Hierarchical Bayesian models for learning linguistic structure

Computational Cognitive Science, Lecture 15

Stella Frank, stella.frank@ed.ac.uk

November 7, 2019

Cognition 118 (2011) 306–338

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The learnability of abstract syntactic principles

Amy Perfors [a,*], Joshua B. Tenenbaum [b], Terry Regier [c]

[a] Department of Psychology, University of Adelaide, Australia
[b] Department of Brain & Cognitive Science, Massachusetts Institute of Technology, United States
[c] Department of Linguistics, Cognitive Science Program, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Received 30 June 2010
Accepted 1 November 2010
Available online 24 December 2010

Keywords:
Poverty of stimulus
Bayesian modeling
Language learnability

ABSTRACT

Children acquiring language infer the correct form of syntactic constructions for which they appear to have little or no direct evidence, avoiding simple but incorrect generalizations that would be consistent with the data they receive. These generalizations must be guided by some inductive bias – some abstract knowledge – that leads them to prefer the correct hypotheses even in the absence of directly supporting evidence. What form do these inductive constraints take? It is often argued or assumed that they reflect innately specified knowledge of language. A classic example of such an argument moves from the phenomenon of auxiliary fronting in English interrogatives to the conclusion that children must innately know that syntactic rules are defined over hierarchical phrase structures rather than linear sequences of words (e.g., Chomsky, 1965, 1971, 1980; Crain & Nakayama, 1987). Here we use a Bayesian framework for grammar induction to address a version of this argument and show that, given typical child-directed speech and certain innate domain-general capacities, an ideal learner could recognize the hierarchical phrase structure of language without having this knowledge innately specified as part of the language faculty. We discuss the implications of this analysis for accounts of human language acquisition.

Cognition 118 (2011) 306–338

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The learnability of abstract syntactic principles

Amy Perfors [a,*], Joshua B. Tenenbaum [b], Terry Regier [c]

[a] Dep
[b] Dep
[c] Dep

Use hierarchical Bayesian models to settle a longstanding debate in linguistics: how can children learn syntax?

A R

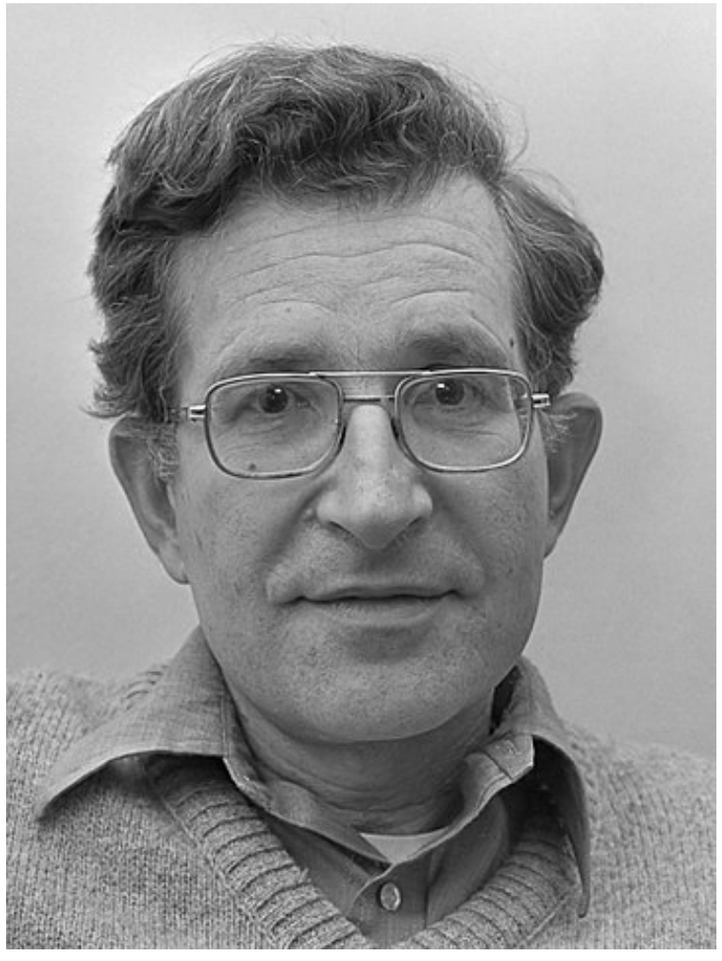
Article
Received 30 June 2010
Accepted 1 November 2010
Available online 24 December 2010

Keywords:
Poverty of stimulus
Bayesian modeling
Language learnability

hich they appear to have little or no direct evidence, avoiding simple but incorrect generalizations that would be consistent with the data they receive. These generalizations must be guided by some inductive bias – some abstract knowledge – that leads them to prefer the correct hypotheses even in the absence of directly supporting evidence. What form do these inductive constraints take? It is often argued or assumed that they reflect innately specified knowledge of language. A classic example of such an argument moves from the phenomenon of auxiliary fronting in English interrogatives to the conclusion that children must innately know that syntactic rules are defined over hierarchical phrase structures rather than linear sequences of words (e.g., Chomsky, 1965, 1971, 1980; Crain & Nakayama, 1987). Here we use a Bayesian framework for grammar induction to address a version of this argument and show that, given typical child-directed speech and certain innate domain-general capacities, an ideal learner could recognize the hierarchical phrase structure of language without having this knowledge innately specified as part of the language faculty. We discuss the implications of this analysis for accounts of human language acquisition.

"Poverty of Stimulus" argument (Chomsky et alia)

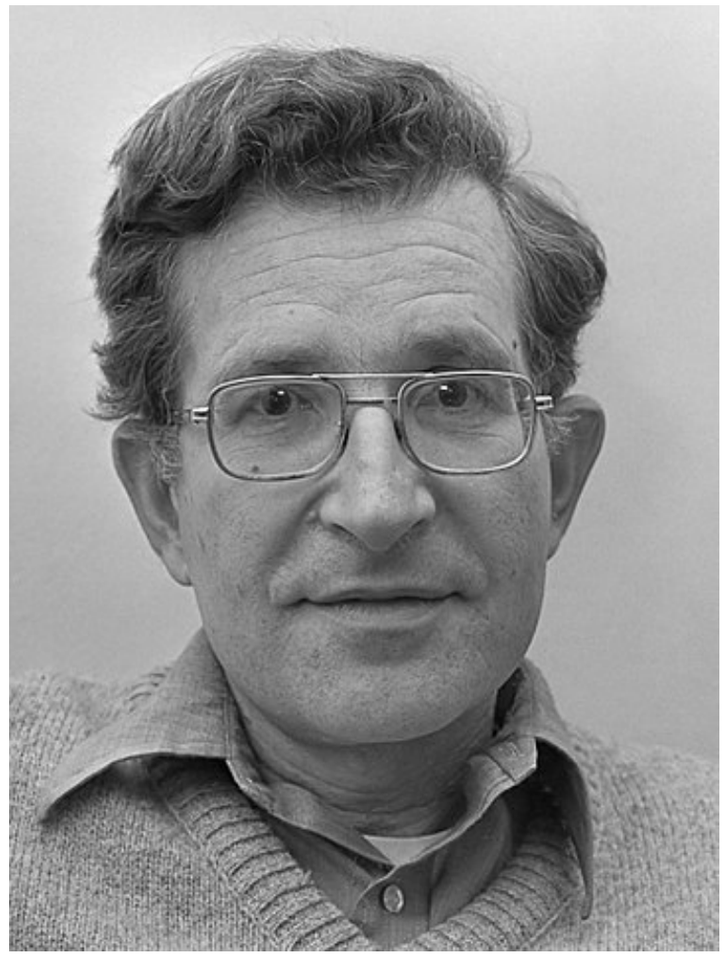
1. Children speak languages
2. Language is complex (see example)
3. Children do not hear enough instances of [complex example] to possibly learn it from data
4. Ergo children must have innate linguistic knowledge about [complex example]



"Poverty of Stimulus" argument (Chomsky et alia)

1. Children speak languages

2. Language is structured hierarchically (e.g. phrase structures, trees)



3. Children do not hear enough instances of [complex example] to possibly learn it from data

4. Ergo children must have innate knowledge about linguistic structure, i.e. that it requires hierarchical representations

Complex example: Aux-raising in English questions

"The bear is eating the fish's breakfast"

"Is the bear eating the fish's breakfast?"

Complex example: Aux-raising in English questions

"[[The bear] is [eating [the fish's breakfast]]]"

"[Is [the bear] [eating [the fish's breakfast]]?]"

Complex example: Aux-raising in English questions

"The students who are in the classroom are still awake"

"Are the students who in the classroom are still awake?"

"Are the students who are in the classroom still awake?"

Complex example: Aux-raising in English questions

"The students who are in the classroom are still awake"

Linear rule: move the first auxiliary verb

"Are the students who in the classroom are still awake?"

"Are [[the students] who are in the classroom] still awake?"

Hierarchical rule: move the aux in main clause

Aux-raising in other languages?

Aux raising in other languages

German:

Finnish:

[[Poika], joka on onnellinen,] on leikkimässä
(The boy who is happy is playing)

Language "acquisition device"

- Any infant who can learn any other language can also learn English, so:

- (Learning) the aux-raising rule (and thus: hierarchical structure of language) has to be part of the linguistic capabilities of all infants.

- But: English child-directed speech contains only ~0.05% complex interrogatives - can't be enough to learn from

- Ergo hierarchical structure must be specified innately

Counter "Poverty of Stimulus" (Perfors et al)

1. Language is complex

2. Sure - and children learn it as a system

3. Arguing from a single (type of) example is silly

4. Overhypotheses!

Counter "Poverty of Stimulus" (Perfors et al)

To put this point another way, while it may be sensible to ask what a rational learner can infer about language as a whole without any language-specific biases, it is less sensible to ask what a rational learner can infer about any single specific linguistic rule (such as auxiliary-fronting). The need to acquire a whole system of linguistic rules together imposes constraints among the rules, so that an *a priori* unbiased learner may acquire constraints that are based on the other linguistic rules it must learn at the same time.

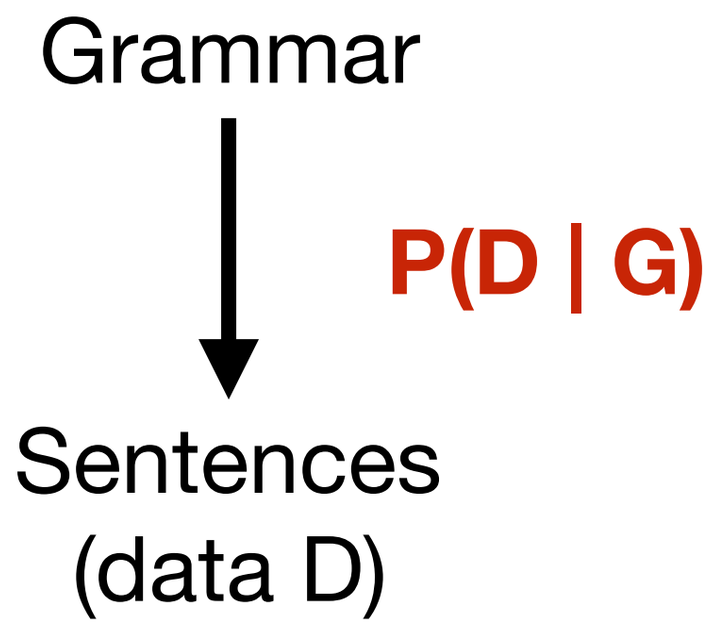
Problem statement

Learners have to learn how to produce questions,
i.e., learn a grammar that

1. accounts for the observed data (e.g. child-directed speech from CHILDES corpus)
2. generates questions correctly

How to model this learning task in a Bayesian framework?
- in a way that allows us to distinguish between linear and hierarchical structures?

Grammars give rise to utterances



Infer grammar from data using posterior

Grammar **P(G | D) ∝ P(D | G) P(G)**

Sentences
(data D)

Requires a hypothesis space over grammars G
(as well as a prior distribution over G

One space of grammars:
Probabilistic Context-Free Grammars

"The cat who is happy is purring"
Pos tags: (det n wh aux vbg aux vbg)

S -> NP VP
VP -> aux vbg
NP -> det n
NP -> NP RelCl
RelCl -> wh VP

This toy grammar is far too small to generalise!
Need to add more rules to cover more sentences

Probabilistic Context-Free Grammars

- Space of grammars: all possible subsets of all possible (binary) rules + probabilities,
  given terminals and some set of non-terminals;

- A good grammar, with high posterior P(G | D)

  - has high P(D | G): assigns high probability to the data (and conversely low probability to unseen phenomena)

  - and high P(G): depends on setup; usually is "simpler" in some way (e.g. fewer rules)

PCFG incorporates hierarchical assumption

Data: "The cat who is happy is purring"
(in pos tags: det n wh aux vbg aux vbg)

S -> NP VP
VP -> aux vbg
NP -> det n
NP -> NP RelCl
RelCl -> wh VP

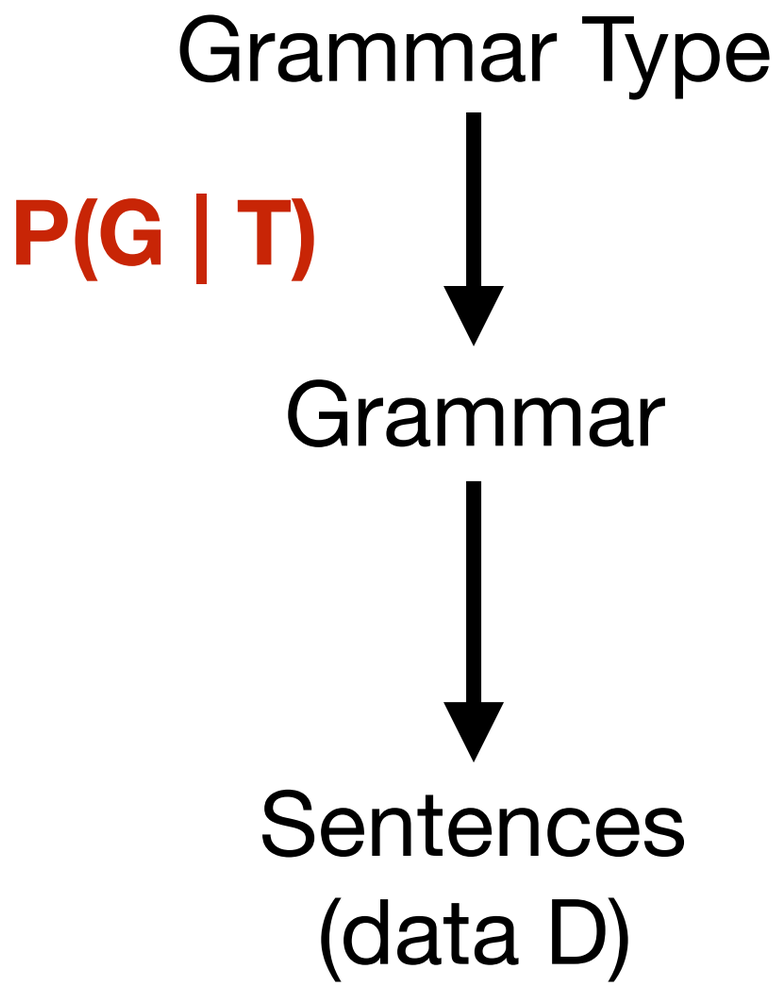
PCFG parse:

$[[\text{det n}]_{\text{NP}}\ [\text{wh}\ [\text{aux vbg}]_{\text{VP}}]_{\text{RelCl}}\ \text{aux vbg}]_{\text{VP}}]_{\text{S}}$

Flat parse isn't possible: isn't within the space of possible grammars

Can't evaluate linear hypothesis!

Hierarchical model of linguistic structure



$$P(T, G \mid D) \propto P(D \mid G)\, P(G \mid T)\, P(T)$$

Space of Grammar Types

- 'one state' grammar: X -> word X; can generate all possible sentences
- flat grammar: memorises all sentences in the corpus
- regular grammars: represent sentences linearly (different numbers of non-terminals, rules)
- context-free grammars: represent sentences hierarchically

All sentences represented as sequences of syntactic categories (POS tags)

All grammars are *probabilistic*: can assign a probability to a sentence.

Context-free grammar CFG-S

NP → NP PP | NP CP | NP C | N | det N | adj N | pro | prop

N → n | adj N

Context-free grammar CFG-L

NP → NP PP | NP CP | NP C | N PP | N CP | N C | pro PP | pro CP | pro C |
prop PP | prop CP | prop C | N | det N | adj N | pro | prop

N → n | adj N

Flat grammar

S → pro aux part

S → adj n aux n prep det n

S → det n v n

S → pro aux adj n comp pro v

Regular grammar REG-N

NP → pro | prop | n | det N | adj N | pro PP | prop PP | n PP | det N_{PP} | adj N_{PP} |
pro CP | prop CP | n CP | det N_{CP} | adj N_{CP} | pro C | prop C | n C | det
N_C | adj N_C

N → n | adj N

N_{PP} → n PP | adj N_{PP}

N_{CP} → n CP | adj N_{CP}

N_C → n C | adj N_C

Hierarchical model of linguistic structure

P(T) ~ uniform

Grammar Type

P(G|T) prefers simpler grammars of a given type

Grammar

Sentences (data D)

P(T, G| D) ∝ P(D|G) P(G|T) P(T)

Approach: find ***best*** grammar of each type, and evaluate its posterior probability, given plausible data D (from Childes)

Results by sentence frequency

Table 2
Log prior, likelihood, and posterior probabilities of each hand-designed grammar for each level of evidence. Because numbers are negative, smaller absolute values correspond to higher probability. If two grammars have log probabilities that differ by n, their actual probabilities differ by e^n; thus, the best hierarchical phrase-structure grammar CFG-L is e^{101} ($\sim 10^{43}$) times more probable than the best linear grammar REG-M. Bold values indicate the highest posterior score at each level.

Corpus	Probability	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Level 1	Prior	−99	−148	−124	−117	−94	−155	−192
	Likelihood	−17	−20	−19	−21	−36	−27	−27
	Posterior	**−116**	−168	−143	−138	−130	−182	−219
Level 2	Prior	−630	−456	−442	−411	−201	−357	−440
	Likelihood	−134	−147	−157	−162	−275	−194	−177
	Posterior	−764	−603	−599	−573	**−476**	−551	−617
Level 3	Prior	−1198	−663	−614	−529	−211	−454	−593
	Likelihood	−282	−323	−333	−346	−553	−402	−377
	Posterior	−1480	−986	−947	−875	**−764**	−856	−970
Level 4	Prior	−5839	−1550	−1134	−850	−234	−652	−1011
	Likelihood	−1498	−1761	−1918	−2042	−3104	−2078	−1956
	Posterior	−7337	−3311	−3052	−2892	−3338	**−2730**	−2967
Level 5	Prior	−10,610	−1962	−1321	−956	−244	−732	−1228
	Likelihood	−2856	−3376	−3584	−3816	−5790	−3917	−3703
	Posterior	−13,466	−5338	−4905	−4772	−6034	**−4649**	−4931
Level 6	Prior	−67,612	−5231	−2083	−1390	−257	−827	−1567
	Likelihood	−18,118	−24,454	−25,696	−27,123	−40,108	−27,312	−26,111
	Posterior	−85,730	−29,685	−27,779	−28,513	−40,365	−28,139	**−27,678**

Data from higher levels include more infrequent sentence types

Complex Aux-questions

Table 7
Ability of each grammar to parse specific sentences. The complex declarative sentence "Eagles that are alive can fly" occurs in the Adam corpus. Only the context-free grammars can parse the corresponding complex interrogative sentence.

Type	In input?	Example	Can parse?						
			FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Decl Simple	Y	Eagles can fly. (n aux vi)	Y	Y	Y	Y	Y	Y	Y
Int Simple	Y	Can eagles fly? (aux n vi)	Y	Y	Y	Y	Y	Y	Y
Decl Complex	Y	Eagles that are alive can fly. (n comp aux adj aux vi)	Y	Y	Y	Y	Y	Y	Y
Int Complex	N	Can eagles that are alive fly? (aux n comp aux adj vi)	N	N	N	N	Y	Y	Y
Int Complex	N	*Are eagles that alive can fly? (aux n comp adj aux vi)	N	N	N	N	Y	N	N

- 1-state can parse everything (by construction)
- Only CFGs parse the correct form of the question and fail to parse the incorrect form

Summary

- A learner with the representational capacity for both flat (regular) and hierarchical (context-free) grammars can infer, from child-directed speech data, that hierarchical structures capture the data better.

- Such a grammar can also correctly generalise to new structures, such as complex questions.

- No initial bias towards hierarchy or particular linguistic structures is necessary: data provides enough evidence.

Next week:

Words as high-dimensional objects (not discrete atomic categories), capturing semantics, syntax, phonology, etc.

- Is this representation cognitively realistic?
- How can we discover these representations?