Statistical properties of probabilistic context-sensitive grammars

Kai Nakaishi [1] and Koji Hukushima [1,2]

[1]*Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan*
[2]*Komaba Institute for Science, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan*

Probabilistic context-free grammars (PCFGs), which are commonly used to generate trees randomly, have been well analyzed theoretically, leading to applications in various domains. Despite their utility, the distributions that the grammar can express are limited to those in which the distribution of a subtree depends only on its root and not on its context. This limitation presents a challenge for modeling various real-world phenomena, such as natural languages. To overcome this limitation, a probabilistic context-sensitive grammar (PCSG) is introduced, where the distribution of a subtree depends on its context. Numerical analysis of a PCSG reveals that the distribution of a symbol does not constitute a qualitative difference from that in the context-free case, but mutual information does. Furthermore, a novel metric introduced to directly quantify the breaking of this limitation detects a distinct difference between PCFGs and PCSGs. This metric, applicable to an arbitrary distribution of a tree, allows for further investigation and characterization of various tree structures that PCFGs cannot express.

I. INTRODUCTION

Hierarchical structures underlie many real-world phenomena, including natural languages. A context-free grammar (CFG), a fundamental concept in formal language theory, was originally introduced to analyze hierarchical syntactic structures in natural languages [1]. Furthermore, it provides a basis for describing more general hierarchical structures that are not limited to natural languages. A CFG, defined by a set of production rules, generates strings with trees in a formal way. The strings correspond to sentences, whereas the trees describe the hierarchical syntactic structures behind the sentences. A probabilistic extension of a CFG, known as a probabilistic context-free grammar (PCFG) or stochastic context-free grammar [2], introduces probabilities into the production rules. According to the rules, this model generates trees in a probabilistic manner. This probabilistic grammar has been used to model syntactic structures of a natural language [3] or a programming language [4], and to study many other phenomena with tree or hierarchical structures in fields such as music [5,6], human cognition [7], a long-short-term-memory network [8], RNA [9], cosmic inflation [10], or a more abstract model [11–13]. Additionally, other frameworks are closely related to a PCFG, including a branching process and a Lindenmayer system (or L-system) [14–16].

An essential property of a PCFG is that the distribution of a subtree depends only on its root, not on the context, which we will designate as context-free independence. This property allows an exact mathematical analysis of the statistical properties of PCFGs. Indeed, earlier studies have analyzed and resolved various aspects of PCFGs, including the probability of symbol occurrence [17], the correlation function [11,12], mutual information between nodes [8], the mean sentence length [18], entropy [19,20], branching rates [19,20], tree size [20], and the conditions for sentence generation to terminate with probability 1 [18,21]. At the same time, this property is too strict to impose on real-world phenomena. Particularly, it is well known in linguistics that some languages in the real world cannot be described using a CFG because of its inability to represent cross-serial dependencies [22,23]. In natural language processing, empirical evidence suggests that a naive parser relying on a PCFG is insufficient for inferring syntactic structures [3]. Moreover, certain parsers that relax context-free independence in technical manners can express more complex distributions and can achieve higher accuracy [24,25]. Outside of language-related domains, the possibility that introducing context sensitivity is useful for describing music is also discussed [5]. Therefore, the distributions that a PCFG can express are regarded as severely limited.

To understand more realistic phenomena with hierarchical structures, it is necessary to introduce and analyze a model that captures the distribution of a tree beyond context-free independence. For this purpose, one can naturally consider context-sensitive grammars (CSGs) [26], which form the class one level higher than CFGs in the hierarchy of expressive power: the so-called Chomsky hierarchy. Similarly to a PCFG, a probabilistic context-sensitive grammar (PCSG) can be formulated as a probabilistic extension of a CSG. A PCSG explicitly relaxes context-free independence. Consequently, the theoretical analyses developed for a PCFG are not applicable to a PCSG. The statistical properties of a PCSG have only rarely been analyzed, either theoretically or numerically.

To address this point, for this work we defined a simple PCSG and investigated its statistical properties by numerical

simulations systematically, mainly examining whether a qualitative difference from a PCFG exists, or not. To be more precise, we implemented a PCSG to measure the distribution of a symbol, mutual information between two nodes, and mutual information between two pairs of children of nodes on which the symbols are fixed. Here, we present a comparison of the observed similarities and differences between PCFG and PCSG: No qualitative difference was found in the distribution of a symbol between a PCSG and a PCFG. This result suggests that the properties observed in PCFGs are likely to be preserved in PCSGs. Given that the absence of a singularity in the distribution in an ensemble of PCFGs has been proven [17], it is reasonable to infer that PCSGs would not exhibit the singularity, similarly to PCFGs. This singularity is relevant for the discussion on a phase transition in the random language model (RLM) [27,28], which might be analogous to discontinuity in human language acquisition according to earlier research.

However, the behaviors of mutual information between two nodes differ between a PCFG and a PCSG. The mutual information in a PCFG decays exponentially with the distance between two nodes, i.e., the path length in a tree graph. In contrast, in a PCSG, the mutual information decays exponentially with the effective distance, which is defined by considering the effect of context sensitivity.

Additionally, a more pronounced difference concerns the mutual information between pairs of children of symbol-fixed nodes. This novel metric, proposed in this research, quantifies context-free independence breaking. From a theoretical physics perspective, this metric represents the degree to which the network of interactions deviates from a tree structure. Linguistically, it represents the strength of mutual dependence between the structures of two constituents or phrases of given types. This metric not only detects whether context-free independence is broken, it also quantifies where and how strongly the breaking occurs. In a PCFG, the context-free independence breaking is always zero. By contrast, in a PCSG, it is positive and decays similarly to the mutual information between nodes. As a result, the most striking difference between a PCFG and a PCSG is in this metric. This quantification is intuitive and is definable for any distribution of a tree. Measuring this metric in other mathematical models or real-world phenomena will help deepen the understanding of them by investigating how their behavior differs from that of a PCFG.

Here, we provide a brief summary of the main contributions made in this paper. Our first main contribution is the systematic investigation of a PCSG, which is a simple model for generating hierarchical structures beyond those produced by PCFGs. A key distinction between a PCFG and a PCSG is in the distance that determines the decay of mutual information. Second, we propose a novel metric for the context-free independence breaking, which has not been quantified previously. This metric allows for further quantitative investigation of various hierarchical structures that violate the context-free independence. Our results show that this metric decays exponentially for a PCSG while it remains zero for a PCFG, demonstrating the usefulness of this metric.

This paper is structured as follows: The models, a PCFG and a PCSG, are introduced in Sec. II. The analysis of the distribution of a symbol in a PCSG and the argument about the phase transition in the RLM are presented in Sec. III. Then, in Sec. IV, a numerical analysis of the mutual information between two nodes is presented, including the definition of the effective distance. The introduction and analysis of the quantification of the context-free independence breaking are given in Sec. V. Finally, we summarize the results and briefly discuss future works in the last section.

II. MODEL

A. Probabilistic context-free grammar

In formal language theory [26], a grammar G consists of a vocabulary V and a finite set R of rules. A vocabulary V, a finite set of symbols, is divided into nonterminal symbols $A, B, \ldots \in V_N$ and terminal symbols $a, b, \ldots \in V_T$. Each rule in R is of the form $\varphi \to \psi$, meaning that a finite string φ in V is rewritten as another finite string ψ. Also, the left-hand side φ of the rule must include at least one nonterminal symbol. The grammar G generates a sentence by the following process: Initially, a special symbol $S \in V_N$, called the starting symbol, is given. Next, S is rewritten by a rule $S \to \varphi$. When a substring ψ of φ includes a nonterminal symbol, φ can be rewritten by replacing ψ with another string ω according to a rule $\psi \to \omega$. This process is repeated. Finally, if the string has no nonterminal symbol, it can no longer be rewritten by any rule. The final string is called a sentence. The whole process of generating a sentence is called a derivation. The set of sentences generated using a grammar G is a language of G. The importance of the finiteness of symbols and rules is noteworthy. If infinite sets V and R are allowed, then it becomes trivially possible to construct a grammar that generates an arbitrary language by introducing a symbol A and a rule $A \to \varphi$ for each sentence φ in the language. The infinite number of symbols or rules would make the concept of characterizing and classifying languages in terms of grammars irrelevant.

A grammar G is a CFG [1] if every rule of G is of the form $A \to \varphi$ with A being nonterminal. The derivation in a CFG can be represented as a tree, which is analogous to the syntactic structure of a sentence in a natural language analyzed by immediate constituent analysis, as shown in Fig. 1. In fact, any CFG can be transformed to an equivalent CFG where every rule is of the form $A \to BC$ or $A \to a$ for $A, B, C \in V_N$ and $a \in V_T$, ensuring that the generated language remains unchanged. This transformed form is referred to as the Chomsky normal form (CNF) [29].

A PCFG [2] is a probabilistic version of a CFG. It is introduced by assigning a probabilistic weight $M_{A \to \varphi}$ to each CFG rule $A \to \varphi$, meaning that a nonterminal symbol A is rewritten as φ with probability $M_{A \to \varphi}$. The PCFG specified by the set of weights $M_{A \to \varphi}$ determines the probability of a derivation, which is the product of the weights of all rules applied in the derivation. If we adopt the idea of simplifying a speaker or a group of speakers of a language as an agent that generates strings with syntactic structures probabilistically, then a PCFG can be a simple mathematical model for a language. Indeed, a PCFG has been used for modeling a natural language [3] and a programming language [4]. In addition, because a PCFG can be regarded as a simple mathematical model for randomly generating trees or hierarchical structures, many studies have
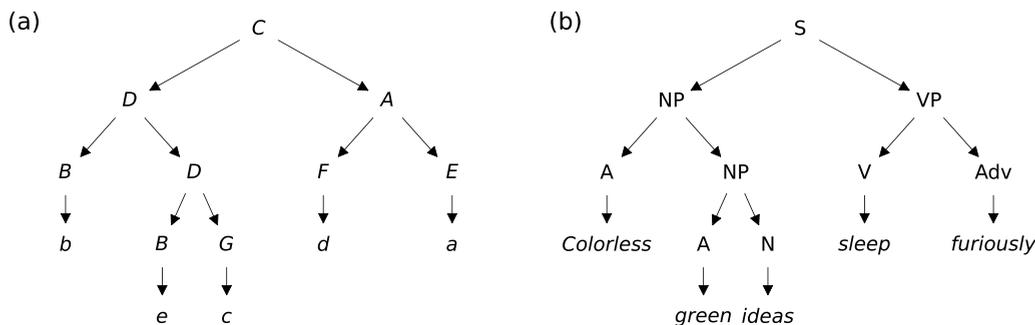
FIG. 1. (a) Example of a derivation generated using a CFG in CNF. A node with its children means that the node is rewritten as the children. In this example, the initial string is C. Applying the first rule $C \to DA$, the string becomes DA. The next rule $D \to BD$ (or $A \to FE$) rewrites the string as BDA (or DFE). The remainder of the derivation is similar. The final string, i.e., the sentence, is *becda*. (b) Syntactic structure behind the sentence *Colorless green ideas sleep furiously* in terms of immediate constituent analysis. This diagram means, for instance, that the noun phrase (NP) *green ideas* consists of the adjective (A) *green* and the noun (N) *ideas*. Roughly speaking, a nonterminal symbol in a CFG corresponds to a constituent in syntax; a terminal symbol corresponds to a word.

used it as a model not only for a natural or formal language but also for other phenomena [5–9,11–13]. A PCFG also has a close relation to other physical and mathematical frameworks [14–16].

By definition, the distribution of a subtree in a PCFG depends only on the root. It is unaffected by the context, i.e., the neighboring symbols of the root. Because of this context-free independence, many properties of a PCFG can be analyzed theoretically. For instance, the distribution of a symbol or the joint distribution of several symbols at arbitrary nodes can be computed recursively from the root of the entire tree, similarly to a Markov chain. Indeed, many earlier studies have analyzed properties of a PCFG theoretically and exactly [8,11,12,17–21]. The context-free independence allows for the theoretical analysis of various properties of PCFGs, but it also severely restricts the range of distributions that a PCFG can express. In general, it is not reasonable to expect a natural phenomenon to satisfy such a restriction. Linguistically, for instance, some real-world languages cannot be described by CFGs [22,23]. Moreover, natural language processing researchers have found it necessary, empirically, to relax the context-free independence for modeling the syntactic structures of natural languages [24,25]. However, no report of the relevant literature describes a systematic investigation of a simple mathematical model that goes beyond the independence or a quantitative analysis of the degree to which context-free independence is broken in any model or phenomenon. This need for study prompts us to consider such a model and to quantify how far the model is from the independence.

B. Probabilistic context-sensitive grammar

A model introduced by allowing each rule in a CFG to refer to the context is a CSG, which has one level higher expressive power than a CFG in formal language theory [26]. In a CSG, a rule is of the form $\varphi A \psi \to \varphi \omega \psi$. In other words, the result ω of rewriting A can depend on the substrings φ and ψ next to A, i.e., the context of A. The class of languages generated by CSGs is believed to be larger than the class of possible natural languages [30]. Additionally, we can naturally define a

probabilistic version of a CSG, namely a PCSG, by assigning a probabilistic weight to each rule, similar to the introduction of a PCFG. A PCSG relaxes the context-free independence, meaning that the distribution of a subtree in a PCSG depends not only on its root but also on the context. The theoretical analyses of a PCFG described above [8,11,12,17–21], all of which impose the independence, are not applicable to a PCSG. Consequently, the behavior of a PCSG and its characteristics, such as which of its properties are similar to or different from those of a PCFG, are unknown.

The class of all possible grammars defined as a probabilistic extension of a CSG is too large and complicated to analyze. We focus, therefore, on a simpler model within a CSG to examine its behavior. First, we consider a CSG with a vocabulary consisting of binary nonterminal symbols, $V_N = \{0, 1\}$. We do not consider terminal symbols. In the following, a symbol simply means a nonterminal symbol unless otherwise noted. Additionally, we restrict rules to the form of $A \to BC$ or $LAR \to LBCR$. The former is a nonterminal rule of a CFG in CNF, whereas the latter is a CSG rule with context sensitivity that refers only to the two symbols next to the rewritten symbol. Consequently, the cause of the difference between our model and the binary CFG or PCFG in CNF is, in essence, the context sensitivity to L and R. In our notation, A, B, and C represent symbols, whereas L and R can be symbols or nulls λ. For example, if the rule $\lambda 01 \to \lambda 111$ is applied to the leftmost 0 in the string 0110, then the string turns to 11110.

Our PCSG is defined as the probabilistic extension of this CSG. The probabilistic weight M_{ABC}^{CF} is assigned to each CFG rule $A \to BC$, and $M_{LAR,BC}^{\mathrm{CS}}$ to each CSG rule $LAR \to LBCR$. Next, we introduce the probability q that a CSG rule is chosen to control the degree of context sensitivity. More precisely, symbol A in the context LAR is rewritten as BC by a CFG rule $A \to BC$ with probability $(1 - q)M_{ABC}^{\mathrm{CF}}$, or as DE by a CSG rule $LAR \to LDER$ with probability $qM_{LAR,DE}^{\mathrm{CS}}$. A PCSG with $q = 0$ is a PCFG. Additionally, we must determine the order in which rules are applied to a string because, in a PCSG, unlike a PCFG, a derivation depends on the order. For this study, we choose to apply rules in a uniformly random manner as a neutral alternative. If the length of a present string is l, we first generate a random permutation τ of $\{0, \ldots, l-1\}$ according
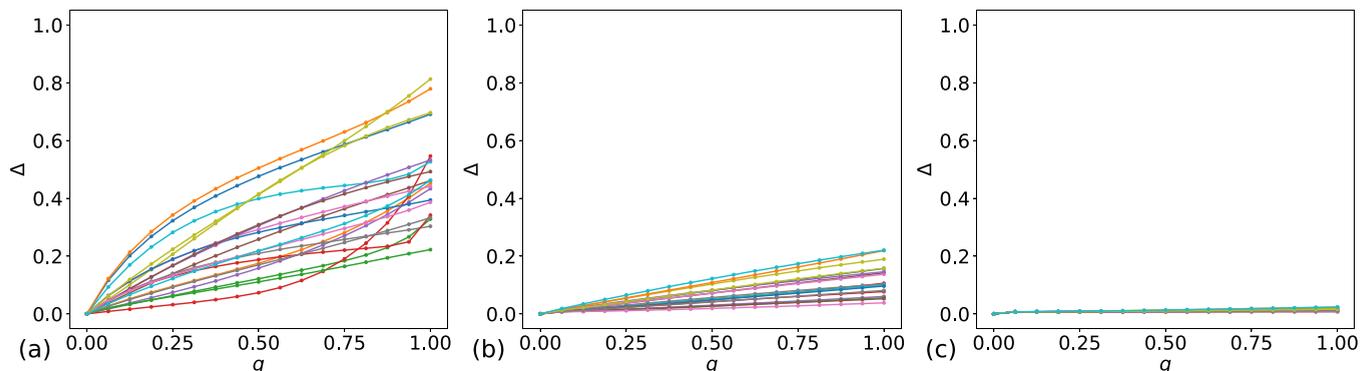
FIG. 2. Differences $\Delta(q, M)$ between $\pi_{A,i}$ with $q = 0$ and that with $q > 0$ as functions of q, computed from 10^4 sampled trees of depth 10. Different colors represent different M's. Panels (a), (b), and (c), respectively, present results for M's generated from the log-normal distribution with $\epsilon = 10^{-2}$, 10^0, and 10^2.

to a uniform distribution, and then apply rules to the symbols sequentially, from the $\tau(0)$th one to the $\tau(l-1)$th one. After all symbols of the preceding string are rewritten, the length becomes $2l$. The whole procedure to generate a tree is as follows: The first step in the derivation is to choose a symbol from a uniform distribution over V_N. Subsequently, a string is rewritten recursively. For each step, the order of application of rules and each rewriting are determined randomly in the manner we explained above. Because no terminal symbol exists in this setting, a rule can always be applied to the string no matter how many steps the derivation goes through. Consequently, we stop the process when the step is repeated D times, which is a value determined in advance [31].

Although the discussion in the remainder of this paper is based on the above setting, we have found that the properties of a PCSG remain qualitatively unchanged under alternative settings. For example, the model exhibits similar behavior when each rule refers to two left neighbors and two right neighbors, or when symbols are rewritten in a different order, such as left-to-right or inside-to-outside.

This type of PCSG is specified by the probability q and the weights $M = (M^{\mathrm{CF}}, M^{\mathrm{CS}})$, where $M^{\mathrm{CF}} = \{M^{\mathrm{CF}}_{ABC}\}_{ABC}$ and $M^{\mathrm{CS}} = \{M^{\mathrm{CS}}_{LAR,BC}\}_{LAR,BC}$. The probabilistic weights are sampled according to the log-normal distributions with normalization conditions

$$M^{\mathrm{CF}}_{ABC} = \frac{\tilde{M}^{\mathrm{CF}}_{ABC}}{\sum_{B',C'} \tilde{M}^{\mathrm{CF}}_{AB'C'}}, \quad P(\tilde{M}^{\mathrm{CF}}_{ABC}) \propto e^{-\epsilon \ln^2 \tilde{M}^{\mathrm{CF}}_{ABC}},$$

$$M^{\mathrm{CS}}_{LAR,BC} = \frac{\tilde{M}^{\mathrm{CS}}_{LAR,BC}}{\sum_{B',C'} \tilde{M}^{\mathrm{CS}}_{LAR,B'C'}}, \quad P(\tilde{M}^{\mathrm{CS}}_{LAR,BC}) \propto e^{-\epsilon \ln^2 \tilde{M}^{\mathrm{CS}}_{LAR,BC}}.$$

Therein, ϵ is the parameter used to control the width of the log-normal distributions.

For this study, we are interested in how the introduction of context sensitivity affects the statistical properties of PCFGs. Specifically, we implement PCSGs and conduct numerical analyses of three statistical quantities. The first involves the distribution of a symbol at a node, analogous to magnetization in a spin model. This quantity is related to the phase transition in the RLM [27,28]. The second specifically examines the mutual information between two nodes, which is associated with a two-point correlation. Finally, we introduce the mutual information between the children of two symbol-fixed nodes.

This metric, which is zero for $q = 0$ by definition, reflects how strongly the independence is broken.

III. DISTRIBUTION OF A SYMBOL

A. Distribution of a symbol

Primary emphasis should be on the distribution of a symbol on a single node. We denote the probability that symbol A occurs on node i as

$$\pi_{A,i}(q, M) \equiv \langle \delta_{A,\sigma_i} \rangle_{q,M},$$

where σ_i is a symbol on node i, and $\langle \cdots \rangle_{q,M}$ represents the average over trees under a PCSG with parameters (q, M). This quantity corresponds to the magnetization in the Potts spin model [32], where each site i has a spin σ_i, and the magnetization along the direction A is defined by the ratio of sites with $\sigma_i = A$. In the case of $q = 0$, i.e., a PCFG, the context-free independence enables us to apply the concept of Markov chains. Because of this, the probability $\pi_{A,i}$ can be computed. If node i is the left child of node j, then $\pi_{B,i} = \sum_A (\sum_C M^{\mathrm{CF}}_{ABC}) \pi_{A,j}$. If node i is the right child, then it is the same except that $\pi_{B,i}$ and $\sum_C$ are replaced, respectively, with $\pi_{C,i}$ and $\sum_B$. However, this no longer holds in the case of $q > 0$ because of the broken independence.

To see the degree to which the distribution of a symbol changes with the context sensitivity, we measured the Euclidean distance Δ between $\{\pi_{A,i}\}_{A,i}$ with $q = 0$ and that with $q > 0$, expressed as

$$\Delta(D, q, M) \equiv \sqrt{\frac{\sum_{i,A} [\pi_{A,i}(q, M) - \pi_{A,i}(0, M)]^2}{2(2^{D+1} - 1)}}. \quad (1)$$

Figure 2 presents the distances Δ as functions of q for $\epsilon = 10^{-2}$, 10^0, and 10^2. We sampled 20 M's for each ϵ, and 10^4 complete trees for each PCSG, with depth D of a tree set to 10. These figures show that Δ increases monotonically and continuously for any M. It can also be observed that the increase is slower with larger ϵ. If ϵ is larger, most of the generated M^{CF}_{ABC}'s and $M^{\mathrm{CS}}_{LAR,BC}$'s are near $1/2^2$. As a result, $\pi_{A,i}$'s are near $1/2$ for any A and i with most M's. This fact leads to the slower increase. This behavior of Δ implies that the context sensitivity drives $\{\pi_{A,i}(q, M)\}_{A,i}$ farther away, monotonically and continuously, from that for $q = 0$, and that no singularity
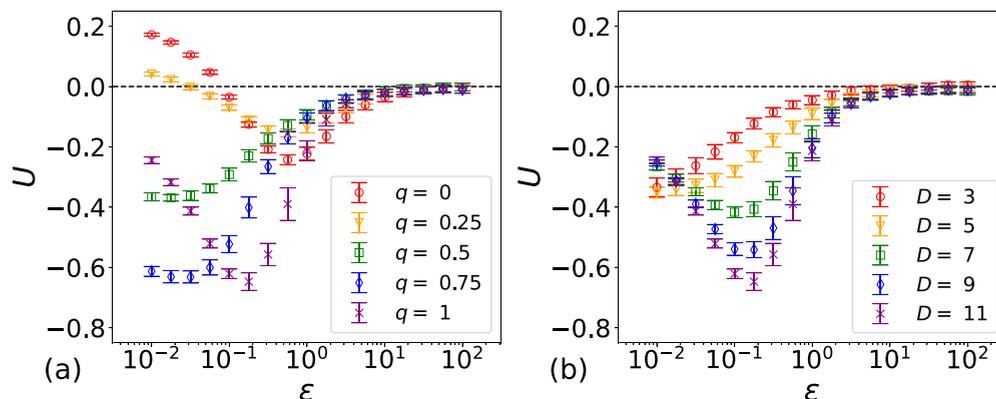
FIG. 3. Binder parameter U of the mean ratio π_A of symbol A as a function of the parameter ϵ (a) for depth $D = 11$ and context sensitivity $q = 0$, 0.25, 0.5, 0.75, or 1, and (b) for $D = 3$, 5, 7, 9, or 11 with $q = 1$.

occurs at any point in $0 < q < 1$. It is noteworthy that the context sensitivity is not the only factor that contributes to this behavior, at least qualitatively. Suppose we interpolate between a PCFG M^{CF} and another independently generated PCFG $M^{CF'}$, instead of an M^{CS}. Even in this case, Δ will grow similarly with q. It is not possible to see any qualitative difference between a PCFG and a PCSG in terms of the distribution of a symbol.

The observations presented here are for finite trees. However, for most of the 20 M's, Δ with $D = 10$ seems to converge almost to that in the limit $D \to \infty$. Consequently, it is unlikely that Δ has a singularity, even in the limit of infinite trees. Supplemental material [33] provides numerical observations of how Δ converges as D increases.

B. Order parameter for the random language model

In our case, because the tree topology is always the same, the mean ratio π_A of symbol A in a whole tree is the average of $\pi_{A,i}$ over nodes i. We denote it as

$$\pi_A(D, q, M) \equiv \frac{\sum_i \pi_{A,i}(q, M)}{2^{D+1} - 1}.$$

The probability density of π_A attributable to the randomness of M, defined as

$$P(\pi_A | D, q, \epsilon) \equiv \int dM P(M) \delta(\pi_A - \pi_A(q, M)),$$

plays a crucially important role in the discussion of the phase transition in the RLM, proposed in [27,28]. The RLM is defined as an ensemble of PCFGs generated according to the log-normal distribution, which is equivalent to the $q = 0$ case in our model. An earlier study investigated the possibility of a phase transition characterized by the singularity of an *order parameter* as the parameter ϵ varies. This earlier study suggested that the phase transition can be interpreted as a possible discontinuity in human language acquisition. However, recent findings in [17] have revealed that the singularity of their order parameter, if any, is reduced to that of the probability density of π_A and that the probability density is an analytic function of ϵ with finite vocabulary. In other words, the phase transition does not exist as long as the number of types of symbols is finite. This conclusion holds true for any analytic distribution

of M, irrespective of whether it follows the log-normal distribution or whether the sizes of trees are finite or infinite. Because the proof relies on the assumption of context-free independence, it cannot be extended to a context-sensitive case with $q > 0$. Therefore, whether a phase transition exists in the context-sensitive RLM remains a nontrivial question.

To investigate whether the distribution of π_A in the context-sensitive RLM has a singularity, we measured the Binder parameter of π_A, defined as

$$U(D, q, \epsilon) \equiv 1 - \frac{[(\Delta \pi_A)^4]_\epsilon}{3[(\Delta \pi_A)^2]_\epsilon^2},$$

where $\Delta \pi_A \equiv \pi_A - 1/2$, and $[\cdots]_\epsilon$ means the average over M's according to the log-normal distribution determined by ϵ. This parameter has been used to detect the transition in various statistical-mechanical models numerically [34,35]. This parameter is zero when π_A follows a Gaussian distribution and nonzero when the distribution of π_A is multimodal or non-Gaussian. To compute the Binder parameter, we sampled 10^4 M's for each ϵ and 10^3 trees for each M. Error bars were computed using the bootstrap method [36,37] with 10^2 bootstrap sets.

Figure 3(a) shows the result obtained when the tree depth is fixed at $D = 11$ and the context sensitivity is $q = 0$, 0.25, 0.5, 0.75, and 1. From these findings, the Binder parameter seems to change analytically, but it changes more dramatically if the context sensitivity is larger. Consequently, if the singularity exists, it might occur for $q = 1$. We also computed the Binder parameters for $q = 1$ while varying the depth D of a tree, the result of which is shown in Fig. 3(b). For all previously known cases of phase transitions detected by this parameter, a discontinuous jump from zero to nonzero is found at the transition temperature in the thermodynamic limit. However, it is unlikely that such a transition occurs for the limit $D \to \infty$ because the Binder parameter for large ϵ becomes farther away from zero as D increases. Note that we do not rule out the possibility of another phase transition detected by other methods, which remains an open problem.

IV. MUTUAL INFORMATION BETWEEN TWO NODES

As described in the preceding section, we examined the distribution of a symbol on a node, but we could find no
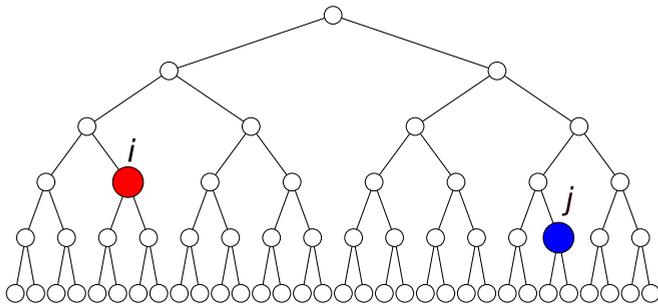
FIG. 4. Mutual information I defined by Eq. (2) is the mutual information between the red node i and the blue node j.

FIG. 5. Structural and horizontal distances between node i and j with $i = (1, 0, 0, 0)$ and $j = (0, 0, 0, 0, 0)$, $(0, 1, 1, 0, 1)$, $(1, 0, 0, 1, 1)$, or $(1, 1, 1, 1, 0)$. Different colors and lines represent different j's. The structural distance is the path length from i to j, denoted by the line along the edges. The horizontal distance is the number of nodes lying horizontally between the higher node and the lower node's ancestor of the same depth as the former, as indicated by the horizontal arrows. Nodes $j = (0, 0, 0, 0, 0)$ and $(0, 1, 1, 0, 1)$ are in the left subtree. The horizontal distance is 8 in the former case and 2 in the latter case, whereas the structural distance is 9 in both cases. Node $j = (1, 0, 0, 1, 1)$ belongs to the right subtree. The structural and horizontal distances between this node and i are, respectively, 3 and 1. Node $j = (1, 1, 1, 1, 0)$ belongs to the right subtree, too. The structural and horizontal distances are 7. When j belongs to the right branch, the horizontal distance grows exponentially as the structural distance increases.

significant difference between a PCFG and a PCSG. For the discussion in this section, we turn our interest to mutual information, which has a close relation to the two-point correlation function [38] and which has been used for measuring correlation in symbolic sequences such as formal and natural languages [8,39,40], music [8], birdsong [41], DNA [42], and so forth. We denote the mutual information between nodes i and j, as depicted in Fig. 4, as

$$I_{i,j}(q, M) \equiv \sum_{\sigma_i, \sigma_j} P(\sigma_i, \sigma_j) \ln \frac{P(\sigma_i, \sigma_j)}{P(\sigma_i) P(\sigma_j)}. \quad (2)$$

This measures the dependence between the two nodes. Although the behavior of mutual information in a PCFG is well known through theoretical analysis [8], this analysis is also based on context-free independence. Consequently, understanding what occurs in a PCSG regarding the mutual information, where the independence is broken, is nontrivial again.

Before presenting the results of the numerical analysis, we introduce some notations and quantities. In the following, we designate a node by a binary sequence that represents the path from the root to the node by assigning 0 and 1, respectively, to a left and right child. For example, nodes (), (0), and (0, 1) represent the root, the left child of the root, and the right child of the left child of the root, respectively. To characterize the relative position of two nodes, we introduce the two distinct distances described in Fig. 5. The first is the structural distance, i.e., the length of the path between the two nodes. The second, designated as the horizontal distance, is the number of nodes lying horizontally between the two nodes. If the depths of the two nodes differ, then the horizontal distance is the number of nodes between the higher node and the lower node's ancestor of the same depth as the former.

One of the two nodes was fixed at $i = (1, 0, 0, 0, 0, 0)$, which is the leftmost node of depth 6 in the subtree whose root is the right child of the root of the whole tree. The other node j could be any node in the whole tree. The relation between structural and horizontal distances differs based on whether node j belongs to the left or right subtree, as presented in Fig. 5. Presuming that the depth of node j is fixed, then when j is in the left subtree, i.e., $j = (0, \dots)$, the structural distance is the same, irrespective of the horizontal distance. However, when j is in the right subtree, i.e., $j = (1, \dots)$, the horizontal distance is roughly exponential of the structural distance.

In the context-free case with $q = 0$, the dependence of the mutual information on the two distances is already known. Lin

and Tegmark [8] have proved that the mutual information decays exponentially with the structural distance. Recalling that the mutual information in a Markov chain decays exponentially with the chain length, this result is intuitively reasonable when considering context-free independence. When j is in the left subtree, the mutual information is the same for any node j of the same depth because the mutual information depends only on the structural distance, which is independent of the horizontal distance. However, when j is in the right subtree, the mutual information decays according to a power law of the horizontal distance because the horizontal distance grows exponentially in the structural distance. One main claim of Lin and Tegmark [8] was that this power law might be the mechanism of the power-law decay of mutual information in natural language texts.

In the context-sensitive case with $q > 0$, we examined the behavior of the mutual information. We sampled 10^8 complete trees of depth $D = 7$ and estimated I. Because the mutual information between X and Y is decomposed into $S(X) + S(Y) - S(X, Y)$ where $S(\cdot)$ is Shannon entropy, we computed the mutual information by estimating the entropy from the empirical distribution. This estimate has a bias from the entropy of the true distribution, resulting in biased mutual information, which is not negligible in the region of the small mutual information. Consequently, to compute the entropy in the present and the subsequent sections, we used the bias-reduced estimator proposed by Ref. [43]. This estimator is represented by

$$\hat{S}(X) \equiv \Psi(N) - \frac{1}{N} \sum_x n_x \Psi(n_x).$$
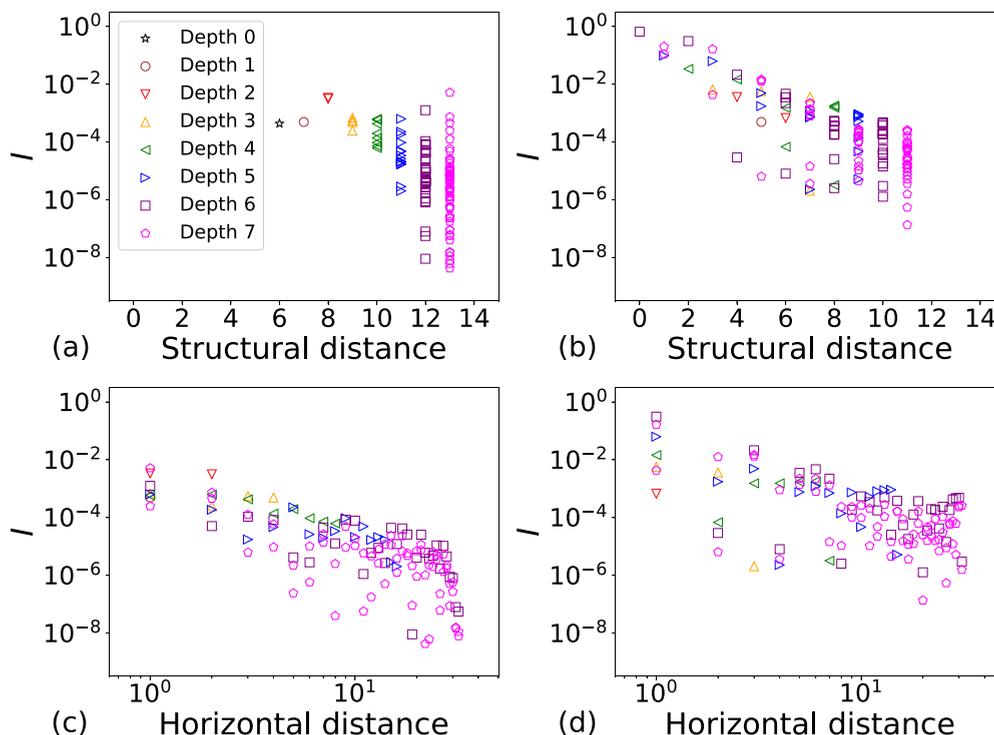
FIG. 6. Mutual information I defined by Eq. (2) against the distance between i and j. Weights M are generated with $\epsilon = 10^{-2}$. The context sensitivity is set to $q = 1$. The position of i is fixed at $(1, 0, 0, 0, 0, 0)$. Mutual information against the structural distance when node j is in the left branch, i.e., $j = (0, \dots)$, is shown in (a). The same quantity when node j is in the right branch, i.e., $j = (1, \dots)$, is in (b). Similarly, plots against the horizontal distance are shown in (c) for $j = (0, \dots)$ and (d) for $j = (1, \dots)$. The result when node j is the root, i.e., $j = ()$, is in (a). Markers and colors are different for different depths.

Therein, Ψ is the digamma function, x represents a state which X takes, n_x denotes the number of samples such that $X = x$, and $N = \sum_x n_x$ is the total number of samples.

Figure 6 shows I's for an M generated with $\epsilon = 10^{-2}$ and $q = 1$, where rewriting always refers to the context. The structural distance dependences of I for j belonging to the left and right branches are shown, respectively, in Figs. 6(a) and 6(b). The horizontal distance dependence is also shown in Figs. 6(c) and 6(d). When j belongs to the right branch, i.e., $j = (1, \dots)$, as shown in the right subfigures (b) and (d), what is observed with a PCFG roughly holds. Here, I decays exponentially in the structural distance and follows a power law in the horizontal distance. However, different behavior is observed when j belongs to the left branch, i.e., $j = (0, \dots)$, as shown in the left subfigures (a) and (c). In Fig. 6(a), I has clearly different values even with the same structural distances, whereas it decays in the power law of the horizontal distance in Fig. 6(c), similarly to the case with $j = (1, \dots)$. This result differs from the behavior found with a PCFG.

The mutual information between nodes in a PCSG depends explicitly on the horizontal distance. This observation can be attributed to the context sensitivity inherent in PCSG rules. If the context-free independence holds, then a node can correlate with other nodes only along the path in the tree graph. This result engenders the exponential decay with the structural distance. In contrast, in a PCSG where each rule involves the context L and R as well as A, a node can correlate with its left and right neighbors directly, even in the absence of a direct

path between them. This horizontal correlation can bypass the long structural distance between two nodes belonging to different subtrees, leading to the effective distance. As shown in Fig. 7, the horizontal distance increases exponentially with the effective distance. If the mutual information does not decay exponentially with the structural distance, but instead with the effective distance, then the mutual information will decay in a power law in the horizontal distance, irrespective of whether node j belongs to the left or the right branch.

The effective distance is definable as explained hereinafter. Presuming that nodes i' and j' are ancestors of i and j, respectively, and that i' and j' are the horizontal neighbors of one another, then the effective distance between nodes i and j
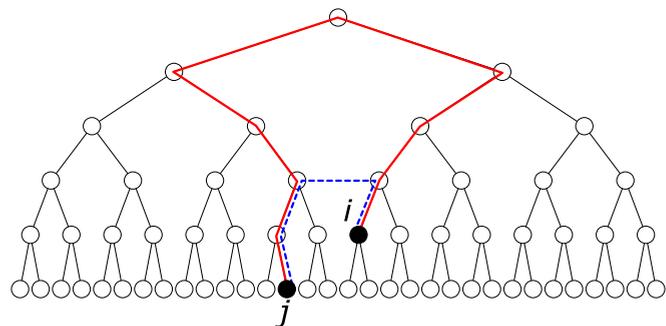
FIG. 7. Structural and effective distances between nodes $i = (1, 0, 0, 0)$ and $j = (0, 1, 1, 0, 1)$. The former is described by the red line, whereas the latter is shown by the blue dashed line.
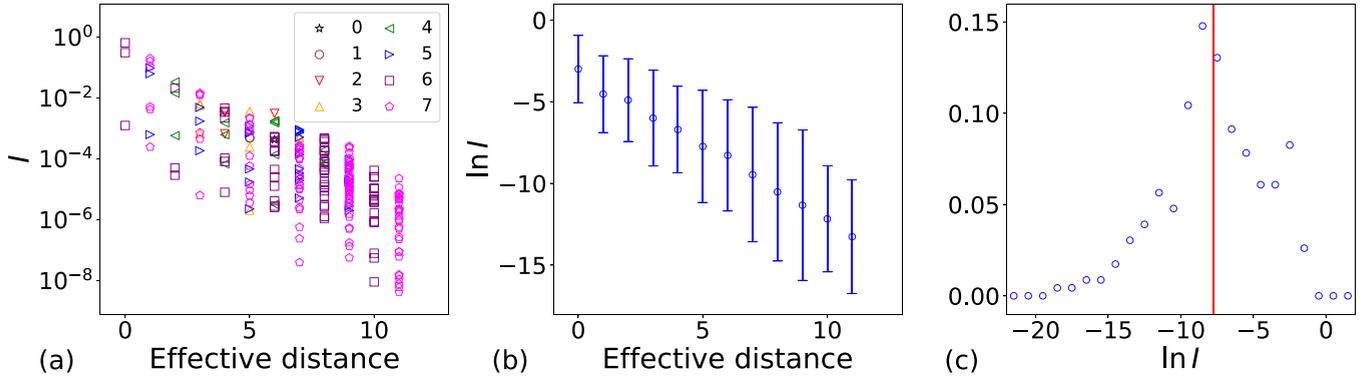
FIG. 8. (a) Mutual information I defined by Eq. (2) against the effective distance between i and j. Weights M are generated from the log-normal distribution. Markers and colors differ for different depths. (b) Averages and standard deviations of $\ln I$ over j's of the same effective distance and over 10 M's generated. (c) Normalized histograms of $\ln I$ for 10 M's for which the effective distance is 5. The red vertical line represents the average. For all (a), (b), and (c), the parameter in the log-normal distribution is $\epsilon = 10^{-2}$, the context sensitivity is $q = 1$, and node i is fixed at $(1, 0, 0, 0, 0, 0)$.

is the sum of the path length from i to i' and from j to j'. Here, we assume that the effective distance is equal to the structural distance if one of the two nodes is the ancestor of the other. We plot the same I as in Fig. 6, but against the effective distance, in Fig. 8(a). From this, it can be confirmed that the mutual information decays exponentially with the effective distance, as expected. This result indicates the existence of a typical effective distance that corresponds to a correlation length, which is the inverse of the decay rate. The mutual information is small beyond this typical distance.

It is intuitively reasonable to infer that the mutual information decays exponentially with the effective distance. Joint probability $P(\sigma_0, \ldots, \sigma_{2^{D+1}-2})$ of all nodes is the product of all $2^D - 1$ rewriting weights. For two nodes i and j, marginalizing the remaining nodes yields the joint probability $P(\sigma_i, \sigma_j)$ of the two nodes. The greatest contribution to this is the product of the weights on the shortest effective path, described by the blue dashed line in Fig. 7. Although an effective path and its corresponding weights depend on the order of application of rules at each step, the length of the shortest effective path asymptotically equals the effective distance. Therefore, the joint probability of two nodes scales as an exponential function of the effective distance. This result implies that the mutual information scales in the same manner [8].

What we describe here is not unique to this instance. It is typically observed across the M's sampled. We measured I for 10 M's under the same settings and computed the averages and the standard deviations of $\ln I$ over j's of each effective distance and over M's. Whereas mutual information is always non-negative, the estimate by the method in [43] sometimes takes negative values when the true value is small. We simply excluded nonpositive estimates to compute the logarithm. This exclusion caused the average to be biased upward, but this bias was slight in this case. The results presented in Fig. 8(b) show that the exponential decay in the effective distance discussed above for a single M is observed across 10 M's. Figure 8(c) also presents histograms of $\ln I$ for the effective distance 5, where the frequencies are normalized. The points are distributed around the average. The deviations in Fig. 8(b) and the distribution in Fig. 8(c) originate from differences in j's and M's rather than from sample fluctuations.

The rate of decay and the correlation length depend on weights M, causing the average rate to change as the parameter ϵ varies. One can infer that, as ϵ increases, the distribution of trees under generated weights M tends to approach the uniform distribution. Therefore, the mutual information is expected to decay faster, meaning that the correlation length will become smaller. Additionally, the rate of decay depends on the context sensitivity q. With larger q, rewriting operations depends not only on the rewritten symbol but also on the context, with higher probability. This dependence seems to engender slower decay. The numerically computed mutual information with different ϵ and q, as presented in the supplemental material [33], follows these expectations.

V. QUANTIFICATION OF CONTEXT-FREE INDEPENDENCE BREAKING

Finally, we investigate the effect of context sensitivity more directly by quantifying the extent to which the context-free independence is broken. This independence means that two subtrees are mutually independent under the condition that the symbols of their roots are fixed. Therefore, quantifying the breakage of the context-free independence involves the measurement of the mutual information between the subtrees under this condition. However, it requires extremely large
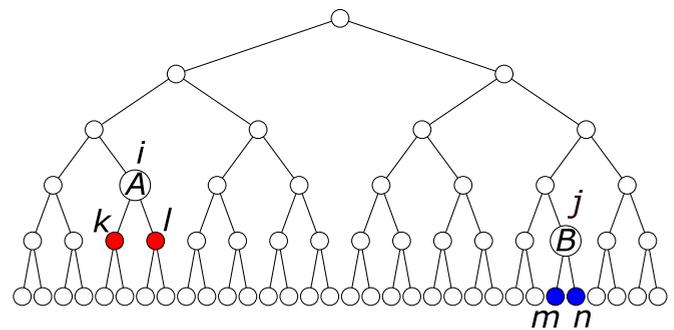
FIG. 9. Context-free independence breaking J defined by Eq. (3) is the mutual information between the red nodes k and l and the blue nodes m and n.
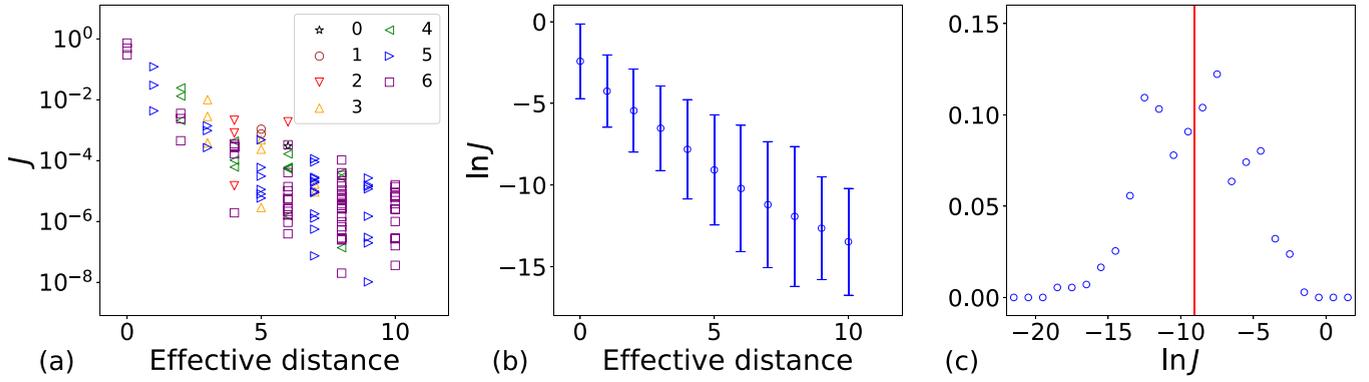
FIG. 10. (a) Degree of the context-free independence breaking, or parent-fixed mutual information J defined by Eq. (3) for $A = B = 0$ against the effective distance between i and j. Weights M are generated from the log-normal distribution. Markers and colors differ for different depths. (b) Averages and standard deviations of $\ln J$ over j's of the same effective distance, over the symbols A of node i and B of j, and over 10 M's generated. (c) Normalized histograms of $\ln J$ for 10 M's with effective distance 5. The red vertical line represents the average. For all (a), (b), and (c), the parameter in the log normal distribution is $\epsilon = 10^{-2}$, the context sensitivity is $q = 1$, and node i is fixed at $(1, 0, 0, 0, 0, 0)$.

amounts of data to obtain the distribution of a subtree when the subtree is large. To overcome this difficulty, we instead specifically examine the mutual information between the children of their roots, as shown in Fig. 9. We denote this mutual information as

$$
\begin{aligned}
& J_{i,j;A,B}(q, M) \\
& \equiv \sum_{\sigma_k, \sigma_l, \sigma_m, \sigma_n} P(\sigma_k, \sigma_l, \sigma_m, \sigma_n | \sigma_i = A, \sigma_j = B) \\
& \times \ln \frac{P(\sigma_k, \sigma_l, \sigma_m, \sigma_n | \sigma_i = A, \sigma_j = B)}{P(\sigma_k, \sigma_l | \sigma_i = A, \sigma_j = B)P(\sigma_m, \sigma_n | \sigma_i = A, \sigma_j = B)}.
\end{aligned}
$$
(3)

Therein, k and l, respectively, represent the left and the right children of i; m and n are the children of j. This quantity is always zero for any i and j in a PCFG because of the context-free independence. This is the requirement that must be met for this quantity to be a meaningful metric of the breaking of independence.

In addition to measuring the degree of context-free independence breaking, the metric J has other interpretations. One interpretation derives from theoretical physics. If the network of interactions forms a tree, where every interaction in the system is between a node and its child, then J is zero. In the presence of loops in the network, as seen in a PCSG, J can take a positive value. In this sense, J represents the degree to which the network of interactions deviates from a tree. Another interpretation is linguistic: Suppose that two constituents or phrases, i.e., subtrees of a derivation, are, for example, a noun phrase and a verb phrase. Under this condition, the structures of the noun phrase and the verb phrase are mutually dependent; J represents the strength of this dependence.

We measured the context-free independence breaking J in the same manner as for the mutual information I in the preceding section, under the same setting, where $i = (1, 0, 0, 0, 0, 0)$, $q = 1$, $\epsilon = 10^{-2}$, and 10^8 trees of the depth 7 were sampled for each M. Our observations revealed that J behaves very similarly to I. Figure 10(a) shows J for $A = B = 0$ against the effective distance for an M generated from

the log-normal distribution. It is evident that J exhibits exponential decay with the effective distance. Again, this finding indicates that there exists a correlation length, or a typical effective distance beyond which the dependence between two subtrees is small. We computed the averages and the standard deviations of $\ln J$ over j's of each effective distance, over M's, and over A and B, using the data size, i.e., the number of generated trees satisfying $\sigma_i = A$ and $\sigma_j = B$, as the weights. Additionally, we simply discarded nonpositive estimates of J, which only led to a small bias. Figure 10(b) presents the results, suggesting that the exponential decay of J with the effective distance occurs across different M's, as well as different A's and B's. Figure 10(c) shows the normalized histogram of $\ln J$ obtained for the effective distance 5, where the data sizes were used as the weights. The distribution of $\ln J$'s centers around the red vertical line representing the average.

The dependence of J on the parameter ϵ and the context sensitivity q exhibits similar tendencies to those observed for I. Particularly, as ϵ increases or q decreases, the decay rate becomes more pronounced whereas the correlation length becomes smaller. Supplemental material [33] provides additional results for different values of ϵ and q. In a general system, I and J do not necessarily behave similarly. Indeed, in a PCFG, I is positive and decays exponentially with the structural distance, whereas J is always zero. It is somewhat nontrivial that both I and J decay exponentially with the effective distance in a PCSG.

VI. CONCLUSION

A PCFG, a simple mathematical model for randomly generating a tree, has been used to model various hierarchical phenomena, including natural languages. This model satisfies the assumptions of context-free independence. Although this feature allows for the theoretical analysis of various properties of a PCFG, the restriction is too strong for a PCFG to be expressive of distributions.

We introduced the simple PCSG by relaxing the context-free independence, and we analyzed its statistical properties

systematically. First, we specifically examined the distribution of a symbol on a single node. This distribution is to a PCSG what magnetization is to a spin system. Although the context sensitivity affects the distribution, its effect brings only continuous and quantitative changes. Such changes can occur even without context sensitivity, for example in the interpolation between two PCFGs. Our numerical investigation also shows that the Binder parameter of the mean ratio of a symbol, which is an analytic function of ϵ in the context-free RLM, is unlikely to be discontinuous in a context-sensitive case.

The second quantity of interest is the mutual information between two nodes, which is related closely to the two-point correlation function [38]. It is noteworthy that mutual information decays exponentially with the effective distance between two nodes, which is a consequence of the horizontal correlation because of context sensitivity. This feature contrasts with the fact that the decay of the mutual information in a PCFG is exponential with respect to the structural distance, i.e., the path length.

In addition, to quantify the degree to which context-free independence is broken, we proposed the use of mutual information between two pairs of nodes under the condition that the parent symbols are fixed. This metric can also indicate the degree to which the network of interactions deviates from a tree in theoretical physics, and it can indicate the mutual dependence between the structures of two constituents in linguistics. This quantity emphasizes the most distinct difference between a PCFG and a PCSG. The context-free independence breaking decays exponentially with the effective distance in a PCSG, similar to the mutual information between two nodes, whereas the breaking always remains zero in a PCFG.

Possible future issues, in our view, are divisible into four main directions. First, it is necessary to develop methods for theoretical analysis and efficient numerical approximation to confirm and further investigate the behaviors of PCSGs observed in this study, such as the exponential decay of the mutual information and the context-free independence breaking. The main challenges are the exponential growth of tree sizes and the complex interactions due to context sensitivity.

Second, another important approach would be to examine specific PCSGs, particularly those exhibiting atypical behavior, in contrast to our analysis of the typical properties of randomly sampled PCSGs. It might be true that PCSGs with low probabilistic measures exhibit nonanalytic behavior in Δ as a function of the context sensitivity q, or nonexponential decay of the mutual information or context-free independence breaking. The existence of such PCSGs and the mechanism underlying their atypical behavior are left as intriguing open problems.

Third, CSG is not the only linguistic framework beyond CFG. Although the CSG framework makes tree structures context-sensitive in a straightforward manner, modern linguists do not consider a CSG to be a relevant model of a natural language. This skepticism arises because a CSG can generate a set of sentences extending beyond natural languages [30]. Also, formal language theory predominantly addresses surface sentences rather than syntactic structures [44]. Conversely, several alternative models have been proposed as grammars closer to natural languages, such as tree adjoining grammar [45], combinatory categorial grammar [46], and minimalist grammar [47]. The natural progression is to introduce probabilistic extensions to these grammars and to investigate their statistical properties, as examined in this study. Particularly, all probabilistic extensions of a CSG and the three grammars described above will violate the context-free independence, but their independence breaking J might decay exponentially, polynomial, or nonmonotonically, depending on the grammar. If the decay is, for example, exponential in every model, then their decay rates might differ. These probabilistic grammars can be characterized by emphasizing the distinctions in their independence breaking J, thereby contributing to a comprehensive understanding of the grammars from a physical perspective.

As a fourth point, we discuss the application of our metric J for the context-free independence breaking, which is applicable not only to probabilistic grammars such as PCFGs but also to any distribution of a tree, including those underlying human languages and birdsongs. Earlier research has demonstrated that the behavior of mutual information in PCFGs, human languages, and birdsongs is similar in that it decays as a power-law function of the horizontal distance or the sequence length [8,41]. However, J will allow us to detect and quantify the distinction between human languages and PCFGs, given the empirical knowledge that context-free independence breaking occurs in natural languages [24,25]. It might also be possible to identify characteristics unique to human languages, which are not present in birdsongs, using J. By quantifying the degree of independence breaking, we can more deeply compare tree structures among different mathematical models or natural phenomena.

[1] N. Chomsky, *Syntactic Structures* (Mouton & Co., Berlin, 1957).

[2] F. Jelinek, J. D. Lafferty, and R. L. Mercer, Basic methods of probabilistic context free grammars, in *Speech Recognition and Understanding*, edited by P. Laface and R. De Mori (Springer, Berlin, 1992), pp. 345–360.

[3] E. Charniak, Statistical techniques for natural language parsing, AI Magazine **18**, 33 (1997).

[4] K. Ellis, A. Solar-Lezama, and J. B. Tenenbaum, Unsupervised learning by program synthesis, Adv. Neural Inf. Process. Syst. **28** (2015).

[5] P. Worth and S. Stepney, Growing music: Musical interpretations of l-systems, in *Applications of Evolutionary Computing*, edited by F. Rothlauf, J. Branke, S. Cagnoni, D. W. Corne, R. Drechsler, Y. Jin, P. Machado, E. Marchiori, J. Romero, G. D. Smith, and G. Squillero (Springer, Berlin, 2005), pp. 545–550.

[6] É. Gilbert and D. Conklin, A probabilistic context-free grammar for melodic reduction, in *Proceedings of the International Workshop on Artificial Intelligence and Music, 20th International Joint Conference on Artificial Intelligence, Hyderabad, India* (2007), pp. 83–94.

[7] P. Tano, S. Romano, M. Sigman, A. Salles, and S. Figueira, Towards a more flexible language of thought: Bayesian grammar updates after each concept exposure, Phys. Rev. E **101**, 042128 (2020).

[8] H. W. Lin and M. Tegmark, Critical behavior in physics and probabilistic formal languages, Entropy **19**, 299 (2017).

[9] B. Knudsen and J. Hein, RNA secondary structure prediction using stochastic context-free grammars and evolutionary history, Bioinformatics **15**, 446 (1999).

[10] D. Harlow, S. H. Shenker, D. Stanford, and L. Susskind, Tree-like structure of eternal inflation: A solvable model, Phys. Rev. D **85**, 063516 (2012).

[11] W. Li, Spatial 1/f spectra in open dynamical systems, Europhys. Lett. **10**, 395 (1989).

[12] W. Li, Expansion-modification systems: A model for spatial 1/f spectra, Phys. Rev. A **43**, 5240 (1991).

[13] R. Lieck and M. Rohrmeier, Recursive bayesian networks: Generalising and unifying probabilistic context-free grammars and dynamic bayesian networks, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, New York, 2021), Vol. 34, pp. 4370–4383.

[14] A. Lindenmayer, Mathematical models for cellular interactions in development I. Filaments with one-sided inputs, J. Theor. Biol. **18**, 280 (1968).

[15] A. Lindenmayer, Mathematical models for cellular interactions in development II. Simple and branching filaments with two-sided inputs, J. Theor. Biol. **18**, 300 (1968).

[16] G. T. Herman and A. Walker, Context free languages in biological systems, Int. J. Comput. Math. **4**, 369 (1974).

[17] K. Nakaishi and K. Hukushima, Absence of phase transition in random language model, Phys. Rev. Res. **4**, 023156 (2022).

[18] T. L. Booth and R. A. Thompson, Applying probability measures to abstract languages, IEEE Trans. Comput. **C-22**, 442 (1973).

[19] M. I. Miller and J. A. O'Sullivan, Entropies and combinatorics of random branching processes and context-free languages, IEEE Trans. Inf. Theor. **38**, 1292 (1992).

[20] Z. Chi, Statistical properties of probabilistic context-free grammars, Comput. Linguist. **25**, 131 (1999).

[21] J. Esparza, A. Gaiser, and S. Kiefer, A strongly polynomial algorithm for criticality of branching processes and consistency of stochastic context-free grammars, Inf. Process. Lett. **113**, 381 (2013).

[22] S. M. Shieber, Evidence against the context-freeness of natural language, Linguist. Philos. **8**, 333 (1985).

[23] C. Culy, The complexity of the vocabulary of bambara, Linguist. Philos. **8**, 345 (1985).

[24] M. Johnson, T. Griffiths, and S. Goldwater, Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models, Adv. Neural Inf. Process. Syst. **19** (2006).

[25] T. J. O'Donnell, J. B. Tenenbaum, and N. D. Goodman, Fragment grammars: Exploring computation and reuse in language, Technical Report No. MIT-CSAIL-TR-2009-013, Massachusetts Institute of Technology, Cambridge, MA, 2009.

[26] N. Chomsky, Three models for the description of language, IRE Trans. Inf. Theor. **2**, 113 (1956).

[27] E. DeGiuli, Random language model, Phys. Rev. Lett. **122**, 128301 (2019).

[28] E. DeGiuli, Emergence of order in random languages, J. Phys. A **52**, 504001 (2019).

[29] J. E. Hopcroft, R. Motwani, and J. Ullman, *Introduction to Automata Theory, Languages, and Computation*, 3rd ed. (Addison-Wesley, Boston, 2007).

[30] G. Jäger and J. Rogers, Formal language theory: refining the chomsky hierarchy, Philos. Trans. R. Soc., B **367**, 1956 (2012).

[31] Although the process of rewriting a symbol as a function of the symbol and its neighbors is similar to that of an elementary cellular automaton, our model has several features. In our model, rewriting operations is asynchronous and random. Furthermore, the number of cells or symbols grows exponentially because a single symbol becomes two symbols. A cellular automaton with asynchronous and random updates is called an asynchronous cellular automaton [48]. Our PCSG can be regarded as a modified version of an asynchronous cellular automaton such that the system size grows.

[32] F. Y. Wu, The Potts model, Rev. Mod. Phys. **54**, 235 (1982).

[33] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.6.033216 for additional information.

[34] K. Binder, Finite size scaling analysis of ising model block distribution functions, Z. Phys. B **43**, 119 (1981).

[35] K. Binder and D. P. Landau, Finite-size scaling at first-order phase transitions, Phys. Rev. B **30**, 1477 (1984).

[36] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC Monographs on Statistics and Applied Probability (Chapman and Hall, London, 1993).

[37] P. Young, Everything you wanted to know about data analysis and fitting but were afraid to ask, arXiv:1210.3781.

[38] W. Li, Mutual information functions versus correlation functions, J. Stat. Phys. **60**, 823 (1990).

[39] W. Li, Mutual information functions of natural language texts (1989), SFI Working Paper, https://www.santafe.edu/research/results/working-papers/mutual-information-functions-of-natural-language-t.

[40] K. Tanaka-Ishii, *Statistical Universals of Language: Mathematical Chance vs. Human Choice* (Springer, Cham, 2021).

[41] T. Sainburg, B. Theilman, M. Thielk, and T. Q. Gentner, Parallels in the sequential organization of birdsong and human speech, Nat. Commun. **10**, 3636 (2019).

[42] W. Li and K. Kaneko, Long-Range correlation and partial 1/f spectrum in a noncoding DNA sequence, Europhys. Lett. **17**, 655 (1992).

[43] P. Grassberger, Entropy estimates from insufficient samplings, arXiv:physics/0307138.

[44] N. Fukui, A note on weak vs. strong generation in human language, Stud. Chin. Linguist. **36**, 59 (2015).

[45] A. K. Joshi, Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? in *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, Studies in Natural Language Processing, edited by D. R. Dowty, L. Karttunen, and A. M. Zwicky (Cambridge University Press, Cambridge, 1985), pp. 206–250.

[46] M. Steedman, Combinatory grammars and parasitic gaps, Nat. Lang. Linguist. Theor. **5**, 403 (1987).

[47] E. Stabler, Derivational minimalism, in *Logical Aspects of Computational Linguistics*, edited by C. Retoré (Springer, Berlin, 1997), pp. 68–95.

[48] N. Fatès, Asynchronous cellular automata, *Encyclopedia of Complexity and Systems Science* (Springer, New York, 2018), p. 21.