72,73]. There exist simple processes such that $M_U = \infty$ and $M_{HM} < \infty$ [74], e.g., the Golden Mean process [45] or the Simple Nonunifilar Source [75]. In fact, these two processes have only two hidden states in their minimal nonunifilar presentations but their minimal unifilar presentations have uncountably many hidden states. These processes are very simple examples of processes with $M_U = \infty$ but they have no linguistic interpretation.

In the second turn, we can propose another simple example of a process that does not have a finite-state presentation, even a nonunifilar one, and can be considered an idealized model of the unbounded accumulation of randomly accessed factual knowledge. As we mentioned in Section 1, we model the factual knowledge as a compressed infinite sequence of bits that becomes gradually revealed in text. There are two obvious choices: We can consider a fixed sequence $(z_k)_{k\in\mathbb{N}}$ where $z_k \in \{0,1\}$, or putting on a Bayesian hat, when we do not know this sequence a priori, we can model the factual knowledge as an IID process $(Z_k)_{k\in\mathbb{N}}$ over alphabet $\{0,1\}$ with the uniform distribution, i.e., $P(Z_k = 0) = P(Z_k = 1) = 1/2$. We can consider the **Santa Fe processes**, which are sequences of random variables $(X_i)_{i\in\mathbb{N}}$ that consist of either pairs

$$X_i = (K_i, Z_{K_i}) \tag{7}$$

or pairs

$$X_i = (K_i, z_{K_i}), \tag{8}$$

where $(K_i)_{i\in\mathbb{N}}$ is an IID process over alphabet $\mathbb{N}$ with **Zipf's distribution** $P(K_i = k) \propto k^{-\alpha}$ for a parameter $\alpha > 1$. These processes were discovered by us in August 2002 during our visit at the Santa Fe Institute, but they were first published in [28,29].

From a linguistic point of view, we can interpret Santa Fe processes as a toy model of a stochastic process that conveys an infinite number of elementary meanings in a repetitive way. Namely, these processes can be interpreted as sequences of random statements $X_i = (k, z)$ that assert for a randomly chosen index k that the kth fact, i.e., the kth item of the factual knowledge, equals z: $Z_k = z$ or $z_k = z$. This stochastic description, although indices K_i are scattered at random, is never contradictory: If statements $X_i = (k, z)$ and $X_i = (k', z')$ describe the same fact, i.e., $k = k'$, then both statements assign the same value to it, i.e., $z = z'$. Moreover, since random variables $(K_i)_{i\in\mathbb{N}}$ constitute an IID process and $P(K_i = k) > 0$ for all $k \in \mathbb{N}$, ultimately, every fact is described in a sufficiently long text X_1^n **almost surely**. "Almost surely" is a mathematical quantifier that means "with probability 1". Moreover, the Zipf distribution of random variables K_i allows us to deduce a stronger property: The number of distinct facts Z_k or z_k described by a random text X_1^n is asymptotically proportional to $n^{1/\alpha}$ almost surely [36]. That is, the facts follow a sort of Herdan–Heaps' law, originally formulated as a power-law growth of the number of distinct words [31–34]. A generalization of this property is called **perigraphic processes** in Section 5, applying the concept of Hilberg exponents developed in Section 3 and the notion of algorithmic randomness. What is interesting for linguistic discussions is that the non-IID Santa Fe processes (7) are not finite-state processes. We have $M_{HM} = \infty$ for them since the Shannon mutual information between the past and future is infinite, as we discuss in Section 3. In this article, we show that perigraphic processes, such as the IID Santa Fe processes (8), cannot be finite-state processes either.

Looking for more realistic models of language, we can proceed in the hierarchy of discrete stochastic processes further. Two important notions in the theory of stochastic processes are stationary and ergodic processes. Usuallym they are defined by applying measure theory, but for discrete processes, the respective conditions can be expressed using finite-dimensional distributions. In particular, process $(X_i)_{i\in\mathbb{N}}$ is a **stationary process** if and only if the probabilities are shift invariant, i.e.,

$$P(X_1^{|w|} = w) = P(X_{t+1}^{t+|w|} = w) \tag{9}$$

for all strings $w \in \mathbb{X}^*$ and shifts $t \in \mathbb{N}$. Every one-sided stationary process $(X_i)_{i\in\mathbb{N}}$ can be extended to the stationary sequence of random variables $(X_i)_{i\in\mathbb{Z}}$ extending into two directions. An important result, the **Birkhoff ergodic theorem** states that, for a stationary process $(X_i)_{i\in\mathbb{N}}$, relative frequencies converge almost surely, i.e., if we define event

$$\Omega_S := \bigcap_{w\in\mathbb{X}^*} \left(\liminf_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbf{1}\left\{X_{t+1}^{t+|w|} = w\right\} = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbf{1}\left\{X_{t+1}^{t+|w|} = w\right\} \right) \tag{10}$$

then $P(\Omega_S) = 1$, cf. [36] (Section 4.2) and [76,77]. Moreover, a stationary process $(X_i)_{i\in\mathbb{N}}$ is an **ergodic process** if and only if the relative frequencies of all strings converge to their probabilities almost surely, i.e., when $P(\Omega_P) = 1$ for

$$\Omega_P := \bigcap_{w\in\mathbb{X}^*} \left(\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbf{1}\left\{X_{t+1}^{t+|w|} = w\right\} = P(X_1^{|w|} = w) \right). \tag{11}$$

The Birkhoff ergodic theorem is a generalization of the law of large numbers for IID processes. There are a few more effective criteria of ergodicity, cf. [77] and [36] (Section 4.3). In particular, it can be shown that the non-IID Santa Fe processes (7) are stationary but not ergodic, whereas the Santa Fe processes (8) are IID and, hence, ergodic.

Not all stochastic processes are stationary, ergodic, computable, or perigraphic. It is important to note that these conditions interact not only with each other but also with a particular interpretation that we ascribe to the concept of probability, as applied to language modeling in particular. There are two main distinct interpretations of probability: subjective and objective—as we call them in this paper. The **subjective probabilities** represent subjective odds of a language user—or of an effective predictor, speaking more generally. As such, the subjective probabilities should be computable, but they can be nonergodic—since there may be some prior random variables in the mental state of a language user such as variables Z_k in the Santa Fe process (7). Upon the conditioning of subjective probabilities on the previously seen text, the prior random variables becomes more and more concentrated on some particular fixed values. This concentration process can be equivalently named the process of learning of the unknown parameters. The **objective probabilities** represent an arbitrary limit of this learning process, where all prior random variables become instantiated by some fixed values such as values z_k in the Santa Fe process (8). Miraculously, it turns out that objective probabilities of strings are exactly the asymptotic relative frequencies of these strings in the particularly generated infinite text. As such, the objective probabilities should be ergodic by the Birkhoff ergodic theorem if the generating subjective odds form a stationary process but they can be uncomputable since the limit of computable functions need not be computable.

This difference in desiderata for subjective (computable but not ergodic) and objective (ergodic but not computable) statistical language models is formally reconciled by the **ergodic decomposition theorem**, which says that, for any stationary distribution P, there exists a unique prior ν supported on stationary ergodic distributions such that

$$P(A) = \int F(A) d\nu(F) \tag{12}$$

for all events A, cf. [36] (Section 4.4) and [62,63,77]. That is, the computable subjective distribution P is the average of ergodic objective distributions F taken with a computable prior ν, whereas ergodic objective distributions F can be interpreted as so-called ergodic components of the computable subjective distribution P. In some sense, the set of measures

F is given uniquely for a given measure P. In particular, for the Santa Fe processes (7), which are computable and nonergodic, the ergodic components take the form of processes (8), where $(z_k)_{k\in\mathbb{N}}$ are fixed infinite binary sequences. The prior ν is simply the uniform measure on these sequences, i.e., the probability that $Z_1^k = z_1^k$ equals 2^{-k}. Processes $(X_i)_{i\in\mathbb{N}}$ given by (8) are ergodic, but for almost all sequences $(z_k)_{k\in\mathbb{N}}$, they are not computable for the simple reason that individual sequences $(z_k)_{k\in\mathbb{N}}$ are not computable themselves.

3. Hilberg's Hypothesis

The relaxed Hilberg hypothesis for natural language states that the Shannon mutual information between two blocks of random variables for a reasonable statistical language model should grow roughly as a power of the block length [40,42–45,78]. Considering this hypothesis, we can be seriously concerned with how to identify the right statistical language model. To address this problem, in this section, we adjust the statement of Hilberg's hypothesis for natural language to make it independent of the distinction between subjective and objective probabilities. We note that the ergodic decomposition is a technically difficult theorem in the general stationary case but the distinction between subjective and objective probabilities affects the values of Shannon mutual information. For nonergodic Santa Fe processes (7), the Shannon mutual information between a finite past and a finite future diverges as a power law, whereas it equals zero for their ergodic components (8) since those are obviously IID processes. Thus, when stating Hilberg's hypothesis, we must be careful whether we work with subjective or with objective probabilities. Either we must specify what kind of statistical language model we speak of or we should make our statement of Hilberg's hypothesis invariant with respect to choosing a particular interpretation of probability. In this article, we apply the second solution, an invariant statement, by using algorithmic mutual information instead of Shannon mutual information.

First, let us fix the notation and basic concepts. Symbol $\ln x$ denotes the natural logarithm, in contrast with the binary logarithm $\log x$. Applying the measure-theoretic formalism, $\mathbf{E}\, X := \int X dP$ is the **expectation** of a real random variable X with respect to a probability measure P. The **Shannon entropy** of a discrete random variable X is $H(X) := \mathbf{E}[-\log P(X)]$, where $P(X) = P(X = x)$ if $X = x$, whereas **conditional entropy** of X given random variable Y is $H(X|Y) := \mathbf{E}[-\log P(X|Y)]$, where $P(X|Y) = P(X = x|Y = y)$ if $X = x$ and $Y = y$. Subsequently, the **Shannon mutual information** for random variables X and Y is $I(X;Y) := H(X) + H(Y) - H(X,Y)$.

Let $(X_i)_{i\in\mathbb{Z}}$ be a stationary process over a finite alphabet $\mathbb{X}$. We denote the conditional entropies

$$h_k := H(X_0|X_{-k}^{-1}). \tag{13}$$

It is well known that we can define the **entropy rate** h as the limiting amount of information produced by a single random variable,

$$h := \lim_{n\to\infty} \frac{H(X_1^n)}{n} = \inf_{k\geq 1} h_k = H(X_0|X_{-\infty}^{-1}). \tag{14}$$

As discussed in [36,45], we can also equivalently define the **excess entropy** E as the mutual information between infinite past and infinite future of the process,

$$E := \lim_{n\to\infty} [H(X_1^n) - nh] = \lim_{n\to\infty} I(X_{-n+1}^0; X_1^n) = I(X_{-\infty}^0; X_1^\infty). \tag{15}$$

(The proof in [45] contains a gap, whereas a correct proof can be found in [36] (Theorem 5.13).)

The **data-processing inequality** states that $I(X;Y) \geq I(X;Z)$ if random variables X and Z are conditionally independent given Y. This holds in particular if Z is a function of Y, $Z = f(Y)$, hence the name of this inequality: The information decreases as we process it deterministically. Consequently, if $(X_i)_{i\in\mathbb{Z}}$ is a hidden Markov process with respect to a

Markov process $(Y_i)_{i\in\mathbb{Z}}$, then by the data-processing inequality and the Markov condition, we obtain

$$E = I(X_{-\infty}^0; X_1^\infty) \le I(Y_{-\infty}^0; Y_1^\infty) = I(Y_0; Y_1) \le H(Y_0) \le \log M_{HM}. \tag{16}$$

In particular, the excess entropy of a finite-state process is finite. By contrast, the relaxed Hilberg hypothesis in a variant introduced in [42–45,78] that states that mutual information $I(X_{-n+1}^0; X_1^n)$ grows similar to a power law. Such unbounded growth is clearly impossible for finite-state processes but can be achieved for the nonergodic Santa Fe processes (7). In fact, every stationary nonergodic process with a continuous prior on the ergodic components has infinite excess entropy via the ergodic decomposition of excess entropy, cf. [28] and [36] (Theorems 5.35 and 5.40).

For the sake of further considerations concerning the power-law growth of various quantities, let us introduce so-called **Hilberg exponents**

$$\operatorname*{hilb}_{n\to\infty} s(n) := \limsup_{n\to\infty} \frac{\log \max\{1, s(n)\}}{\log n} \tag{17}$$

for real functions $s(n)$ of natural numbers, cf. [27,36,79] (Definition 8.1), where we gradually approach the above definition. The Hilberg exponents capture the asymptotic power-law growth of the respective functions, such as

$$\operatorname*{hilb}_{n\to\infty} n^\beta = \beta \text{ for } \beta \ge 0. \tag{18}$$

Let us strengthen a simple observation from [27,36]. Our improvement is also very simple and it consists of replacing condition $\mathfrak{J}(n) \ge -C$ with $\mathfrak{S}(n) - n\mathfrak{s} \ge -C$ as sufficient for equality of the respective Hilberg exponents. It is surprising that we have not noticed this earlier.

Theorem 1 (cf. [27] and [36] (Theorem 8.2))**.** *For a function $\mathfrak{S} : \mathbb{N} \to \mathbb{R}$, define $\mathfrak{J}(n) := 2\mathfrak{S}(n) - \mathfrak{S}(2n)$. If $\lim_{n\to\infty} \mathfrak{S}(n)/n = \mathfrak{s}$ for a $\mathfrak{s} \in \mathbb{R}$ then*

$$\operatorname*{hilb}_{n\to\infty} (\mathfrak{S}(n) - n\mathfrak{s}) \le \operatorname*{hilb}_{n\to\infty} \mathfrak{J}(n) \tag{19}$$

with an equality if $\mathfrak{S}(n) - n\mathfrak{s} \ge -C$ for all but finitely many n and some $C > 0$.

By Theorem 1 and identity $I(X_{-n+1}^0; X_1^n) = 2H(X_1^n) - H(X_1^{2n})$ following from stationarity, we can define the Hilberg exponent

$$\beta_H := \operatorname*{hilb}_{n\to\infty} [H(X_1^n) - nh] = \operatorname*{hilb}_{n\to\infty} I(X_{-n+1}^0; X_1^n) \in [0,1]. \tag{20}$$

In particular, $\beta_H = 1/\alpha$ for the nonergodic Santa Fe processes (7). That is, in some particular mathematical model, an unbounded accumulation of factual knowledge can be a reason for the relaxed Hilberg hypothesis. If we infer repeatable information from the process at a power law rate, so must grow the mutual information between the past and the future. We make this intuition precise in Section 5.

The **relaxed Hilberg hypothesis** for natural language in the variant introduced in references [42–45,78] could be simply expressed as condition $\beta_H > 0$ for a reasonable statistical language model. However, such a formulation is ambiguous since, as we mentioned in the beginning of this section, there are two main interpretations of probability, nonergodic subjective and ergodic objective, and this distinction affects the estimates of power-law growth of mutual information. As we indicated, the guiding example are the subjective nonergodic Santa Fe processes (7), where $\beta_H = 1/\alpha$ is an arbitrary number in the range $(0,1)$, whereas $\beta_H = 0$ holds for their objective ergodic components (8), since they are IID. Additionally, for natural language, the estimates of the Hilberg exponent vary depending on the estimation method. Universal coding estimates yield an upper bound of $\beta_H \le 0.8$

[46–48,51,52], whereas methods based on guessing by human subjects seem to yield an upper bound of $\beta_H \le 0.5$ [40,41]. Thus, imposing a condition on the subjective probability Hilberg exponent β_H may differ greatly from imposing a similar condition on the objective probability Hilberg exponent β_H. This is the main conceptual difficulty about Hilberg's hypothesis that researchers in this topic should be aware of.

Some solution to this problem may be using a yardstick that is independent of a concrete probability distribution. In particular, we may apply the algorithmic information theory, where the information content of a particular text is defined in terms of the minimal length of a computer program that outputs this text. In particular, the **prefix Kolmogorov complexity** of a string w, denoted $K(w)$, is the length of the shortest self-delimiting program for a universal computer for which the output is w. Note that the prefix Kolmogorov complexity is in general uncomputable but can be effectively approximated from above. The **algorithmic mutual information** between strings u and w is $J(u;w) := K(u) + K(w) - K(u,w)$. Many results from the Shannon information theory carry on to the algorithmic information theory, but the respective proofs are often more difficult [38,64,65]. Let us observe that the typical difference between expected Kolmogorov complexity $\mathbf{E}\, K(X_1^n)$ and Shannon entropy $H(X_1^n)$ is of the order $\log n$ if the probability measure P is computable. For uncomputable measure P, which holds also if some parameters of a computable formula for P are uncomputable real numbers, this difference can be somewhat greater or even substantially greater, which complicates the transfer of results from one sort of information theory to another.

Let us inspect which of our claims survive in the algorithmic setting. Let $(X_i)_{i\in\mathbb{Z}}$ be a stationary process over a finite alphabet $\mathbb{X}$. Since $\mathbf{E}\, K(X_1^n) \ge H(X_1^n)$ by the prefix-free property of Kolmogorov complexity and $K(w) \le LZ(w)$, where $LZ(w)$ is the length of a self-delimiting universal Lempel–Ziv code [39], then we obtain

$$\lim_{n\to\infty} \frac{K(X_1^n)}{n} = h \text{ almost surely,} \tag{21}$$

$$\lim_{n\to\infty} \frac{\mathbf{E}\, K(X_1^n)}{n} = h. \tag{22}$$

(These equalities were originally shown by Brudno [80] using a much more involved technique.) Hence, by Theorem 1, we can define another Hilberg exponent:

$$\beta_K := \operatorname*{hilb}_{n\to\infty} \mathbf{E}[K(X_1^n) - nh] = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_{-n+1}^0; X_1^n) \in [0,1], \tag{23}$$

where $\beta_K \ge \beta_H$. The difference between exponents β_H and β_K can be as large as 1, depending on the probability distribution of process $(X_i)_{i\in\mathbb{Z}}$. If the probability distribution is computable, then there holds $\beta_H = \beta_K$, since besides $\mathbf{E}\, K(X_1^n) \ge H(X_1^n)$, we also have that $K(X_1^n) \le -\log P(X_1^n) + 2\log n + K(P)$ by the Shannon–Fano coding, where $K(P)$ is the Kolmogorov complexity of measure P [79]. Thus, if we think that Hilberg's hypothesis should be stated for a computable subjective probability, then we can simply express it as $\beta_K > 0$, which has a greater chance of remaining valid also under the objective probability interpretation. (Let us note that, for different probability interpretations, we have expectations of the same random variables but with respect to different probability measures.)

However, this is not the end of the detachment from a probability measure. Let us define a random variable

$$\gamma_K := \operatorname*{hilb}_{n\to\infty} J(X_{-n+1}^0; X_1^n), \tag{24}$$

which is independent of the distribution of the process. As also shown in [79], for any stochastic process $(X_i)_{i\in\mathbb{Z}}$, we have

$$\gamma_K \le \beta_K \text{ almost surely.} \tag{25}$$

Additionally, if the process is ergodic, then Hilberg exponent γ_K is constant almost surely, as shown in [79]—but we do not know whether $\gamma_K = \beta_K$ holds in so general case, cf. [36]

(Section 8.2). Notice also that, for the ergodic decomposition $P(A) = \int F(A) d\pi(F)$, any event A has a subjective probability $P(A) = 1$ if and only if we have $F(A) = 1$ for π, almost every objective probability F. Hence, condition $\gamma_K > 0$ holds almost surely for a subjective distribution if and only if $\gamma_K > 0$ holds almost surely for almost all objective distributions supported by the subjective distribution.

Consequently, to make the statement of Hilberg's hypothesis invariant with respect to switching between subjective and objective perspectives or to adopt an intermediate perspective—as it arises in actual experiments with texts and human subjects—we should express it rather as condition $\gamma_K > 0$ using the algorithmic mutual information. The above paragraphs are the motivation for the following formal definition.

Definition 1 (Hilberg condition)**.** *We say that a stationary process $(X_i)_{i\in\mathbb{Z}}$ satisfies the Hilberg condition if $\gamma_K > 0$ holds almost surely.*

This is our working understanding of the relaxed Hilberg hypothesis, which could be applied both to statistical language models and to more abstract stochastic processes. Using the uncomputable algorithmic information is the price that we pay for working with an underspecified probability model.

4. Finite-State Processes

The aim of this section is to show that no finite-state process satisfies the Hilberg condition. That is, if we believe that natural language satisfies the relaxed Hilberg hypothesis, we cannot expect that it can be reasonably modeled by a hidden Markov process with a finite number of hidden states. However, our intended claim, stated in the algorithmic fashion and detached from the probability measure as far as possible, is not so trivial as claiming that no finite-state has infinite excess entropy. The reason is that algorithmic mutual information $J(X_{-n+1}^0; X_1^n)$ may diverge for some finite-state processes if their transition and emission matrices contain uncomputable real numbers. We want to show that $J(X_{-n+1}^0; X_1^n)$ in this case can only grow quite slow, namely, not faster than $\log n$ multiplied by the number of hidden states. Since we do not know the number of hidden states beforehand, we need to recall some theory of consistent estimation of the number of hidden states and adjust it to our particular needs. This section is a journey through mathematical statistics and information theory.

To prove that no finite-state process satisfies the Hilberg condition, we put together a few ideas that are well known in information theory: normalized maximum likelihood, universal codes in the spirit of Ryabko, strongly consistent order estimators, as well as our own ideas developed for the theorems about facts and words mentioned in Section 1. We work with unifilar processes to obtain a stronger result than we need for the mere refutation of finite-state language models. We translate this result into finite-state processes by the end of this section and we apply it to Oracle processes in Section 6. We recall from Section 2 that a unifilar process is a hidden Markov process with an arbitrary (possibly infinite) number of hidden states that is deterministic in the automata sense, i.e., the next hidden state is a fixed function of the previous hidden state and the previous emitted symbol.

In this section, we consider a family of unifilar process distributions where the number $k = 1, 2, 3, \dots$ of hidden states is finite and the emitted symbols belong to a fixed finite alphabet $\mathbb{X}$. That is, for a given sequence of symbols x_1^n and states y_1^n, our **unifilar distributions** take the following form:

$$\mathbb{P}(x_1^n, y_1^n | k, \pi, \tau, \varepsilon) := \pi(y_1)\varepsilon(x_1|y_1) \prod_{i=2}^{n} \mathbf{1}\{y_i = \tau(y_{i-1}, x_{i-1})\}\varepsilon(x_i|y_i), \tag{26}$$

where $\pi : \{1, .., k\} \to [0, 1]$ with $\sum_y \pi(y) = 1$ is the initial hidden state distribution, $\tau : \{1, .., k\} \times \mathbb{X} \to \{1, .., k\}$ is the transition table, and $\varepsilon : \mathbb{X} \times \{1, .., k\} \to [0, 1]$ with $\sum_x \varepsilon(x|y) = 1$ is the emission matrix. We also denote the marginal distribution

$$\mathbb{P}(x_1^n|k,\pi,\tau,\varepsilon) := \sum_{y_1^n} \mathbb{P}(x_1^n, y_1^n|k,\pi,\tau,\varepsilon) \tag{27}$$

and the conditional distribution

$$\mathbb{P}(x_1^n|k,y_1,\tau,\varepsilon) := \frac{1}{\pi(y_1)} \sum_{y_2^n} \mathbb{P}(x_1^n, y_1^n|k,\pi,\tau,\varepsilon). \tag{28}$$

Subsequently, we define three distributions of the shape well-known in minimum description length theory [81]: the **maximum likelihood** (ML)

$$\hat{\mathbb{P}}(x_1^n|k) := \max_{y,\tau,\varepsilon} \mathbb{P}(x_1^n|k,y,\tau,\varepsilon); \tag{29}$$

the **normalized maximum likelihood** (NML) in the spirit of Shtarkov [82]

$$\mathbb{P}(x_1^n|k) := \frac{\hat{\mathbb{P}}(x_1^n|k)}{\sum_{z_1^n \in \mathbb{X}^n} \hat{\mathbb{P}}(z_1^n|k)} \le \hat{\mathbb{P}}(x_1^n|k); \tag{30}$$

and the **Ryabko mixture**, cf. [58,59],

$$\mathbb{P}(x_1^n) := \sum_{k=1}^{\infty} w_k \mathbb{P}(x_1^n|k), \quad w_k := \frac{1}{k} - \frac{1}{k+1}. \tag{31}$$

We notice that the maximum likelihood satisfies $\hat{\mathbb{P}}(x_1^n|k) = 1$ for $k \ge n$, since having as many hidden states as the string length, we can put $\pi(1) = 1$, $\tau(i, x_i) = i + 1$, and $\varepsilon(x_i|i) = 1$. Consequently, the NML equals $\mathbb{P}(x_1^n|k) = |\mathbb{X}|^{-n}$ for $k \ge n$ and the Ryabko mixture $\mathbb{P}(x_1^n)$ is a computable function of x_1^n since the defining infinite series can be truncated. We stress that the maximum likelihood, the NML, and the Ryabko mixture are computable in the sense of computability theory, which suffices for our needs of bounding algorithmic mutual information in Theorem 5, but they are computationally intractable since we need to perform an exhaustive search over all transition tables τ combined with summation over exponentially growing domains $\mathbb{X}^n$.

Subsequently, such as in [81], we introduce the **family complexity** of the unifilar family:

$$\mathbb{C}(n|k) := -\log \mathbb{P}(x_1^n|k) + \log \hat{\mathbb{P}}(x_1^n|k) = \log \sum_{z_1^n \in \mathbb{X}^n} \hat{\mathbb{P}}(z_1^n|k) \le n \log|\mathbb{X}|. \tag{32}$$

This family complexity is a different concept than the **statistical complexity** of a stochastic process discussed in [72–74]. The family complexity (32) is a property of a class of processes, roughly related to the number of distinguishable distributions in the class. By contrast, the statistical complexity by [72–74] is the entropy of the hidden state distribution in the minimal unifilar presentation of a given process. The statistical complexity is smaller than or equal to $\log M_U$ but greater than or equal to excess entropy (15). Unlike excess entropy, it can be infinite for some finite-state nonunifilar sources such as the Golden Mean process [45] or the Simple Nonunifilar Source [75]. By contrast, it is a rule of thumb that the family complexity of a distribution family with exactly k real parameters is roughly $k \log n$. There also exist more exact expressions assuming some particular conditions [81]. Here, we only need a very rough bound for $\mathbb{C}(n|k)$ but assuming that we have not only a real-parameter emission matrix ε but also an integer-parameter transition table τ. We can observe a small correction up to the aforementioned rule of thumb.

Theorem 2. *For the unifilar family, the family complexity satisfies*

$$\mathbb{C}(n|k) \le [k|\mathbb{X}| + 1] \log[k(n+1)]. \tag{33}$$

The next fact that we present is the **universality of the Ryabko mixture**, i.e., the Ryabko mixture yields a strongly consistent and asymptotically unbiased estimator of the entropy rate. For distribution families that contain Markov chain distributions of all orders and for which the family complexity $\mathbb{C}(n|k)$ grows sublinearly with the sample size n for any order k, the Ryabko mixture is a universal distribution by a reasoning following the ideas of papers [58,59]. It turns out that this is the case for the unifilar hidden Markov family. As a consequence, the Ryabko mixture can be used for universal compression of data generated by any stationary ergodic process, i.e., there is a computable procedure that takes text X_1^n and compresses it losslessly as a string of $-\log \mathbb{P}(X_1^n) \approx hn$ bits, and this compression cannot be substantially improved. The following theorem states the universality of the Ryabko mixture:

Theorem 3. *For a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet,*

$$\lim_{n\to\infty} \frac{1}{n}[-\log \mathbb{P}(X_1^n)] = h \text{ almost surely}, \tag{34}$$

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{E}[-\log \mathbb{P}(X_1^n)] = h. \tag{35}$$

For completeness, we present the proof in Appendix A but we do not claim originality of the idea. Additionally, as we have mentioned, this particular Ryabko mixture is computable in the sense of computability theory, but it is intractable and highly impractical as a universal compression procedure. We need it only for further theoretical applications.

As we announced in the beginning, all this is needed to estimate the unifilar order of the process, i.e., the number of hidden states, and to link this estimate with the algorithmic mutual information for an unknown process, being a statistical language model in particular. Thus, subsequently, we consider a unifilar order estimator that is a certain modification of estimators of the Markov order and the hidden Markov order proposed by Merhav, Gutman, and Ziv [83] and by Ziv and Merhav [84], respectively. The idea of [83,84] is that the estimator returns the smallest order for which the maximum likelihood is larger than a penalized universal probability. Consequently, we will define the **unifilar order estimator**:

$$\mathbb{M}(x_1^n) := \min\left\{k : \hat{\mathbb{P}}(x_1^n|k) \ge w_n \mathbb{P}(x_1^n)\right\}, \quad w_n := \frac{1}{n} - \frac{1}{n+1}. \tag{36}$$

We can see that the estimator is nicely bounded by $\mathbb{M}(x_1^n) \le n$ since $\hat{\mathbb{P}}(x_1^n|k) = 1$ for $k \ge n$. In the literature on Markov order estimation [83,85–95], sublinear penalty $-\log w_n = o(n)$ in estimators resembling (36) can be traced in [88,90,94]. In the literature on hidden Markov order estimation [84,96–104], the majority of articles consider very similar ideas and prove the strong consistency of related estimators. Thus, we do not claim a particular originality of estimator (36).

The unifilar order estimator (36) is computable in the sense of computability theory, but it is intractable since it applies exact maximum likelihood and normalized maximum likelihood. We need it as is since it yields the most elegant upper bound for the algorithmic mutual information. Ignoring the question of obtaining this bound for a while, we note that we can make the estimator somewhat computationally simpler while preserving strong consistency if we replace universal distribution $\mathbb{P}(x_1^n)$ with a simpler universal compression procedure such as the Lempel–Ziv code [39]. This idea was proposed by Merhav, Gutman, and Ziv [83] and by Ziv and Merhav [84] themselves. This substitution, however, breaks the simple upper bound for mutual information to be stated in Theorem 5 while not solving the problem of computing the maximum likelihood, which requires an exhaustive search over all transition tables τ. By contrast, some practical estimators of the hidden Markov order can be found in [102,104].

The following theorem states a **strong consistency** and **asymptotic unbiasedness** of unifilar order estimator (36), which makes use of the universality of the Ryabko mixture claimed in Theorem 3.

Theorem 4. *For a stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet,*

$$\lim_{n\to\infty} \mathbb{M}(X_1^n) = M_U \text{ almost surely,} \tag{37}$$

$$\lim_{n\to\infty} \mathbf{E}\,\mathbb{M}(X_1^n) = M_U, \tag{38}$$

and we have the overestimation bound $P\big(\mathbb{M}(X_1^n) > M_U\big) \le w_n$.

The proof is quite complicated and deferred to Appendix A. Our proof technique for the impossibility of overestimation is taken from Markov order estimation proof ideas such as [90,94]. We suppose that our proof of the impossibility of underestimation is more original—although some expressions in it superficially resemble some results by Gassiat and Boucheron [101]. In contrast with [101], we also prove consistency in the case of $M_U = \infty$. That is, estimator $\mathbb{M}(X_1^n)$ grows unboundedly almost surely if the process does not have a finite unifilar presentation—which may be the case of natural language. Since we apply a result about asymptotically mean stationary channels by Kieffer and Rahe [105], we suspected that Kieffer [98] might have used a similar technique in the context of hidden Markov order estimation but we did not find it there.

What remains is to link the mutual information with the unifilar order estimator. First, we compare the algorithmic mutual information with the Ryabko mixture mutual information. By nonnegativity of the Kullback–Leibler divergence, $\mathbf{E}\big[-\log\mathbb{P}(X_1^n)\big] \ge H(X_1^n) \ge hn$, so in view of Theorem 1, we define the Hilberg exponent for the Ryabko mixture mutual information:

$$\begin{aligned}\beta_{\mathbb{P}} &:= \operatorname*{hilb}_{n\to\infty} \mathbf{E}[-\log\mathbb{P}(X_1^n) - nh] \\ &= \operatorname*{hilb}_{n\to\infty} \mathbf{E}\Big[-\log\mathbb{P}(X_1^n) - \log\mathbb{P}(X_{n+1}^{2n}) + \log\mathbb{P}(X_1^{2n})\Big] \in [0,1].\end{aligned} \tag{39}$$

By the universality of the Ryabko mixture proven in Theorem 3 and inequality

$$K(x_1^n) \le -\log\mathbb{P}(X_1^n) - \log w_n + K(\mathbb{P}) \tag{40}$$

stemming from the computability of the Ryabko mixture and the Shannon–Fano coding [38,106], we also obtain

$$\beta_H \le \beta_K \le \beta_{\mathbb{P}}. \tag{41}$$

Thus, our first goal of relating algorithmic mutual information to Ryabko mixture is accomplished.

Next, we relate the Ryabko mixture mutual information to a unifilar order estimator, which as we recall, was also defined in terms of the Ryabko mixture. The next theorem, which provides the requested link, resembles the second part of the theorems about facts and words, which bound the growth of Shannon and algorithmic mutual information in terms of the growth of the number of distinct words detectable in a random text.

Theorem 5. *For a stationary process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet,*

$$\beta_{\mathbb{P}} \le \beta_{\mathbb{M}} := \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,\mathbb{M}(X_1^n). \tag{42}$$

The proof of Theorem 5 applies a simple subadditivity technique, which is the essence of the proofs of the second part of the theorems about facts and words from [27,29] and [36] (Section 8.4). Seen from this perspective, we may interpret the unifilar order estimator

$\mathbb{M}(X_1^n)$ as an approximation of the number of distinct words that may be detected in text X_1^n. It may be interesting to investigate whether $\mathbb{M}(X_1^n)$ can actually be related to grammar-based coding, which was the original technique for proving the theorems about facts and words, cf. [29,107].

Let us observe that, in view of asymptotic unbiasedness (38) of the unifilar order estimator, we obtain $\beta_{\mathbb{M}} = 0$ for stationary ergodic finite-state unifilar processes over a finite alphabet. Consequently, in view of Theorem 5, all such processes satisfy $\beta_K = 0$. Using the data-processing inequality for algorithmic mutual information and the finite ergodic decomposition of finite-alphabet Markov processes, we may generalize this result to arbitrary finite-state processes.

Theorem 6. *For a finite-state stationary process $(X_i)_{i\in\mathbb{Z}}$ over a finite alphabet, we have $\beta_K = 0$.*

Hence, by inequality (25), we obtain $\gamma_K = 0$ almost surely, i.e., no finite-state stationary process over a finite alphabet satisfies the Hilberg condition. There is some technical detail here that may inspire some future research: Whereas $\beta_K = 0$ holds for finite-state processes in general, some of these processes, such as the Golden Mean process [45] or the Simple Nonunifilar Source [75], have the unifilar order $M_U = \infty$, cf. [72–74]. Thus, it is an interesting open problem whether $\beta_{\mathbb{M}} = 0$ holds also in the general case of nonunifilar finite-state processes. We suppose that it does.

Resuming this section, Hilberg's hypothesis refutes finite-state models also when we formulate it as an almost sure power law for algorithmic mutual information. If we believe in Hilberg's hypothesis seriously, we cannot defend finite-state language models.

5. Perigraphic Processes

Now, we are in a position that we need to justify Hilberg's hypothesis itself. Is it true in general that a power-law-rate accumulation of factual knowledge in an agent that reads a random text implies Hilberg's hypothesis? Well, this seems the first part of the theorems about facts and words discussed at length in papers [27,29] and book [36]. However, those discussions pertain to the expected number of facts and expected mutual information. Here, we strengthen these results a bit to relate them to the almost sure growth of algorithmic mutual information stated as the Hilberg condition in Definition 1. Along the way, we formally introduce the concept of perigraphic processes as defined in [27,36], which captures a power-law-rate accumulation of factual knowledge for stationary stochastic processes. We show that, perigraphic processes satisfy $\beta_K > 0$, i.e., the classes of perigraphic processes and finite-state processes are disjoint.

To approach these topics, first, we can ask whether there exist processes such that $\gamma_K > 0$ almost surely, i.e., ones that satisfy the Hilberg condition. In fact, as it was evaluated in [79,108], for the nonergodic Santa Fe processes (7), we obtain

$$\gamma_K = \beta_K = \beta_H = 1/\alpha \in (0,1) \text{ almost surely.} \tag{43}$$

Equality $\gamma_K = 1/\alpha$ almost surely transfers to almost all but not all ergodic Santa Fe processes (8). In fact, if we fix the sequence $(z_k)_{k\in\mathbb{N}}$ as $(0,0,0,\dots)$, we obtain $J(X_{-n+1}^0; X_1^n) \le J(K_{-n+1}^0; K_1^n) + C$ by the **data-processing inequality for algorithmic mutual information**, where $(K_i)_{i\in\mathbb{Z}}$ is an IID process. In turn, we may suspect that algorithmic mutual information $J(K_{-n+1}^0; K_1^n)$ is low and that the main contribution to high algorithmic mutual information $J(X_{-n+1}^0; X_1^n)$ for almost all ergodic components (8) may come from the high Kolmogorov complexity of the fixed sequence $(z_k)_{k\in\mathbb{N}}$.

In fact, there is an important concept in the algorithmic information theory, called algorithmic randomness, that allows us to deal with that intuition at ease. Precisely, a binary sequence $(z_k)_{k\in\mathbb{N}}$ is called **algorithmically random (in the Martin-Löf sense)** if it is incompressible in the sense that

$$K(z_1^n) \ge n - c \tag{44}$$

for all n and a constant $c < \infty$ [37,38]. Since almost all binary sequences $(z_k)_{k\in\mathbb{N}}$ with respect to the uniform measure $P(Z_1^k = z_1^k) = 2^{-k}$ are algorithmically random, we may suppose that $\gamma_K = 1/\alpha \in (0,1)$ holds almost surely for an ergodic Santa Fe process (8) if $(z_k)_{k\in\mathbb{N}}$ is algorithmically random.

To show that it is actually the case, we may use another important observation. Namely, some prefix of sequence $(z_k)_{k\in\mathbb{N}}$ can be computed from both blocks X_{-n+1}^0 and X_1^n for the ergodic Santa Fe process (8). Let us denote random variables

$$U_{m,n} := \min\{k \ge 1 : K_i \ne k \text{ for all } i \text{ such that } m \le i \le n\}. \tag{45}$$

Then, exactly string $z_1^{L_n-1}$ can be computed from both blocks X_{-n+1}^0 and X_1^n given the random number $L_n := \min\{U_{-n+1,0}, U_{1,n}\}$. Hence, by the data-processing inequality and algorithmic randomness of $(z_k)_{k\in\mathbb{N}}$, we obtain

$$\begin{aligned} J(X_{-n+1}^0; X_1^n) &\overset{+}{>} J(z_1^{L_n-1}; z_1^{L_n-1}) - 2K(L_n) \overset{+}{=} K(z_1^{L_n-1}) - 2K(L_n) \\ &\overset{+}{>} L_n - 1 - c - 4\log L_n \end{aligned} \tag{46}$$

Consequently, applying the techniques resumed in [36] (Theorem 8.14), we can show that inequality

$$\gamma_K \ge \operatorname*{hilb}_{n\to\infty} L_n = 1/\alpha \in (0,1) \text{ almost surely} \tag{47}$$

holds for all ergodic Santa Fe processes (8) with an algorithmically random sequence $(z_k)_{k\in\mathbb{N}}$. Thus, each of these processes taken individually satisfies the Hilberg condition.

In information theory, there is a formal construction for what we have shown above. It is called the **common information in the sense of Gács and Körner** [109]. Staying within the framework of Shannon information theory, if we have two random variables X and Y and a random variable Z that is a function of X and Y each, $Z = f(X) = g(Y)$, then the Shannon mutual information between X and Y is bounded as $I(X;Y) \ge I(Z;Z) = H(Z)$ by the data-processing inequality. The Gács–Körner common information $C_{GK}(X;Y)$ is the supremum of entropies $H(Z)$ taken over all random variable Z such that $Z = f(X) = g(Y)$. What is surprising is that inequality $C_{GK}(X;Y) \le I(X;Y)$ can be strict also if we perform the analogous construction in the algorithmic information theory [109]. There is also a related concept called the **common information in the sense of Wyner** $C_W(X;Y)$ [110], which satisfies a reversed inequality $C_W(X;Y) \ge I(X;Y)$. The theorems about facts and words discussed in [27,29,36] can be regarded as a certain application or generalization of inequalities $C_{GK}(X;Y) \le I(X;Y) \le C_W(X;Y)$.

Hence, the technique for Santa Fe processes can be generalized a bit. Consider an arbitrary computable function $g : \mathbb{N} \times \mathbb{X}^* \to \{0,1,2\}$, which we call a **knowledge extractor**, and an arbitrary fixed algorithmically random binary sequence $z = (z_k)_{k\in\mathbb{N}}$. Define random variables

$$U_{m,n}^{g,z} := \min\{k \ge 1 : g(k, X_m^n) \ne z_k\}. \tag{48}$$

If we put $L_n^{g,z} := \min\left\{U_{-n+1,0}^{g,z}, U_{1,n}^{g,z}\right\}$, then our preceding reasoning for Santa Fe processes carries over and we obtain

$$\gamma_K \ge \gamma_{g,z} := \operatorname*{hilb}_{n\to\infty} L_n^{g,z}, \tag{49}$$

$$\beta_K \ge \beta_{g,z} := \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, L_n^{g,z} \tag{50}$$

for an arbitrary process $(X_i)_{i\in\mathbb{Z}}$. For symmetry, let us define also

$$\gamma^+_{g,z} := \operatorname*{hilb}_{n\to\infty} U^{g,z}_{1,n} \ge \gamma_{g,z}, \tag{51}$$

$$\beta^+_{g,z} := \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U^{g,z}_{1,n} \ge \beta_{g,z}. \tag{52}$$

In [27], we have shown a seemingly stronger statement than (50), namely,

$$\beta_K \ge \beta^+_{g,z}, \tag{53}$$

which is the first part of the theorems about facts and words and holds for arbitrary stationary processes over a finite alphabet. To provide some further order in this zoo of Hilberg exponents, let us show that (50) and (53) can often boil down to the same statement since equality $\beta^+_{g,z} = \beta_{g,z}$ holds under mild conditions:

Theorem 7. *Let $(X_i)_{i\in\mathbb{N}}$ be a stationary process. If there hold inequalities $U^{g,z}_{m,n} \le U^{g,z}_{m,n+1}, U^{g,z}_{m-1,n}$, then*

$$\beta^+_{g,z} \ge \gamma^+_{g,z} \text{ almost surely}, \tag{54}$$

$$\beta_{g,z} \ge \gamma_{g,z} \text{ almost surely}. \tag{55}$$

If additionally we have $\lim_{n\to\infty} \mathbf{E}\, U^{g,z}_{1,n}/n = 0$, then

$$\beta^+_{g,z} = \beta_{g,z}. \tag{56}$$

Let us resume these constructions by giving them a name and by relating them to previous results.

Definition 2 (perigraphic process [27])**.** *A stationary process such that $\beta^+_{g,z} > 0$ for a certain computable knowledge extractor g and a certain algorithmically random sequence z is called a perigraphic process.*

For the obvious choice of knowledge extractor $g(k, x_1^n)$ for the Santa Fe processes that reads off the value of bit z_k, if it appears in sequence x_1^n and returns 2 otherwise [27], the assumptions of Theorem 7 are satisfied. Hence by (47), (55) and (56), an example of a perigraphic process is the ergodic Santa Fe process (8) with an algorithmically random sequence $(z_k)_{k\in\mathbb{N}}$. In the conclusion of the book in [36], we stated open problem no. 4 asking whether the classes of perigraphic and finite-state processes are disjoint. We supposed that this is true. According to Theorem 6 and inequality (53), these classes are disjoint indeed, so our conjecture was correct.

Now, it is time for a short break from maths to comment on a linguistic interpretation for the above considerations. As we announced in the Introduction, perigraphic processes may be a probabilistic model of texts that admit a power-law-rate accumulation of factual knowledge in agents that try to predict them. The role of factual knowledge in this model is played by the algorithmically random sequence z, i.e., the factual knowledge is compressed as much as possible, infinite, and time-independent. By contrast, the computable knowledge extractor g plays the role of language competence, which allows us to effectively and ultimately infer all factual knowledge from the infinite text regardless of where the agent starts observing the infinite text. Of course, these are quite strong assumptions from the point of view of what we may speculate about the real human language but we think that perigraphic processes may be an interesting linking model between the fields of linguistics and of stochastic processes.

We should be also aware that perigraphic processes can be much more complex than Santa Fe processes and that there are some not fully understood interactions between perigraphicness, nonergodicity, and computability of prior ν in the ergodic decomposition (12). Necessarily perigraphic processes must be uncomputable since their probability distributions encode an algorithmically random sequence [27]. As a more complicated example, we also found stochastic processes called **random hierarchical association (RHA) processes**, cf. [30] and [36] (Section 11.4), which seem to exhibit not only the Hilberg condition but also a bottom-up hierarchical structure of an infinite height. These processes are nonergodic, and we suspect that their ergodic components are perigraphic with quite a nontrivial knowledge extractor g and algorithmically random sequences z, which are different for different ergodic components. From our point of view, it is interesting that some seemingly abstract mathematical concepts such as nonergodicity or uncomputability acquire an idealized linguistic interpretation. There is a great opportunity to exhibit further examples of processes and to pursue further modeling ideas. One such idea is the transience of factual knowledge, which seems to correspond to the phenomenon of mixing in stationary stochastic processes [108]. We comment on this a bit in Section 7.

6. Oracle Processes

Let us note that Theorem 5 pertains to processes over a finite alphabet, whereas Santa Fe processes are processes over a countably infinite alphabet $\mathbb{X} \times \{0,1\}$. We need a comparably simple example of a perigraphic stationary process over a finite alphabet. For this goal, we present a novel example, which we call Oracle processes. The construction of these processes builds on Benoît Mandelbrot's and George Miller's **monkey-typing explanations** of Zipf's law [24,67]. These researchers observed that, if the characters on the type-writer keyboard are pressed at random, then the resulting text approximately obeys Zipf's law for words understood as random strings of letters delimited by spaces.

The Oracle processes are uncomputable unifilar processes with a countable number of states, which can be thought of as encoding the ergodic Santa Fe processes (8) into a finite alphabet. Similar perigraphic processes over a finite alphabet can be constructed directly through stationary variable-length coding of the Santa Fe processes, cf. [36] (Section 11.3) and [66,108], but that construction leads to much more complicated and approximate calculations. Thanks to the simplicity of Oracle processes, we prove for them an equality of all different Hilberg exponents discussed in this article—except for β_H, which is probably 0. That is, the bounds given by Theorem 5 and inequality (53) can be tight.

The construction of an Oracle process applies a memoryless source over alphabet $\{0,1,2\}$, a binary code for natural numbers $\psi : \mathbb{N} \to \{0,1\}^*$, and an oracle containing an algorithmically random sequence $z = (z_k)_{k\in\mathbb{N}}$. Using these, the Oracle source first applies the memoryless source to emit some random string $y2$, where $y \in \{0,1\}^*$ is a binary string and then it emits the corresponding bit $z_{\psi^{-1}(y)}$ read off from the oracle. Once this bit is emitted, the procedure is repeated ad infinitum. The formal definition of an Oracle process is as follows.

Definition 3 (Oracle process)**.** *Let $\psi : \mathbb{N} \to \{0,1\}^*$, where $\psi(k)$ is the binary expansion of number k stripped of the initial digit 1: $\psi(1) = \lambda$, $\psi(2) = 0$, $\psi(3) = 1$, $\psi(4) = 00$, etc. Let $\phi = \psi^{-1}$ be the inverse function. Let $z = (z_k)_{k\in\mathbb{N}}$ be an algorithmically random binary sequence. The Oracle(θ) process with a parameter $\theta \in [0,1]$ is the unifilar process defined by*

- *The set of symbols $\mathbb{X} = \{0,1,2\}$;*
- *The set of states $\mathbb{Y} = \{a,b\} \times \{0,1\}^*$;*
- *$\varepsilon(x|ay) = \theta/2$ and $\tau(ay,x) = ayx$ for $x \in \{0,1\}$ and $y \in \{0,1\}^*$;*
- *$\varepsilon(2|ay) = (1-\theta)$ and $\tau(ay,2) = by$ for $y \in \{0,1\}^*$; and*
- *$\varepsilon(z_{\phi(y)}|by) = 1$ and $\tau(by, z_{\phi(y)}) = a$ for $y \in \{0,1\}^*$.*

As we can see, the above presentation is, by definition, unifilar so realizations of the Oracle(θ) process are recognized by a deterministic push-down automaton combined with an algorithmically random oracle: First, the random output binary string y is pushed onto the stack, and upon producing symbol 2, the stack is emptied with an expectation of symbol $z_{\phi(y)}$ as a next output. Having met this expectation, the stack is ready to be refilled. The simple connection between Oracle processes and Santa Fe processes is that, if we sort random strings $y2$ according to their frequencies, then the rank-frequency distribution is approximately Zipf's distribution with exponent $\alpha = 1 - \log \theta$. This observation dates back to famous articles [24,67].

Since we have not discussed Oracle processes before, as a warm-up, let us compute the entropy rate of an Oracle process. Our proof makes use of unifilarity of this process.

Theorem 8. *The entropy rate of the stationary Oracle(θ) process equals*

$$h = \frac{h(\theta) + \theta}{2 - \theta}, \tag{57}$$

where $h(\theta) := -\theta \log \theta - (1-\theta)\log(1-\theta)$.

Now, let us proceed to the main result of this section, i.e., computing Hilberg exponents for Oracle processes and showing that they are equal and can take arbitrary values in the range $(0, 1)$. To determine Hilberg exponents $\gamma_{g,z}$, $\beta_{g,z}$, $\gamma^+_{g,z}$, and $\beta^+_{g,z}$, we use quite an obvious knowledge extractor

$$g(k, x_1^n) := \begin{cases} 0 & \text{if } 2\_\psi(k)20 \sqsubseteq x_1^n \text{ and } 2\_\psi(k)21 \not\sqsubseteq x_1^n, \\ 1 & \text{if } 2\_\psi(k)21 \sqsubseteq x_1^n \text{ and } 2\_\psi(k)20 \not\sqsubseteq x_1^n, \\ 2 & \text{else.} \end{cases} \tag{58}$$

In the above definition, symbol '_' matches any symbol.

Theorem 9. *For knowledge extractor (58) and the stationary Oracle(θ) process,*

$$\gamma_{g,z} = \beta_{g,z} = \gamma^+_{g,z} = \beta^+_{g,z} = \gamma_K = \beta_K = \beta_{\mathbb{P}} = \beta_{\mathbb{M}} = \frac{1}{1 - \log \theta} \text{ almost surely.} \tag{59}$$

As we can see by the above theorem, Oracle processes can have arbitrary Hilberg exponents in the range $(0, 1)$. In particular, they satisfy the Hilberg condition. Moreover, the unifilar order estimator (36) can diverge as a power law even for as simple unifilar processes as Oracle processes and it diverges at the slowest possible rate prescribed by the bound in Theorem 5. That is, this bound can be nontrivially tight. This tightness seems to be a new result in our little theory of perigraphic processes.

7. Discussion

In the previous sections, we stated the Hilberg hypothesis in terms of algorithmic mutual information and we showed that no finite-state statistical language model is compatible even with so generalized hypothesis, whereas there exist simple perigraphic processes, called Santa Fe and Oracle processes, which are fully compatible with Hilberg's hypothesis. Obviously, Santa Fe and Oracle processes are toy models, i.e., simple mathematical examples that are specially tailored to possess certain properties while simultaneously being easy to analyze. However, we can ask seriously how such idealized models can help with the fundamental problem of statistical language modeling. There are several specific questions that we address in the following, which entail further research hypotheses.

7.1. Is It Possible to Decide by Computation that a Given Empirical Stream of Data Satisfies the Hilberg Condition or Was Generated by a Perigraphic Source?

As we have stated informally in Section 1.1, an important open problem in the theory of perigraphic processes is whether we can consistently estimate exponent

$$\beta^+ := \sup_{g,z} \beta^+_{g,z}, \tag{60}$$

where the supremum is taken over all computable knowledge extractors g and over all algorithmically random sequences z. An analogous question pertains to exponent

β_K. Namely, the open questions are whether there exist computable functions of finite texts that return some estimates converging to β^+ or to β_K almost surely. We suppose that such functions may not exist unless we restrict ourselves to a certain subclass of stationary processes. The main difficulty here is not extracting a sort of recurrent factual knowledge from a random text but warranting that this extracted factual knowledge cannot be compressed too much or that this knowledge does not evolve slowly in time. Thus, we suppose that the property of perigraphicness or the Hilberg condition can be empirically verified only for a special subclass of stationary processes for which a certain incompressibility of extracted recurrent factual knowledge is warranted by definition. It is a matter of future research to determine whether a certain process in this subclass could resemble natural language.

7.2. Is It Plausible That Human Speech Not Only Satisfies the Hilberg Condition in a Certain Approximation but Also Resembles a Perigraphic Process?

Since we cannot give fully convincing empirical arguments, let us resort to rational ones. The motivation for perigraphic processes is of a semantic rather than a formal nature. We can probably agree that an important aim of human speech is gathering and sharing factual knowledge. We can further agree that there should be an effective procedure for extracting the factual knowledge from speech that resembles the computable knowledge extractor g from the definition of a perigraphic process. Moreover, at each time instant, we can compress the finite factual knowledge that we already have to a finite string z_1^k, which has a higher density of Kolmogorov complexity, i.e., string z_1^k is closer to being algorithmically random. Thus the doubt remains whether factual knowledge can be unbounded and whether it is possible to extract a compressed representation of factual knowledge from a spoken or written text at a power-law rate.

The answer to these questions depends on the exact nature of factual knowledge. If the factual knowledge entails the knowledge of an immutable state of a physical world that has a very high Kolmogorov complexity, then communication about this state of the physical world may be a stochastic process with a practically unbounded acquisition of factual knowledge, but the essential power-law rate of the acquisition is not secured. We suppose, however, that the vast part of the factual knowledge that we communicate about is the conventional knowledge accumulated in the gradual development of human culture. Culture can be a sort of a random virtual world that fosters conventional knowledge for the sake of itself and creates an environment in which fast accumulation of knowledge by individuals can be not only possible but also rewarded.

As we can see, the justification of perigraphic processes brings linguistics in touch with fundamental questions about the presence of algorithmic randomness in nature and in culture as well as with interactions between culture and nature. Let us also state clearly that the possible presence of algorithmic randomness in culture does not debunk its value necessarily. There are sorts of algorithmically random sequences that contain highly useful information. In the realm of mathematics, some example thereof is **Chaitin's halting probability** Ω, which is an infinite algorithmically random sequence encoding which mathematical statements are true or false [111,112]. All knowledge can be squeezed to a certain randomness but not every randomness is a useful knowledge.

Another important phenomenon that we have to face is the transience of factual knowledge transmitted by culture, i.e., there are conventional facts that become gradually forgotten. If this transience pertains to all facts transmitted through language, then the stochastic process describing language communication becomes a **mixing process** (from a subjective probability perspective) rather than a perigraphic process. However, even in this mixing case, the process may satisfy the Hilberg condition and may differ from a finite-state process. In fact, we investigated such a mixing phenomenon in the framework of Shannon information theory in [36] (Section 11.2) and [108], but it may be interesting to translate the respective phenomenon into algorithmic information theory.

7.3. What Kind of Linguistic Structures or Phenomena Do Perigraphic Processes Account for by Their Very Definition?

Stationary perigraphic processes are examples of stochastic sources in which the semantic function implies a certain formal structure. Our contribution in this domain was proving the theorems about facts and words, which state inequality $\beta^+_{g,z} \le \beta_K \le \beta_V$, where β_V is the Hilberg exponent for the expected total length of distinct words detectable in the text using the **shortest grammar-based compression**, cf. [29] and [36] (Problem 7.4). Hence, the perigraphicness of a stochastic process implies Hilberg's hypothesis and this implies discernibility of discrete words, i.e., the **double articulation**, and a Zipfian distribution of words. Since in this article, we have shown equality $\beta^+_{g,z} = \beta_{\mathbb{M}}$ for the Oracle processes, we may expect that some nice class of perigraphic processes exhibits also equality $\beta^+_{g,z} = \beta_V$. Does this mean that, in that case, we may have an approximate computable one-to-one correspondence between elementary statements (k, z_k) and words given by the shortest grammar-based compression?

It would be interesting to investigate the above question in the future since it may shed light onto origins of lexical semantics. The question matters also for a construction of **knowledge extractors** for practical statistical language models. Namely, if the number of independent elementary facts described by a text is approximately equal to the number of automatically detectable words, then an appearance of a new word in the predicted text can be a heuristic prompt for the predicting agent that a new fact needs to be added to the agent's database of acquired factual knowledge. However, the added fact need not be necessarily a description of the new word.

As for syntax, we may easily notice on the example of Santa Fe and Oracle processes that perigraphic processes need not exhibit nested hierarchical structures. All syntactic structures that we can observe in Santa Fe or Oracle processes are elementary statements (k, z_k), in which we can seek out a primitive **sentence information structure**—theme k and rheme z_k—at the very best weather. Perigraphicness, which is a sort of Zipf's law for algorithmic information, seems to be a different cause against finite-state language models than context-free syntax of an unbounded height. A mathematically plausible language system with an infinitely complex semantics can be just an infinite set of meaningful words or rather meaningful commands applied in texts at random. However, we must be a bit careful with such statements. The lack of a hierarchical structure does not mean that Oracle processes can be recognized by a finite-state automaton. To recognize Oracle processes, we need a push-down automaton with an oracle. In this simple wording, there is also a pretty complicated computer hidden that allows to look up a particular bit of the oracle corresponding to a given string on the stack.

7.4. Are There Competing Refutations of Finite-State Language Models Based on Other Quantitative Linguistic Observations?

Let us restrict ourselves to quantitative linguistic observations that can be easily operationalized by computational means and checked empirically also for abstract stochastic processes. See [46] for a justification of such a naturalistic approach to language and further examples of statistical universals in this sense.

We could probably agree that texts in natural language strongly diverge from typical outcomes of IID processes and that memory is a preformal concept that partly captures this difference. The standard way of formally defining long memory in numerical time series goes through the **power-law decay of autocorrelations** [113]. This condition can be partly adapted to categorical times series as the power-law decay of Shannon mutual information $I(X_0; X_n)$. Lin and Tegmark [114] claimed to observe such a power-law decay of mutual information for texts in natural language. Moreover, they proved that this power-law decay is incompatible with finite-state processes, and they argued that it may be be compatible with processes that exhibit hierarchical structures of an unbounded height; see also [46] for more computational experiments.

Another argument against finite-state processes applies the scaling of the **maximal repetition length** in a given text. For many mixing sources, which include finite-state processes and probably also Oracle processes, the maximal repetition length grows asymptotically similar to the logarithm of the text length [115,116]. For texts in natural language, however, it seems that the maximal repetition length grows roughly similar to the cube of the logarithm of the text length [117], which begs for an explanation, cf. [36] (Chapter 9) and [118]. We think that the cube-logarithmic scaling of the maximal repetition length is a phenomenon that may inspire interesting mathematical models of cohesive narration rather than of unbounded accumulation of factual knowledge. However, cohesive narration and knowledge accumulation can be coupled phenomena both in language and in some mathematical models thereof. There may be a common underlying mechanism for both of them.

7.5. Are There Perigraphic Processes That Satisfy All of These Quantitative Linguistic Laws and Exhibit Hierarchical Structures of an Unbounded Height?

We can meaningfully ask whether there exist simple processes that combine all statistical phenomena mentioned above and exhibit hierarchical structures. In fact, we constructed certain stochastic processes called **random hierarchical association (RHA) processes**, which seem to simultaneously exhibit the Hilberg condition, the power-law logarithmic growth of the maximal repetition length, and a bottom-up hierarchical structure of an infinite height, cf. [30] and [36] (Section 11.4). We suppose that the ergodic components of RHA processes are also perigraphic and satisfy the power-law decay of mutual information $I(X_0; X_n)$, but we have not demonstrated it yet. In [30], it was also shown that RHA processes are nonergodic and have an infinite entropy of the invariant algebra, which would be a very promising symptom since perigraphicness and strong nonergodicity are similar conditions, cf. [27] and [36] (Section 8.3). Our definition of RHA processes is quite complicated, however, which makes them difficult to analyze, and we are not sure whether all results in [30] are correct. Probably the construction should be somewhat simplified in order to obtain more conclusive and convincing results.

7.6. How Can We Improve Practical Statistical Language Models Using Ideas Borrowed from Perigraphic Processes?

Since perigraphic processes satisfy the power-law growth of algorithmic mutual information, the expected conditional Kolmogorov complexity of the next symbol given a finite past tends toward the entropy rate very slowly with respect to the length of the past. This means than the optimal predicting agent never stops learning from a perigraphic process and its memory load grows unboundedly. If natural language resembles a perigraphic process, the pretty obvious message for practitioners of statistical language models is that they should never switch off their training. The power-law tails of learning curves, observed in [48,50–53], may be something more fundamental than just an accidental empirical law. With each new input, a portion of factual knowledge may come that may be useful for the prediction of subsequent inputs. However, obviously, not all input information should be memorized since most of it is random noise. Here, the theorems about facts and words [27,29] may help us. As we suggested in Section 7.3, a simple heuristic prompt for a statistical model to add a new fact to the database of factual knowledge may be the appearance of a new word type or rather of a new term—since "words" in this context are defined by the **shortest grammar-based compression** [29,60,61] and they can be morphemes or multiword expressions [55]. Moreover, this new fact need not be a description of the new term but rather a sort of reaction to it.

The detailed mechanism of factual knowledge extraction may be different for different perigraphic processes. Hence, while constructing practical statistical language models, it may be useful to draw various inspirations from information theory, probability, logic, statistical laws of language, and neuroscience. Let us stress, however, that the problem of factual knowledge extraction is closely linked to the problem of estimating exponent β^+, discussed in Section 7.1. In particular, if we had a single **knowledge extractor** that works for a reasonable subclass of stationary processes, then by compressing the extracted knowledge, we could find a desired lower bound for the number of distinct time-independent facts necessary to verify the perigraphicness property. We notice that finding such a universal knowledge extractor is a different problem than constructing the minimal unifilar representation of the process, called the ϵ-**machine** in [72–74], but there may be some connections between these two tasks, cf. [103]. The relationship between the universal knowledge extractor and the ϵ-machine may be analogous to the difference between the Gács and Körner common information [109] and the Wyner common information [110]. The former is a lower bound for the learning problem, whereas the latter is an upper bound.

8. Conclusions

Recapitulating this article, we suppose that refuting finite-state language models through various power laws for algorithmic information yields some fresh insight into human (and maybe not necessarily only human) language. We hypothesize this despite dealing explicitly with some abstract mathematical models. Our novel refutation is of a semantic rather than a syntactic nature and rests on a hypothetical Zipf law for independent elementary meanings. We think that this is an interesting feature since semantics precedes syntax in communication whereas advanced syntax is a later evolved mechanism that makes the mapping between signals and complex meanings more fault tolerant (and more redundant, by the way). In fact, syntax can be also investigated using ideas from information theory [119,120].

We hope that perigraphic processes can be an important mathematical model that may bring information theory and linguistics closer. Even if perigraphic processes turn out not to be realistic models of human language in the course of future investigations, they point out a research direction in which formal semantics and the structure of human languages can be fruitfully combined with information theory. What seems also interesting in this framework is that we may also ask metalinguistic questions such as what kind of theories of language can be potentially finite—such as unbounded lexicons vs. finite universal grammars. The Chomskyan linguistics stressed the importance of finite theories of language learning, which is a great interdisciplinary research question, but from the perspective of an ever-learning language user, divergent language theories such as bloated dictionaries or imprecise school grammars can be very useful, too—and they should not be abandoned.

Funding: The research work was partly supported by the National Science Centre Poland grant No. 2018/31/B/HS1/04018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Jan Mielniczuk, Damian Niwiński, Elisabeth Gassiat, Peter Grünwald, Marcelo Weinberger, Richard Futrell, Adam Przepiórkowski, Magdalena Derwojedowa, and Joanna Rączaszek-Leonardi for providing feedback concerning the mathematical results presented in this article and their presentation in the linguistic context. The article has also benefited from the suggestions of anonymous reviewers.

Conflicts of Interest: The present article is based on two earlier manuscripts: (a) *On a Class of Markov Order Estimators Based on PPM and Other Universal Codes* (https://arxiv.org/abs/2003.04754 (accessed on 29 August 2021)) and (b) *Bounds for Algorithmic Mutual Information and a Unifilar Order Estimator* (https://arxiv.org/abs/2011.12845 (accessed on 29 August 2021)). Due to their various defects in acknowledging the state-of-the-art in information theory, we decided not to publish them in a journal or at a conference. In particular, manuscript (a) discusses an apparently novel consistent Markov (not hidden Markov) order estimator, which is roughly known in the information-theoretic literature. What is more novel is that manuscript (a) proves an upper bound for algorithmic mutual information in terms of the Markov order estimator and the respective number of distinct subwords. By contrast, manuscript (b) builds an analogous theory for unifilar hidden Markov processes, which is partly witnessed in the information-theoretic literature as well. The additional, more novel contribution of manuscript (b) concerns perigraphic and Oracle processes. In terms of mathematical content, manuscript (b) is almost equivalent to the present article but it contains much fewer language-oriented passages. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Proofs

Proof of Theorem 1. Write $\delta = \operatorname{hilb}_{n\to\infty} \mathfrak{I}(n)$. The proof of

$$\operatorname*{hilb}_{n\to\infty}(\mathfrak{S}(n) - n\mathfrak{s}) \le \delta \tag{A1}$$

can be found in [36] (Theorem 8.2). (Notice that, in [27], which proves apparently the same statement, the Hilberg exponent is defined on a sparse subsequence.) Now assume that $\mathfrak{S}(n) - n\mathfrak{s} \ge -C$ for all but finitely many n. We have then

$$\mathfrak{S}(n) - n\mathfrak{s} = \frac{\mathfrak{I}(n)}{2} + \frac{\mathfrak{S}(2n) - 2n\mathfrak{s}}{2} \ge \frac{\mathfrak{I}(n) - C}{2} \tag{A2}$$

for a sufficiently large n. Hence, $\delta \le \operatorname{hilb}_{n\to\infty}(\mathfrak{S}(n) - n\mathfrak{s})$. Thus, we obtain the equality in (19). □

Proof of Theorem 2. Let us denote the set of distinct maximum likelihood parameters

$$\mathcal{P}_{kn} := \left\{(y,\tau,\varepsilon) : \exists_{x_1^n} \mathbb{P}(x_1^n|k,y,\tau,\varepsilon) = \hat{\mathbb{P}}(x_1^n|k)\right\}. \tag{A3}$$

As explained in [81], we can bound

$$\mathbb{C}(n|k) = \log\left(\sum_{x_1^n\in\mathbb{X}^n} \hat{\mathbb{P}}(x_1^n|k)\right) \le \log|\mathcal{P}_{kn}|. \tag{A4}$$

Given a fixed (y,τ), likelihood $\mathbb{P}(x_1^n|k,y,\tau,\varepsilon)$ is maximized for the empirical distribution

$$\varepsilon(b|a) = \frac{\sum_{i=1}^n \mathbf{1}\{(y_i,x_i) = (a,b)\}}{\sum_{i=1}^n \mathbf{1}\{y_i = a\}}, \tag{A5}$$

where $y_1 = y$ and $y_i = \tau(y_{i-1}, x_{i-1})$ for $i \ge 2$. Since there are $k \cdot k^{k|\mathbb{X}|}$ possible values of pairs (y,τ), whereas for a fixed pair (y,τ) there are less than $(n+1)^{k|\mathbb{X}|}$ distinct empirical distributions ε, we can bound the family complexity of the unifilar hidden Markov family as

$$\mathbb{C}(n|k) \le \log|\mathcal{P}_{kn}| \le \log\left(k^{k|\mathbb{X}|+1}(n+1)^{k|\mathbb{X}|}\right) \le [k|\mathbb{X}|+1]\log[k(n+1)]. \tag{A6}$$

□

Proof of Theorem 3. Letting $\tau(x_1^k, x_{k+1}) = x_2^{k+1}$ and $\varepsilon(x_{k+1}|x_1^k) = P(X_{k+1} = x_{k+1}|X_1^k = x_1^k)$, we obtain the conditional probability bound

$$\hat{\mathbb{P}}(X_1^n||\mathbb{X}|^k) \ge \mathbb{P}(X_1^n||\mathbb{X}|^k, X_{-k+1}^0, \tau, \varepsilon) = \prod_{i=1}^n P(X_i|X_{i-k}^{i-1}). \tag{A7}$$

Hence, by the upper bound (33) for the family complexity and by the Birkhoff ergodic theorem, we obtain

$$\begin{aligned}&\limsup_{n\to\infty} \frac{1}{n}\left[-\log\mathbb{P}(X_1^n||\mathbb{X}|^k)\right] \\ &\quad= \limsup_{n\to\infty} \frac{1}{n}\left[-\log\hat{\mathbb{P}}(X_1^n||\mathbb{X}|^k) + \mathbb{C}(n||\mathbb{X}|^k)\right] \le h_k \text{ almost surely.}\end{aligned} \tag{A8}$$

Thus, by the upper bound

$$-\log \mathbb{P}(x_1^n) \leq -\log \mathbb{P}(x_1^n|k) - \log w_k. \tag{A9}$$

and by the Barron lemma [121] (Theorem 3.1), we obtain (34). Noticing that $\mathbb{P}(X_1^n) \geq w_n \mathbb{P}(X_1^n|n) = w_n |\mathbb{X}|^{-n}$, we hence obtain (35) by dominated convergence. □

Proof of Theorem 4. Our proof of consistency (37) is split into two separate impossibility proofs for overestimation and for underestimation.

The bound for the overestimation probability (considering the case of $M_U < \infty$ is sufficient) is received by inequality $\hat{\mathbb{P}}(X_1^n|M_U) \geq P(X_1^n|Y_1)$, where Y_1 is the hidden state emitting X_1, and by the Barron lemma [121] (Theorem 3.1). Hence,

$$\begin{aligned} P(\mathbb{M}(X_1^n) > M_U) &\leq P\left(\hat{\mathbb{P}}(X_1^n|M_U) < w_n \mathbb{P}(X_1^n)\right) \\ &\leq P\left(\frac{w_n \mathbb{P}(X_1^n)}{P(X_1^n|Y_1)} > 1\right) \leq w_n. \end{aligned} \tag{A10}$$

Since $\sum_{n=1}^{\infty} w_n = 1$, the impossibility of overestimation follows by the Borel–Cantelli lemma.

Now, we demonstrate the impossibility of underestimation for both $M_U < \infty$ and $M_U = \infty$, which is more involved. Since the Ryabko mixture is universal in the sense of (34) and the penalty is $-\log w_n = o(n)$, it is sufficient to show that

$$\liminf_{n\to\infty} \frac{1}{n}\left[-\log \hat{\mathbb{P}}(X_1^n|k)\right] > h \text{ almost surely for } k < M_U. \tag{A11}$$

Our reasoning proceeds by showing that the left-hand side of the above inequality equals almost surely a sort of conditional entropy $h_{[k]}$, which is strictly greater than h if $k < M_U$.

We observe first that, for any finite set $\mathcal{M}$,

$$\liminf_{n\to\infty} \min_{m\in\mathcal{M}} a_{nm} = \min_{m\in\mathcal{M}} \liminf_{n\to\infty} a_{nm}. \tag{A12}$$

It is so since we can take sufficiently large n on both sides to interchange the infimums. In our case, the set of pairs (y,τ) for a fixed k is finite. Hence,

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n}\left[-\log \hat{\mathbb{P}}(X_1^n|k)\right] &= \liminf_{n\to\infty} \min_{y,\tau,\varepsilon} \frac{1}{n}[-\log \mathbb{P}(X_1^n|k,y,\tau,\varepsilon)] \\ &= \min_{y,\tau} \liminf_{n\to\infty} \min_{\varepsilon} \frac{1}{n}[-\log \mathbb{P}(X_1^n|k,y,\tau,\varepsilon)]. \end{aligned} \tag{A13}$$

Notice that, here, we cannot apply Lemmas 1–7 from Gassiat and Boucheron [101] although they relate to an expression partly resembling (A13) for $M_U < \infty$ (the order of the limit on n and the minimization on ε are interchanged).

To deal also with $M_U = \infty$ and to circumvent the difficulty of directly using results from [101], we apply a technically difficult but beautiful result by Kieffer and Rahe [105], which says that an ergodic Markov channel applied to an ergodic asymptotically mean stationary process yields a jointly ergodic asymptotically mean stationary process. Denote $Y_1^{y,\tau} := y \in \{1,\ldots,k\}$ and $Y_{i+1}^{y,\tau} := \tau(Y_i^{y,\tau}, X_i)$. We can see that the distribution of $(Y_i^{y,\tau})_{i\in\mathbb{N}}$ given $(X_i)_{i\in\mathbb{N}}$ is an ergodic Markov channel, whereas process $(X_i)_{i\in\mathbb{N}}$ is stationary ergodic. Thus, process $(X_i, Y_i^{y,\tau})_{i\in\mathbb{N}}$ is asymptotically mean stationary ergodic. Let process $(\bar{X}_i, \bar{Y}_i^{y,\tau})_{i\in\mathbb{Z}}$ be distributed according to the stationary mean of $(X_i, Y_i^{y,\tau})_{i\in\mathbb{N}}$. Since $(X_i)_{i\in\mathbb{N}}$ is stationary, we can assume without loss of generality that $(\bar{X}_i)_{i\in\mathbb{N}} = (X_i)_{i\in\mathbb{N}}$. Moreover, by definition of the stationary mean, recursion $\bar{Y}_{i+1}^{y,\tau} = \tau(\bar{Y}_i^{y,\tau}, X_i)$ holds by recursion $Y_{i+1}^{y,\tau} = \tau(Y_i^{y,\tau}, X_i)$. (Notice, however, that we cannot assume $\bar{Y}_i^{y,\tau} = \sigma_{y,\tau}(X_{-\infty}^{i-1})$ since there is a simple counterexample: a periodic process $(Y_i^{y,\tau})_{i\in\mathbb{N}}$ with a constant process $(X_i)_{i\in\mathbb{N}}$.)

The beauty of asymptotically mean stationary processes lies in the fact that we have a generalization of the Birkhoff ergodic theorem [122]. The claim is that the Cesàro averages converge almost surely to expectations with respect to the stationary mean. Hence, by the application of the Birkhoff ergodic theorem to empirical counts in the most likely distribution ε given (y, τ), we obtain

$$\begin{aligned}
&\min_{y,\tau} \liminf_{n\to\infty} \min_{\varepsilon} \frac{1}{n}[-\log \mathbb{P}(X_1^n|k,y,\tau,\varepsilon)] \\
&\quad = \min_{y,\tau} \liminf_{n\to\infty} \min_{\varepsilon} \frac{1}{n} \sum_{i=1}^{n} \left[-\log \varepsilon(X_i|Y_i^{y,\tau})\right] \\
&\quad = h_{[k]} := \min_{y,\tau} \mathbf{E}\left[-\log P(X_i|\bar{Y}_i^{y,\tau})\right] \text{ almost surely.}
\end{aligned} \tag{A14}$$

Let y and τ be some minimizing parameters, and let us abbreviate $\bar{Y}_i := \bar{Y}_i^{y,\tau}$. What happens if $h_{[k]} := H(X_i|\bar{Y}_i) = h$? Since $\bar{Y}_{i+1} = \tau(\bar{Y}_i, X_i)$, we can write

$$H(X_i|\bar{Y}_i) - H(X_i|X_1^{i-1}, \bar{Y}_1) = I(X_i; X_1^{i-1}, \bar{Y}_1^{i-1}|\bar{Y}_i). \tag{A15}$$

Hence, by stationarity of $(X_i, \bar{Y}_i^{y,\tau})_{i\in\mathbb{Z}}$,

$$\sum_{i=1}^{n} H(X_i|\bar{Y}_i) = H(X_1^n|\bar{Y}_1) + \sum_{i=0}^{n-1} I(X_j; X_{j-i}^{j-1}, \bar{Y}_{j-i}^{j-1}|\bar{Y}_j). \tag{A16}$$

Dividing by n and letting $n \to \infty$ yields

$$h_{[k]} := H(X_i|\bar{Y}_i) = h + I(X_j; X_{-\infty}^{j-1}, \bar{Y}_{-\infty}^{j-1}|\bar{Y}_j), \tag{A17}$$

where we freely apply Shannon information measures for arbitrary σ-fields. (Their properties were described in [123,124].) That is, if $h_{[k]} = h$, then $I(X_i; X_{-\infty}^{i-1}, \bar{Y}_{-\infty}^{i-1}|\bar{Y}_i) = 0$. Since also $\bar{Y}_{i+1} = \tau(\bar{Y}_i, X_i)$, then $(X_i)_{i\in\mathbb{N}}$ is a unifilar hidden Markov process with $\leq k$ hidden states distributed according to $\bar{Y}_i$. Consequently, we have $M_U \leq k$.

Subsequently, let us prove the asymptotic unbiasedness (38). By $\mathbb{M}(x_1^n) \leq n$ and by the overestimation bound (A10), we have

$$\mathbf{E}\,\mathbb{M}(X_1^n) \leq M_U + nP(\mathbb{M}(x_1^n) > M_U) = M_U + \frac{1}{n+1}. \tag{A18}$$

On the other hand, by the Fatou lemma,

$$M_U = \mathbf{E} \liminf_{n\to\infty} \mathbb{M}(X_1^n) \leq \liminf_{n\to\infty} \mathbf{E}\,\mathbb{M}(X_1^n). \tag{A19}$$

Hence, claim (38) follows. □

Proof of Theorem 5. The maximum log-likelihood is subadditive:

$$-\log \hat{\mathbb{P}}(x_1^n|k) - \log \hat{\mathbb{P}}(x_{n+1}^{n+m}|k) + \log \hat{\mathbb{P}}(x_1^{n+m}|k) \leq 0. \tag{A20}$$

Denoting $k := \mathbb{M}(X_1^{2n}) \leq 2n$, we observe by (36), (A9), and (A20) that

$$\begin{aligned}
&-\log \mathbb{P}(X_1^n) - \log \mathbb{P}(X_{n+1}^{2n}) + \log \mathbb{P}(X_1^{2n}) \\
&\quad \leq -\log \mathbb{P}(X_1^n|k) - \log \mathbb{P}(X_{n+1}^{2n}|k) - 2\log w_k + \log \hat{\mathbb{P}}(X_1^{2n}|k) - \log w_{2n} \\
&\quad \leq 2\mathbb{C}(n|k) - 2\log w_k - \log w_{2n} - \log \hat{\mathbb{P}}(X_1^n|k) - \log \hat{\mathbb{P}}(X_{n+1}^{2n}|k) + \log \hat{\mathbb{P}}(X_1^{2n}|k) \\
&\quad \leq 2\mathbb{C}(n|k) - 2\log w_k - \log w_{2n}.
\end{aligned} \tag{A21}$$

Hence, by bound (33) for family complexity, we obtain

$$\beta_{\mathbb{P}} \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, \mathbb{C}(n|\mathbb{M}(X_1^{2n})) \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, \mathbb{M}(X_1^n). \tag{A22}$$

□

Proof of Theorem 6. Let $(X_i)_{i\in\mathbb{Z}}$ be a hidden Markov process over a finite alphabet $\mathbb{X}$ with respect to a Markov process $(Y_i)_{i\in\mathbb{Z}}$ over a finite alphabet $\mathbb{Y}$. Then, process $(\tilde{X}_i)_{i\in\mathbb{Z}}$ given by $\tilde{X}_i = (X_i, Y_i)$ is also a Markov process over a finite alphabet $\mathbb{X}\times\mathbb{Y}$. Since $X_i = f(\tilde{X}_i)$ for a computable function f, the data-processing inequality for algorithmic mutual information yields

$$J(X_{-n+1}^0; X_1^n) \le J(\tilde{X}_{-n+1}^0; \tilde{X}_1^n) + C, \tag{A23}$$

where constant C does not depend on n. Process $(\tilde{X}_i)_{i\in\mathbb{Z}}$ is a Markov process over a finite alphabet so it has finitely many ergodic components, which can be written as $P(A) = \sum_{i=1}^k \nu_i F_i(A)$. Thus, by asymptotic unbiasedness (38) of the unifilar order estimator on each ergodic component, we obtain

$$\begin{aligned}\lim_{n\to\infty} \mathbf{E}\, \mathbb{M}(\tilde{X}_1^n) &= \lim_{n\to\infty} \int \sum_{i=1}^k \nu_i \mathbb{M}(\tilde{X}_1^n) dF \\ &= \sum_{i=1}^k \nu_i \lim_{n\to\infty} \int \mathbb{M}(\tilde{X}_1^n) dF = \sum_{i=1}^k \nu_i \tilde{M}_U^i \le \tilde{M}_U < \infty,\end{aligned} \tag{A24}$$

since $\tilde{M}_U^i \le \tilde{M}_U$, where $\tilde{M}_U^i$ is the unifilar order of the ith ergodic component and $\tilde{M}_U$ is the unifilar order of process $(\tilde{X}_i)_{i\in\mathbb{Z}}$. Hence, by (A23), Theorem 5, and (A24), we obtain

$$\beta_K \le \tilde{\beta}_K \le \tilde{\beta}_{\mathbb{P}} \le \tilde{\beta}_{\mathbb{M}} = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, \mathbb{M}(\tilde{X}_1^n) = 0, \tag{A25}$$

where the quantities with the tilde pertain to process $(\tilde{X}_i)_{i\in\mathbb{Z}}$. □

Proof of Theorem 7. Inequalities $\beta_{g,z}^+ \ge \gamma_{g,z}^+$ and $\beta_{g,z} \ge \gamma_{g,z}$ follow by the general property of Hilberg exponents: If $S_n \le S_{n+1}$ holds for a sequence of random variables S_n, then $\operatorname{hilb}_{n\to\infty} S_n \le \operatorname{hilb}_{n\to\infty} \mathbf{E}\, S_n$ almost surely [36] (Theorem 8.4). As for claim $\beta_{g,z}^+ = \beta_{g,z}$, let us observe that

$$\begin{aligned}U_{1,n}^{g,z} \ge L_n^{g,z} = \min\left\{U_{-n+1,0}^{g,z}, U_{1,n}^{g,z}\right\} &= U_{-n+1,0}^{g,z} + U_{1,n}^{g,z} - \max\left\{U_{-n+1,0}^{g,z}, U_{1,n}^{g,z}\right\} \\ &\ge U_{-n+1,0}^{g,z} + U_{1,n}^{g,z} - U_{-n+1,n}^{g,z}.\end{aligned} \tag{A26}$$

Applying expectations and stationarity yields

$$\mathbf{E}\, U_{1,n}^{g,z} \ge \mathbf{E}\, L_n^{g,z} \ge 2\,\mathbf{E}\, U_{1,n}^{g,z} - \mathbf{E}\, U_{1,2n}^{g,z}. \tag{A27}$$

Consequently, $\beta_{g,z}^+ = \beta_{g,z}$ follows by Theorem 1 if $\lim_{n\to\infty} \mathbf{E}\, U_{1,n}^{g,z}/n = 0$. □

Proof of Theorem 8. A unifilar process $(X_i)_{i\in\mathbb{N}}$ is stationary and extendable to a stationary process $(X_i, Y_i)_{i\in\mathbb{Z}}$ if

$$\sum_{x_1,y_1} \pi(y_1)\varepsilon(x_1|y_1)\mathbf{1}\{y_2 = \tau(y_1, x_1)\} = \pi(y_2). \tag{A28}$$

Using Equation (A28), we can easily determine the stationary initial distribution π as $\pi(ay) = \pi(a)\left(\frac{\theta}{2}\right)^{|y|}$, $\pi(by) = \pi(ay)(1-\theta)$, and $\pi(a) = (1-\theta)/(2-\theta)$. Subsequently, we recall that the entropy rate of a process $(X_i)_{i\in\mathbb{N}}$ unifilar with respect to a stationary process $(Y_i)_{i\in\mathbb{N}}$ with entropy $H(Y_i) = -\sum_{y\in\mathbb{Y}} \pi(y) \log \pi(y) < \infty$ equals

$$h = \sum_{y\in\mathbb{Y}} \pi(y)\left[-\sum_{x\in\mathbb{X}} \varepsilon(x|y)\log\varepsilon(x|y)\right], \tag{A29}$$

cf. [125–127]. (The entropy rate of a nonunifilar hidden Markov process is much more difficult to compute [128–131].) In our case, we have

$$\begin{aligned} H(Y_i) &= -\sum_{y\in\{a,b\}\times\{0,1\}^*} \pi(y)\log\pi(y) \\ &= \sum_{y\in\{0,1\}^*} [-(2-\theta)\pi(ay)\log\pi(ay) - (1-\theta)\pi(ay)\log(1-\theta)] \\ &= \frac{1-\theta}{2-\theta}\sum_{k=0}^{\infty}\left[-(2-\theta)\theta^k k\log\frac{\theta(1-\theta)}{2(2-\theta)} - (1-\theta)\theta^k\log(1-\theta)\right] \\ &= \frac{(1-\theta)}{(2-\theta)}\left[-\frac{(2-\theta)\theta}{(1-\theta)^2}\log\frac{\theta(1-\theta)}{2(2-\theta)} - \log(1-\theta)\right] \\ &= \frac{(2-\theta)\theta[-\log\theta + 1 + \log(2-\theta)] - \log(1-\theta)}{(1-\theta)(2-\theta)}. \end{aligned} \tag{A30}$$

Since this entropy is finite, we can compute the entropy rate by (A29) as

$$\begin{aligned} h &= \sum_{y\in\{0,1\}^*} \pi(ay)\left[-\theta\log\frac{\theta}{2} - (1-\theta)\log(1-\theta)\right] \\ &= \pi(a)\sum_{n=0}^{\infty}\theta^n[h(\theta)+\theta] = \frac{h(\theta)+\theta}{2-\theta}. \end{aligned} \tag{A31}$$

□

Proof of Theorem 9. Let us observe that, for knowledge extractor (58), we have $U^{g,z}_{m,n} \le U^{g,z}_{m,n+1}, U^{g,z}_{m-1,n}$, so inequalities (54)–(55) apply. Thus, by Theorem 5 and inequalities (49)–(52), it suffices to show $\beta \le \gamma_{g,z}$ and $\beta_{\mathbb{M}} \le \beta$, where

$$\beta := \frac{1}{1-\log\theta}. \tag{A32}$$

The proof of $\beta \le \gamma_{g,z}$ applies techniques developed in [27] for Santa Fe processes. The proof of $\beta_{\mathbb{M}} \le \beta$ uses some ideas from [108] derived also for Santa Fe processes. For both goals of the proof, we apply random variables $W_i \in \{0,1\}^*$ and $Z_i \in \{0,1\}$ constructed through parsing

$$X_1^\infty = R_0 W_0 2 Z_0 W_1 2 Z_1 W_2 2 Z_2 W_3 2 Z_3 \dots, \tag{A33}$$

$$X_{-\infty}^0 = \dots W_{-3} 2 Z_{-3} W_{-2} 2 Z_{-2} W_{-1} 2 Z_{-1} R_{-1}, \tag{A34}$$

where R_{-1} and R_0 are the shortest random strings such that these equalities are satisfied. Obviously, $Z_i = z_{\phi(W_i)}$ for the Oracle(θ) process. By contrast, by the strong Markov property, random variables $(W_i)_{i\in\mathbb{Z}\setminus\{0\}}$ form an IID process, where

$$P(W_i = y) = (1-\theta)\left(\frac{\theta}{2}\right)^{|y|}. \tag{A35}$$

Write $N_n^\pm := 2n + \sum_{i=1}^n |W_{\pm i}|$. Since

$$\mathbf{E}|W_i| = (1-\theta)\sum_{y\in\{0,1\}^*}|y|\left(\frac{\theta}{2}\right)^{|y|} = (1-\theta)\sum_{n=0}^{n} n\theta^n = \frac{\theta}{1-\theta}, \tag{A36}$$

we have $\mathbf{E}\, N_n^\pm = \rho\, n$, where $\rho := \frac{2-\theta}{1-\theta}$. Since $\mathbf{E}|W_i| < \infty$, then by the Birkhoff ergodic theorem, we obtain $\lim_{n\to\infty} N_n^\pm / n = \rho$ almost surely. Additionally, if we define $\tilde{N}_n^\pm := \max\{m : N_m^\pm \le n\}$, then $\lim_{n\to\infty} \tilde{N}_n^\pm / n = \rho^{-1}$ almost surely.

To demonstrate the first goal of the proof, we define

$$U_n^{\pm} := \min\{k \geq 1 : \psi(k) \notin \{W_{\pm i}\}_{i=1}^{n}\}. \tag{A37}$$

We see that $L_n^{g,z} = \min\left\{U^-_{\tilde{N}_n^-}, U^+_{\tilde{N}_n^+}\right\}$. Since $\lim_{n\to\infty} \tilde{N}_n^{\pm}/n = \rho^{-1}$ almost surely and $U_n^{\pm}$ is a nondecreasing function of n, we obtain

$$\gamma_{g,z} = \operatorname*{hilb}_{n\to\infty} L_n^{g,z} \geq \liminf_{n\to\infty} \frac{\log \min\{U_n^-, U_n^+\}}{\log n} \text{ almost surely.} \tag{A38}$$

Consequently, so as to bound $U_n^{\pm}$, we observe

$$\begin{aligned} P(U_n^{\pm} < 2^m) &\leq \sum_{k=1}^{2^m-1} P(\psi(k) \notin \{W_i\}_{i=1}^n) = \sum_{y\in\{0,1\}^{<m}} P(y \notin \{W_i\}_{i=1}^n) \\ &= \sum_{k=0}^{m-1} 2^k \left(1 - (1-\theta)\left(\frac{\theta}{2}\right)^k\right)^n \leq 2^m \left(1 - (1-\theta)\left(\frac{\theta}{2}\right)^m\right)^n \\ &\leq 2^m \exp\left(-(1-\theta)n\left(\frac{\theta}{2}\right)^m\right) = 2^m \exp\left(-(1-\theta)n2^{-m/\beta}\right). \end{aligned} \tag{A39}$$

Putting $m_n = \beta(1-\epsilon)\log n$ for an arbitrary $\epsilon > 0$, we obtain

$$\sum_{n=1}^{\infty} P(U_n^- < 2^{m_n} \text{ or } U_n^+ < 2^{m_n}) \leq 2\sum_{n=1}^{\infty} n^{\beta(1-\epsilon)} \exp(-(1-\theta)n^{\epsilon}) < \infty. \tag{A40}$$

Hence, by the Borel–Cantelli lemma, we obtain

$$\beta \leq \liminf_{n\to\infty} \frac{\log \min\{U_n^-, U_n^+\}}{\log n} \leq \gamma_{g,z} \text{ almost surely.} \tag{A41}$$

Thus, we accomplished the first goal of the proof.

To demonstrate the second goal of the proof, let us define

$$M_n := \sum_{y\in\{0,1\}^*} (|y|+2)\mathbf{1}\{y \in \{W_i\}_{i=1}^n\}. \tag{A42}$$

We observe $\hat{\mathbb{P}}(X_1^n|M_{\tilde{N}_n^+} + C_n) \geq P(X_1^n|Y_1)$, where Y_1 is the hidden state emitting X_1 and $C_n = |R_0W_02Z_0| + \left|W_{\tilde{N}_n^+}2Z_{\tilde{N}_n^+}\right|$. It is so since we can express probability $P(X_1^n|Y_1)$ as probability of a unifilar process with $M_{\tilde{N}_n^+} + C_n$ hidden states. Thus, by the Barron lemma [121] (Theorem 3.1), we obtain

$$\begin{aligned} P\left(\mathbb{M}(X_1^n) > M_{\tilde{N}_n^+} + C_n\right) &\leq P\left(\hat{\mathbb{P}}(X_1^n|M_{\tilde{N}_n^+} + C_n) < w_n\mathbb{P}(X_1^n)\right) \\ &\leq P\left(\frac{w_n\mathbb{P}(X_1^n)}{P(X_1^n|Y_1)} > 1\right) \leq w_n. \end{aligned} \tag{A43}$$

Since $\mathbb{M}(X_1^n) \leq n$ holds uniformly,

$$\mathbf{E}\,\mathbb{M}(X_1^n) \leq \mathbf{E}\,M_{\tilde{N}_n^+} + \mathbf{E}\,C_n + nw_n. \tag{A44}$$

We observe that $\mathbf{E}\,C_n + nw_n$ is bounded as a function of n, so it suffices to take care of $\mathbf{E}\,M_{\tilde{N}_n^+}$. Since M_n is a nondecreasing function of n and $\tilde{N}_n^+ < n$, we may further bound

$$\mathbf{E}\,M_{\tilde{N}_n^+} \leq \mathbf{E}\,M_n. \tag{A45}$$

Hence,

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\,\mathbb{M}(X_1^n) \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,M_{\tilde{N}_n^+} \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,M_n. \tag{A46}$$

Consequently, so as to bound $\mathbf{E}\, M_n$, we notice

$$\begin{aligned} \mathbf{E}\, M_n &= \sum_{y \in \{0,1\}^*} (|y|+2) P(y \in \{W_i\}_{i=1}^n) \\ &= \sum_{k=0}^{\infty} (k+2) 2^k \left(1 - \left(1 - (1-\theta)\left(\frac{\theta}{2}\right)^k\right)^n\right) \\ &\le \sum_{k=0}^{\infty} (k+2) 2^k \left(1 - \left(1 - 2^{-k/\beta}\right)^n\right). \end{aligned} \tag{A47}$$

Hence, adapting the computations from the proof of Proposition 1 by [108], we obtain up to a small constant

$$\begin{aligned} \mathbf{E}\, M_n &\lesssim \int_0^{\infty} (k+2) 2^k \left(1 - \left(1 - 2^{-k/\beta}\right)^n\right) dk \\ &= \frac{1}{\ln 2} \int_1^{\infty} (\log p + 2) \left(1 - \left(1 - p^{-1/\beta}\right)^n\right) dp \quad \left\{p := 2^k\right\} \\ &= \frac{\beta^2}{\ln 2} \int_0^1 \frac{(1-u)(\log(1-u^{1/n})^{-1} + 2) du}{u^{1-1/n} n (1-u^{1/n})^{\beta+1}} \quad \left\{u := \left(1 - p^{-1/\beta}\right)^n\right\} \\ &= \frac{\beta^2 n^\beta (\log n + 2)}{\ln 2} \int_0^1 f_n(u) du + \frac{\beta^2 n^\beta}{\ln 2} \int_0^1 g_n(u) du, \end{aligned} \tag{A48}$$

where we denote functions

$$f_n(u) := \frac{(1-u)}{u^{1-1/n} [n(1-u^{1/n})]^{\beta+1}}, \quad g_n(u) := f_n(u) \log[n(1-u^{1/n})]^{-1}. \tag{A49}$$

These functions tend to limits

$$\lim_{n\to\infty} f_n(u) = f(u) := \frac{(1-u)}{u(-\ln u)^{\beta+1}}, \quad \lim_{n\to\infty} g_n(u) = g(u) := f(u) \log(-\ln u)^{-1}. \tag{A50}$$

We notice upper bounds $f_n(u) \le f(u)$ and $g_n(u) \le g_1(u)$ for $u \in (0,1)$. Moreover, functions $f(u)$ and $g_1(u)$ are integrable on $u \in (0,1)$. Indeed putting $t := -\ln u$ and integrating by parts yields

$$\begin{aligned} \int_0^1 f(u) du &= \int_0^{\infty} (1-e^{-t})\, t^{-\beta-1} dt \\ &= (1-e^{-t})(-\beta^{-1}) t^{-\beta} |_0^{\infty} + \int_0^{\infty} e^{-t} \beta^{-1} t^{-\beta} dt = \beta^{-1} \Gamma(1-\beta), \end{aligned} \tag{A51}$$

whereas putting $t = 1 - u$ and integrating by parts yields

$$\begin{aligned} \int_0^1 g_1(u) du &= -\int_0^1 \frac{\log t}{t^\beta} dt \\ &= -(\log t)(1-\beta)^{-1} t^{1-\beta} |_0^1 + \int_0^1 (1-\beta)^{-1} t^{-\beta} dt = (1-\beta)^{-2}. \end{aligned} \tag{A52}$$

Hence, we derive

$$\underset{n\to\infty}{\mathrm{hilb}}\, \mathbf{E}\, \mathbb{M}(X_1^n) \le \underset{n\to\infty}{\mathrm{hilb}}\, \mathbf{E}\, M_n \le \beta. \tag{A53}$$

This completes the second goal of the proof. □

References

1. Skinner, B.F. *Verbal Behavior*; Prentice Hall: Englewood Cliffs, NJ, USA, 1957.
2. Chomsky, N. Three models for the description of language. *IRE Trans. Inf. Theory* **1956**, *2*, 113–124. [CrossRef]

3. Chomsky, N. *Syntactic Structures*; Mouton & Co.: The Hague, The Netherland, 1957.
4. Chomsky, N. A Review of B. F. Skinner's Verbal Behavior. *Language* **1959**, *35*, 26–58. [CrossRef]
5. Chomsky, N.; Miller, G. Finite State Languages. *Inf. Control.* **1959**, *1*, 91–112. [CrossRef]
6. Pereira, F. Formal Grammar and Information Theory: Together Again? *Philos. Trans. R. Soc. Lond. Ser. A* **2000**, *358*, 1239–1253. [CrossRef]
7. Jelinek, F. Continuous speech recognition by statistical methods. *Proc. IEEE* **1976**, *64*, 532–556. [CrossRef]
8. Jelinek, F. *Statistical Methods for Speech Recognition*; MIT Press: Cambridge, MA, USA, 1997.
9. Kupiec, J. Robust part-of-speech tagging using a hidden Markov model. *Comput. Speech Lang.* **1992**, *6*, 225–242. [CrossRef]
10. Charniak, E. *Statistical Language Learning*; MIT Press: Cambridge, MA, USA, 1993.
11. Chi, Z.; Geman, S. Estimation of probabilistic context-free grammars. *Comput. Linguist.* **1998**, *24*, 299–305.
12. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
14. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the 2013 Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 5–10 December 2013.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 2017 Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019.
17. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. Available online: https://openai.com/blog/better-language-models/ (accessed on 29 August 2021).
18. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the 2020 Conference on Neural Information Processing Systems (NIPS), virtual meeting, 6–12 December 2020.
19. Chomsky, N. *Aspects of the Theory of Syntax*; The MIT Press: Cambridge, MA, USA, 1965.
20. Ahn, S.; Choi, H.; Pärnamaa, T.; Bengio, Y. A Neural Knowledge Language Model. A Rejected but Interesting Paper. Available online: https://openreview.net/forum?id=BJwFrvOeg (accessed on 29 August 2021).
21. Khmaladze, E. *The Statistical Analysis of Large Number of Rare Events*; Technical Report MS-R8804; Centrum voor Wiskunde en Informatica: Amsterdam, The Netherlands, 1988.
22. Baayen, R.H. *Word Frequency Distributions*; Kluwer Academic Publishers: Dordrecht, The Netherland, 2001.
23. Zipf, G.K. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*; Houghton Mifflin: Boston, MA, USA, 1935.
24. Mandelbrot, B. Structure formelle des textes et communication. *Word* **1954**, *10*, 1–27. [CrossRef]
25. Bar-Hillel, Y.; Carnap, R. An Outline of a Theory of Semantic Information. In *Language and Information: Selected Essays on Their Theory and Application*; Addison-Wesley: Reading, UK, 1964; pp. 221–274.
26. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *30*, 379–423.623–656. [CrossRef]
27. Dębowski, Ł. Is Natural Language a Perigraphic Process? The Theorem about Facts and Words Revisited. *Entropy* **2018**, *20*, 85. [CrossRef]
28. Dębowski, Ł. A general definition of conditional information and its application to ergodic decomposition. *Stat. Probab. Lett.* **2009**, *79*, 1260–1268. [CrossRef]
29. Dębowski, Ł. On the Vocabulary of Grammar-Based Codes and the Logical Consistency of Texts. *IEEE Trans. Inf. Theory* **2011**, *57*, 4589–4599. [CrossRef]
30. Dębowski, Ł. Regular Hilberg Processes: An Example of Processes with a Vanishing Entropy Rate. *IEEE Trans. Inf. Theory* **2017**, *63*, 6538–6546. [CrossRef]
31. Kuraszkiewicz, W.; Łukaszewicz, J. The number of different words as a function of text length. *Pamiętnik Literacki* **1951**, *42*, 168–182. (In Polish)
32. Guiraud, P. *Les Caractères Statistiques du Vocabulaire*; Presses Universitaires de France: Paris, France, 1954.
33. Herdan, G. *Quantitative Linguistics*; Butterworths: London, UK, 1964.
34. Heaps, H.S. *Information Retrieval—Computational and Theoretical Aspects*; Academic Press: New York, NY, USA, 1978.
35. Kornai, A. How many words are there? *Glottometrics* **2002**, *4*, 61–86.
36. Dębowski, Ł. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*; Wiley & Sons: New York, NY, USA, 2021.
37. Martin-Löf, P. The definition of random sequences. *Inf. Control.* **1966**, *9*, 602–619. [CrossRef]
38. Li, M.; Vitányi, P.M.B. *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed.; Springer: New York, NY, USA, 2008.
39. Ziv, J.; Lempel, A. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory* **1977**, *23*, 337–343. [CrossRef]
40. Hilberg, W. Der bekannte Grenzwert der redundanzfreien Information in Texten—eine Fehlinterpretation der Shannonschen Experimente? *Frequenz* **1990**, *44*, 243–248. [CrossRef]
41. Shannon, C. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64. [CrossRef]

42. Ebeling, W.; Nicolis, G. Entropy of Symbolic Sequences: The Role of Correlations. *Europhys. Lett.* **1991**, *14*, 191–196. [CrossRef]
43. Ebeling, W.; Pöschel, T. Entropy and long-range correlations in literary English. *Europhys. Lett.* **1994**, *26*, 241–246. [CrossRef]
44. Bialek, W.; Nemenman, I.; Tishby, N. Complexity through nonextensivity. *Phys. A Stat. Mech. Appl.* **2001**, *302*, 89–99. [CrossRef]
45. Crutchfield, J.P.; Feldman, D.P. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos* **2003**, *15*, 25–54. [CrossRef]
46. Tanaka-Ishii, K. *Statistical Universals of Language: Mathematical Chance vs. Human Choice*; Springer: New York, NY, USA, 2021.
47. Takahira, R.; Tanaka-Ishii, K.; Dębowski, Ł. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy* **2016**, *18*, 364. [CrossRef]
48. Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M.; Ali, M.; Yang, Y.; Zhou, Y. Deep Learning Scaling Is Predictable, Empirically. *arXiv* **2017**, arXiv:1712.00409.
49. Hahn, M.; Futrell, R. Estimating Predictive Rate-Distortion Curves via Neural Variational Inference. *Entropy* **2019**, *21*, 640. [CrossRef]
50. Braverman, M.; Chen, X.; Kakade, S.M.; Narasimhan, K.; Zhang, C.; Zhang, Y. Calibration, Entropy Rates, and Memory in Language Models. In Proceedings of the 2020 International Conference on Machine Learning (ICML), virtual meeting, 12–18 July 2020.
51. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling Laws for Neural Language Models. *arXiv* **2020**, arXiv:2001.08361.
52. Henighan, T.; Kaplan, J.; Katz, M.; Chen, M.; Hesse, C.; Jackson, J.; Jun, H.; Brown, T.B.; Dhariwal, P.; Gray, S. Scaling Laws for Autoregressive Generative Modeling. *arXiv* **2020**, arXiv:2010.14701.
53. Hernandez, D.; Kaplan, J.; Henighan, T.; McCandlish, S. Scaling Laws for Transfer. *arXiv* **2021**, arXiv:2102.01293.
54. Dębowski, Ł. On Hilberg's law and its links with Guiraud's law. *J. Quant. Linguist.* **2006**, *13*, 81–109. [CrossRef]
55. de Marcken, C.G. Unsupervised Language Acquisition. Ph.D Thesis, Massachussetts Institute of Technology, Cambridge, MA, USA, 1996.
56. Dębowski, Ł. On processes with hyperbolically decaying autocorrelations. *J. Time Ser. Anal.* **2011**, *32*, 580–584. [CrossRef]
57. Cleary, J.G.; Witten, I.H. Data compression using adaptive coding and partial string matching. *IEEE Trans. Commun.* **1984**, *32*, 396–402. [CrossRef]
58. Ryabko, B.Y. Prediction of random sequences and universal coding. *Probl. Inf. Transm.* **1988**, *24*, 87–96.
59. Ryabko, B. Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series. In Proceedings of the 2008 IEEE Information Theory Workshop, Porto, Portugal, 5–9 May 2008; pp. 271–275.
60. Kieffer, J.C.; Yang, E. Grammar-based codes: A new class of universal lossless source codes. *IEEE Trans. Inf. Theory* **2000**, *46*, 737–754. [CrossRef]
61. Charikar, M.; Lehman, E.; Lehman, A.; Liu, D.; Panigrahy, R.; Prabhakaran, M.; Sahai, A.; Shelat, A. The Smallest Grammar Problem. *IEEE Trans. Inf. Theory* **2005**, *51*, 2554–2576. [CrossRef]
62. Rokhlin, V.A. On the fundamental ideas of measure theory. *Am. Math. Soc. Transl.* **1962**, *10*, 1–54.
63. Gray, R.M.; Davisson, L.D. Source coding theorems without the ergodic assumption. *IEEE Trans. Inf. Theory* **1974**, *20*, 502–516. [CrossRef]
64. Gács, P. On the symmetry of algorithmic information. *Dokl. Akad. Nauk. SSSR* **1974**, *15*, 1477–1480.
65. Chaitin, G.J. A theory of program size formally identical to information theory. *J. ACM* **1975**, *22*, 329–340. [CrossRef]
66. Dębowski, Ł. Variable-length Coding of Two-sided Asymptotically Mean Stationary Measures. *J. Theor. Probab.* **2010**, *23*, 237–256. [CrossRef]
67. Miller, G.A. Some effects of intermittent silence. *Am. J. Psychol.* **1957**, *70*, 311–314. [CrossRef] [PubMed]
68. Billingsley, P. *Probability and Measure*; Wiley & Sons: New York, NY, USA, 1979.
69. Markov, A.A. Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain. *Bulletin l'Académie Impériale Sci. St.-Pétersbourg* **1913**, *7*, 153–162.
70. Markov, A.A. An Example of Statistical Investigation of the Text 'Eugene Onegin' Concerning the Connection of Samples in Chains. *Sci. Context* **2006**, *19*, 591–600. [CrossRef]
71. Miller, M.I.; O'Sullivan, J.A. Entropies and Combinatorics of Random Branching Processes and Context-Free Languages. *IEEE Trans. Inf. Theory* **1992**, *38*, 1292–1310. [CrossRef]
72. Crutchfield, J.P.; Young, K. Inferring statistical complexity. *Phys. Rev. Lett.* **1989**, *63*, 105–108. [CrossRef] [PubMed]
73. Löhr, W. Properties of the Statistical Complexity Functional and Partially Deterministic HMMs. *Entropy* **2009**, *11*, 385–401. [CrossRef]
74. Jurgens, A.M.; Crutchfield, J.P. Divergent Predictive States: The Statistical Complexity Dimension of Stationary, Ergodic Hidden Markov Processes. *arXiv* **2021**, arXiv:2102.10487.
75. Marzen, S.E.; Crutchfield, J.P. Informational and Causal Architecture of Discrete-Time Renewal Processes. *Entropy* **2015**, *17*, 4891–4917. [CrossRef]
76. Birkhoff, G.D. Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. USA* **1932**, *17*, 656–660. [CrossRef]
77. Gray, R.M. *Probability, Random Processes, and Ergodic Properties*; Springer: New York, NY, USA, 2009.
78. Dębowski, Ł. The Relaxed Hilberg Conjecture: A Review and New Experimental Support. *J. Quant. Linguist.* **2015**, *22*, 311–337. [CrossRef]

79. Dębowski, Ł. Hilberg Exponents: New Measures of Long Memory in the Process. *IEEE Trans. Inf. Theory* **2015**, *61*, 5716–5726. [CrossRef]
80. Brudno, A.A. Entropy and the complexity of trajectories of a dynamical system. *Trans. Moscovian Math. Soc.* **1982**, *44*, 124–149.
81. Grünwald, P.D. *The Minimum Description Length Principle*; The MIT Press: Cambridge, MA, USA, 2007.
82. Shtarkov, Y.M. Universal sequential coding of single messages. *Probl. Inf. Transm.* **1987**, *23*, 3–17.
83. Merhav, N.; Gutman, M.; Ziv, J. On the estimation of the order of a Markov chain and universal data compression. *IEEE Trans. Inf. Theory* **1989**, *35*, 1014–1019. [CrossRef]
84. Ziv, J.; Merhav, N. Estimating the Number of States of a Finite-State Source. *IEEE Trans. Inf. Theory* **1992**, *38*, 61–65. [CrossRef]
85. Csiszar, I.; Shields, P.C. The Consistency of the BIC Markov Order Estimator. *Ann. Stat.* **2000**, *28*, 1601–1619. [CrossRef]
86. Csiszar, I. Large-scale typicality of Markov sample paths and consistency of MDL order estimator. *IEEE Trans. Inf. Theory* **2002**, *48*, 1616–1628. [CrossRef]
87. Morvai, G.; Weiss, B. Order estimation of Markov chains. *IEEE Trans. Inf. Theory* **2005**, *51*, 1496–1497. [CrossRef]
88. Peres, Y.; Shields, P. Two new Markov order estimators. *arXiv* **2005**, arXiv:math/0506080.
89. Dalevi, D.; Dubhashi, D. The Peres-Shields Order Estimator for Fixed and Variable Length Markov Models with Applications to DNA Sequence Similarity. In *Algorithms in Bioinformatics*; Casadio, R., Myers, G., Eds.; Springer,: New York, NY, USA 2005, pp. 291–302.
90. Ryabko, B.; Astola, J. Universal Codes as a Basis for Time Series Testing. *Stat. Methodol.* **2006**, *3*, 375–397. [CrossRef]
91. Csiszar, I.; Talata, Z. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inf. Theory* **2006**, *52*, 1007–1016. [CrossRef]
92. Talata, Z. Divergence rates of Markov order estimators and their application to statistical estimation of stationary ergodic processes. *Bernoulli* **2013**, *19*, 846–885. [CrossRef]
93. Baigorri, A.R.; Goncalves, C.R.; Resende, P.A.A. Markov chain order estimation based on the chi-square divergence. *Can. J. Stat.* **2014**, *42*, 563–578. [CrossRef]
94. Ryabko, B.; Astola, J.; Malyutov, M. *Compression-Based Methods of Statistical Analysis and Prediction of Time Series*; Springer: New York, NY, USA, 2016.
95. Papapetrou, M.; Kugiumtzis, D. Markov chain order estimation with parametric significance tests of conditional mutual information. *Simul. Model. Pract. Theory* **2016**, *61*, 1–13. [CrossRef]
96. Finesso, L. Order Estimation for Functions of Markov Chains. Ph.D Thesis, University of Maryland, College Park, MD, USA, 1990.
97. Weinberger, M.J.; Lempel, A.; Ziv, J. A Sequential Algorithm for the Universal Coding of Finite Memory Sources. *IEEE Trans. Inf. Theory* **1992**, *38*, 1002–1014. [CrossRef]
98. Kieffer, J.C. Strongly Consistent Code-Based Identification and Order Estimation for Constrained Finite-State Model Classes. *IEEE Trans. Inf. Theory* **1993**, *39*, 893–902. [CrossRef]
99. Weinberger, M.J.; Feder, M. Predictive stochastic complexity and model estimation for finite-state processes. *J. Stat. Plan. Inference* **1994**, *39*, 353–372. [CrossRef]
100. Liu, C.C.; Narayan, P. Order Estimation and Sequential Universal Data Compression of a Hidden Markov Source bv the Method of Mixtures. *IEEE Trans. Inf. Theory* **1994**, *40*, 1167–1180.
101. Gassiat, E.; Boucheron, S. Optimal Error Exponents in Hidden Markov Models Order Estimation. *IEEE Trans. Inf. Theory* **2003**, *49*, 964–980. [CrossRef]
102. Lehéricy, L. Consistent order estimation for nonparametric Hidden Markov Models. *Bernoulli* **2019**, *25*, 464–498. [CrossRef]
103. Shalizi, C.R.; Shalizi, K.L.; Crutchfield, J.P. An Algorithm for Pattern Discovery in Time Series. *arXiv* **2002**, arXiv:cs/0210025.
104. Zheng, J.; Huang, J.; Tong, C. The order estimation for hidden Markov models. *Phys. A Stat. Mech. Appl.* **2019**, *527*, 121462. [CrossRef]
105. Kieffer, J.C.; Rahe, M. Markov Channels are Asymptotically Mean Stationary. *Siam J. Math. Anal.* **1981**, *12*, 293–305. [CrossRef]
106. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley & Sons: New York, NY, USA, 2006.
107. Ochoa, C.; Navarro, G. RePair and All Irreducible Grammars are Upper Bounded by High-Order Empirical Entropy. *IEEE Trans. Inf. Theory* **2019**, *65*, 3160–3164. [CrossRef]
108. Dębowski, Ł. Mixing, Ergodic, and Nonergodic Processes with Rapidly Growing Information between Blocks. *IEEE Trans. Inf. Theory* **2012**, *58*, 3392–3401. [CrossRef]
109. Gács, P.; Körner, J. Common information is far less than mutual information. *Probl. Control. Inf. Theory* **1973**, *2*, 119–162.
110. Wyner, A.D. The Common Information of Two Dependent Random Variables. *IEEE Trans. Inf. Theory* **1975**, *IT-21*, 163–179. [CrossRef]
111. Chaitin, G. *Meta Math!: The Quest for Omega*; Pantheon Books: New York, NY, USA, 2005.
112. Gardner, M. The random number Ω bids fair to hold the mysteries of the universe. *Sci. Am.* **1979**, *241*, 20–34. [CrossRef]
113. Beran, J. *Statistics for Long-Memory Processes*; Chapman & Hall: New York, NY, USA, 1994.
114. Lin, H.W.; Tegmark, M. Critical Behavior in Physics and Probabilistic Formal Languages. *Entropy* **2017**, *19*, 299. [CrossRef]
115. Szpankowski, W. Asymptotic Properties of Data Compression and Suffix Trees. *IEEE Trans. Inf. Theory* **1993**, *39*, 1647–1659. [CrossRef]

116. Szpankowski, W. A generalized suffix tree and its (un)expected asymptotic behaviors. *Siam J. Comput.* **1993**, *22*, 1176–1198. [CrossRef]
117. Dębowski, Ł. Maximal Repetitions in Written Texts: Finite Energy Hypothesis vs. Strong Hilberg Conjecture. *Entropy* **2015**, *17*, 5903–5919. [CrossRef]
118. Dębowski, Ł. Maximal Repetition and Zero Entropy Rate. *IEEE Trans. Inf. Theory* **2018**, *64*, 2212–2219. [CrossRef]
119. Futrell, R.; Qian, P.; Gibson, E.; Fedorenko, E.; Blank, I. Syntactic dependencies correspond to word pairs with high mutual information. In Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, Syntaxfest 2019), Paris, France, 27–28 August 2019; pp. 3–13.
120. Hahn, M.; Degen, J.; Futrell, R. Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychol. Rev.* **2021**, *128*, 726–756. [CrossRef]
121. Barron, A.R. Logically Smooth Density Estimation. Ph.D Thesis, Stanford University, Stanford, CA, USA, 1985.
122. Gray, R.M.; Kieffer, J.C. Asymptotically mean stationary measures. *Ann. Probab.* **1980**, *8*, 962–973. [CrossRef]
123. Wyner, A.D. A definition of conditional mutual information for arbitrary ensembles. *Inf. Control* **1978**, *38*, 51–59. [CrossRef]
124. Dębowski, Ł. Approximating Information Measures for Fields. *Entropy* **2020**, *22*, 79. [CrossRef] [PubMed]
125. Travers, N.F.; Crutchfield, J.P. Exact synchronization for finite-state sources. *J. Stat. Phys.* **2011**, *145*, 1181–1201. [CrossRef]
126. Travers, N.F.; Crutchfield, J.P. Asymptotic synchronization for finite-state sources. *J. Stat. Phys.* **2011**, *145*, 1202–1223. [CrossRef]
127. Travers, N.F.; Crutchfield, J.P. Infinite Excess Entropy Processes with Countable-State Generators. *Entropy* **2014**, *16*, 1396–1413. [CrossRef]
128. Blackwell, D. The entropy of functions of finite-state Markov chains. In *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*; Czechoslovak Academy of Sciences: Prague, Czech Republic, 1957; pp. 13–20.
129. Ephraim, Y.; Merhav, N. Hidden Markov processes. *IEEE Trans. Inf. Theory* **2002**, *48*, 1518–1569. [CrossRef]
130. Han, G.; Marcus, B. Analyticity of entropy rate of hidden Markov chain. *IEEE Trans. Inf. Theory* **2006**, *52*, 5251–5266. [CrossRef]
131. Jacquet, P.; Seroussi, G.; Szpankowski, W. On the entropy of a hidden Markov process. *Theor. Comput. Sci.* **2008**, *395*, 203–219. [CrossRef] [PubMed]