Title: What is the landscape of natural language? Insights from a random language model

Speakers: Eric De Giuli

Series: Colloquium

Date: November 06, 2019 - 2:00 PM

URL: http://pirsa.org/19110050

Abstract: Many complex systems have a generative, or linguistic, aspect: life is written in the language of DNA; protein structure is written in a language of amino acids, and human endeavour is often written in text. Are there universal aspects of the relationship between sequence and structure? I am trying to answer this question using models of random languages. Recently I proposed a model of random context-free languages [1] and showed using simulations that the model has a transition from an unintelligent phase to an ordered phase. In the former, produced sequences look like noise, while in the latter they have a nontrivial Shannon entropy; thus the transition corresponds to the emergence of information-carrying in the language.

In this talk I will explain the basics of natural language syntax, without assuming any prior knowledge of linguistics. I will present the results from the model above, and explain how the model is related to complex matrix models with disorder [2].

[1] https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.122.128301

[2] https://arxiv.org/abs/1902.07516

What is the landscape of natural language?

Insights from a random language model

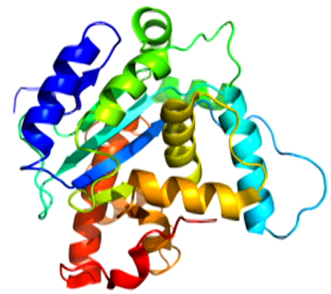
Eric De Giuli

Department of Physics
Ryerson University

motivation — complex generative systems

sequences encode structure

amino acid sequence



folded protein

Are there universal features of the sequence → structure map?

natural language as a model system

natural language is a complex generative system

& has been studied for 100+ years

Can we use it as a model system?

→ quantitative

→ foundation for richer problems

rigidity of language

1. Is John the man who is tall?

2. *Is John is the man who tall?

3. Colorless green ideas sleep furiously.

4. *Furiously sleep ideas green colorless.

syntax = logical structure
semantics = `meaning' = connection to `truth'

Chomsky 1950s

formal grammars

(Pāṇini 400BC, Chomsky, Backus 1950s)

grammar[1] = set of string rewriting rules

a,b,c,…. hidden[2] symbols

A,B,C,…. observable[3] symbols

begin with start symbol, s

repeatedly apply rules until string of observables

e.g. s → ss
s → AsB
s → AB

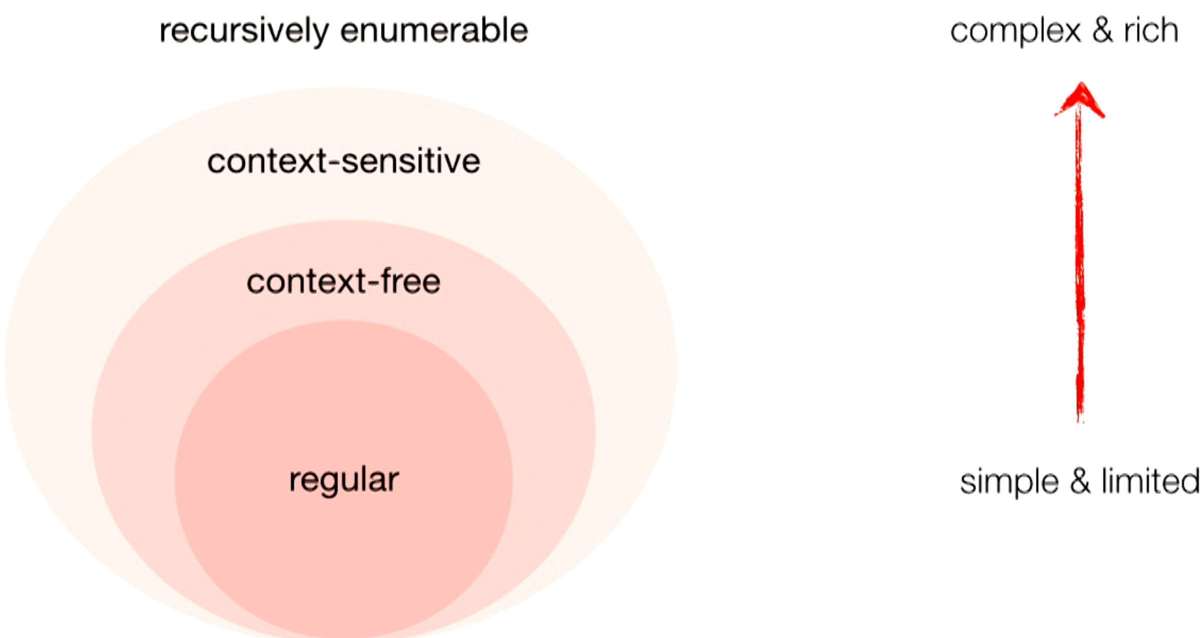
s → ss → AsBs → AABBs → AABBAB

equivalent to (( )) ( )

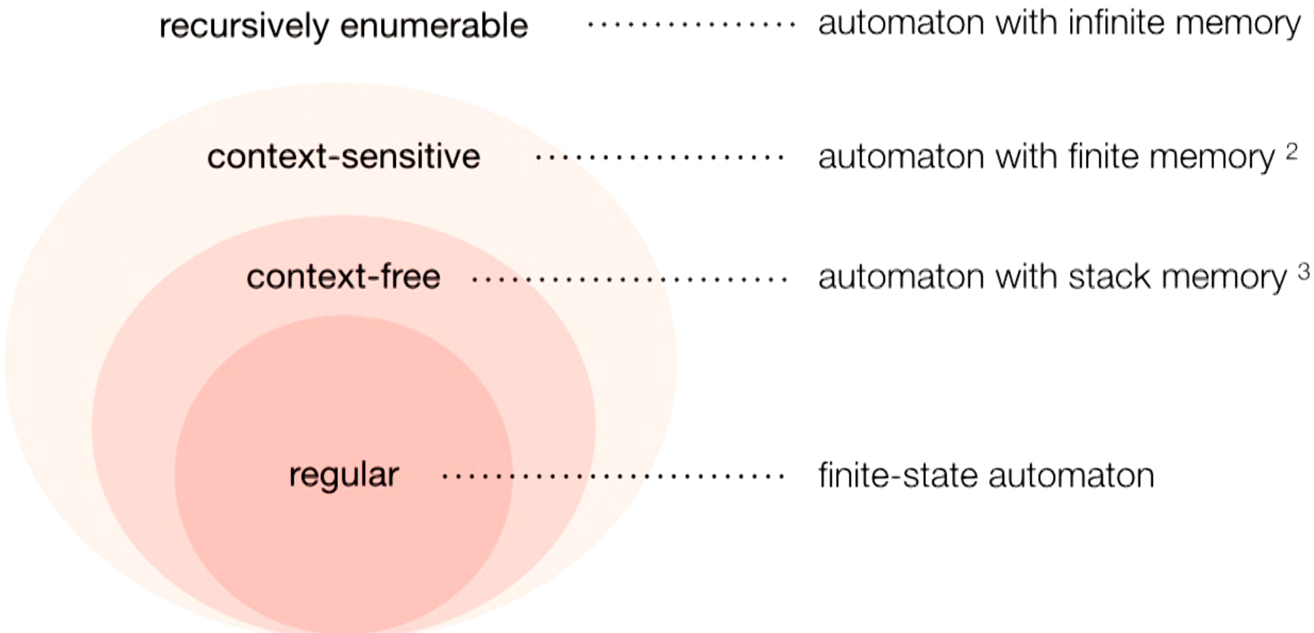
language = set of observable strings

[1] grammar = 'generative grammar' [2] 'nonterminal' [3] 'terminal'

Chomsky hierarchy (1950's)



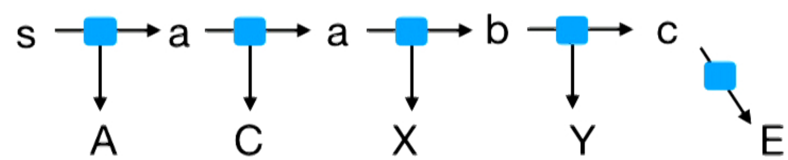
Chomsky hierarchy (1950's)



[1] Turing machine
[2] linear-bounded non-deterministic Turing machine
[3] non-deterministic pushdown automaton

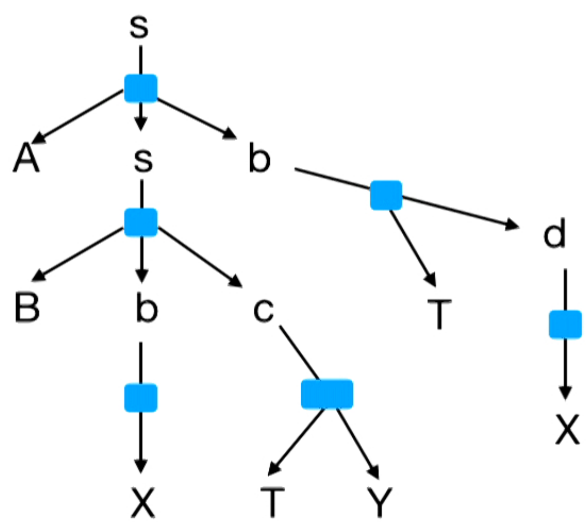
structure of derivations

regular grammar:



- always linear
- used in computer science (e.g. search patterns)
- structure of hidden Markov models (used in protein sequence analysis)

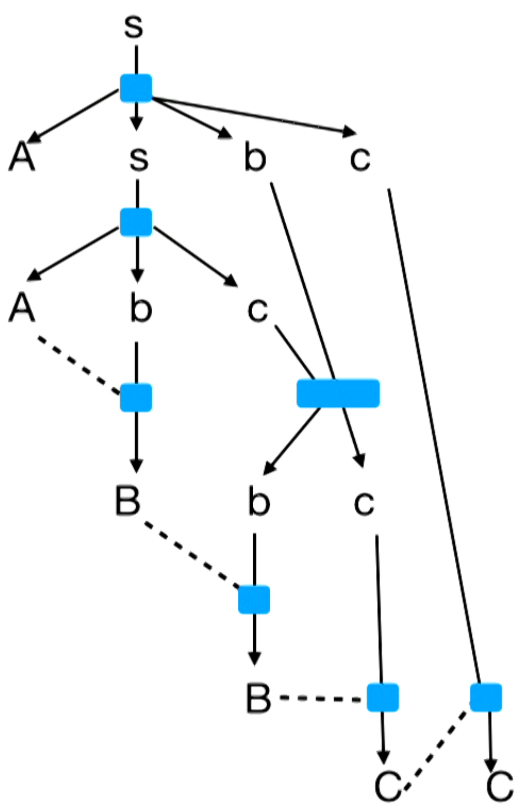
context-free grammar:



- always a tree
- used in linguistics for phrase structure (Chomsky 1956)
- central to computer science since Backus-Naur works ~1960

structure of derivations

context-sensitive grammar:



s ⇒ Asbc ⇒ AAbcbc ⇒ AABcbc

⇒ AABbcc ⇒ AABBcc ⇒ AABBCc

⇒ AABBCC

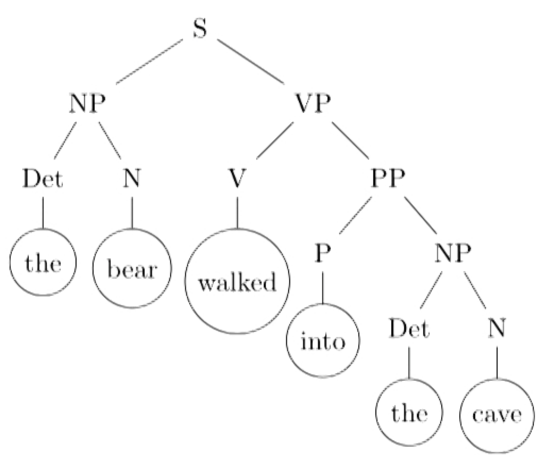
grammar:

s→ Asbc
s → Abc
cb → bc
Ab → AB
Bb → BB
Bc → BC
Cc →CC

what about natural languages?

- ~7000 existing languages
- only 2 have confirmed non-context-free features (Swiss-German, Bambara)

i.e. context-free languages define an *ensemble* for natural language syntax



content of the tree?

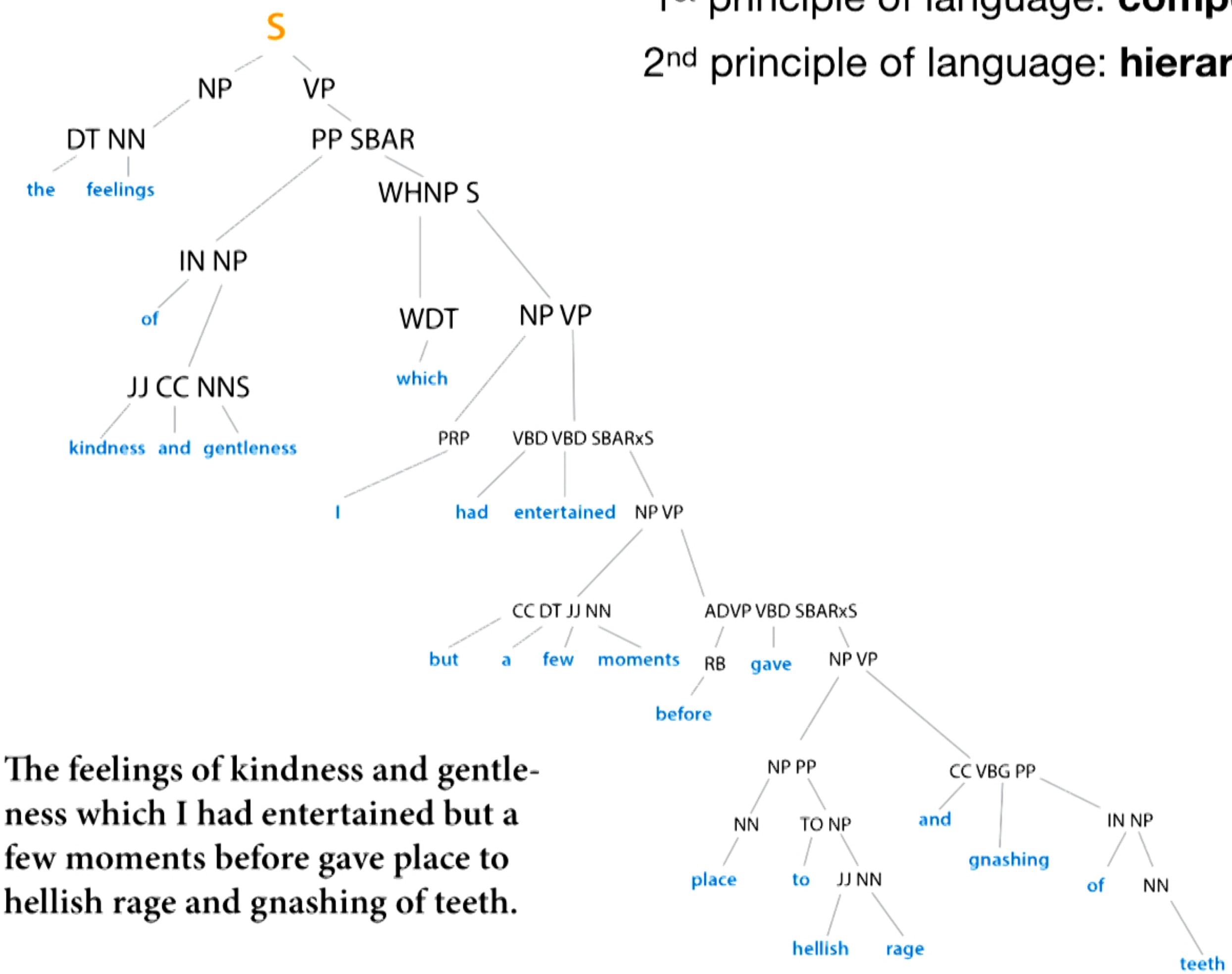
'the cave' behaves like 'cave'

'into the cave' behaves like 'into—noun'

Pullum & Gazdar 1982, Shieber 1985, Culy 1985

1st principle of language: **composition**

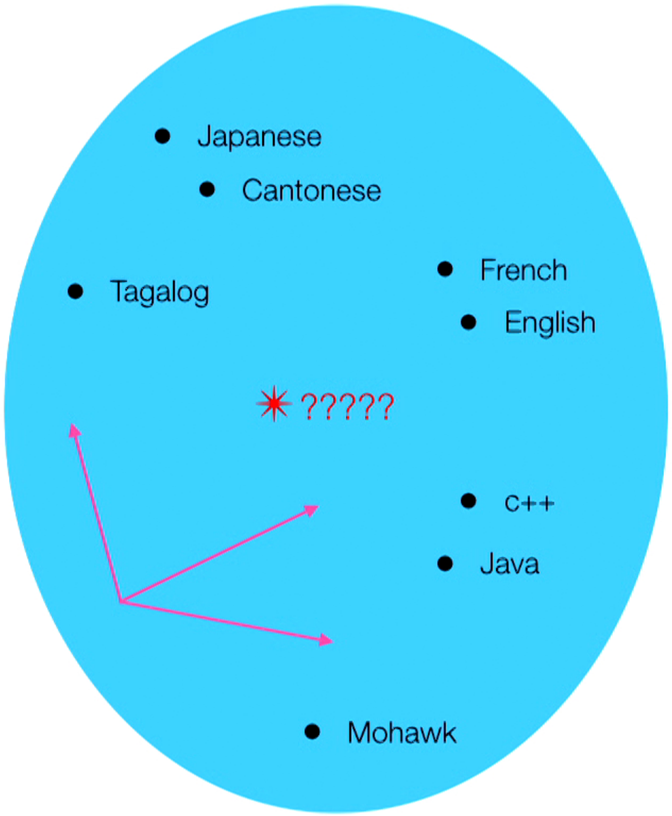
2nd principle of language: **hierarchy**



The feelings of kindness and gentleness which I had entertained but a few moments before gave place to hellish rage and gnashing of teeth.

W. Gilpin, online 2017

language ensemble



Consider ensemble of CFGs

Mathematical theorems
$\Longleftrightarrow$ borders of CFG space

How do typical CFGs behave?

$\Rightarrow$ statistical mechanics of language !

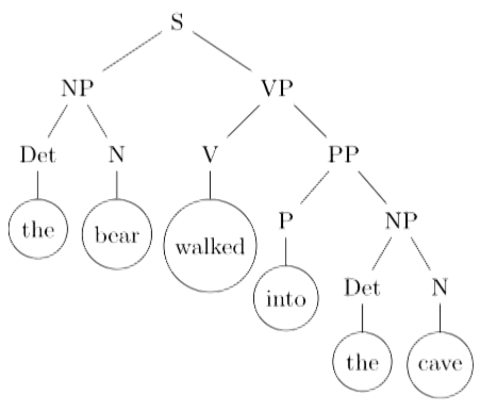
random language model – strategy

1. Quantify grammar with weights ⇒ \`energy' for trees

   low energy ⟺ grammatical

2. Define ensemble of grammars

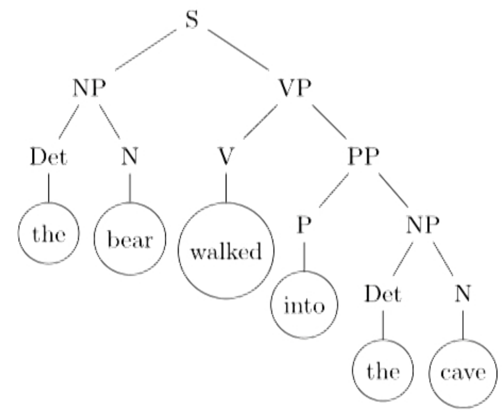
   ⇒ \`temperature' of a grammar



random language model

1. can assume binary tree[1]

all rules either a ➞ bc or a ➞ B



2. so far, rules have been yes/no. let rules ➞ conditional probabilities

then a grammar is defined by

$$M_{abc} = \mathbb{P}(a \to bc \mid a \to \text{hidden}),$$
$$O_{aB} = \mathbb{P}(a \to B \mid a \to \text{observable}),$$

[1] binary tree = 'Chomsky normal form'

random language model

for simplicity, fix tree topology T

σ = hidden symbols,

o = observables

$$M_{abc} = \mathbb{P}(a \to bc \mid a \to \text{hidden}),$$
$$O_{aB} = \mathbb{P}(a \to B \mid a \to \text{observable}),$$

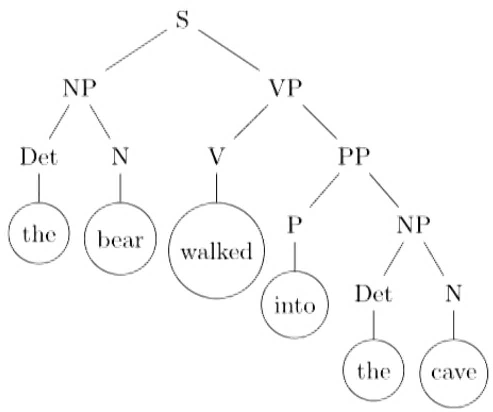
$$E = -\sum_{a,b,c} \pi_{abc}(\sigma) \log M_{abc} - \sum_{a,B} \rho_{aB}(\sigma, o) \log O_{aB}$$

$$\mathbb{P}(\{\sigma_i, o_t\} | M, O, \mathcal{T}) = \frac{1}{Z} e^{-E}$$

note: M,O are probabilities for a fixed grammar, then we have an ensemble of grammars

random language model

1. can assume binary tree[1]



all rules either $a \to bc$ or $a \to B$

2. so far, rules have been yes/no. let rules → conditional probabilities

then a grammar is defined by

$$M_{abc} = \mathbb{P}(a \to bc \mid a \to \text{hidden}),$$
$$O_{aB} = \mathbb{P}(a \to B \mid a \to \text{observable}),$$

[1] binary tree = 'Chomsky normal form'

random language model

what is the measure on grammars?

M,O act multiplicatively ⇒ lognormal

$$s_d = \frac{1}{N^3}\sum_{a,b,c}\log^2\left[\frac{M_{abc}}{\overline{M}}\right], \quad s_s = \frac{1}{NT}\sum_{a,B}\log^2\left[\frac{O_{aB}}{\overline{O}}\right]$$

deep sparsity — surface sparsity

small sparsity
⇒ uniform sampling of \`rules'
⇒ unintelligent

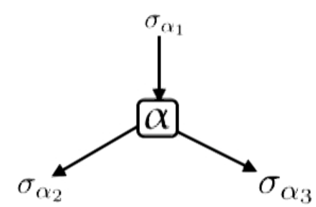
$$\mathbb{P}_G(M,O) \equiv Z_G^{-1}\; J\; e^{-\epsilon_d s_d} e^{-\epsilon_s s_s}$$

ϵ_d deep temperature

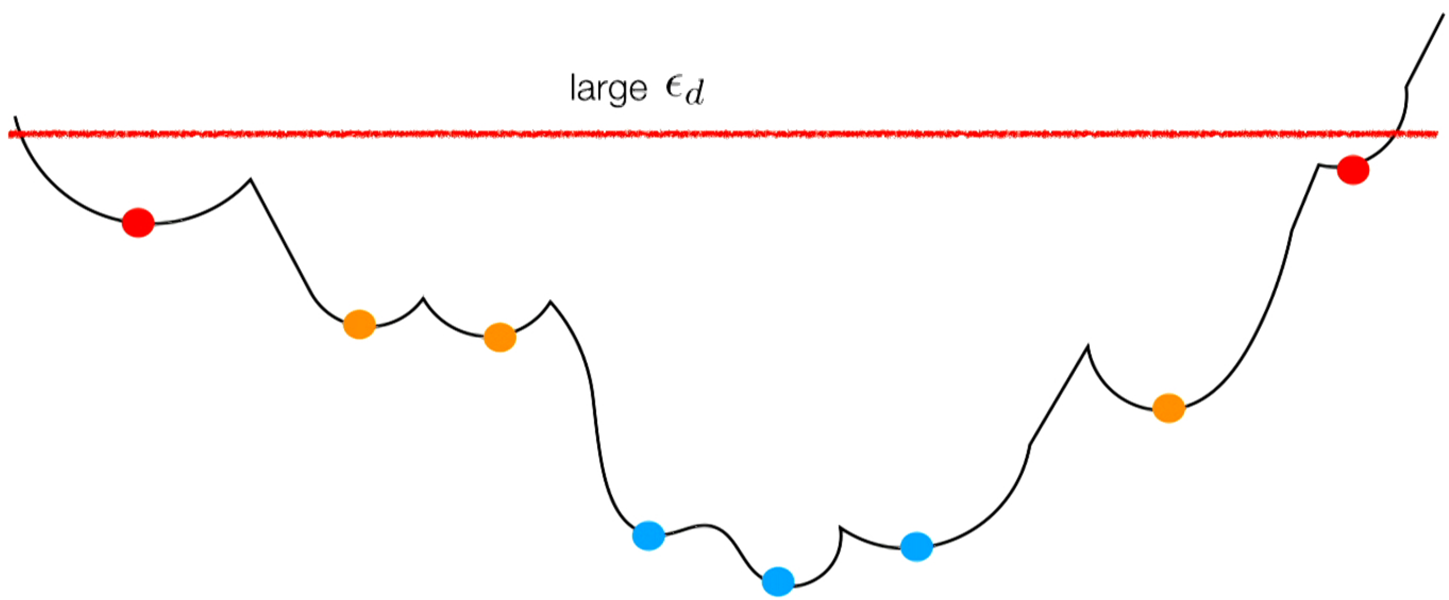
ϵ_s surface temperature

$$\overline{s_d} \sim \frac{N^3}{\epsilon_d} \qquad \overline{s_s} \sim \frac{NT}{\epsilon_s}$$

random language model



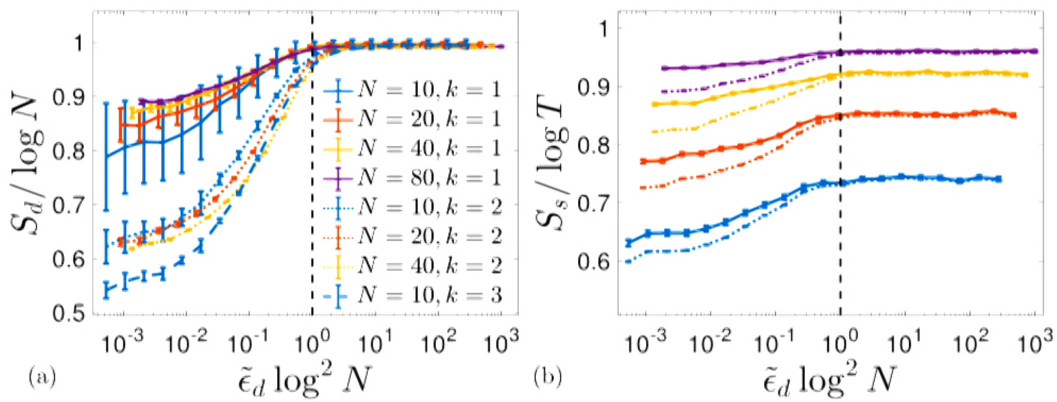
intuition for `temperatures'



random language model – Numerical results

Fix T=27, ε_s = 0.01 N T

Vary N, ε_d. Sample ~ 25 000 languages



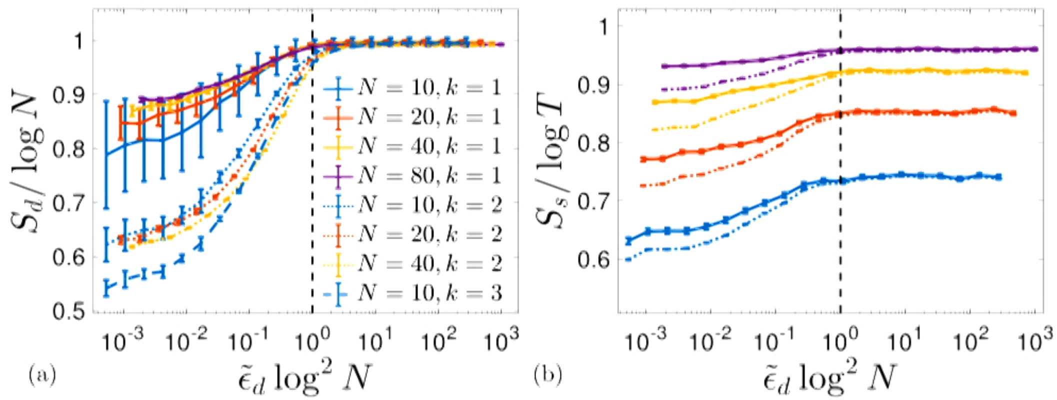
Shannon entropies

$$\tilde{\epsilon}_d = \epsilon_d / N^3$$

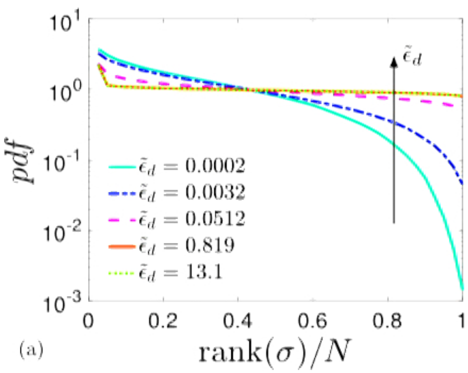
$$S_d(\mathcal{G}; k) = \frac{1}{k} \left\langle \log 1/\mathbb{P}(\sigma_1, \sigma_2, \ldots, \sigma_k | \mathcal{G}) \right\rangle \qquad S_s(\mathcal{G}; k) = \frac{1}{k} \left\langle \log 1/\mathbb{P}(o_1, o_2, \ldots, o_k | \mathcal{G}) \right\rangle$$

emergence of deep structure at $\epsilon_* \sim N^3 / \log^2 N$

random language model – Numerical results



$$\tilde{\epsilon}_d = \epsilon_d/N^3$$

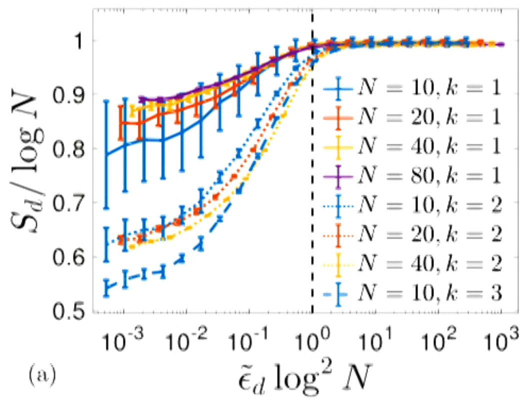


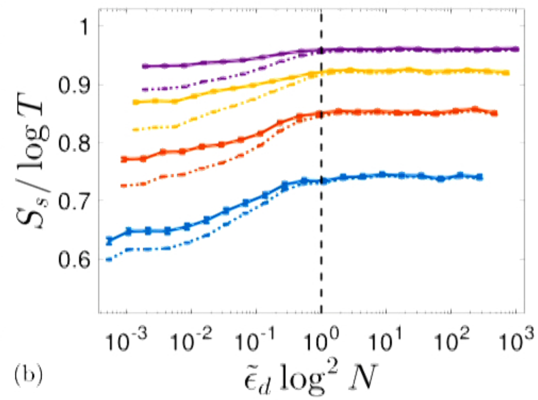
Zipf plot

Permutation symmetry spontaneously broken at $\boldsymbol{\epsilon}_*$

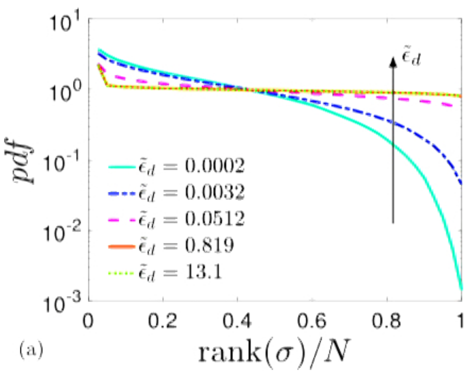
Can understand $\boldsymbol{\epsilon}_*$ from balancing `energy' and entropy

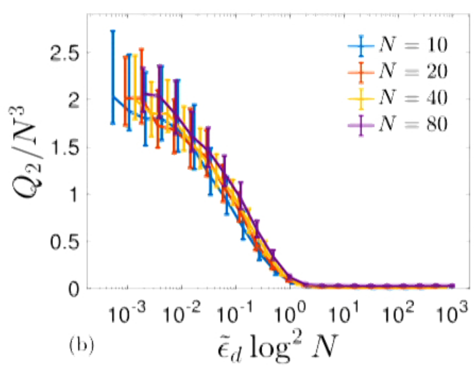
random language model – Numerical results





$$\tilde{\epsilon}_d = \epsilon_d / N^3$$





$$Q_2 \equiv \overline{\sum_{a,b,c} Q_{abc}^2},$$

Zipf plot

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

how does a child learn syntax?

"principles & parameters" theory [1]

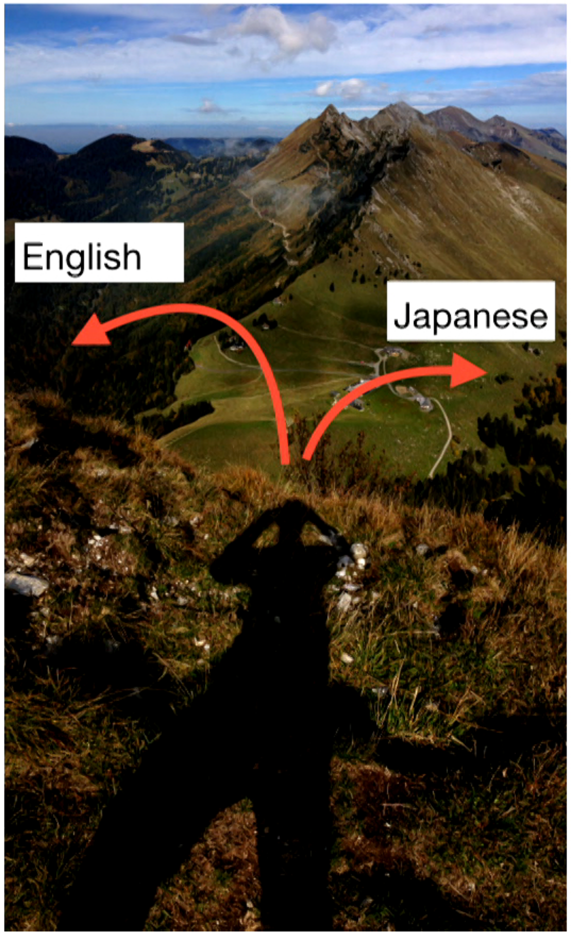
Child endowed with principles of grammar

Syntax controlled by parameters, e.g.
verbs come before objects, or vice versa

but apparently many parameters are needed! [2]

[1] Chomsky 1993 [2] Ramchand & Svenonius 2014

how does a child learn syntax?



"principles & parameters" theory [1]

Child endowed with principles of grammar

Syntax controlled by parameters, e.g.
verbs come before objects, or vice versa

but apparently many parameters are needed! [2]

RLM: learning = `energy' descent in grammar space

parameters = symmetry breaking transitions

key point: transitions are emergent properties of model

theoretical phase diagram
⇒ syntax of human languages?

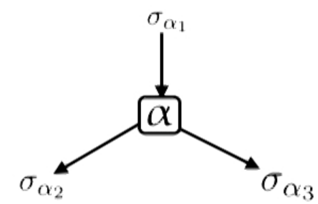
[1] Chomsky 1993 [2] Ramchand & Svenonius 2014

theory

-energy $\log \mathbb{P}(\{\sigma_i, o_t\}|M, O, \mathcal{T}) = \log P_{\sigma_0} + \sum_{\alpha \in \Omega} \log M_{\sigma_{\alpha_1}\sigma_{\alpha_2}\sigma_{\alpha_3}} + \sum_{\alpha \in \partial\Omega} \log O_{\sigma_{\alpha_1} o_{\alpha_2}}$

looks a bit like a spin model … except

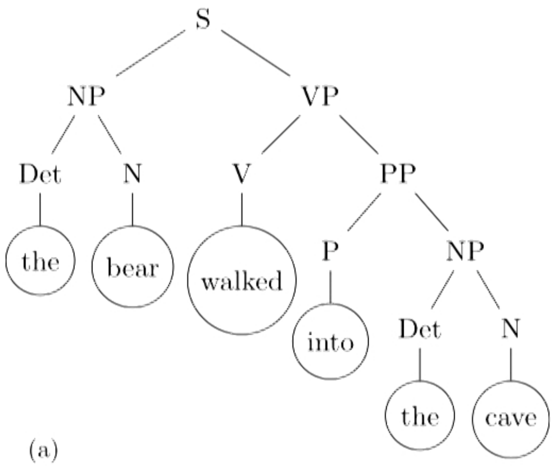
$J_{ijk} \quad vs. \quad M_{\sigma_i \sigma_j \sigma_k}$



Is there a more natural representation?

Idea of theory:

Write down model whose Feynman diagrams
generate trees with correct weights



(a)

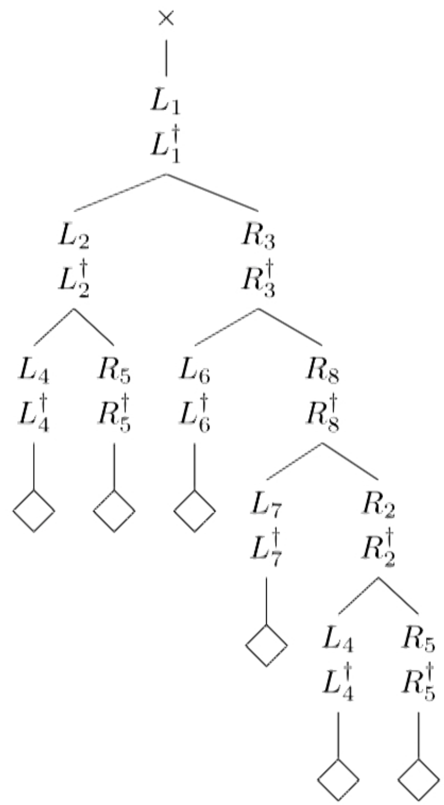


Figure 3. Feynman diagram corresponding to derivation tree in Figure 1a. Alphabet of hidden symbols is $\chi_d = (S, NP, VP, Det, N, V, P, PP)$ and alphabet of surface symbols is $\chi_s = (the, bear, walked, into, cave)$. Vertices are represented by $\wedge$ with heads at the tip. The diagram has a weight $2h\xi^6\eta^5 g^{11} M_{123} M^2_{245} M_{368} M_{872} O^2_{41} O_{52} O_{63} O_{74} O_{55}$.

theory

Feynman diagrams of F generate graphs with correct weights

$$\mathbb{F}(\mathcal{G}) = \int DL \int DR\; e^{-\frac{1}{g}\sum_a \left[L_a L_a^\dagger + R_a R_a^\dagger\right]} e^I$$

$$I = \zeta h(L_1 + R_1) + \xi \sum_a O_a(L_a^\dagger + R_a^\dagger) + \eta \sum_{a,b,c} M_{abc}(L_a^\dagger + R_a^\dagger) L_b R_c.$$

Extract m trees with l leaves

$$\mathbb{Z}(\mathcal{G}; m, \ell) = m! \oint' \frac{d\zeta}{\zeta^{1+m}} \oint' \frac{d\xi}{\xi^{1+\ell}} \oint' \frac{d\eta}{\eta^{1+\ell-m}}\, \mathbb{F}(\mathcal{G}),$$

Disorder (grammar) average

$$\overline{\log \mathbb{Z}(\mathcal{G})} = \overline{\left.\frac{\partial \mathbb{Z}(\mathcal{G})^n}{\partial n}\right|_{n=0}} = \left.\frac{\partial}{\partial n}\right|_{n=0} \overline{\mathbb{Z}(\mathcal{G})^n},$$

theory — replica symmetric ansatz

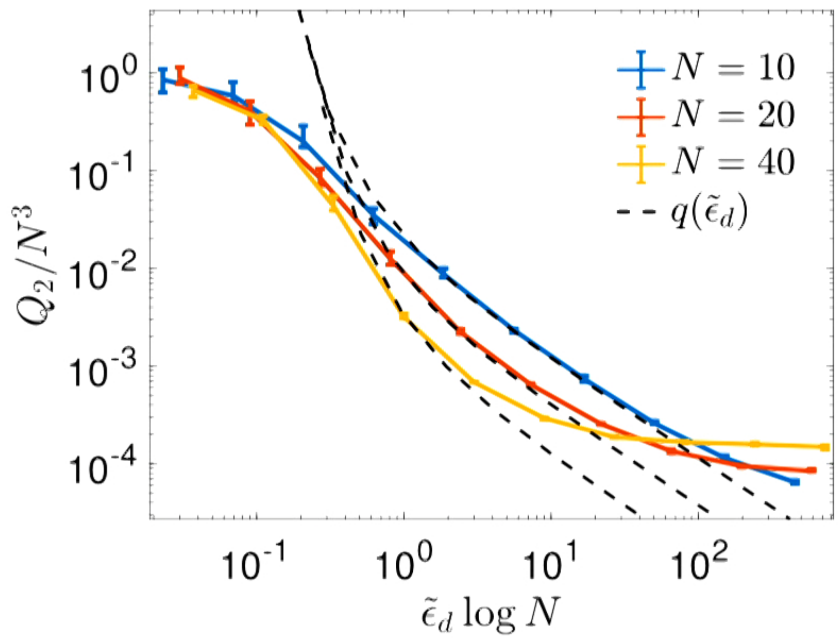


Figure 2. Order parameter Q_2 on logarithmic axes. Solid lines show numerical data from random grammars with N as indicated and $\ell \approx 10^5$. The plateau at large $\tilde{\epsilon}_d$ is a finite-ℓ effect; empirically it scales as $Q_2^\infty \sim N^4/\ell$. The function $q(\tilde{\epsilon}_d) = (e^{1/(2\tilde{\epsilon}_d)} - 1)(N^2 - 1)/N^4$ is the theoretical prediction, Eq.45.

perspectives

phase diagram:

What is the phase diagram of CF languages (with fields)?

Where are human languages?

Is this robust in merge grammars? (c.f. Piatelli-Palmarini & Vitiello)

Is there a language that parses foldable proteins?

perspectives

semantics: syntax isn't everything..

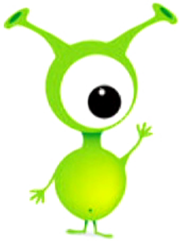
e.g. who is 'he' in this dialogue: [1]

Alice: I'm leaving you.
Bob: Who is he?!

Is there a physical approach to semantics?
c.f. lambda calculus, proof nets, …

[1] from S Pinker, The Language Instinct

conclusions



- complex systems can be generative
- natural language as a model system (compositional and hierarchical)
- context-free grammars define a simple model for these properties
- ensemble of grammars = random language model
- RLM has a glass transition
- the statistical mechanical problem is not trivial, but not intractable

Mathematical linguistics has been around for 60 years.

It's time for physical linguistics!

(numerics) Phys. Rev. Lett 2019
(theory) J.Phys A 2019

Thanks to:

Remi Monasson, Jorge Kurchan, Francesco Zamponi, Guilhem Semerjian, Pierfrancesco Urbani, Giorgio Parisi