The Formal Complexity of Natural Language Grammars Revisited

Perspectives from Neural Networks

Bob Frank
Yale University

ICGI 2020/2021

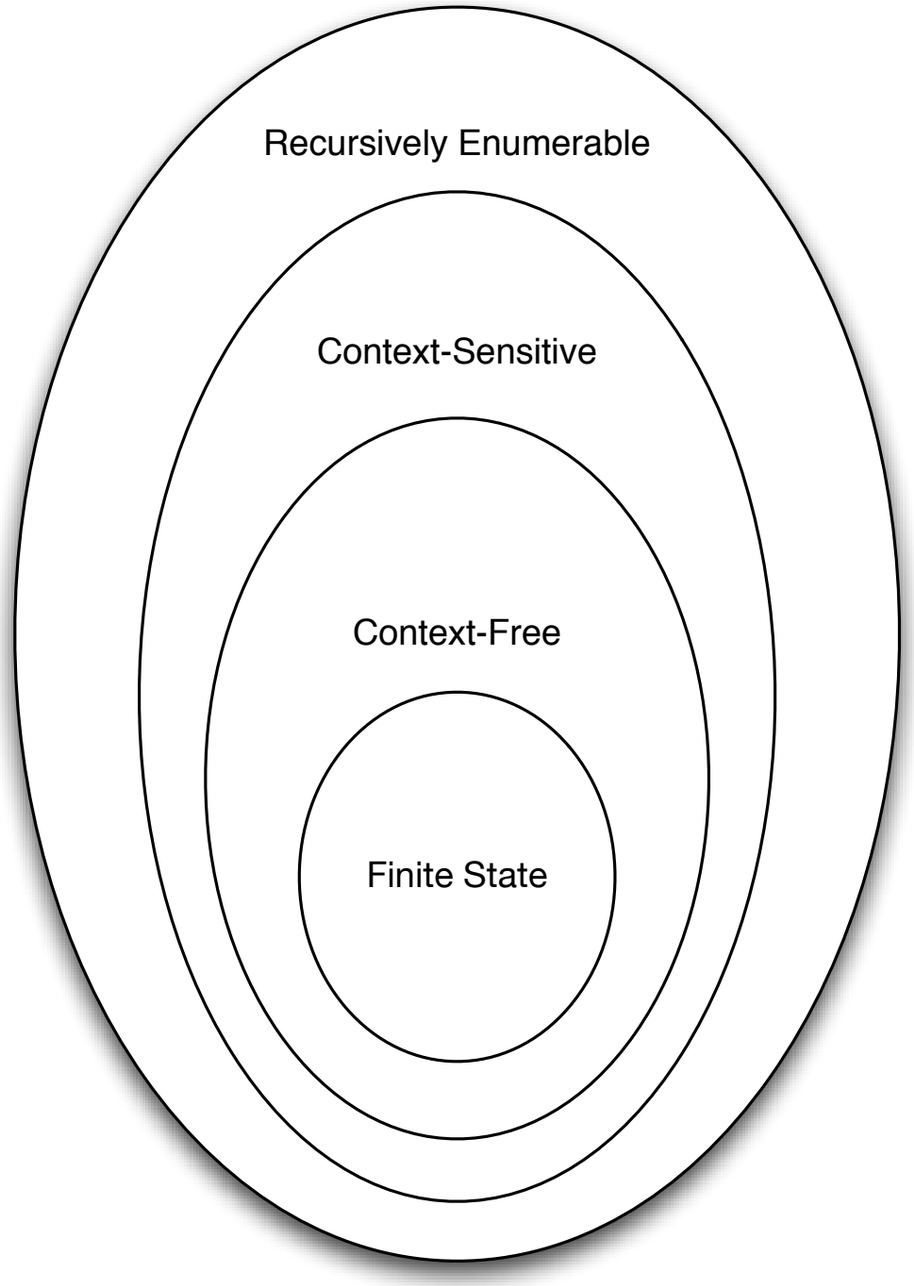
The Goals of Linguistic Theory

Chomsky (1965)

- **Descriptive Adequacy**: Is the theory sufficiently rich to provide a description of a natural language? (Lower bound on complexity)

Comparing formal grammars

- **Chomsky Hierarchy**:
  - Unrestricted Grammar/ Turing Machine
  - Context-Sensitive Grammar/Linear Bounded Automaton
  - Context-Free Gramma/ Push-Down Automaton
  - Right Linear Grammar/ Finite Automaton

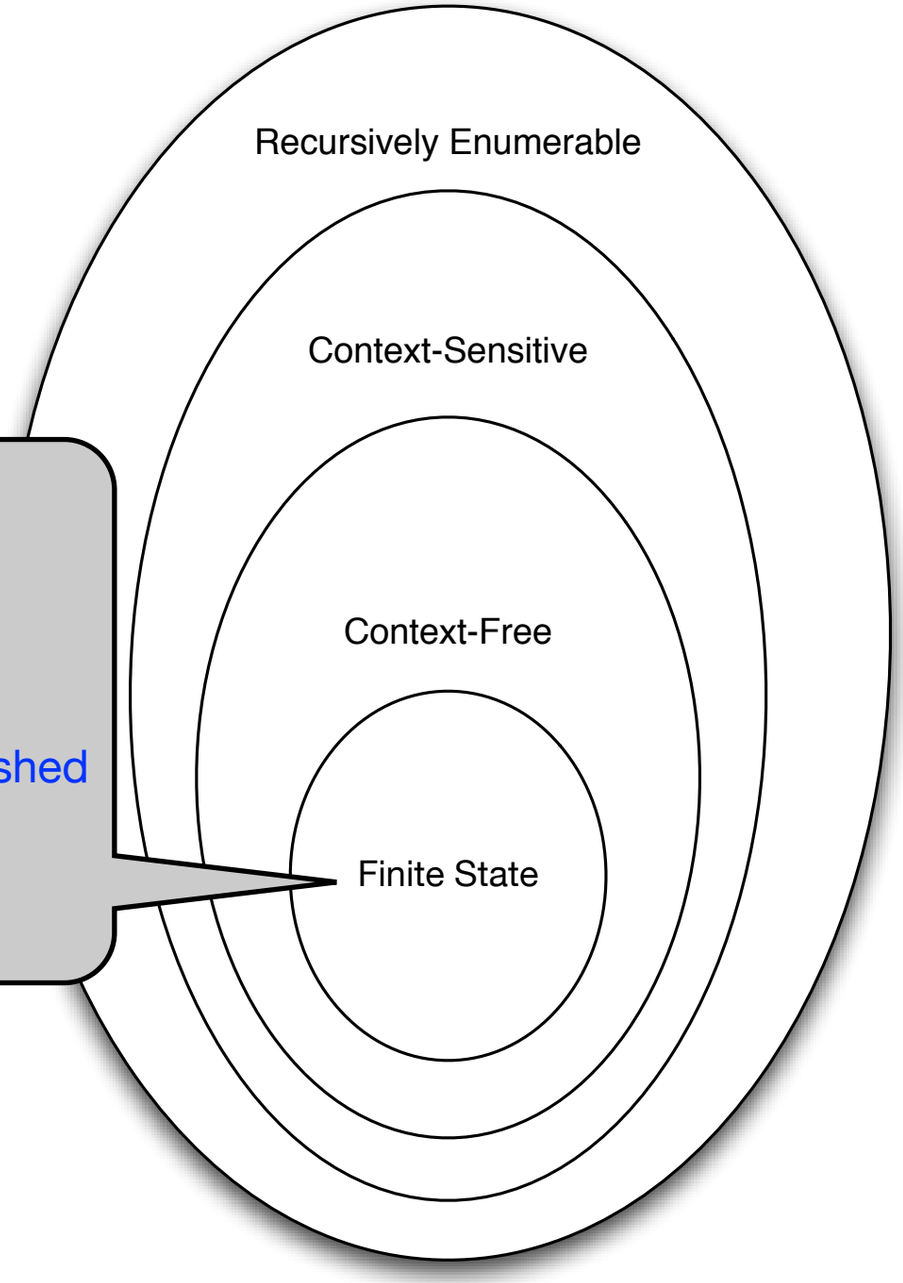


Comparing Formal Grammars

Descriptive Accuracy

Finite State is not sufficiently powerful:

The book an italian I never heard of wrote got published

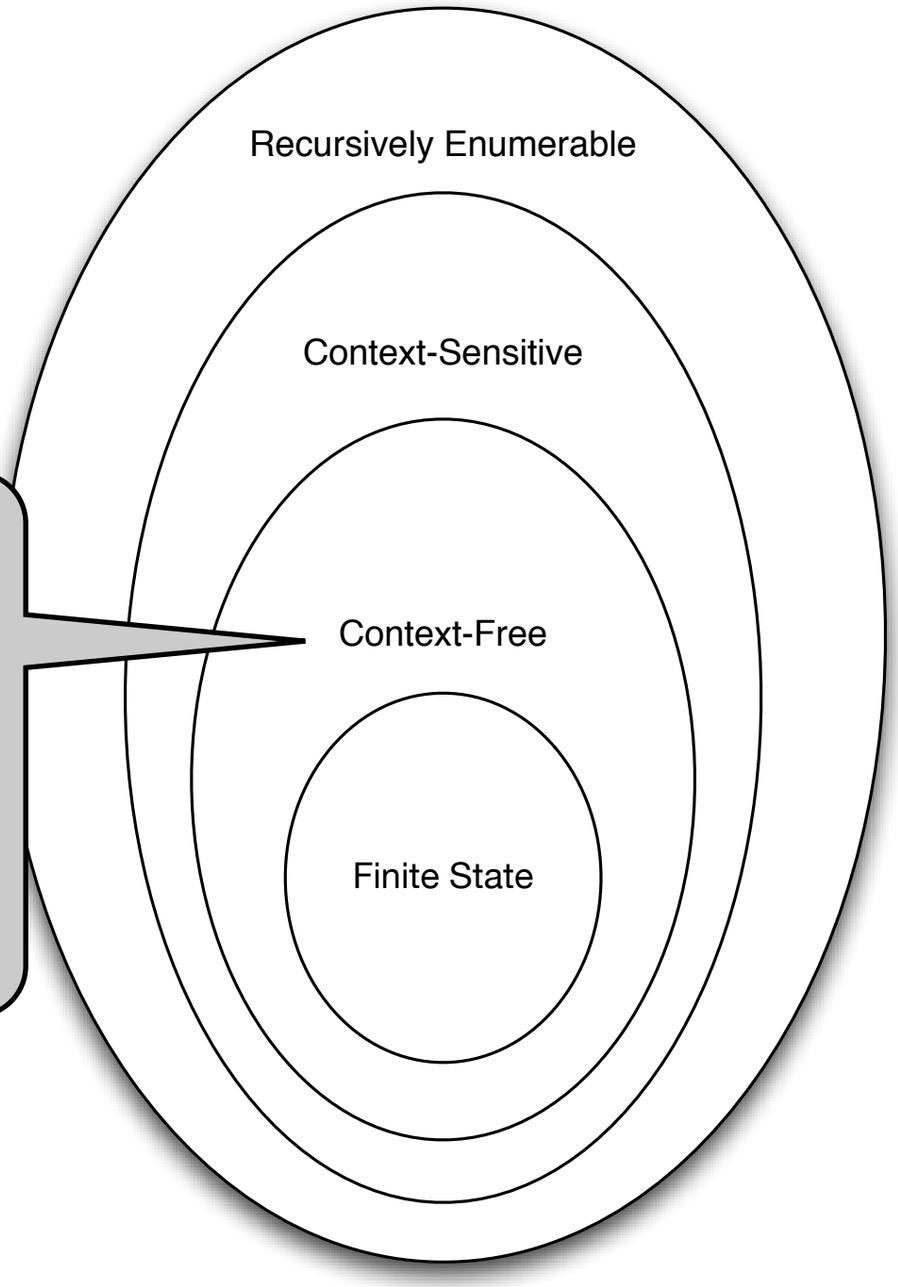


Comparing Formal Grammars

Descriptive Accuracy

Context-free is not sufficiently powerful:

mer d’chind em Hans es huus lönd hälfe aastrüche
we the children Hans the house let help paint
``we let the children help Hans paint the house.'



The Goals of Linguistic Theory

Chomsky (1965)

- **Descriptive Adequacy**: Is the theory sufficiently rich to provide a description of a natural language? (Lower bound on complexity)

- **Explanatory Adequacy**: Does the theory provide an account of how a learner projects from a finite sample of data into a grammar?

- **Chomsky's Idea (Inductive Bias)**: Theory provides a "ranking" of grammatical hypotheses, which determines their preference for the learner.

Comparing Formal Grammars

Explanatory Adequacy

- Question formation:
  - The zebra is resting ⇒
    Is the zebra resting?
  - The zebra is hoping that the lion has eaten ⇒
    Is the zebra hoping that the lion has eaten?

Move First (linear):
move the first verb's auxiliary verb to the front of the sentence

Move Main (structural):
move the main verb's auxiliary verb to the front of the sentence

Comparing Formal Grammars

Explanatory Adequacy

- These hypotheses are distinguishable on the basis of data that is (plausibly) not part of a child's experience:

  - [The zebra that is resting] has eaten a bug ⇒
    MAIN: Has [the zebra that is resting] eaten a bug?
    FIRST: *Is [the zebra that resting] has eaten a bug?

- A grammar formalism that favors the right generalization (i.e., makes the right rule simpler) is preferable

Move First (linear): move the first verb's auxiliary verb to the front of the sentence

Move Main (structural): move the main verb's auxiliary verb to the front of the sentence

Comparing Formal Grammars

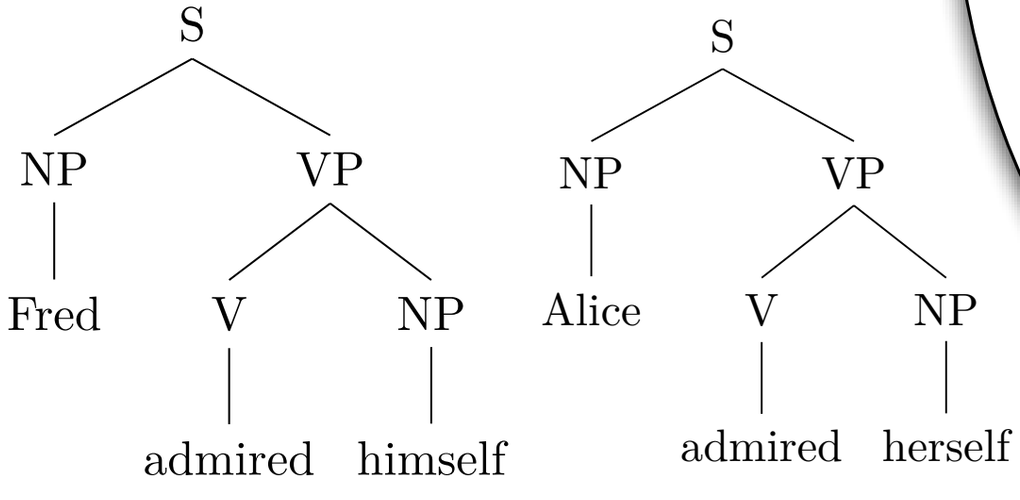
Explanatory Adequacy and CFGs

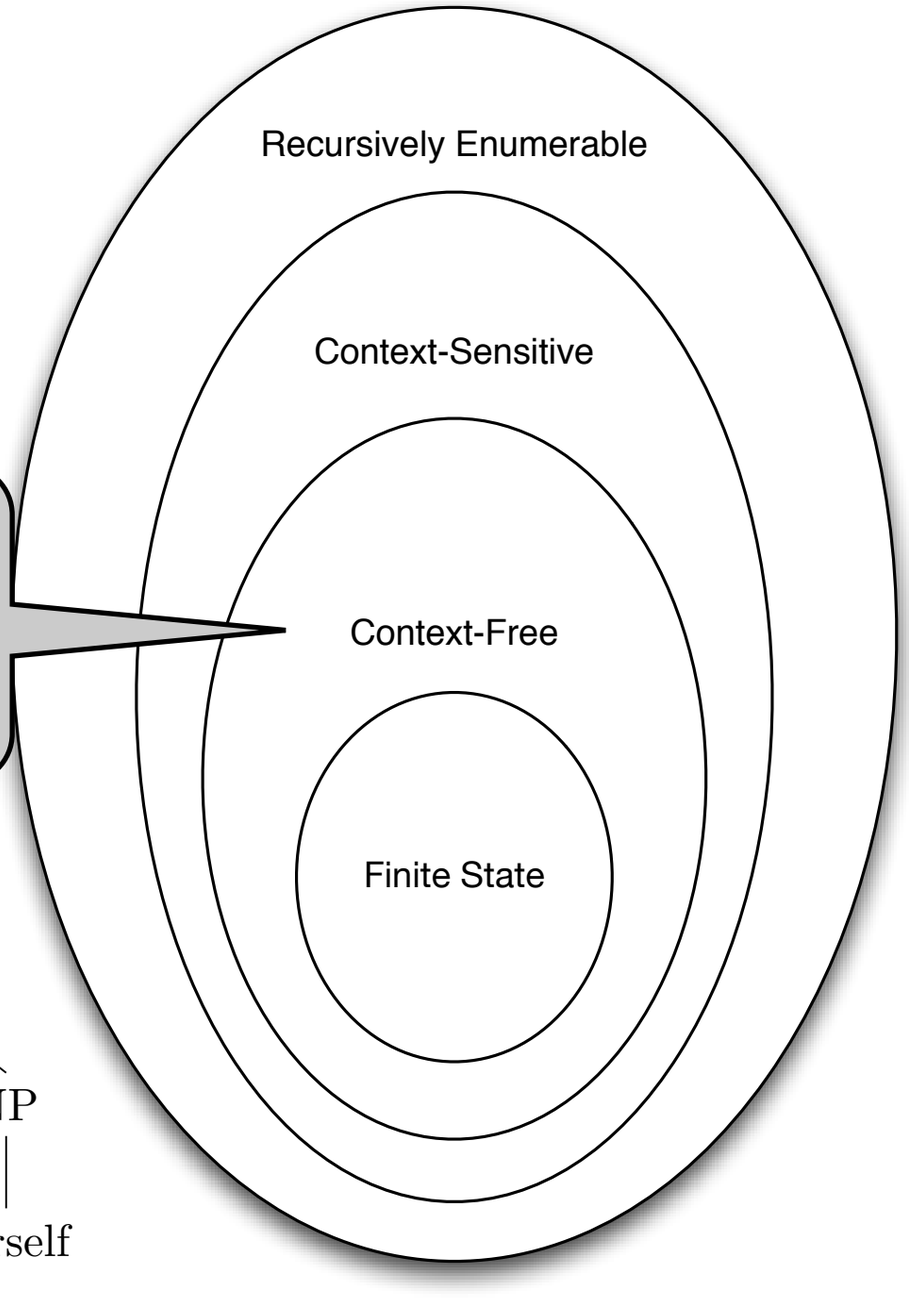
Context-free does not make the right things simple:

Fred admired himself. Alice admired herself

$S \rightarrow NP\ VP$

$VP \rightarrow V\ NP$





Comparing Formal Grammars

Explanatory Adequacy

What counts as “simple” in CFG does not correspond to grammatical naturalness!

Context-free does not make the right things simple:

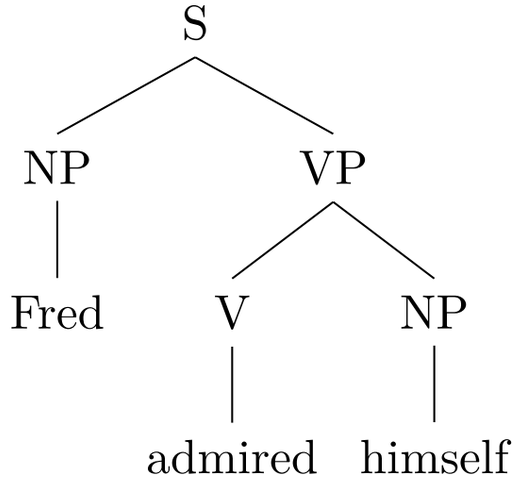
Fred admired himself. Alice admired herself

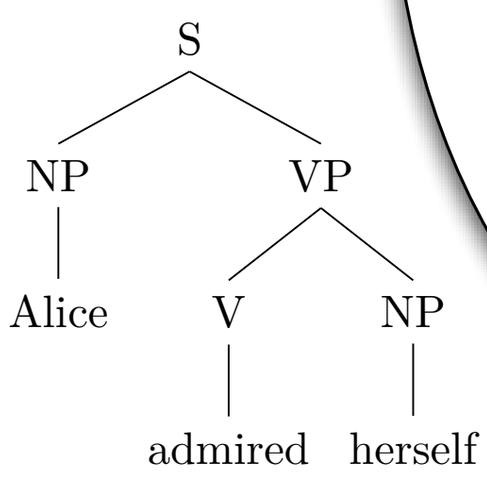
$S \rightarrow NP\text{-}m\ VP\text{-}m$

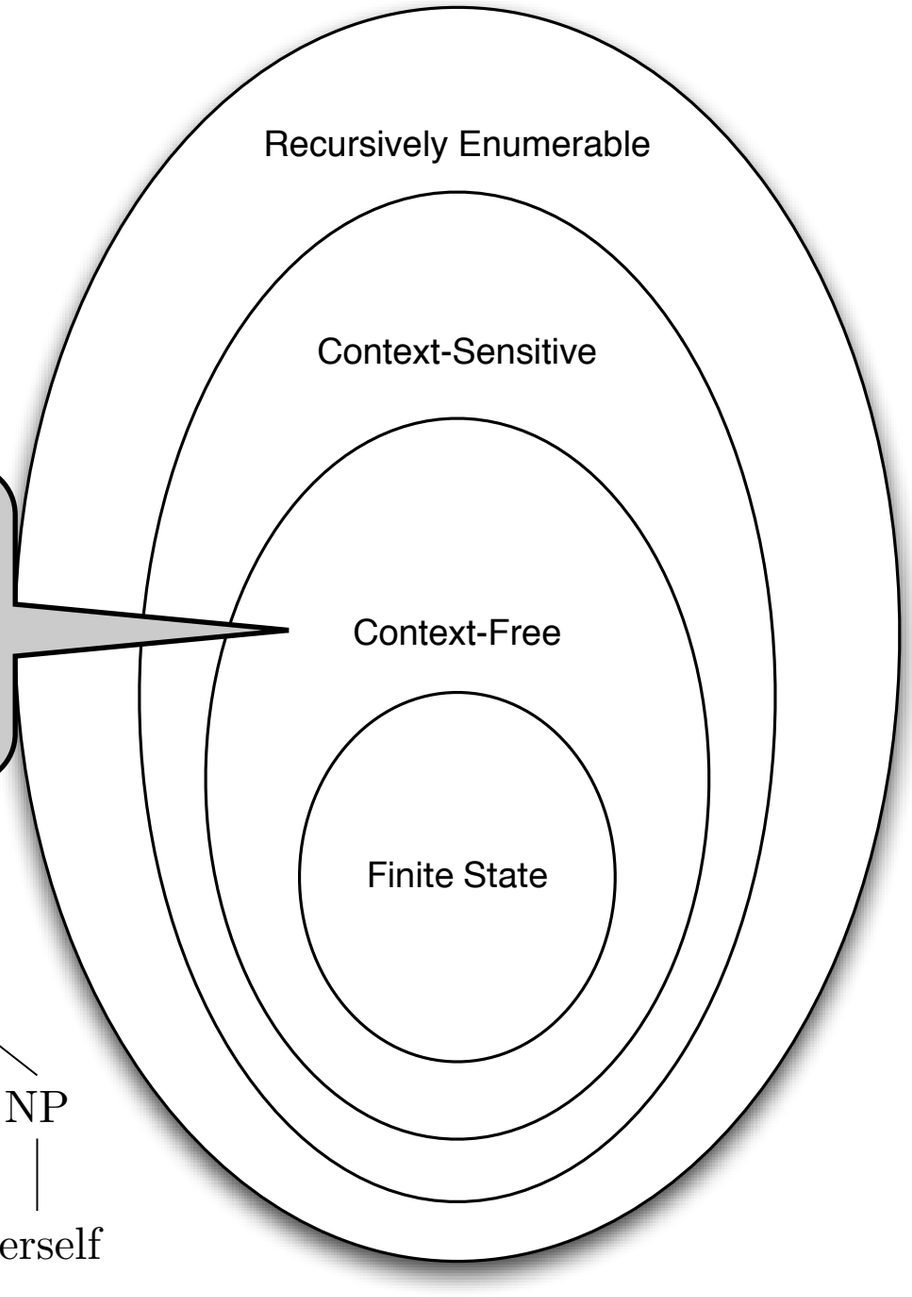
$VP\text{-}m \rightarrow V\ NP\text{-}m$

$S \rightarrow NP\text{-}f\ VP\text{-}f$

$VP\text{-}f \rightarrow V\ NP\text{-}f$







Comparing Formal Grammars

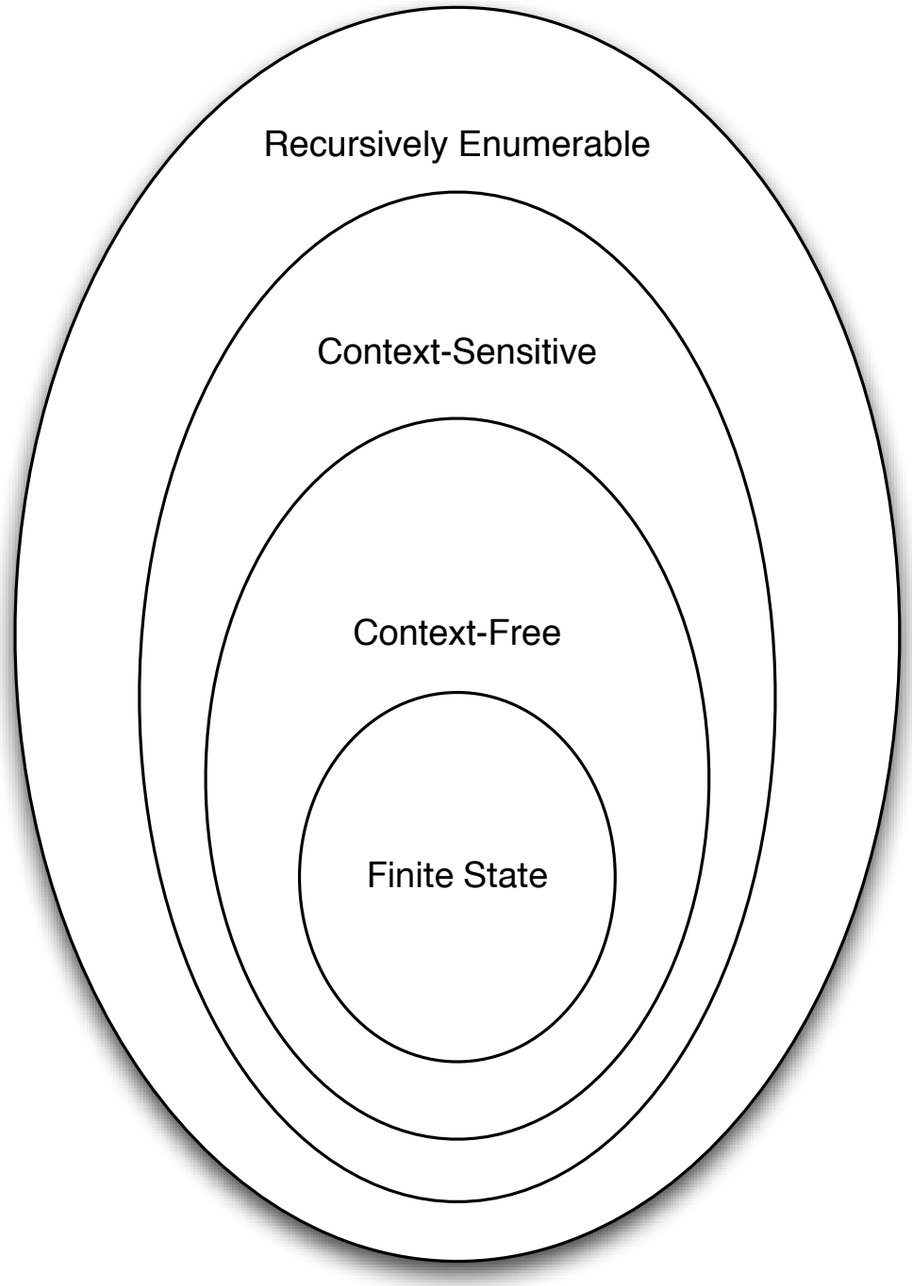
Explanatory Adequacy

Transformational grammars
(CFG + structure-dependent modifications)
provide a natural account of grammatical phenomena

Non-structural generalization are harder to state

$$S \rightarrow NP\ VP$$
$$VP \rightarrow V\ NP$$

$$X - Nom - Y - Nom' \quad - Z \rightarrow$$
$$X - Nom - Y - Nom'+Self - Z$$



The Goals of Linguistic Theory

Chomsky (1965)

- **Descriptive Adequacy**: Is the theory sufficiently rich to provide a description of a natural language? (*Lower bound on complexity*)

- **Explanatory Adequacy**: Does the theory provide an account of how a learner projects a finite sample of data into the right grammar?

  - **von Humboldt's idea**: Theory limits grammatical hypotheses to those that are compatible with linguistic universals. (*Upper bound on complexity*)

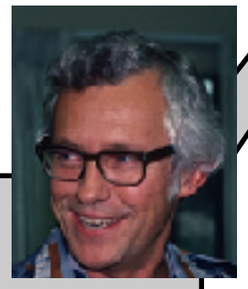
  - Constraint imposed by a grammatical theory limits hypotheses (grammatical universals), and should/could make the grammar learning easier.

Comparing Formal Grammars

Explanatory Adequacy

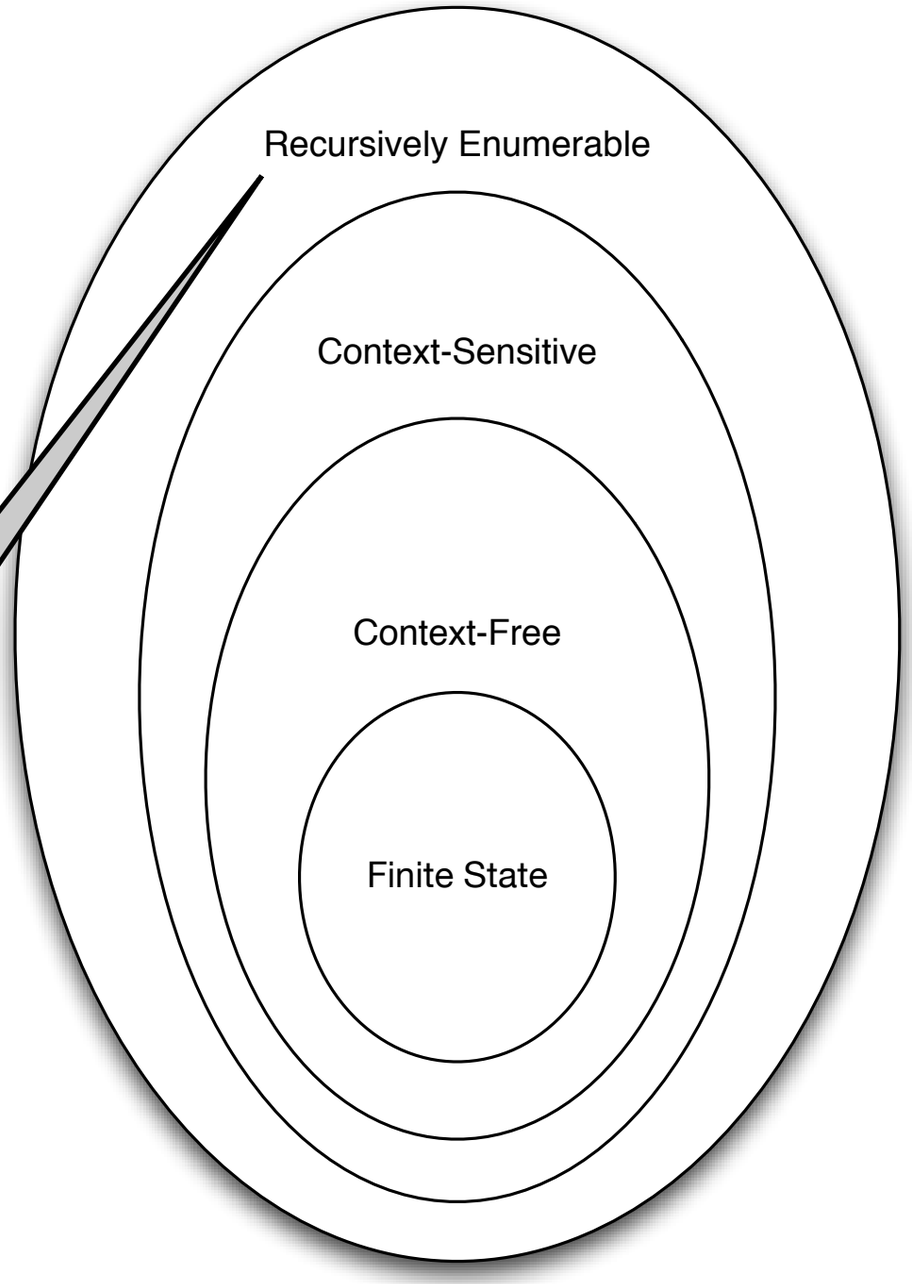
Transformational grammars
(CFG + structure-dependent modifications)
provide a natural account of grammatical phenomena

Non-structural generalization are harder to state



But TGs are very powerful (Peters and Ritchie)!

They do not limit the expected range of
grammatical patterns or analytical options



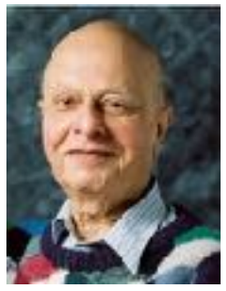
What are Language Universals?

- Most work in generative grammar: stipulate properties of grammars:
  - **Final over final condition**: no structures where a head-initial phrase is contained in a head-final phrase in the same extended projection/domain
  - **Williams cycle**: a phrase cannot move into a position that is “less prominent” than the one it is currently in
  - **Subjacency**: No movement across more than one bounding node.
- An alternative: derive constraints on what is possible from formal (computational) properties of grammars

The Changing Tide

- Chomsky (1980)
  - We are now asking **whether there is some reason of principle why these grammars must generate recursive sets. No serious argument has ever been advanced in support of such a claim**.
- Chomsky (2005)
  - The Principles-and-Parameters approach opened the possibility for serious investigation of the third factor, and **the attempt to account for properties of language in terms of general considerations of computational efficiency**, eliminating some of the technology postulated as specific to language and providing more principled explanation of linguistic phenomena.

Computational Universals of Grammar



- Joshi (1985) introduces the **Mildly Context-Sensitive Languages (MCSLs)**
  - **Limited crossing dependencies:** the class of natural languages should be sufficiently expressive to characterize the cross-serial word order patterns found in languages like Swiss German (Shieber 1985)

    ww but not $MIX = \{w \in \{a, b, c\}^* \mid \#_a(w) = \#_b(w) = \#_c(w)\}$

  - **Constant Growth property**: the sentences in any natural language should not have gaps of unbounded length (i.e., if we order the strings in a language by length, the gaps should be bounded in size).

    $\{a^{2n} \mid n \in \mathbb{N}\}$ but not $\{a^{n^2} \mid n \in \mathbb{N}\}$

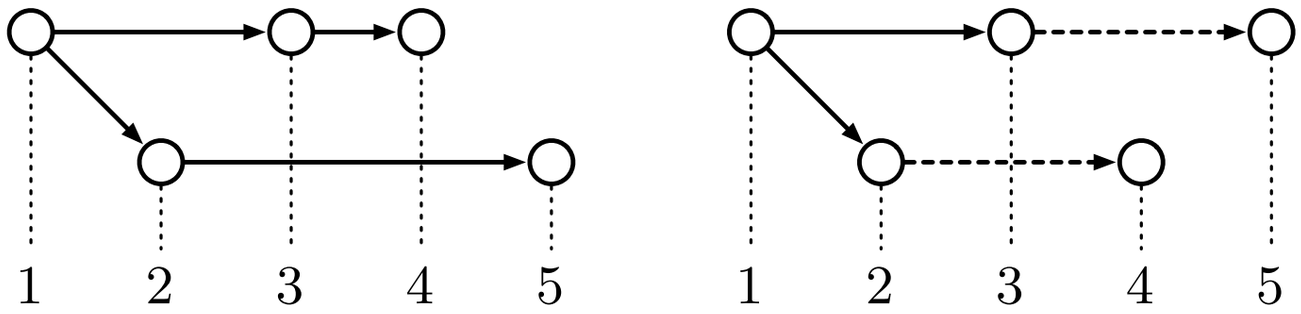
  - **Polynomial parsing**: structural descriptions can be efficiently assigned to strings

- Not a formal definition, but several weakly equivalent instantiations: Tree Adjoining Grammars, Combinatory Categorial Grammars, Linear Indexed Grammars (MCFGs and MGs are richer)

Computational Universals of Grammar

- **Well-Nesting** (Kuhlman 2007):

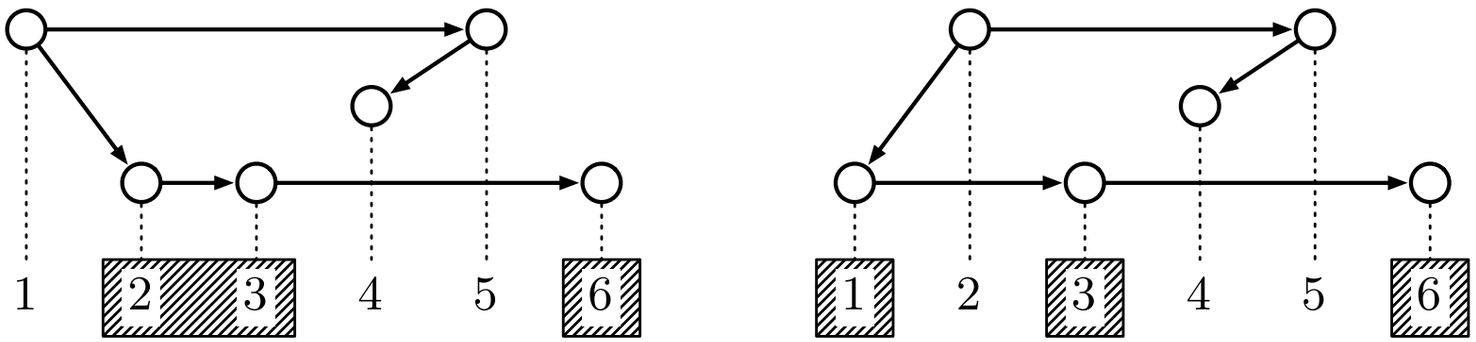
  For all edges v1 → v2 , w1 → w2 in D, if [v1, v2] partially overlaps [w1, w2] then v1 →* w1 or w1 →* v1.



	DDT		PDT 1.0		PDT 2.0	
projective	3 730	84.95%	56 168	76.85%	52 805	77.02%
weakly non-proj.	3 794	86.40%	60 048	82.16%	56 367	82.21%
well-nested	4 386	99.89%	73 010	99.89%	68 481	99.88%

Computational Universals of Grammar

- **Block Degree** (Kuhlman 2007):



(a) D_1, block-degree 2 (b) D_2, block-degree 3

Well-nested dependencies of block degree at most 2 are exactly those generable by TAG/CCG

block-degree	DDT		PDT 1.0		PDT 2.0	
1 (projective)	3 730	84.95%	56 168	76.85%	52 805	77.02%
2	654	14.89%	16 608	22.72%	15 467	22.56%
3	7	0.16%	307	0.42%	288	0.42%
4	–	–	4	0.01%	1	< 0.01%
5	–	–	1	< 0.01%	1	< 0.01%
TOTAL	4 391	100.00%	73 088	100.00%	68 562	100.00%

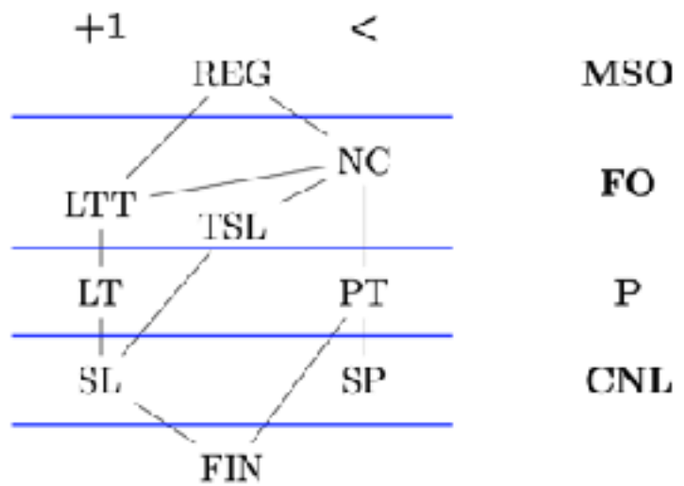
Well nested 2-MCFGs (weakly) equivalent to TAG/CCG

Computational Universals of Grammar

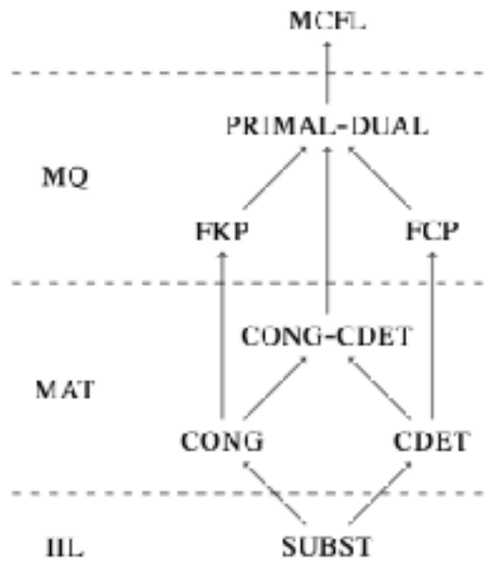
- **The good news**: computational properties of MCSLs derive a number of linguistic patterns:
  - Subjacency (Kroch and Joshi)
  - Williams cycle (Frank)
  - Constraints on ordering (Steedman, Kroch and Santorini)
- **The bad news**: The constraints imposed by MCS don’t help learning
  - Gold: even the lowest rung on the Chomsky hierarchy doesn’t constitute a learnable class

Formal grammars and learning

- **Another path**: look for learnable subclasses
  - Heinz, Chandlee, Jardine, Rogers: subregular languages (for phonological patterns)



- Clark, Eyraud, Yoshinaka, Kanazawa: substitutable and congruential grammars (for syntactic patterns)



Formal grammars and learning

- Do the resulting classes match properties of natural language? Is there a convergence between learning and typology?
- In the domain of Phonology, things looks promising (Heinz et al.)
  - Attested properties admit an elegant typology:
    - Phonotactics/cluster restrictions: Strictly local
    - Long-distance interactions (Chumash sibilants): Strictly piecewise
    - Stress: Tier-based Strictly Local
  - No way to state “impossible” constraints:
    - First/Last Harmony
    - Even number of sibilants

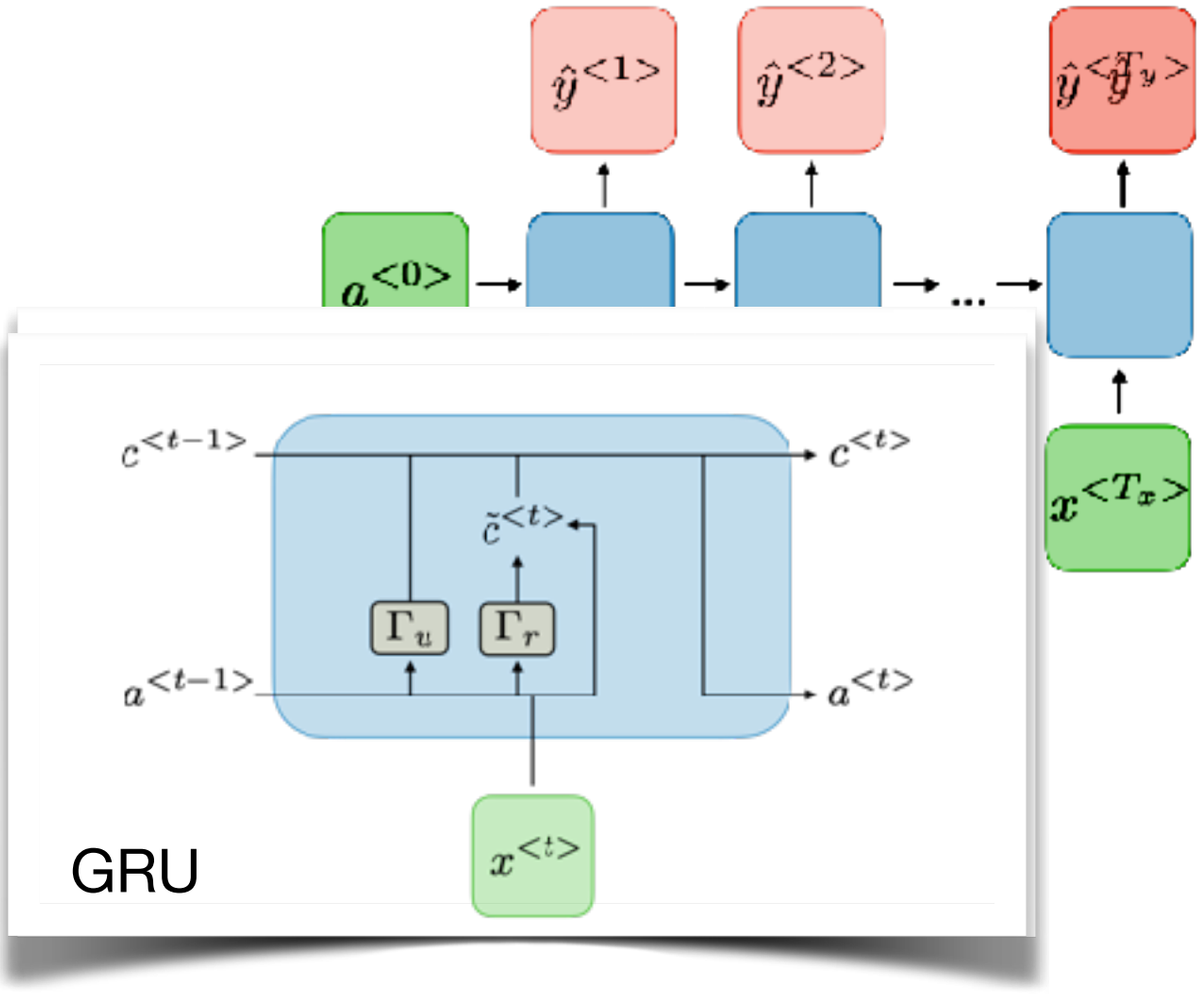
Formal grammars and learning

- Do the resulting classes match properties of natural language? Is there a convergence between learning and typology?
  - Syntax: Less promising (Hao, 2019, Graf 2013)
  - *Non-structural properties*
    $\text{MOD}_n = \{x \mid |x| = 0 \bmod n\}$ is k-substitutable for all k
  - *Unconstrained crossing dependencies*
    MIX = $\{w \in \{a, b, c\}^* \mid \#_a(w) = \#_b(w) = \#_c(w)\}$ is k-substitutable for all k
  - *Unbounded copying*
    $\text{COPY}_n(L) = \{(x\#)^n \mid x \in L\}$ with *L* context-free is generated by a 3-MCFG with the 2-FKP and 1-FCP

The world of neural networks

Recurrent networks

- A model of transduction or acceptance:



$$a^{<t>} = \tanh(W_0 a^{<t-1>} + W_1 x^{<t>} + b_a)$$

$$y^{<t>} = \text{softmax}(W_y a^{<t>} + b_y)$$

$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r \odot a^{<t-1>}, x^{<t>}] + b_c)$$

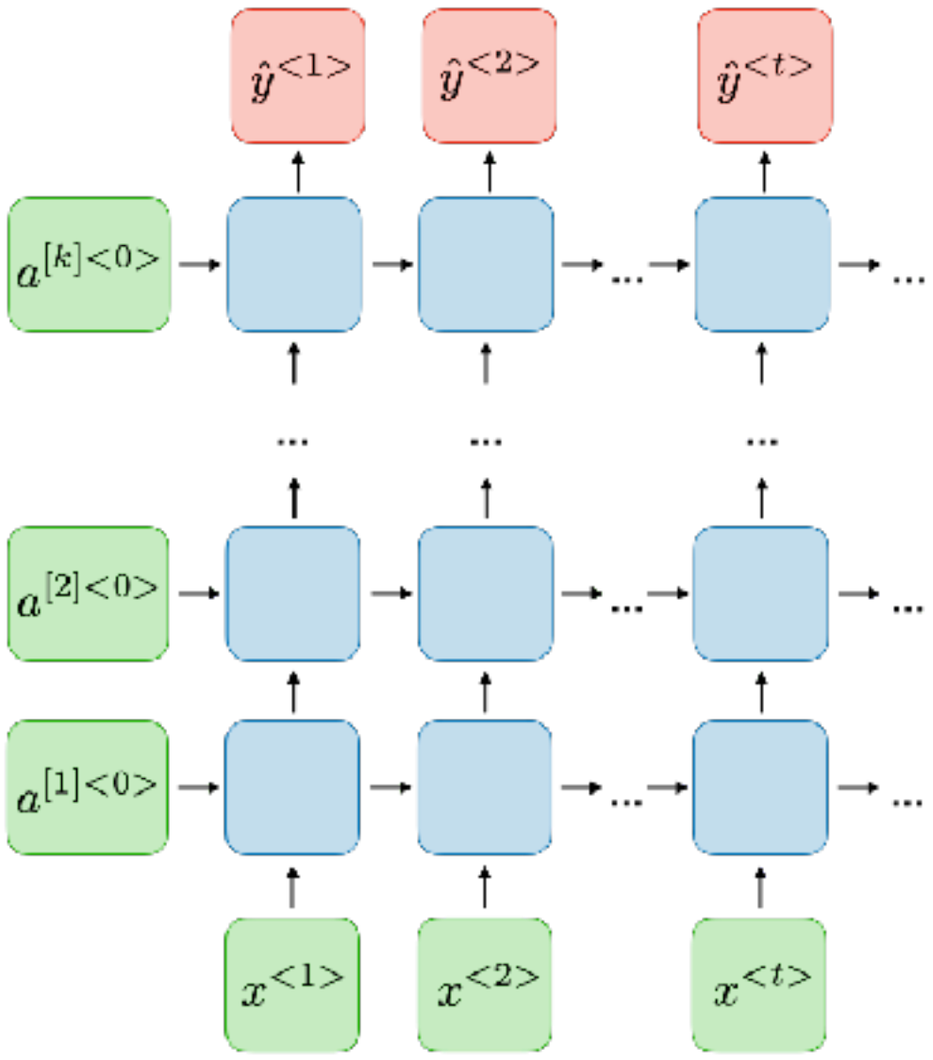
$$c^{<t>} = \Gamma_u \odot \tilde{c}^{<t>} + (1 - \Gamma_u) \odot c^{<t-1>}$$

$$a^{<t>} = c^{<t>}$$

$$\Gamma = \sigma(Wx^{<t>} + Ua^{<t-1>} + b)$$

The world of neural networks

Deep RNNs

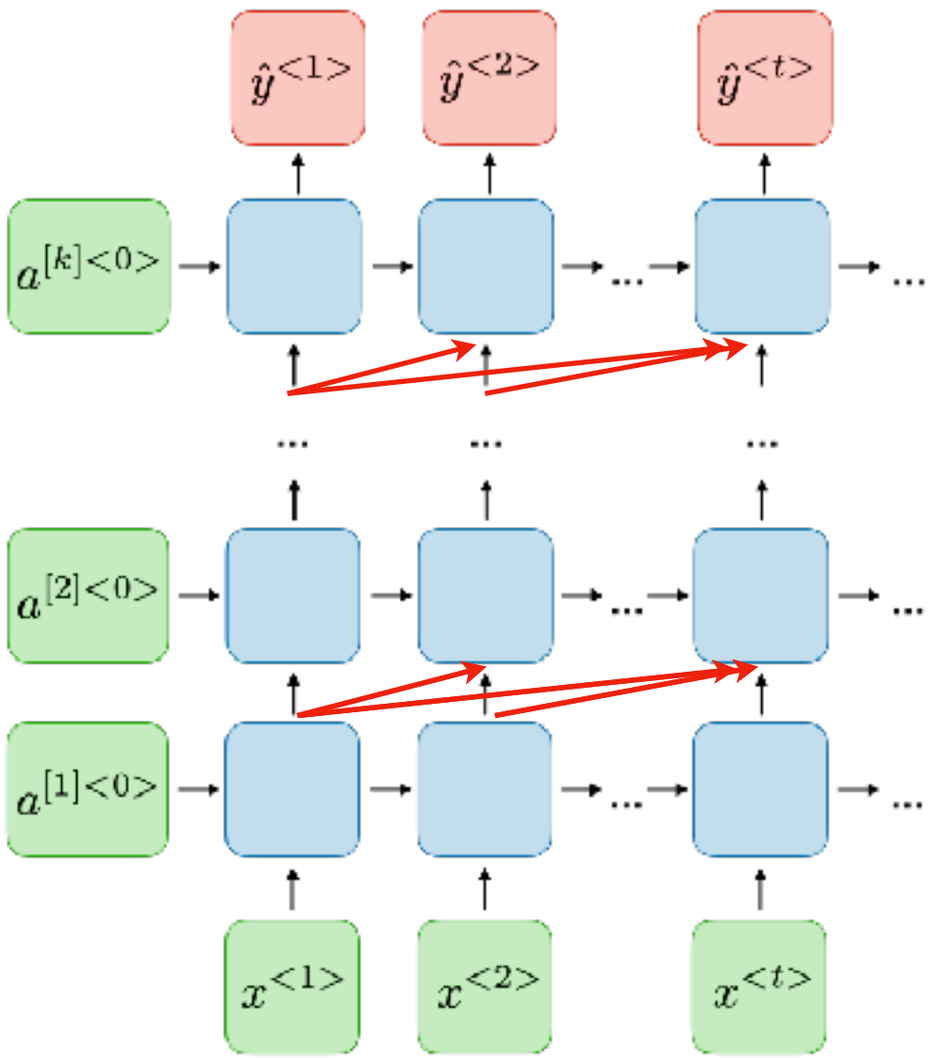


The world of neural networks

Attention

Give access to all (previous) time steps (attention).

Take sum of all these inputs weighted by mask that is computed by comparing representations.

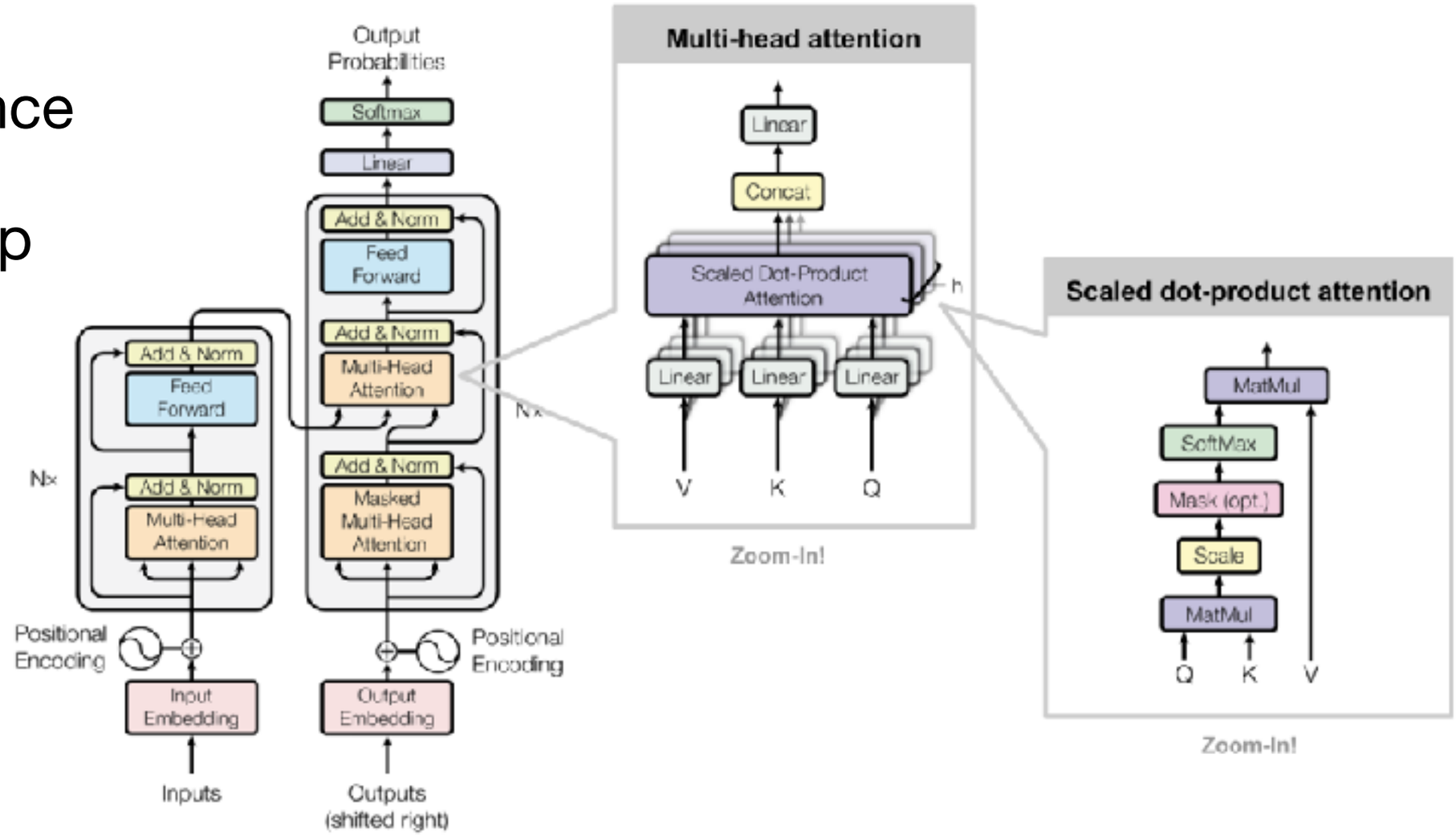


The world of neural networks

Transformers

Eliminate recurrence and use only attention and deep networks to pass information.

Need for explicit representation of position of input token.



Descriptive Adequacy of NNs

- Are networks sufficiently rich to capture the properties of natural language?
  - Siegelmann and Sontag (1992, 1994), Siegelmann (1998): RNNs can compute all Turing machine computable functions.
  - Pérez et al. (2019), Bhattamishra et al. (2020): Transformers are Turing complete too.
  - However both of these results require arbitrary (rational) precision and unbounded computation time

Descriptive Adequacy of RNNs

- Weiss et al. (2019), Merrill (2020), Merrill et al. (2020) impose bounded precision (saturated values):
  - SRNLs = GRULs = Regular Languages
  - LSTMLs ⊆ CLs (real time counter languages)
- Kirov and Frank (2011): With bounded time and o(log n) precision, SRNs can represent nested and crossing dependencies
- As yet unknown:
  - Under what conditions can RNNs encode MIX or unbounded copying?
  - Do they enforce constant growth and well-nesting?

Descriptive Adequacy of Transformers

Hard Attention

- Hahn (2020):
  - Parity $\{w \in \{0,1\}^* \mid \#_1(w) = 0 \mod 2\}$
  - Dyck-2
    S → ϵ
    S → (S)
    S → [S]
    S → S S

    Why Dyck?
    "Pure" representation of hierarchical structure

    (([( )])) [ ]
    - Parity, Dyck-2 ∉ HardTLs
- Angluin and Hao (2021) generalize this result Hard attention Transformer languages are bounded by AC^0 (languages accepted by constant-depth polynomial-size Boolean circuits)
- But Yao et al. (2021) show a D+1-layer Transformer with Hard Attention can recognize Dyck-k with bounded depth D (which requires o(log n) precision)

Descriptive Adequacy of Transformers

Soft Attention

- Yao et al. (2021): A 2-layer soft-attention Transformer can recognize Dyck-k
    - Note 1: Their construction employs hard attention alongside soft attention
    - Note 2: It requires a certain sort of non-standard positional embedding: i/n
- These facts points to the conclusion, recently championed by Baroni (2021): we need to take the details of NN models more seriously.
    - How does expressive capacity of Transformers change with different attention models and different positional encodings?

Approaching Explanatory Adequacy

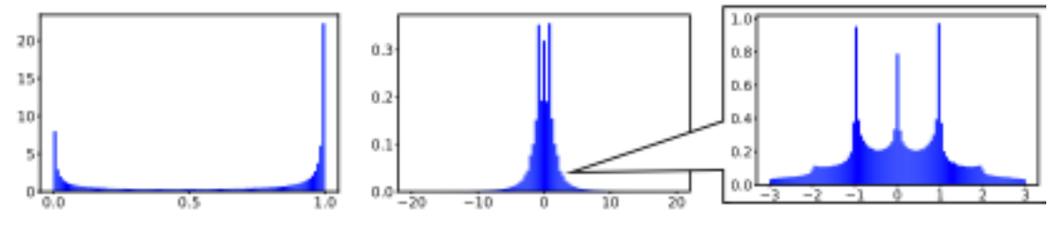
- Do Neural Networks provide a basis for the learning of grammatical patterns?
  - **Chomsky's Idea**: Theory provides a "ranking" of grammatical hypotheses, which determines their preference for the learner.
- How do we characterize grammatical simplicity or the inductive bias of a Gradient Descent learner when our model has millions or billions of parameters?

Approaching Explanatory Adequacy

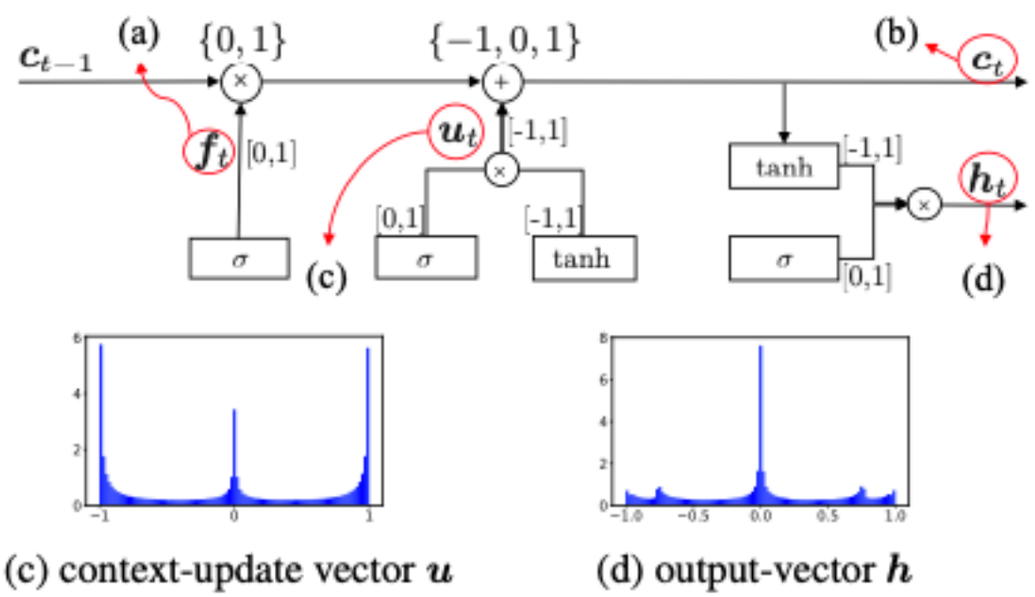
- Razin and Cohen (2020): inductive bias of deep neural networks + Gradient Descent is not characterizable as minimization of norms (for the task of matrix factorization).
  - R&C suggest matrix rank as an alternative possible characterization of inductive bias
- Gradient Descent during network training instead induces norm growth.
- What is the effect of that?

Approaching Explanatory Adequacy

- Merrill et al. (2021): Norm growth leads to saturated models (hard attention), and these are associated with weak expressivity noted above.
  - LSTMLs = CLs
- Shibata et al. (2020): PTB-trained LSTM language models show saturation in activations



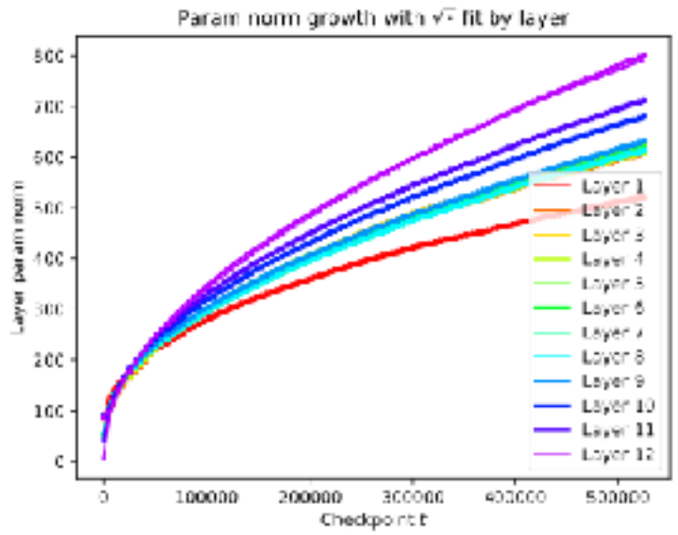
(a) forget vector f (b) context vector c



(c) context-update vector u (d) output-vector h

Approaching Explanatory Adequacy

- Merrill et al. (2021): norm growth empirically found in a big Transformer models (T5)



- Activations of pre-trained (but not untrained) Transformers are correlated with “saturated” versions