Editors' Suggestion Featured in Physics

Random Language Model

E. DeGiuli
Institut de Physique Théorique Philippe Meyer, École Normale Supérieure, PSL University, Sorbonne Universités, CNRS, 75005 Paris, France

(Received 11 September 2018; revised manuscript received 5 February 2019; published 29 March 2019)

Many complex generative systems use languages to create structured objects. We consider a model of random languages, defined by weighted context-free grammars. As the distribution of grammar weights broadens, a transition is found from a random phase, in which sentences are indistinguishable from noise, to an organized phase in which nontrivial information is carried. This marks the emergence of deep structure in the language, and can be understood by a competition between energy and entropy.

DOI: 10.1103/PhysRevLett.122.128301

It is a remarkable fact that structures of the most astounding complexity can be encoded into sequences of digits from a finite alphabet. Indeed, the complexity of life is written in the genetic code, with an alphabet $\{A, T, C, G\}$, proteins are coded from strings of 20 amino acids, and human-written text is composed in small, fixed alphabets. This "infinite use of finite means" [1] was formalized by Post and Chomsky with the notion of generative grammar [2,3], and has been elaborated upon since, both by linguists and computer scientists [4]. A generative grammar consists of an alphabet of hidden symbols, an alphabet of observable symbols, and a set of rules, which allow certain combinations of symbols to be replaced by others. From an initial start symbol S, one progressively applies the rules until only observable symbols remain; any sentence produced this way is said to be "grammatical," and the set of all such sentences is called the language of the grammar. The sequence of rule applications is called a derivation. For example, the grammar $\{S \to SS, S \to (S), S \to ()\}$ has a single hidden symbol S and two observable symbols, (and), and produces the infinite set of all strings of well-formed parentheses. A simple derivation in this grammar is $S \to SS \to (S)S \to (())S \to (())()$. Besides their original use in linguistics, where the observable symbols are typically taken to be words, and grammars produce sentences [Fig. 1(a)] [3,5], generative grammars have found applications in manifold domains: in the secondary structure of ribonucleic acid (RNA) [Fig. 1(b)] [6,7], in compiler design [4], in self-assembly [8], in protein sequence analysis [9], and in quasicrystals [10], to name a few.

The complexity of a language is limited by conditions imposed on its grammar, as described by the Chomsky hierarchy, which, in increasing complexity, distinguishes regular, context-free, context-sensitive, and recursively enumerable grammars [11]. Each class of grammar has a characteristic graphical structure of its derivations: regular grammars produce linear derivations, context-free grammars produce trees (Fig. 1), and context-sensitive and recursively enumerable grammars produce more elaborate graphs. Associated with an increase in complexity is an increased difficulty of parsing [4]. Because biological instantiations of grammars must have been discovered by evolution, there is a strong bias toward simpler grammars; we consider context-free grammars (CFGs), which are the lowest order of the Chomsky hierarchy that supports hierarchical structure.

Despite their ubiquity in models of complex generative systems, grammars have hitherto played a minor role in physics, and most known results on grammars are theorems regarding worst-case behavior [12], which need not represent the typical case. Human languages show Zipf's law [13–15], a power-law dependence of word frequency on its rank, and many sequences, including human text, show long-range information-theoretic correlations [16–18], which can be created by a CFG [18]; but are these *typical* features of some ensemble of grammars? In this work we initiate this research program by proposing and simulating an ensemble of CFGs, so that grammars can be considered as physical systems [19]. We will find that CFGs possess two natural "temperature" scales that control grammar complexity, one at the surface interface, and another in

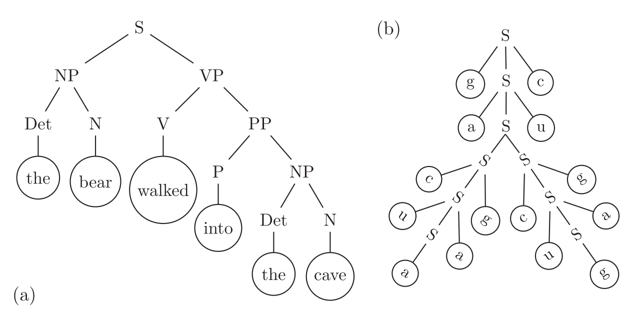

FIG. 1. Illustrative derivation trees for (a) simple English sentence, and (b) the RNA secondary structure (after [6]). The latter is a derivation of the sequence "gacuaagcugaguc" and shows its folded structure. Terminal symbols are encircled.

the tree interior. As either of these temperatures is lowered, there is a phase transition, which corresponds to the emergence of nontrivial information propagation. We characterize this phase transition using results from simulations, and understand its location by a balance between energy and entropy.

Generative grammars.—A generative grammar is defined by an alphabet χ and a set of rules $\mathcal{R}$. The alphabet has N hidden, "nonterminal" symbols χ_H, and T observable, "terminal" symbols χ_O. The most general rule is of the form $a_1 a_2 \ldots a_n \to b_1 b_2 \ldots b_m$, where $a_i \in \chi_H$, $b_i \in \chi = \chi_H \cup \chi_O$. In a CFG the rules are specialized to the form $a_1 \to b_1 b_2 \ldots b_m$, and we will insist that $m \geq 1$, so that there is no "empty" string. Without loss of generality, we consider CFGs in the Chomsky normal form, in which case all rules are of the form [4] $a \to bc$ or $a \to A$, where a, b, $c \in \chi_H$ and $A \in \chi_O$. Note that we may have $b = a$, or $b = c$, or $a = b = c$. Any derivation in the Chomsky reduced form can be drawn on a binary tree. Beginning from the start symbol $S \in \chi_H$, rules are applied until the string contains only observable symbols. Such a string is called a sentence. The set of all sentences is the language of the grammar. Given a string of observables $\mathcal{S} = A_1 \ldots A_\ell$ and a grammar $\mathcal{G}$, one can ask whether there exists a derivation that produces $\mathcal{S}$ from the start symbol S; if so, $\mathcal{S}$ is said to be grammatical.

A formal grammar as defined above can only distinguish grammatical from ungrammatical sentences. A richer model is obtained by giving each rule a non-negative real valued weight. Such a weighted grammar is useful in applications, because weights can be continuously driven by a learning process, and can be used to define probabilities of parses. Moreover, a weighted grammar can be put into the Gibbs form, as shown below. For CFGs, to every rule of the form $a \to bc$ we assign a weight M_{abc}, and to every rule of the form $a \to A$ we assign a weight O_{aA}.

Each candidate derivation of a sentence has two different types of degrees of freedom. There is the topology $\mathcal{T}$ of the tree, namely the identity (terminal or nonterminal) of each node, as well as the variables, both terminal and nonterminal, on the nodes. We write $\Omega_\mathcal{T}$ for the set of internal factors, i.e., factors of the form $a \to bc$, and $\partial\Omega_\mathcal{T}$ for the boundary factors, i.e., those associated to $a \to A$ rules. The number of boundary factors is written $\ell_\mathcal{T}$, which is also the number of leaves. Since derivations are trees, the number of internal factors is $\ell_\mathcal{T} - 1$. We will write σ for nonterminal symbols, and o for terminals; these can be enumerated in an arbitrary way $1, \ldots, N$ and $1, \ldots, T$, respectively. Given $\mathcal{T}$, we can write σ_i for the value of the nonterminal on site i, and similarly o_j for the terminal on site j. The number of σ_i is $2\ell_\mathcal{T} - 1$, while the number of o_j is $\ell_\mathcal{T}$. We write $\mathcal{G}$ for the pair M, O, σ for $\{\sigma_i\}$, and o for $\{o_t\}$.

To define a probability measure on derivations, it is convenient to factorize it into the part specifying $\mathcal{T}$, and the remainder. In this way, we separate the tree shape from the influence of the grammar on variables. For a fixed $\mathcal{T}$, the weight of a configuration is

$$W(\sigma, o|\mathcal{T}, \mathcal{G}) = \prod_{\alpha\in\Omega_\mathcal{T}} M_{\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}} \prod_{\alpha\in\partial\Omega_\mathcal{T}} O_{\sigma_{\alpha_1} o_{\alpha_2}}, \qquad (1)$$

where each $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ is a factor in the order $\sigma_{\alpha_1} \to \sigma_{\alpha_2}\sigma_{\alpha_3}$. Note that $M_{abc} \neq M_{acb}$ in general; thus the left and right branches are distinguished [20]. We can write $W = e^{-E}$ with

$$E = -\sum_{a,b,c} \pi_{abc}(\sigma) \log M_{abc} - \sum_{a,B} \rho_{aB}(\sigma, o) \log O_{aB}, \quad (2)$$

where π_{abc} is the number of times the rule $a \to bc$ appears in the configuration σ, and likewise, ρ_{aB} is the number of times the rule $a \to B$ appears. This defines a conditional probability measure on configurations $\mathbb{P}(\sigma, o|\mathcal{T}, \mathcal{G}) = e^{-E(\sigma,o|\mathcal{T},\mathcal{G})}/Z(\mathcal{T}, \mathcal{G})$ where

$$Z(\mathcal{T}, \mathcal{G}) = \sum_{\{\sigma_i, o_t\}} e^{-E(\sigma,o|\mathcal{T},\mathcal{G})}. \qquad (3)$$

All configurations have S at the root node. For simplicity, in this work we consider as a model for the tree topology probability $\mathbb{P}(\mathcal{T}|\mathcal{G}) = W_{\text{tree}}/Z_{\text{tree}}$ with $W_{tree}(\mathcal{T}) = p^{|\partial\Omega_\mathcal{T}|}(1-p)^{|\Omega_\mathcal{T}|}$, where p is the emission probability, the probability that a hidden node becomes an observable node. p controls the size of trees; we will choose it such that the tree size distribution is cut off above a length $\xi = 1000$. Some facts about the resulting binary trees are recorded in the Supplemental Material (SM) [22].

A model with weights of the form of Eq. (1) is called a weighted CFG (WCFG). In the particular case where $1 = \sum_{b,c} M_{abc} = \sum_A O_{aA}$ for all a, it is easy to see that M and O are conditional probabilities: $M_{abc} = \mathbb{P}(a \to bc|a \to \text{nonterminal})$ and $O_{aA} = \mathbb{P}(a \to A|a \to \text{terminal})$. In this case the model is called a probabilistic CFG (PCFG). In the main text, we consider a weighted CFG, model W; in the SM, we show that our results are robust in model P, a PCFG. There are tradeoffs between these models: model P is easier to sample, because it has $Z(\mathcal{T}, \mathcal{G}) = 1$ from normalization of probability, and thus is factorized. But model W is more amenable to theory, since it is less constrained.

Random language model.—Each grammar defines probabilities for sentences. To extract the universal properties of grammars, which do not depend on all details of M and O, we need a measure on the space of grammars. What is an appropriate measure? From Eq. (2), $\log M$ and $\log O$ are analogous to coupling constants in statistical mechanics. A simple model is to assume a Gaussian distribution for these, so that M and O are lognormal. This can be motivated as follows: language evolution is a dynamical process, which must be slow in order for language to remain comprehensible at any given moment. If each $\log M_{abc}$ and $\log O_{aB}$ are the accumulation of independent, additive increments [25], these will lead to a lognormal. We define deep and surface sparsities as, respectively,

$$s_d = \frac{1}{N^3}\sum_{a,b,c} \log^2\left(\frac{M_{abc}}{\bar{M}}\right), \quad s_s = \frac{1}{NT}\sum_{a,B} \log^2\left(\frac{O_{aB}}{\bar{O}}\right), \tag{4}$$

where $\bar{M} = 1/N^2$ and $\bar{O} = 1/T$ are the corresponding uniform probabilities; it is convenient to use this normalization even for model W where weights are not strictly normalized. A lognormal distribution of grammar weights is

$$\mathbb{P}_G(M, O) \equiv Z_G^{-1} J e^{-\epsilon_d s_d} e^{-\epsilon_s s_s}, \tag{5}$$

where $J = e^{-\sum_{a,b,c} \log M_{abc} - \sum_{a,B} \log O_{aB}}$, and the space of M and O is defined by appropriate normalization and positivity constraints. We define the random language model (RLM) as the ensemble of grammars drawn from Eq. (5).

An alternative motivation of Eq. (5) is that this is the maximum-entropy measure when the grammar averages $\bar{s_d}$ and $\bar{s_s}$ are constrained. s_d and s_s measure the density of rules about their respective median values $\bar{M}$ and $\bar{O}$. When s_d and s_s are finite, all rules must have a finite probability: this reflects the fact that, given any finite amount of data, one can only put a lower bound on the probability of any particular rule. In model W the Lagrange multipliers ϵ_d and ϵ_s satisfy

$$\overline{s_d} = \frac{N^3}{2\epsilon_d}, \qquad \overline{s_s} = \frac{NT}{2\epsilon_s}. \tag{6}$$

When $\epsilon_d \to \infty$, $\bar{s_d} \to 0$, which is the value corresponding to a completely uniform deep grammar, that is, when for a nonterminal a, all rules $a \to bc$ have the same probability $1/N^2$. This is clearly the limit in which the grammar carries no information. As ϵ_d is lowered, s_d increases, and the grammar carries more information. In terms of how deterministic the rules are, ϵ_d plays the role of temperature, with random $\leftrightarrow$ hot and deterministic $\leftrightarrow$ cold; we will refer to it as the deep temperature. This analogy can also be seen formally: in the SM, we show that if the energy E is replaced by βE, then Eq. (6) is replaced by $\overline{s_d} = \beta^2 N^3/(2\epsilon_d)$, such that lowering ϵ_d is equivalent to increasing β. Similarly, ϵ_s controls information transmission at the surface; we call it the surface temperature.

To investigate the role of ϵ_d on language structure, we sampled grammars from the RLM at fixed values $T = 27$, $\epsilon_s/(NT) = 0.01$. Since the surface sparsity is large, there is already some simple structure at the surface; we will explore how deep structure emerges as N and ϵ_d are varied. For each value of N and ϵ_d, we created 120 distinct grammars, from which we sample 200 sentences (see SM for more details). Altogether, approximately 7200 distinct languages were constructed.

The information content of a grammar $\mathcal{G}$ is naturally encoded by Shannon entropies. For a sequence $o_1, o_2, \ldots, o_k$ the Shannon block entropy rate is

$$H_s(\mathcal{G}; k) = \frac{1}{k}\langle \log 1/\mathbb{P}(o_1, o_2, \ldots, o_k|\mathcal{G})\rangle. \tag{7}$$

For CFGs, we can also consider the block entropy rate of deep configurations,

$$H_d(\mathcal{G}; k) = \frac{1}{k}\langle \log 1/\mathbb{P}(\sigma_1, \sigma_2, \ldots, \sigma_k|\mathcal{G})\rangle, \tag{8}$$

where the symbols are taken from a (leftmost) derivation. In both cases the ensemble average is taken with the actual probability of occurrence, $\mathbb{P}(o|\mathcal{G})$ for H_s, and $\mathbb{P}(\sigma|\mathcal{G})$ for H_d.

The grammar averages $\bar{H}_d(k)$ and $\bar{H}_s(k)$ are shown in Fig. 2, for k as indicated; here and in the following, the bars show the 20th and 80th percentiles, indicating the observable range of H_d and H_s over the ensemble of grammars [26]. The dependence on ϵ_d is striking: for $\epsilon_d \gtrsim N^3/\log^2 N$, both $\bar{H}_s(1)$ and $\bar{H}_d(1)$ are flat. In this regime, $\bar{H}_d(1) \approx \log N$, indicating that although configurations strictly follow the rules of a WCFG, deep configurations are nearly indistinguishable from completely random configurations. However, at $\epsilon_d = \epsilon_* \approx N^3/\log^2 N$ there is a pronounced transition, and both entropies begin to drop. This transition corresponds to the emergence of deep structure.

The first block entropy $H_d(\mathcal{G}; 1)$ measures information in the single-character distribution, while the differential entropies $\delta H_d(\mathcal{G}; k) = (k+1)H_d(\mathcal{G}; k+1) - kH_d(\mathcal{G}; k)$ measure incremental information in the higher-order distributions [17]. The Shannon entropy rate including all correlations can either be obtained from $\lim_{k\to\infty} H_d(\mathcal{G}; k)$, or from $\lim_{k\to\infty} \delta H_d(\mathcal{G}; k)$. These coincide, but the latter converges faster [17]. In the SM, we show that $\delta H_d(\mathcal{G}; k)$, and thus, the limiting rate appears to collapse with $\tilde{\epsilon}_d \log N$. For all entropies, the sample-to-sample fluctuations decrease rapidly with k, suggesting that the limiting rates are self-averaging.

To further investigate the nature of the transition, we show in Fig. 3(a) a Zipf plot: the frequency of each symbol, arranged in decreasing order. Figure 3(a) shows the Zipf plot for deep structure; the Zipf plot for surface structure is

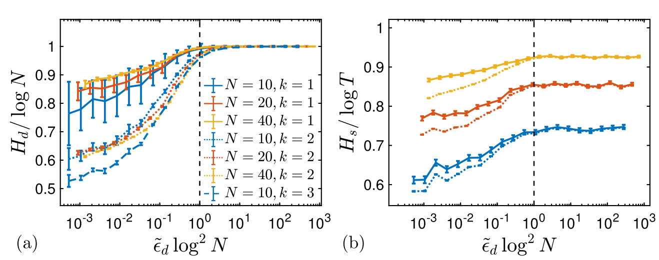

FIG. 2. Shannon entropy of random CFGs as functions of $\tilde{\epsilon}_d = \epsilon_d/N^3$. (a) Block entropy of hidden configurations for indicated k and N. (b) Block entropy of observed strings; symbols as in (a). The constant value for $\epsilon_d > \epsilon_*$ depends on the surface temperature ϵ_s. Bars indicate 20th and 80th percentiles.

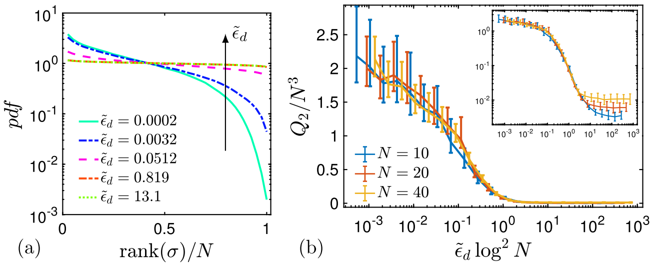

FIG. 3. (a) Zipf plot of hidden symbols for $N = 40$. Here $\tilde{\epsilon}_d = \epsilon_d/N^3$. (b) Order parameter Q_2, with bars indicating 20th and 80th percentile ranges over grammars at each parameter value. Inset: same plot in log-log axes.

similar, but less dramatic (see SM). We see a sharp change at ϵ_*: for $\epsilon_d > \epsilon_*$, the frequencies of hidden symbols are nearly uniform, while below ϵ_*, the distribution is closer to exponential (In the SM, we show that a power-law regime for the observable symbols appears when T is large). The permutation symmetry among hidden symbols is thus spontaneously broken at ϵ_*.

What is the correct order parameter to describe this transition? The ferromagnetic order parameter is $m_r = \overline{\langle N\delta_{\sigma_i,r} - 1\rangle}$, where i is a site. This does not show any signal of a transition, despite the fact that the start symbol explicitly breaks the replica symmetry. A more interesting choice is one of Edwards-Anderson type, such as $Q_{rs}^{EA} = \overline{\langle N\delta_{\sigma_i,r} - 1\rangle\langle N\delta_{\sigma_i,s} - 1\rangle}$ where r and s label different sentences produced from the same grammar, and σ_i is a specified site [27]. However, sentences produced by a CFG do not have fixed derivation trees, so we need to compare symbols in relative position. For each interior rule $a \to bc$ we can define

$$Q_{abc}(\mathcal{G}) = \langle \delta_{\sigma_{\alpha_1},a}(N^2\delta_{\sigma_{\alpha_2},b}\delta_{\sigma_{\alpha_3},c} - 1)\rangle, \qquad (9)$$

averaged over all interior vertices α, and averaged over derivations. Here, σ_{α_1} is the head symbol at vertex α, and σ_{α_2}, σ_{α_3} are the left and right symbols, respectively. Q measures patterns in rule application at each branching of a derivation tree. It is thus an order parameter for deep structure. Upon averaging over grammars in the absence of any fields, the permutation symmetry must be restored: $\bar{Q}_{abc} = q_0 + \delta_{ab}q_l + \delta_{ac}q_r + \delta_{bc}q_h + \delta_{ab}\delta_{ac}q_*$. As shown in the SM, these components show a transition, but there is significant noise below ϵ_*, despite there being 120 replicas at each point. Evidently, Q_{abc} has large fluctuations below ϵ_*. This suggests a definition $Q_2 \equiv \overline{\sum_{a,b,c} Q_{abc}^2}$, plotted in Fig. 3(b). The signal is clear: on the large scale, Q_2 has a scaling form $Q_2 \approx N^3 f(\epsilon_d/\epsilon_*)$ and is small above ϵ_*. The scaling $Q_2 \sim N^3$ suggests that below the transition, all hidden symbols start to carry information in the deep structure.

Theory.—How can we gain some theoretical insight into the RLM? Consider the entropy of an observed string of length ℓ, composed of n sentences of length ℓ_k, $\sum_k \ell_k = \ell$. The entropy of this string derives from three distinct combinatorial levels: (i) each sentence can be represented by a derivation tree with many different topologies, (ii) each derivation tree can host a variety of internal hidden variables, and (iii) given the hidden variables, the observed symbols can themselves vary.

Some scaling considerations are useful. Each derivation tree can have many topologies: the entropy of binary trees scales as $\ell_k \log 4$, so that the total tree entropy scales as $S_t \sim \ell \log 4$. Each derivation tree has $2\ell_k - 1$ hidden variables, so that the total number of hidden d.o.f. is $2\ell - n$, and the corresponding deep entropy scales as $S_d \sim (2\ell - n)\log N$. Finally, the sentences have an entropy $S_o \sim \ell \log T$.

We see that when typical sentences are of length $\langle \ell \rangle \gg 1$, so that $\ell - n \sim \ell$, these numbers are independent of partitioning, to the leading order. For large $\langle \ell \rangle$, we get the scaling $S \sim \ell \log(4N^2T)$.

This must be compared with the "energetic" terms $\log W_{\text{tree}} = (\ell - n)\log(1-p) + \ell \log p \sim -2\ell \log 2$ for p near $1/2$, and E, Eq. (2). In E, π is positively correlated with M, since rules with a higher weight are more frequently used; hence we can obtain a simple scaling estimate $E \sim -N^3\pi \log m - NT\rho \log o$ where π is the mean value of π_{abc}, and $\log m$ is the value of a typical positive fluctuation of $\log M_{abc}$, and is similarly for O. From the sum rules $\sum_{a,b,c}\pi_{abc} = |\Omega| = \ell - n$ and $\sum_{a,B}\rho_{aB} = |\partial\Omega| = \ell$ we have $\pi = (\ell - n)/N^3$, $\rho = \ell/(NT)$. The mean value of $\log M_{abc}$ is $\log \bar{M}$, and the mean value of $\log O_{aB}$ is $\log \bar{O}$. These contributions lead to a constant value of E. The positive fluctuations in $\log M$ and $\log O$ that couple to E scale as $\sqrt{(N^3/2\epsilon_d)}$ and $\sqrt{(NT/2\epsilon_s)}$, respectively, leading to

$$E \sim -\ell\sqrt{\frac{N^3}{2\epsilon_d}} - \ell\sqrt{\frac{NT}{2\epsilon_s}} + \text{const.} \qquad (10)$$

Combining this with S, the effective free energy $F = E - \log W_{\text{tree}} - S$ reflects a competition between energy and entropy. If we consider N and ϵ_d as varying, then there is a scale $\epsilon_* = N^3/\log^2 N$ where the energetic fluctuations balance entropy. For $\epsilon_d \gg \epsilon_*$, the energy of a configuration is unimportant, and the grammar is thus irrelevant: the language produced by the WCFG must then be indistinguishable from random sequences, as found empirically above. In contrast, for $\epsilon_d \ll \epsilon_*$, the language reflects those sequences with high intrinsic weight, and their entropy is less important. The characteristic scale ϵ_* identified by these simple arguments agrees with that found empirically above, and locates the emergence of deep structure.

However, further work is needed to predict the behavior of Q_2, H_s, and H_d.

Learning human languages.—Around 6000 languages are spoken around the world [28]; given fractured and highly sparse input, how does a child come to learn the precise syntax of one of these many languages? This question has a long history in linguistics and cognitive science [29,30]. One scenario for learning is known as the Principles and Parameters (P&P) theory [31]. This posits that the child is biologically endowed with a general class of grammars, the "principles," and by exposure to one particular language, fixes its syntax by setting some number of parameters, assumed to be binary. For example, the head-directionality parameter controls whether verbs come before or after objects, like English and Japanese, respectively. A vast effort has been devoted to mapping out the possible parameters of human languages [28,32]. The richness of the discovered structure has been used as criticism of the approach [33]: if the child needs to set many parameters, then do these all need to be innate? This would be a heavy evolutionary burden, and a challenge to efficient learning.

The RLM can shed some light on this debate. First, since only two living human languages are known to possess syntax beyond CFG [34], we consider WCFGs a valid starting point [37]. Following experimental work [30], we picture the learning process as follows. Initially, the child does not know the rules of the grammar, so it begins with some small number of hidden symbols and assigns uniform values to the weights M and O. To learn is to increase the likelihood of the grammar by adjusting the weights and adding new hidden symbols. As weights are driven away from uniform values, the temperatures ϵ_d and ϵ_s decrease. Eventually, the transition to deep structure is encountered, and the grammar begins to carry information.

In the absence of any bias, this transition would occur suddenly and dramatically, spontaneously breaking all N^3 directions in M space simultaneously, as in Fig. 3(b). However, in realistic child language learning, the child's environment acts as a field on this likelihood ascent, and can cause the structure-emerging transitions to occur at different critical deep temperatures, depending on their coupling to the field. For example, a left-right symmetry breaking could correspond to setting the head directionality parameter.

Although this description is schematic, we insist that the various symmetry-breaking transitions, which could give rise to parameters, are emergent properties of the model. Thus, if there are indeed many parameters to be set, these do not all need to be innate: the child only needs the basic structure of a WCFG, and the rest is emergent. The P&P theory is thus consistent with existence of many parameters. If the RLM can be solved, by which we mean that the partition function Z can be computed, then the series of symmetry-breaking transitions that occur in the presence of a field can be inferred, and a map of syntax in CFGs could be deduced. This is a tantalizing goal for future work.

Conclusion.—We introduced a model of random languages, which captures the generative aspect of complex systems. The model has a transition in parameter space that corresponds to the emergence of deep structure. Since the interaction is long-range, we expect that the RLM, or a variant, is exactly solvable. We hope that this will be clarified in the future.

This work benefited from discussions with C. Callan, J. Kurchan, G. Parisi, R. Monasson, G. Semerjian, P. Urbani, F. Zamponi, A. Zee, and Z. Zeravcic.

[1] W. Von Humboldt, *Humboldt:'On Language': On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species* (Cambridge University Press, Cambridge, England, 1999).

[2] E. L. Post, Formal reductions of the general combinatorial decision problem, Am. J. Math. **65**, 197 (1943).

[3] N. Chomsky, *Syntactic Structures* (Walter de Gruyter, Berlin, 2002).

[4] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, 3rd ed. (Pearson, Boston, Ma, 2007).

[5] N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, Massachusetts, 2014), Vol. 11.

[6] D. B. Searls, The language of genes, Nature (London) **420**, 211 (2002).

[7] B. Knudsen and J. Hein, Pfold: RNA secondary structure prediction using stochastic context-free grammars, Nucleic Acids Res. **31**, 3423 (2003).

[8] E. Winfree, X. Yang, and N. C. Seeman, Universal computation via self-assembly of DNA: Some theory and experiments, in *DNA Based Computers II*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science Vol. 44 (American Mathematical Society, Providence, R.I., 1999), p. 191.

[9] J. P. Barton, A. K. Chakraborty, S. Cocco, H. Jacquin, and R. Monasson, On the entropy of protein families, J. Stat. Phys. **162**, 1267 (2016).

[10] J. G. Escudero, Formal Languages for quasicrystals, in *Symmetries in Science IX* (Springer, Boston, 1997), pp. 139–152.

[11] M. A. Nowak, N. L. Komarova, and P. Niyogi, Computational and evolutionary aspects of language, Nature (London) **417**, 611 (2002).

[12] For example, from Ref. [4], Theorem 7.17 on the size of derivation trees, Theorem 7.31 on the conversion of an automaton to a CFG, and Theorem 7.32 on the complexity of conversion to the Chomsky normal form (see below).

[13] G. K. Zipf, *The Psycho-Biology of Language: An Introduction to Dynamic Philology* (Routledge, Milton Park, 2013).

[14] R. F. i. Cancho and R. V. Solé, Least effort and the origins of scaling in human language, Proc. Natl. Acad. Sci. U.S.A. **100**, 788 (2003).

[15] A. Corral, G. Boleda, and R. Ferrer-i Cancho, Zipf's law for word frequencies: Word forms versus lemmas in long texts, PLoS One **10**, e0129031 (2015).

[16] W. Ebeling and T. Pöschel, Entropy and long-range correlations in literary English, Europhys. Lett. **26**, 241 (1994).

[17] T. Schürmann and P. Grassberger, Entropy estimation of symbol sequence, Chaos **6**, 414 (1996).

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

[19] G. Parisi, Complex systems: A physicist's viewpoint, Physica (Amsterdam) **263A**, 557 (1999).

[20] Indeed if the left-right branches are not distinguished, CFGs do not have any more expressive power than regular grammars [21].

[21] J. Esparza, P. Ganty, S. Kiefer, and M. Luttenberger, Parikh's theorem: A simple and direct automaton construction, Inf. Proc. Lett. **111**, 614 (2011).

[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.122.128301 which includes details on binary trees, sampling methods, robustness in PCFG, differential entropies, and equation derivations, and Refs. [23,24].

[23] S. Chib and E. Greenberg, Understanding the Metropolis-Hastings algorithm, Am. Stat. **49**, 327 (1995).

[24] P. Flajolet and R. Sedgewick, *Analytic Combinatorics* (Cambridge University Press, Cambridge, England, 2009).

[25] D. Sornette and R. Cont, Convergent multiplicative processes repelled from zero: Power laws and truncated power laws, J. Phys. I (France) **7**, 431 (1997).

[26] The error bars in measurements are then smaller by factor approximately $\sqrt{120} \sim 11$.

[27] D. J. Gross, I. Kanter, and H. Sompolinsky, Mean-Field Theory of the Potts Glass, Phys. Rev. Lett. **55**, 304 (1985).

[28] M. C. Baker, *The Atoms of Language: The Mind's Hidden Rules of Grammar* (Basic Books, New York, 2008).

[29] R. C. Berwick, P. Pietroski, B. Yankama, and N. Chomsky, Poverty of the stimulus revisited, Cogn. Sci. **35**, 1207 (2011).

[30] C. Yang, S. Crain, R. C. Berwick, N. Chomsky, and J. J. Bolhuis, The growth of language: Universal Grammar, experience, and principles of computation, Neurosci. Biobehav. Rev. **81**, 103 (2017).

[31] N. Chomsky, *Lectures on Government and Binding: The Pisa Lectures* (Walter de Gruyter, Berlin, 1993), Vol. 9.

[32] U. Shlonsky, The cartographic enterprise in syntax, Lang. Linguist. Compass **4**, 417 (2010).

[33] G. Ramchand and P. Svenonius, Deriving the functional hierarchy, Lang. Sci. **46**, 152 (2014).

[34] Only Swiss-German and Bambara have confirmed features beyond CFG [35,36].

[35] C. Culy, The complexity of the vocabulary of Bambara, Linguistics and philosophy **8**, 345 (1985).

[36] S. M. Shieber, Evidence against the context-freeness of natural language, in *Philosophy, Language, and Artificial Intelligence* (Springer, Cambridge, 1985), pp. 79–89.

[37] Note also that some lexicalized models used for machine learning, such as [38], are WCFGs with multi-indexed hidden variables.

[38] M. Collins, Head-driven statistical models for natural language parsing, Computational linguistics **29**, 589 (2003).