 doi:10.20944/preprints202001.0007.v1

Criticality in Pareto Optimal Grammars?

Luís F Seoane[1] and Ricard Solé[2, 3, 4]
[1]Instituto de Física Interdisciplinar y Sistemas Complejos, IFISC (CSIC-UIB), Campus UIB, 07122, Palma de Mallorca, Spain
[2]ICREA-Complex Systems Lab, Universitat Pompeu Fabra (GRIB), Dr Aiguader 80, 08003 Barcelona, Spain
[3]Institut de Biologia Evolutiva, CSIC-UPF, Pg Maritim de la Barceloneta 37, 08003 Barcelona, Spain
[4]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA

What are relevant levels of description when investigating human language? How are these levels connected to each other? Does one description yield smoothly into the next one such that different models lie naturally along a hierarchy containing each other? Or, instead, are there sharp transitions between one description and the next, such that to gain a little bit accuracy it is necessary to change our framework radically? Do different levels describe the same linguistic aspects with increasing (or decreasing) accuracy? Historically, answers to these questions were guided by intuition and resulted in subfields of study, from phonetics to syntax and semantics. Need for research at each level is acknowledged, but seldom are these different aspects brought together (with notable exceptions). Here we propose a methodology to inspect empirical corpora systematically, and to extract from them, blindly, relevant phenomenological scales and interactions between them. Our methodology is rigorously grounded in information theory, multi-objective optimization, and statistical physics. Salient levels of linguistic description are readily interpretable in terms of energies, entropies, phase transitions, or criticality. Our results suggest a critical point in the description of human language, indicating that several complementary models are simultaneously necessary (and unavoidable) to describe it.

Keywords: syntax; Pareto-optimality; bottleneck method; phase transitions; statistical mechanics

I. INTRODUCTION

What is the "right" level of description for the faculty of human language? What would allow us to properly describe how it operates given the multiple scales involved – from letters and words to whole sentences? This nested character of language organization (figure 1) pervades the great challenge of understanding how it originated and how we could generate it artificially. The standard answer to these and similar questions is given by rules of thumb that have helped us, historically, to navigate the linguistic complexities. We have identified salient aspects (e.g. phonetics, formal grammars, etc.) to which whole fields are devoted. In adopting a level of description, we hope to encapsulate a helpful snippet of knowledge. To guide these choices we must broadly fulfill two goals: i) the system under research (human language) must be somehow simplified and ii) despite that simplification we must still capture as many relevant, predictive features about our system's unfolding as possible. Some simplifications work better than others. In general, opting for a specific level does not mean that another one is not informative.

A successful approach to explore human language is through networks. Nodes of a language web can be letters, syllables, or words; and links can represent co-occurrences, structural similarity, phonology, or syntactic or semantic relations [1–7]. Are these different levels of description nested parsimoniously into each other? Or do sharp transitions exist that establish clear phenomenological realms? Most of the network-level topological analyses suggest potential paths to understand linguistic processing and hint at deeper features of language organization. However, the connection between different levels are seldom explored, with few exceptions based on purely topological patterns [8]; or some ambitious attempts to integrate all linguistic scales from the evolutionary one to the production of phonemes [9, 10].

In this paper we present a methodology to tackle this problem in linguistics: When are different levels of description pertinent? When can we forgo some details and focus on others? For example, when do we need to attend to syntactic constraints, and when do we need to pay attention to phonology? How do the descriptions at different levels come together? This interplay can be far from trivial: note, e.g., how phonetics dictates the grammatical choice of the determiner form 'a' or 'an' in English. Similarly, phonetic choices with no grammatical consequence can evolve into rigid syntactic rules in the long term. Is the description at a higher level always grounded in all previous stages, or do descriptions exist that do not depend on details from other scales? Likely, these are not all or nothing question. So, rather, how many details in a given description do we need to carry on to the next one?

To exemplify how these questions can be approached, we look at written corpora as symbolic series. There are many ways in which a written corpus can be considered a symbolic series. For example, we can study the succession of letters in a text. Then, the available *vocabulary* consists of all letters in the alphabet (often including punctuations marks):

$$\chi^{letters} \equiv \{a, b, \ldots, z, !, ?, \ldots\} \tag{1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2020 doi:10.20944/preprints202001.0007.v1

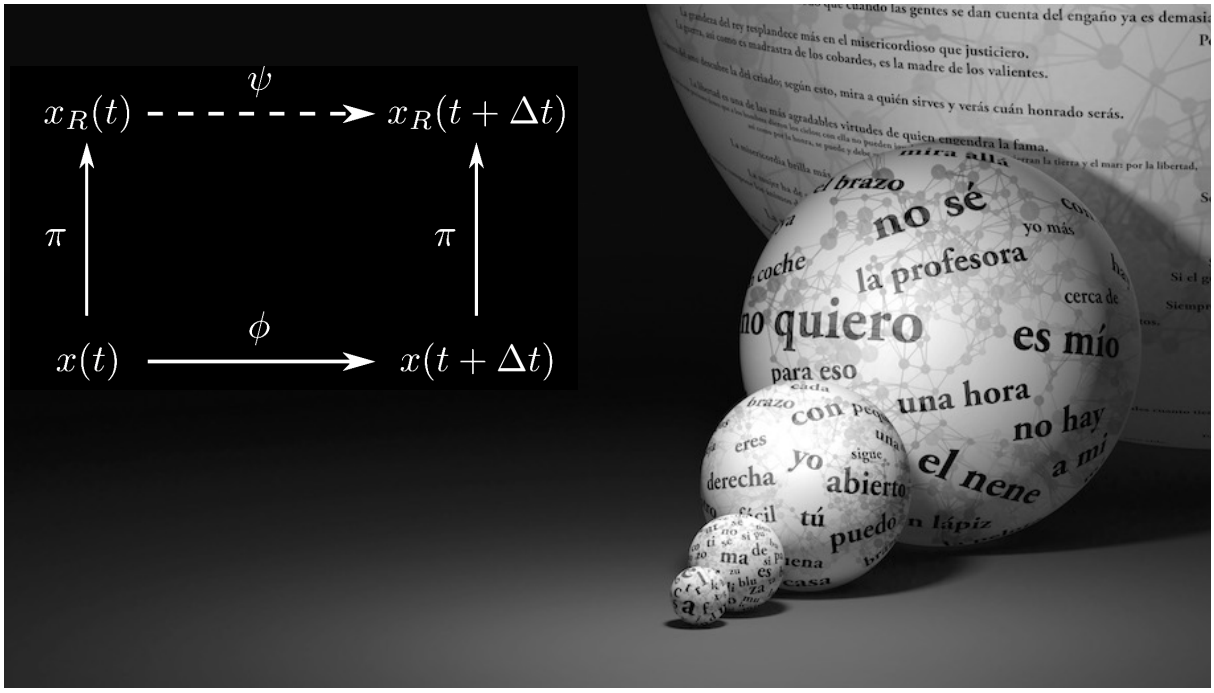

FIG. 1 **Different levels of grammar.** Language contains several layers of complexity that can be gauged using different kinds of measures and are tied to different kinds of problems. The background picture summarizes the enormous combinatorial potential connecting different levels, from the alphabet (smaller sphere) to grammatically correct sentences (larger sphere). On top of this, it is possible to describe each layer by means of a coarse-grained symbolic dynamics approach. One particularly relevant level is the one associated to the way syntax allows generating grammatically correct strings $x(t)$. As indicated in the left diagram, symbols succeed each other following some rules ϕ. A coarse-graining π groups up symbols in a series of classes such that the names of these classes $x_R(t)$ also generate some symbolic dynamics whose rules are captured by ψ. How much information can the dynamics induced by Ψ recover about the original dynamics induced by ϕ? Good choices of π and Ψ will preserve as much information as possible despite being relatively simple.

Alternatively, we can consider words as indivisible. In such case our vocabulary (χ^{words}) would consist of all entries in a dictionary. We can study even simpler symbolic dynamics, e.g., if we group together all words of each given grammatical class and consider words within a class equal to each other. From this point of view we do not gain much by keeping explicit words in our corpora. We can just substitute each one by its grammatical class, for example:

$$\text{green} \quad \text{colorless} \quad \text{ideas} \quad \text{sleep} \quad \text{furiously} \longrightarrow adj \;\; adj \;\; noun \;\; verb \;\; adv. \tag{2}$$

After this, we can study the resulting series that have as symbols elements of the coarse-grained vocabulary:

$$\chi^{grammar} \equiv \{noun, verb, adj, adv, prep, \dots\}. \tag{3}$$

Further abstractions are possible. For example, we can introduce a mapping that retains the difference between nouns and verbs, and groups all other words in an abstract third category:

$$\text{adj} \quad \text{adj} \quad \text{noun} \quad \text{verb} \quad \text{adv} \longrightarrow cat_3 \;\; cat_3 \;\; noun \;\; verb \;\; cat_3. \tag{4}$$

It is fair to ask which of these descriptions are more useful, when to stop our abstractions, whether different levels define complementary or redundant aspects of language, etc. Each of these descriptions introduces an operation that maps the most fine-grained vocabulary into less detailed ones – for example:

$$\pi : \chi^{words} \to \chi^{grammar}. \tag{5}$$

To validate the accuracy of this mapping, we need a second element. At the most fundamental level, some unknown rules ϕ exist. They are the ones connecting words to each other in real language and correspond to the generative mechanisms that we would like to unravel. At the level coarse-grained by a mapping π, we can propose a description Ψ (figure 1) that captures how the less-detailed dynamics advance. How well can we recover the original series depends on our choices of π and Ψ. Particularly good descriptions at different scales conform the answers to the questions raised above. The ϕ and Ψ mappings play roles similar to language grammars – i.e. sets of rules that tell us what words can follow each other. Some rules show up in actual corpora more often than others. Almost every sentence needs to deal with the Subject-Verb-Object (SVO) rule, but only seldom do we find all types of adjectives in a same phrase. If we would infer a grammar empirically by looking at English corpora, we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2020 doi:10.20944/preprints202001.0007.v1

could easily oversee that there is a rule for adjective order too. But since it can be so easily missed, this might not be as important as SVO to understand how English works.

Here we investigate grammars, or sets of rules, that are empirically derived from written corpora. We would like to study as many grammars as possible, and to evaluate numerically how well each of them works. In this approach, a wrong rule (e.g. one proposing that sentence order in English is VSO instead of SVO) would perform poorly and be readily discarded. It is more difficult to test descriptive grammars (e.g. a rule that dictates the adjective order), so instead we adopt abstract models that tell us the probability that classes of words follow each other. For example, in English it is likely to find an adjective or a noun after a determiner; but it is unlikely to find a verb. Our approach is inspired by the *information bottleneck* method [11–15], rate distortion theory [16, 17], and similar techniques [18–22]. In all these studies, arbitrary symbolic dynamics are divided into the observations up to a certain point, $\overleftarrow{x}$, the dynamics from that point onward, $\overrightarrow{x}$, and some coarse-grained model R (which plays the role of our π and Ψ combined) that attempts to conceptualize what has happened in $\overleftarrow{x}$ to predict what will happen in $\overrightarrow{x}$. This scheme allows us to quantify mathematically how good is a choice of $R \equiv \{\pi, \Psi\}$. For example, it is usual to search for models R that maximize the quantity:

$$I(\overleftarrow{x} : R) + \alpha I(\overleftarrow{x} : \overrightarrow{x} | R). \tag{6}$$

The first term captures the information that the model carries about the observed dynamics $\overleftarrow{x}$, the second term captures the information that the past dynamics carry about the future given the filter imposed by the model R, and the metaparameter α weights the importance of each term towards the global optimization.

We will evaluate our probabilistic grammars in a similar (yet slightly different) fashion. For our method of choice we first acknowledge that we are facing a Pareto, or Multi-Objective Optimization (MOO) problem [23–25]. In this kind of problems we attempt to minimize or maximize different traits of the model simultaneously. Such efforts are often in conflict with each other. In our case we want to make our models as simple as possible, but in that simplicity we ask that they retain as much of their predictive power as possible. We will quantify how different grammars perform in both these regards, and rank them accordingly. MOO problems rarely present global optima – i.e. we will not be able to find *the best* grammar. Instead, MOO solutions are usually embodied by Pareto-optimal tradeoffs. These are collections of designs that cannot be improved in both optimization targets simultaneously. In our case these will be grammars that cannot be made simpler without losing some accuracy in their description of a text, or that cannot be made more accurate without making them more complicated.

The solutions to MOO problems are connected with statistical mechanics [25–29]. The geometric representation of the optimal tradeoff reveals phase transitions (similar to the phenomena of water turning into ice or evaporating promptly with slight variations of temperature around 0 or 100 degrees Celsius) and critical points. In our case, Pareto optimal grammars would give us a collection of linguistic descriptions that simultaneously optimize how simply language rules can become while retaining as much of their explanatory power as possible. The different grammars along a tradeoff would become optimal descriptions at different levels, depending on how much detail we wish to track about a corpus. Positive (second order) phase transitions would indicate salient grammars that are adequate descriptions of a corpus at several scales. Negative (first order) phase transitions would indicate levels at which the optimal description of our language changes drastically and very suddenly between extreme sets of rules. Critical points would indicate the presence of somehow irreducible complexity in which different descriptions of a language become simultaneously necessary, and aspects included in one description are not provided by any other. While critical points seem a worst-case scenario towards describing language, they are a favorite of statistical physics. Systems at a critical point often display a series of desirable characteristics, such as versatility, enhanced computational abilities, and optimal handling of memory [30–38].

In section II we explain how we infer our π and Ψ (i.e. our abstract "grammatical classes" and associated grammars), and the mathematical methods used to quantify how simple and accurate they are. In section III we present some preliminary results, always keeping in mind that this paper is an illustration of the intended methodology. More thorough implementations will follow in the future. In section IV we reflect about the insights that we might win with this methods, how they could integrate more linguistic aspects, and how they could be adapted to deal with the complicated, hierarchical nature of language.

II. METHODS

A. Corpus description and preparation

We took a sample of 49 newspaper articles from the Corpus of Contemporary American English [39]. The articles were selected such that they did not contain foreign (non-English) words or symbols. We substituted by a period every punctuation mark that indicated the end of a sentence and removed any other punctuation mark except for the apostrophes indicating a contraction (e.g. "don't") or a genitive (e.g. "someone's"). Ideally, we would like to use raw texts and see Pareto optimal grammars emerging from them. These should also deal with alien symbols and words. But these are rather minor de-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2020 doi:10.20944/preprints202001.0007.v1

Conjunction	Adverb
Cardinal number	Adverb, comparative
Determiner	Adverb, superlative
Existential there	to
Preposition	Interjection
Adjective	Verb, base form
Adjective, comparative	Verb, past tense
Adjective, superlative	Verb, gerund or present participle
Modal	Verb, past participle
Noun, singular	Verb, non-3rd person singular present
Noun, plural	Verb, 3rd person singular present
Proper noun, singular	Wh-determiner
Proper noun, plural	Wh-pronoun
Predeterminer	Possessive wh-pronoun
Possessive ending	Wh-adverb
Personal pronoun	None of the above
Possessive pronoun	'.'

TABLE I Grammatical classes present in the most fine-grained level of our corpora.

tails. Effective grammars should specify first how its own words are articulated.

Our more basic level of analysis will already be a coarse-grained one. Again, ideally, we would present our methods with texts in which each word is explicitly expelled out. Our blind techniques should then infer grammatical classes (if any were useful) based on how different words correlate. For example, we expect that our blind methods would be able, at some point, to group all nouns together based on their syntactic regularities. While this is possible, it is very time- and resource-consuming for the demonstration intended here. Hence, we preprocessed our corpus using Python's Natural Language Processing Toolkit [40] to map every word into one of the $N_G = 34$ grammatical classes shown in table I. We then substituted every word in the corpus by its grammatical class. The resulting texts constitute the symbolic dynamics that we analyze.

B. Word embeddings and coarse-graining

We would like to explore the most general grammars possible. But, as advanced above, to make some headway we restrict ourselves to *grammar models* that encode a tongue's rules in a probabilistic way, telling us how likely it is that words follow each other in a text. Even in this narrower class there is an inscrutably large number of possibilities depending, e.g., on how far back we look into a sentence to determine the next word's likelihood, on whether we build intermediate phrases to keep track of the symbolic dynamics in a hierarchical way, etc. Here, we only attempt to predict the next word given the current one. We will also restrict ourselves to maximum entropy (*MaxEnt*) models, which are the models that introduce less further assumptions provided a series of observations [37, 41–49]. We explain these kind of models in the next subsection. First, we need to introduce some notation and a suitable encoding of our corpus so we can manipulate it mathematically.

We use a one-hot embedding, which substitutes each word in a text by a binary string that consists of all zeros and exactly one 1. The position of the 1 indicates the class of word that we are dealing with. Above, we illustrated several levels of coarse-graining. In a very fundamental one, each word represents a class of its own. Our vocabulary in the simple example sentence "green colorless ideas sleep furiously" consists of:

$$\chi^{words} \equiv \{ideas, sleep, green, colorless, furiously\} \quad (7)$$

which in its binary form becomes:

$$\tilde{\chi}^{words} = \{10000, 01000, 00100, 00010, 00001\}. \quad (8)$$

We also illustrated a level of coarse-graining in which nouns and verbs are retained, but all other words are grouped together in a third category (equation 4). The corresponding vocabulary:

$$\chi \equiv \{noun, verb, cat_3\} \quad (9)$$

becomes, through the one-hot embedding:

$$\tilde{\chi} = \{100, 010, 001\}. \quad (10)$$

Throughout this paper we will note by χ^λ the vocabulary (set of unique symbols) at a description level λ, and we will refer by $\tilde{\chi}^\lambda$ to its one-hot representation. We will name $c_j^\lambda \in \chi^\lambda$, with $j \in \{1, \ldots, N^\lambda\}$, to each of the N^λ unique symbols at description level λ. Each of these symbols stands for an abstract *class* of words, which might or might not correspond to actual grammatical classes in the standard literature. The binary representation of each class is correspondingly noted by $\sigma_j^\lambda \in \tilde{\chi}^\lambda$.

To explore models of different complexity we start with all the grammatical classes outlined in table I and proceed by lumping categories together. We will elaborate a probabilistic grammar for each level of coarse-graining. Later, we will compare the performance of all descriptions. In lumping grammatical classes together there are some choices more effective than others. For example, it seems wise to group comparative and superlative adverbs earlier than nouns and verbs. We expect the former to behave more similarly than the later, and hence to lose less descriptive power when treating both comparative and superlative adverbs as one class. In future versions of this work we intend to explore arbitrary lumping strategies. Here, to produce results within a less demanding computational framework, we use an informed shortcut. We build the maximum entropy model of the least coarse-grained category (which, again, in this paper consists of the grammatical classes in table I). Through

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2020 doi:10.20944/preprints202001.0007.v1

some manipulations explained below, this model allows us to extract correlations between a current word and the next one (illustrated in figure 2). These correlations allow us to build a dendogram (figure 3**a**) based on how similarly different grammatical classes behave.

This dendogram suggests an order in which to merge the different classes, which is just a good guess. There are many reasons why the hierarchy emerging from the dendogram might not be the best coarse-graining. We will explore more exhaustive possibilities in the future. In any case, this scheme defines a series of functions π^λ (which play the role of π in figure 1) that map the elements of the most fine-grained vocabulary $\chi^0 \equiv \chi^{grammar}$ (as defined by the classes in table I) into a series of each time more coarse-grained and abstract categories χ^λ, with $\lambda = 1, \ldots, N_G - 1$ indicating how many categories have been merged at that level.

C. Maximum-entropy models

To build the MaxEnt model at a given level λ of coarse-graining, we substitute every word in our corpus by its binary representation. Our text then becomes a binary string. For example, with the coarse-graining in which nouns and verbs are kept, and all other words are abstracted into cat_3, we have:

$$\text{green} \quad \text{colorless} \quad \text{ideas} \quad \text{sleep} \quad \text{furiously} \longrightarrow 001 \;\; 001 \;\; 100 \;\; 010 \;\; 001 \tag{11}$$

We indicate the i-th word in a text by $w(i)$. Its grammatical class in the description level λ is noted:

$$c^\lambda(i) \equiv \pi^\lambda(w(i)), \tag{12}$$

and its binary representation:

$$\sigma^\lambda(i) \equiv \tilde{\pi}^\lambda(w(i)). \tag{13}$$

Both mappings π^λ and $\tilde{\pi}^\lambda$ contain the same information, and both of them play the role of π in figure 1. Note that $c^\lambda(i) = c^\lambda_j$ for some j, and that while $i \in \{1, \ldots, N_w\}$ indexes words as they happen in a text (of length N_w), $j \in \{1, \ldots, N^\lambda\}$ indexes unique grammatical classes in χ^λ. Each binary representation consists of N^λ bits. When necessary, we will use a subindex k to label $\sigma^\lambda_{j,k}$ as the k-th bit of the j-th class's binary representation at a given coarse-graining level λ.

We next produce binary samples that include each word and the one next to it in a text: $\langle \sigma^\lambda(i) | \sigma^\lambda(i+1) \rangle$, where $\langle \cdot | \cdot \rangle$ indicates concatenation. Thus, the coarse-grained sentence from equation 11 yields the samples:

$$\{001001, 001100, 100010, 010001\}. \tag{14}$$

Each sample has size $2N^\lambda$ (when needed, the index k over bits will also label positions from 1 to $2N^\lambda$). Large corpora will produce huge collections of such samples. We can summarize these collections by giving the empirical frequency $F\left(\langle \sigma^\lambda_j | \sigma^\lambda_{j'} \rangle\right)$ with which each of the $\left(N^\lambda\right)^2$ possible bit strings with length $2N^\lambda$ shows up. These collections behave as samples of what is known as spin glasses in statistical mechanics. We have powerful mathematical tools to infer MaxEnt models for spin glasses – hence all these efforts.

III. RESULTS

In spin glasses, a collection of little magnets (or spins) is arranged in space. We say that a magnet is in state $\sigma = 1$ if its north pole is pointing upwards and in state $\sigma = -1$ if its pointing downwards (these are equivalent to the 1s and 0s in our word samples). Two of these little magnets interact through their magnetic fields. These fields build up a force that tends to align both spins in the same direction, whichever it is, just as two magnets in your hand try to fall along a specific direction with respect to each other. On top of this, the spins can interact with an external magnetic field – think, bringing in a much bigger magnet which orientation cannot be controlled. This external field tends to align the little spins along its fixed, preferred direction. Given the spin states σ_1 and σ_2, the energy of their interaction with the external magnetic field and with each other can be written as:

$$\begin{aligned} E(\sigma_1, \sigma_2) &= -\frac{1}{2}\left(2h_1\sigma_1 + \sigma_1 J_{12}\sigma_2 + \sigma_2 J_{21}\sigma_1 + 2h_2\sigma_2\right) = \\ &= -\frac{1}{2}\left(J_{11}\sigma_1 + \sigma_1 J_{12}\sigma_2 + \sigma_2 J_{21}\sigma_1 + J_{22}\sigma_2\right) \end{aligned} \tag{15}$$

J_{12} and J_{21} (with $J_{12} = J_{21}$) denote the strength of the interaction between the spins, and $J_{11} \equiv 2h_1$ and $J_{22} = 2h_2$ denote the interaction of each spin with the external field. The terms h_1 and h_2 are also known as biases. If the spins are aligned with each other and with the external field, the resulting energy is the lowest possible. Each misalignment increases the energy of the system. In physics, states with less energy are more probable. Statistical mechanics allows us to write precisely the likelihood of finding this system in each of its four ($\{1,1\}$, $\{1,-1\}$, $\{-1,1\}$, and $\{-1,-1\}$) possible states:

$$P(\sigma_1, \sigma_2) = \frac{e^{-\beta E(\sigma_1, \sigma_2)}}{Z}, \tag{16}$$

where $\beta = 1/T$ is the inverse of the temperature. The term

$$\begin{aligned} Z &= e^{-\beta E(1,1)} + e^{-\beta E(1,-1)} + e^{-\beta E(-1,1)} + e^{-\beta E(-1,-1)} \\ &= \sum_{\sigma_1,\, \sigma_2 = \pm 1} e^{-\beta E(\sigma_1, \sigma_2)} \end{aligned} \tag{17}$$

is known as the partition function and is a normalizing factor that guarantees that the probability distribution in equation 16 is well defined.

Back to our text corpus in its binary representation, we know the empirical frequency $F\left(\langle \sigma^\lambda_j | \sigma^\lambda_{j'} \rangle\right)$ with which each of the possible spin configurations shows up – we just need to read it from our corpus. We can treat our

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2020 doi:10.20944/preprints202001.0007.v1

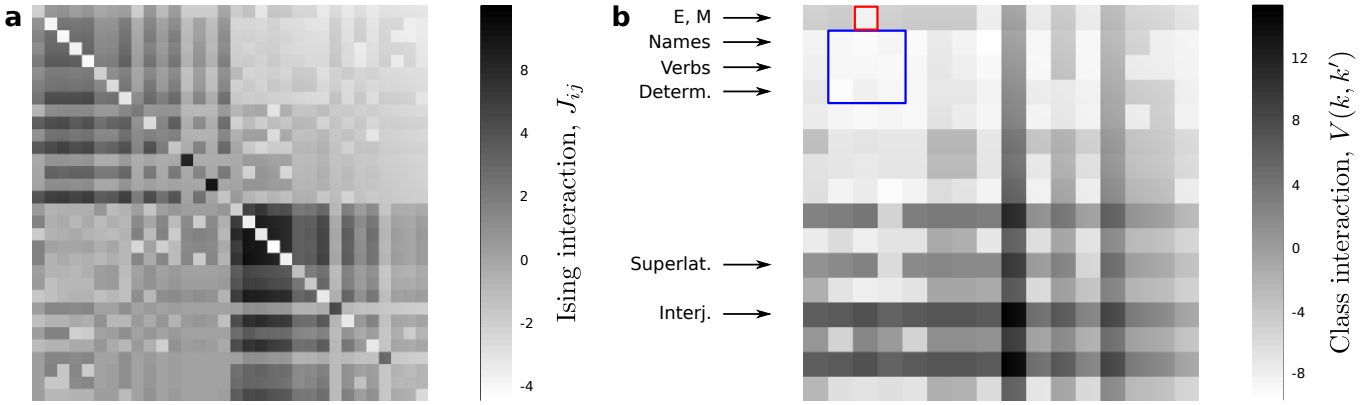

FIG. 2 **Interactions between spins and word classes**. **a** A first crude model with spins encloses more information than we need for the kind of calculations that we wish to do right now. **b** A reduced version of that model gives us an *interaction energy* between words or classes of words. These potentials capture some non-trivial features of English syntax – e.g. the existential 'there' in "there is" or modal verbs (marked E and M respectively) have a lower interaction energy if they are followed by verbs. Interjections present fairly large interaction energy with any other word, perhaps as a consequence of their independence within sentences.

collection of 0s and 1s as if they were ± 1 samples of a spin glass, and attempt to infer the β^λ and J^λ which (through a formula similar to equation 16) more faithfully reproduce the observed sample frequencies. The superindex in β^λ and J^λ indicates that they will change with the level of coarse-graining. Inferring those β^λ and J^λ amounts to finding the MaxEnt model at that coarse-grained level. As advanced above, MaxEnt models are convenient because they are the models that introduce less extra hypotheses given some observations. In other words, if we infer the MaxEnt model for some λ, any other model with the same coarse-graining would be introducing spurious hypotheses that are not suggested by the data. To infer MaxEnt models we used Minimum Probability Flow Learning (MPFL, [50]), a fast and reliable method that infers the J^λ given a sufficiently large sample.

Each grammatical class is represented by N^λ spins at the λ-th coarse-graining. This implies, as we know, that our samples consists of $2N^\lambda$ spints. MPFL returns a matrix J^λ of size $2N^\lambda \times 2N^\lambda$. This matrix embodies our abstract, probabilistic grammar (and plays the role of Ψ in figure 1). Each entry $J^\lambda_{kk'}$ of this matrix tells us the interaction energy between the k-th and k'-th bits in a sample (with $k, k' = 1, \ldots, 2N^\lambda$). However, each grammatical class is represented not by one spin, but by a configuration of spins that has only one 1. To obtain the interaction energies between grammatical classes (rather than between spins) we need to compute:

$$V^\lambda(c^\lambda_j, c^\lambda_{j'}) \;=\; \frac{1}{2}\sum_{k,k'} \sigma^\lambda_{j,k} J^\lambda_{kk'} \sigma^\lambda_{j',k'}. \qquad (18)$$

This energy in turn tells us the frequency with which we should observe each pair of words according to the model:

$$P^\lambda\left(\langle c^\lambda_j | c^\lambda_{j'}\rangle\right) \;=\; \frac{1}{Z^\lambda} e^{\beta V^\lambda(c^\lambda_j, c^\lambda_{j'})}. \qquad (19)$$

We inferred MaxEnt models for the more fine-grained level of description (χ^0 as given by the grammatical classes in table I), as well as for every other intermediate level χ^λ. Figure 2**a** shows the emerging spin-spin interactions for $l = 15$, which consists of only 19 (versus the original 34) grammatical classes. This matrix presents a clear box structure:

$$J^\lambda \;=\; \left[\begin{array}{c|c} 2h^\lambda & \overrightarrow{\partial}^\lambda \\ \hline \overleftarrow{\partial}^\lambda & 2\bar{h}^\lambda \end{array}\right]. \qquad (20)$$

The diagonal blocks ($2h^\lambda$ and $2\bar{h}^\lambda$) represent the interactions between all spins that define, separately, the first and second words in each sample. As our corpus becomes infinitely large, $h^\lambda \to \bar{h}^\lambda$. These terms do not capture the interaction between grammatical classes. In the spin-glass analogy, they are equivalent to the interaction of each word with the external *magnet* that biases the presence of some grammatical classes over others. Such biases affect the frequencies $P^\lambda(c^\lambda_j)$ with which individual classes show up, but not the frequency with which they are paired up. So the h^λ and $\bar{h}^\lambda$ are not giving us much syntactic information.

More interesting for us are the interaction terms stored in $\overrightarrow{\partial}^\lambda$ and $\overleftarrow{\partial}^\lambda$. The inference method used guarantees that $\overrightarrow{\partial}^\lambda = (\overleftarrow{\partial}^\lambda)^T$. It is from these terms that we can compute the part of $V^\lambda(c^\lambda_j, c^\lambda_{j'})$ (shown in figure 2**b**) that pertains to pairwise interaction alone (i.e. the energy of the spin system when we discount the interaction with the external field). $V^\lambda(c^\lambda_j, c^\lambda_{j'})$ encodes the *energy* of two word classes when they are put next to each other in a text. The order in which words appear after each other is relevant – hence that matrix is not symmetric. These energies reflect some of the rules of English. For example, the first row (labeled "E, M") is a class that has lumped together the existential "there" (as in "there is"

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2020 doi:10.20944/preprints202001.0007.v1

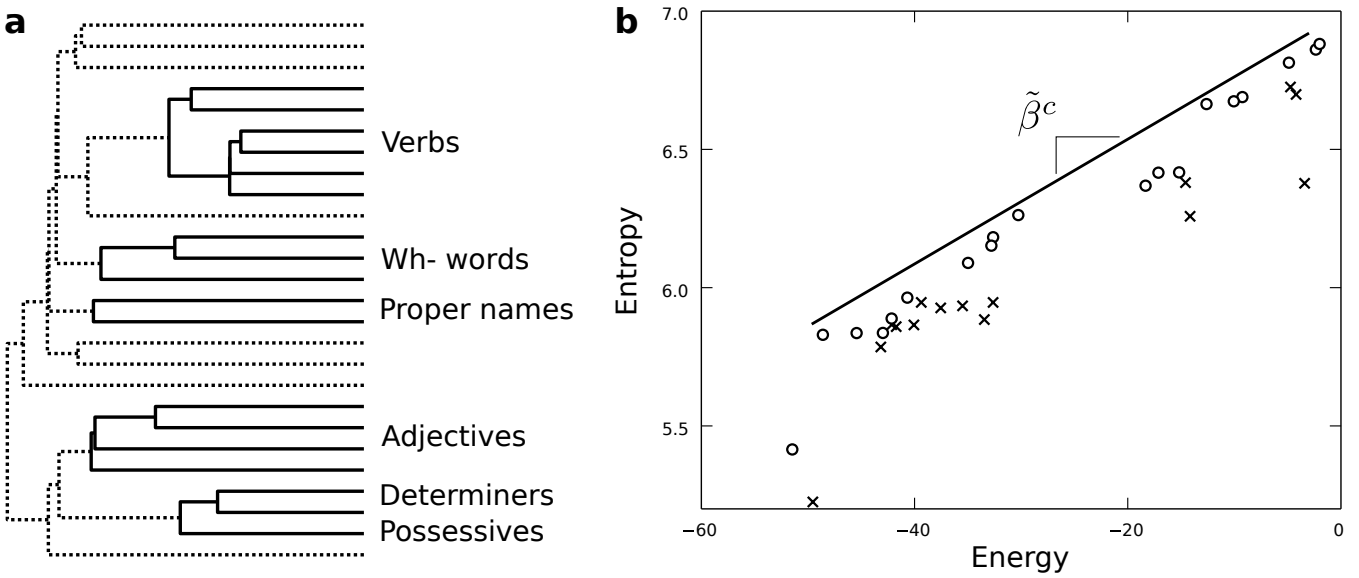

FIG. 3 **Pareto optimal maximum entropy models of human language**. Among all the models that we try out, we prefer those Pareto optimal in energy minimization and entropy maximization. **a** These reveal a hierarchy of models in which different word classes group up at different levels. The clustering reveals a series of grammatical classes that belong together owing to the statistical properties of the symbolic dynamics, such as possessives and determiners which appear near to adjectives. **b** A first approximation to the Pareto front of the problem. Future implementations will try out more grammatical classes and produce better quality Pareto fronts, establishing whether phase transitions or criticality are truly present.

and "there are") with all modal verbs. These tend to be followed by a verb in English, thus the matrix entry coding for $\langle "E, M''| "verb''\rangle$ (marked in red) is much lower than most entries for any other $\langle "E, M''|\cdot\rangle$. The blue square encompasses verbs, nouns, and determiners. While the differences there are very subtle, the energies reflect that it is more likely to see a noun after a determiner and not the other way around, and also that it is less likely to see a verb after a determiner.

It is not straightforward to compare all energies because they are affected by the raw frequency with which pairs of words show up in a text. In that sense, our corpus size might be sampling some pairings insufficiently so that their energies do not reflect proper English use. On the other hand, classes such as nouns, verbs, and determiners happen so often (and so often combined with each other) that they present very low energies as compared with other possible pairs. This makes the comparison more difficult by visual inspection.

It is possible to use $V^\lambda(c_j^\lambda, c_{j'}^\lambda)$ to generate a synthetic text $\tilde{T}^\lambda$ and evaluate its energy $E^0(\tilde{T}^\lambda)$ using the most fine grained model J^0. If the coarse-grained model $V^\lambda(c_j^\lambda, c_{j'}^\lambda)$ retains a lot of the original structure, the generated text will fit gracefully in the rules dictated by J^0 – just as magnets falling into place. Such texts would present very low energy when evaluated by J^0. If the coarse-grained model has erased a lot of the original structure, the synthetic text will present odd pairings. These would feel similar to magnets that we are forcing into a wrong disposition, hence resulting in a large energy when J^0 is used. In other words, this energy reflects how accurate each coarse-grained model is.

That accuracy is one of the targets in our MOO problem, in which we attempt to retain as much information as possible with models as simple as possible. To quantify that second target, simplicity, we turn to entropy. The simplest model possible generates words that fall in either class of χ^0 randomly and uniformly, thus presenting the largest entropy possible. More complex models, in their attempt to remain accurate, introduce constraints as to how the words in the coarse-grained model must be mapped back into the classes available in χ^0. That operation would be the reverse of π^λ. This reverse mapping, however, cannot be undone without error because the coarse-graining erases information. Entropy measures the amount of information that has been erased, and hence how simple the model has been made.

Figure 3**b** shows the energy $E^0(T^\lambda)$ and entropy $S^0(T^\lambda)$ for synthetic texts generated with the whole range of coarse-grainings explored. In terms of Pareto optimality, we expect our models to have as low an energy as possible while having the largest entropy compatible with each energy – just as thermodynamic systems do. Such models would simultaneously optimize their simplicity and accuracy. Within the sample, some of these models are *Pareto dominated* (crosses in figure 3**b**) by some others. This means that for each of those models at least some other one exists that is simpler and more accurate at the same time. These models are suboptimal regarding both optimization targets, so we do not need to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2020 doi:10.20944/preprints202001.0007.v1

bother with them.The non-dominated ones (marked by circles in figure 3**b**) capture better descriptions in both senses (accuracy and simplicity). They are such that we cannot move from one to another without improving an optimization target and worsening the other. They embody the optimal tradeoff possible (of course, limited by all the approximations made in this paper), and we cannot choose a model over the others without introducing some degree of artificial preference either for simplicity or accuracy.

In statistical mechanics the energy and entropy of a system are brought together by the *free energy*:

$$F = E - \hat{T}S = E - S/\hat{\beta}. \tag{21}$$

Here, $\hat{T}$ plays a role akin to a temperature and $\hat{\beta}$ plays the role of its inverse. We noted $\hat{\beta} \neq \beta$ to indicate that these temperature and inverse temperature are different from the ones in equation 19. Those temperatures control how often a word shows up given a model, while $\hat{\beta}$ controls how appropriate each level of description is. When $\hat{\beta}$ is low (and $\hat{T}$ is large), a minimum free energy in equation 21 is attained by maximizing the entropy rather than minimizing the energy. This is, low $\hat{\beta}$ selects for simpler descriptions. When $\hat{\beta}$ is large (and $\hat{T}$ is small), we prefer models with lower energy – i.e. higher accuracy.

By varying $\hat{\beta}$ we visit the range of models available – i.e. we visit the collection of Pareto optimal grammars (circles in figure 3**b**). In statistical mechanics, by varying the temperature of a system we visit a series of states of matter (this is, we put, e.g., a glass of water at different temperatures and observe how its volume and pressure change). At some relevant points, called phase transitions, the states of matter change radically – e.g. water freezes swiftly at 0 degrees Celsius, and evaporates right at 100 degrees Celsius. The geometry of Pareto optimal states of matter tells us when such transitions occur [25–29].

Similarly, the geometric disposition of Pareto optimal models in figure 3**b** tells us when a drastic change in our best description is needed as we vary $\hat{\beta}$. Relevant phase transitions are given by cavities and salient points along the Pareto optimal solutions. In first approach we observe several cavities. More interestingly, perhaps, is the possibility that our Pareto optimal models might fall along a straight line – one has been added as a guideline in figure 3**b**. While there are obvious deviations from it, such description might be feasible at large. Straight lines in this plot are interesting because they indicate the existence of special critical points [28, 37, 46–48]. In the next section we discuss what criticality might mean in this context.

IV. DISCUSSION

In this paper we study how different hierarchical levels in the description of human language are entangled with each other. Our work is currently at a preliminary stage, and this manuscript aims at presenting overall goals and a possible methodological way to tackle relevant questions. Some interesting results are presented as an illustration and discussed in this section to exemplify the kind of debate that this line of research can spark.

Our work puts forward a rigorous and systematic framework to tackle the questions introduced above – namely, what levels of description are relevant to understand human language and how do these different descriptions interact with each other. Historically, we have answered these questions guided by intuition. Some aspects of language are so salient that they demand a subfield of their own. While this complexity and interconnectedness is widely acknowledged, its study is still fairly compartmentalized. The portray of language as a multilayered network system is a recent exception [8], as it is the notable and lasting effort by Christiansen et al. [9, 10] to link all scales of language production, development, and evolution in a unified frame.

We have generated a collection of models that describe a written English corpus. These models trade optimally a decreasing level of accuracy by increasing simplicity. By doing so, they gradually lose track of variables involved in the description at more detailed levels. For example, as we saw above, the existential 'there' is merged with modal verbs. Indeed, these two classes were lumped together before the distinction between all other verbs was erased. While those grammatical classes are conceptually different, our blind methodology found convenient to merge them earlier in order to elaborate more efficient compact grammars.

Remaining as accurate as possible while becoming as simple as possible is a multi-objective optimization problem. The conflicting targets are captured by the energy and entropy that artificial texts generated by a coarse-grained model have when evaluated at the most accurate level of description. We could have quantified these targets in other ways (e.g. counting the number of grammatical classes to quantify complexity, and measuring similarity between synthetic and real texts for accuracy). Those alternative choices should be explored systematically in the future to understand which options are more informative. Our choices, however, make our results easy to interpret in physical terms. For example, improbable (unnatural) texts have high energies in any good model.

The grammars that optimally trade between accuracy (low energy) and simplicity (high entropy) conform the Pareto front (i.e. the solution) of the MOO problem. Its shape in the energy-entropy plane (figure 3) is linked to phase transitions [25–29]. The presence of a positive (second order) phase transition suggests that there is a salient level of description capable of capturing a large amount of linguistic structure in relatively simple terms. For example, if a unique grammatical rule would serve to connect words together disregarding of the grammatical classes in which we have split our vocabulary. We would expect that to be the case, e.g., if a single master rule

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 January 2020 doi:10.20944/preprints202001.0007.v1

such as *merge* would serve to generate all the complexity of human language *without further constraints arising*. This does not seem to be the case. However, this does not rule out the existence of the relevant *merge* operation, nor does it deny its possible fundamental role. Indeed, Chomsky proposes that *merge* is the fundamental operation of syntax, but that it leaves the creative process of language under-constrained [51–53]. As a result, actual implementations (i.e. real languages) see a plethora of further complexities arising in a phenomena akin to symmetry breaking.

The presence of a negative (first order) phase transition would acknowledge several salient levels of description needed to understand human language. These salient descriptions would furthermore present an important gap separating them. This would indicate that discrete approaches would be possible to describe language without missing any detail by ignoring the intermediate possibilities. If that were the case, we would still need to analyze the emerging models and look at similarities between them to understand whether both models capture a same core phenomenology at two relevant (yet distant) scales; or whether each model focuses on a specific, complementary aspect that the other description has no saying about. Some elements in figure 3**b** are compatible with this view.

However, the disposition of several Pareto optimal grammars along a seemingly straight line rather suggests the existence of a special kind of critical phenomenon [28, 37, 46–48]. Criticality is a worst-case scenario in terms of description. It implies that there is no trivial model, nor couple of models, nor relatively small collection that can capture the whole of linguistic phenomenology at any level. A degenerate number of descriptions is simultaneously necessary, and elements trivial in a level can become cornerstones of another. Also, potentially, constraints imposed by a linguistic domain (e.g., phonology) can penetrate all the way and alter the operating rules of other domains (e.g. syntax or semantics). We can list examples of how this happens in several tongues (such as the case of determiners 'a' and 'an' in English mentioned above). The kind of criticality suggested by our results would indicate that such intrusions are the norm rather than the exception. Note that this opportunistic view of grammar appears compatible with Christiansen's thesis that language evolved, as an interface, to make itself useful to our species, necessarily exploiting all kinds of hacks along its way [9].

It has been proved that if the optimization targets of our MOO are an energy and an entropy then the component elements of the models display Zipf's law at the critical point [37, 47]. By now, it is difficult to identify which are the component elements of our models, and in what sense can all of them be put together in a same probability distribution. A possibility is that different constraints need to be invoked with different frequency when generating language, and that these constraints are called up according to Zipf's law. It is a hallmark of linguistics that words appear to follow this distribution. This was already linked to the optimal balance between communicative tensions by Zipf [54–56]. More recent analyses suggest that this distribution might apply to phrases, rather than words, and this seems to offer a principled way to parse a text blindly into relevant component units [57]. It has also been proved mathematically that Zipf's law is an unavoidable feature of open-endedly evolving systems [58]. Language is often counted as open-ended, and its generative rules might reflect (if not enable) this capacity. The possibility of proving the existence of Zipf's law in a complex space of grammatical rules appears as an enticing goal.

Numerous simplifications were introduced to produce the preliminary results in this paper. We start our analysis with words that have already been coarse-grained into 34 grammatical classes, barring the emergence of further intermediate categories dictated, e.g., by semantic use. We know that semantic considerations can condition combinations of words, such as what verbs can be applied to what kinds of agents [59]. The choice of words as units (instead of letters or syllables) is another limiting factor. Words are symbols whose meanings do not depend on physical correlates with the objects signified [60]. In that sense, their association to their constituent letters and phonems is arbitrary. Their meaning is truly emergent and not rooted in their parts. Introducing letters, syllables, and phonetics in our analysis might reveal and allow us to capture that true emergence.

In order to do this it might be necessary to work with hierarchical models that allow correlations beyond the next and previous words considered here. This kind of hierarchy, in general, is a critical aspect of language [53] that our approach should capture in due time. We have excluded it in this work to attain preliminary results in a reasonable time. While hierarchical models are likely to be more demanding (in computational terms), they can be parsimoniously incorporated in our framework. A possibility is to use epsilon machines [61–63], which naturally lump together pieces of symbolic dynamics to find out *causal states*. These causal states act as shielding units that advance a symbolic dynamics in a uniquely determined way – just like phrases or sentences provide a sense of closure at their end, and direct the future of a text in new directions.

Acknowledgments

The authors thank the Santa Fe Institute for hosting our visit where most of this paper was done at the Cormac McCarthy Library. Special thanks to Ephraim Winslow and Thomas Wake for enlightening comments.

References

[1] Ferrer i Cancho, R., Riordan, O., and Bollobás, B. The consequences of Zipf's law for syntax and symbolic reference. *Proc. R. Soc. B* **2005**, *272*(1562), pp.561-565.

[2] Solé, R. Language: syntax for free?. *Nature* **2005**, *434*(7031), pp.289-289.

[3] Corominas-Murtra, B., Valverde, S. and Solé, R. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. *Adv. Complex Syst.* **2009**, 12(03), pp.371-392.

[4] Arbesman, S., Strogatz, S.H. and Vitevitch, M.S. The structure of phonological networks across multiple languages. *Int. J. Bifurcat. Chaos* **2010**, *20*(03), pp.679-685.

[5] Solé, R.V., Corominas-Murtra, B., Valverde, S. and Steels, L. Language networks: Their structure, function, and evolution. *Complexity* **2010**, *15*(6), pp.20-26.

[6] Solé, R.V. and Seoane, L.F. Ambiguity in language networks. *Linguist. Rev.* **2015**, *32*(1), pp.5-35.

[7] Seoane, L.F. and Solé, R. The morphospace of language networks. *Sci. Rep.* **2018**, 8.

[8] Martinčić-Ipšić, S., Margan, D. and Meštrović, A. Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. *Phys. A* **2016**, *457*, pp.117-128.

[9] Christiansen, M.H. and Chater, N. Language as shaped by the brain. *Behav. Brain Sci.* **2008**, *31*(5), pp.489-509.

[10] Christiansen, M.H. and Chater, N. *Creating language: Integrating evolution, acquisition, and processing.* MIT Press **2016**.

[11] Tishby, N., Pereira, F.C., and Bialek, W. The information bottleneck method. arXiv preprint **2000**, physics/0004057.

[12] Still, S., Bialek, W., and Bottou, L. Geometric clustering using the information bottleneck method. In Advances in neural information processing systems **2003**.

[13] Still, S. and Crutchfield, J.P. Structure or Noise? Santa Fe Institute working paper **2007**, #2007-08-020.

[14] Still, S., Crutchfield, J.P., and Ellison, C.J. Optimal causal inference. Santa Fe Institute working paper **2007**, #2007-08-024.

[15] Still, S. Information bottleneck approach to predictive inference. *Entropy* **2014**, *16*(2), pp.968-989.

[16] Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **2001**, *27*(3), 379-423.

[17] Shannon, C.E. and Weaver, W. *The Mathematical Theory of Communication.* Univ of Illinois Press: 1949.

[18] Shalizi, C.R. and Moore, C. What is a macrostate? Subjective observations and objective dynamics. arXiv preprint **2003**, cond-mat/0303625.

[19] Israeli, N. and Goldenfeld, N. Coarse-graining of cellular automata, emergence, and the predictability of complex systems. *Phys. Rev. E* **2006**, *73*(2), p.026203.

[20] Görnerup, O. and Jacobi, M.N. A method for finding aggregated representations of linear dynamical systems. *Adv. Complex Syst.* **2010**, *13*(02), pp.199-215.

[21] Pfante, O., Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. Comparison between different methods of level identification. *Adv. Complex Syst.* **2014**, *17*(02), p.1450007.

[22] Wolpert, D.H., Grochow, J.A., Libby, E., and DeDeo, S. Optimal high-level descriptions of dynamical systems. Santa Fe Institute working paper **2014**, #2015-06-017

[23] Coello, C. Twenty years of evolutionary multi-objective optimization: A historical view of the field. *IEEE Comput. Intel. M.* **2006**, *1*(1), pp.28-36.

[24] Schuster, P. Optimization of multiple criteria: Pareto efficiency and fast heuristics should be more popular than they are. *Complexity* **2012**, *18*(2), pp.5-7.

[25] Seoane, L.F. *Multiobjetive optimization in models of synthetic and natural living systems.* PhD Thesis, Universitat Pompeu Fabra **2016**.

[26] Seoane, L.F. and Solé, R. A multiobjective optimization approach to statistical mechanics. arXiv preprint arXiv:1310.6372 **2013**.

[27] Seoane, L.F. and Solé, R. Phase transitions in Pareto optimal complex networks. *Phys. Rev. E* **2015**, *92*(3), p.032807.

[28] Seoane, L.F. and Solé, R. Systems poised to criticality through Pareto selective forces. arXiv preprint arXiv:1510.08697 **2015**.

[29] Seoane, L.F. and Solé, R. Multiobjective optimization and phase transitions. In *Proceedings of ECCS* **2014** (pp. 259-270). Springer, Cham.

[30] Wolfram, S. Universality and complexity in cellular automata. *Phys. D* **1984**, *10*(1-2), pp.1-35.

[31] Langton, C.G. Computation at the edge of chaos: phase transitions and emergent computation. *Phys. D* **1990**, *42*(1-3), pp.12-37.

[32] Mitchell, M., Hraber, P., Crutchfield, J.P. Revisiting the edge of chaos: Evolving cellular automata to perform computations. arXiv preprint adap-org/9303003 **1993**.

[33] Bak, P. *How nature works: the science of self-organized criticality.* Springer Science & Business Media **1996**.

[34] Kauffman, S. *At home in the universe: The search for the laws of self-organization and complexity.* Oxford university press **1996**.

[35] Legenstein, R., Maass, W. What makes a dynamical system computationally powerful. In *New directions in statistical signal processing: From systems to brain* **2007**, pp.127-154.

[36] Solé, R. *Phase Transitions.* Princeton U. Press. Princeton **2011**.

[37] Mora, T., Bialek, W. Are biological systems poised at criticality? *J. Stat. Phys.* **2011**, *144*(2), pp.268-302.

[38] Muñoz, M.A. *Colloquium*: Criticality and dynamical scaling in living systems. *Rev. Mod. Phys.* **2018**, *90*(3), p.031001.

[39] http://corpus.byu.edu/coca/

[40] http://www.nltk.org/

[41] Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*(4), p.620.

[42] Jaynes, E.T. Information theory and statistical mechanics. II. *Phys. Rev.* **1957**, *108*(2), p.171.

[43] Mora, T., Walczak, A.M., Bialek, W. and Callan, C.G. Maximum entropy models for antibody diversity. *Proc. Nat. Acad. Sci.* **2010**, *107*(12), pp.5405-5410.

[44] Stephens, G.J., Bialek, W. Statistical mechanics of letters in words. *Phys. Rev. E* **2010**, *81*(6), p.066119.

[45] Harte, J. *Maximum entropy and ecology: a theory of abundance, distribution, and energetics.* OUP Oxford **2011**.

[46] Tkačik, G., Marre, O., Mora, T., Amodei, D., Berry II, M.J. and Bialek, W. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.* **2013**, *2013*(03), p.P03011.

[47] Stephens, G.J., Mora, T., Tkačik, G. and Bialek, W.

 doi:10.20944/preprints202001.0007.v1

Statistical thermodynamics of natural images. *Phys. Rev. Let.* **2013**, *110*(1), p.018701.

[48] Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S.E., Berry, M.J. and Bialek, W. Thermodynamics and signatures of criticality in a network of neurons. *Proc. Nat. Acad. Sci.* **2015**, *112*(37), pp.11508-11513.

[49] Lee, E.D., Broedersz, C.P. and Bialek, W. Statistical mechanics of the US Supreme Court. *J. Stat. Phys.* **2015**, *160*(2), pp.275-301.

[50] Sohl-Dickstein, J., Battaglino, P.B., and DeWeese, M.R. New method for parameter estimation in probabilistic models: minimum probability flow. *Phys. Rev. Let.* **2011**, *107*(22), p.220601.

[51] Chomsky, N. An interview on minimalism. *N. Chomsky, On Nature and Language* **2002**, pp.92-161.

[52] Hauser, M.D., Chomsky, N., and Fitch, W.T. The faculty of language: what is it, who has it, and how did it evolve? *Science* **2002**, *298*(5598), pp.1569-1579.

[53] Berwick, R.C. and Chomsky, N. *Why only us: Language and evolution*. MIT press **2016**.

[54] Zipf, G.K., 1949. *Human behavior and the principle of least effort*.

[55] Ferrer i Cancho, R. and Solé, R.V. Least effort and the origins of scaling in human language. *Proc. Nat. Acad. Sci.* **2003**, *100*(3), pp.788-791.

[56] Corominas-Murtra, B., Fortuny, J. and Solé, R.V. Emergence of Zipf's law in the evolution of communication. *Phys. Rev. E* **2011**, *83*(3), p.036115.

[57] Williams, J.R., Lessard, P.R., Desu, S., Clark, E.M., Bagrow, J.P., Danforth, C.M. and Dodds, P.S. Zipf's law holds for phrases, not words. *Sci. Rep.* **2015**, *5*, p.12209.

[58] Corominas-Murtra, B., Seoane, L.F. and Solé, R. Zipf's law, unbounded complexity and open-ended evolution. *J. R. Soc. Interface* **2018**, *15*(149), p.20180395.

[59] Bickerton, D. *Language and species*. University of Chicago Press: 1992.

[60] Deacon, T.W. *The symbolic species: The co-evolution of language and the brain*. WW Norton & Company **1998**.

[61] Crutchfield, J.P. and Young, K. Inferring statistical complexity. *Phys. Rev. Let.* **1989**, *63*(2), p.105.

[62] Crutchfield, J.P. The calculi of emergence: computation, dynamics and induction. *Physica D* **1994**, *75*(1), pp.11-54.

[63] Crutchfield, J.P. and Shalizi, C.R. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E* **1999**, *59*(1), p.275.