The Learnability of Abstract Syntactic Principles

Amy Perfors, Joshua B. Tenenbaum, Terry Regier

Presented by Prutha Deshpande

Poverty of Stimulus (PoS) Argument

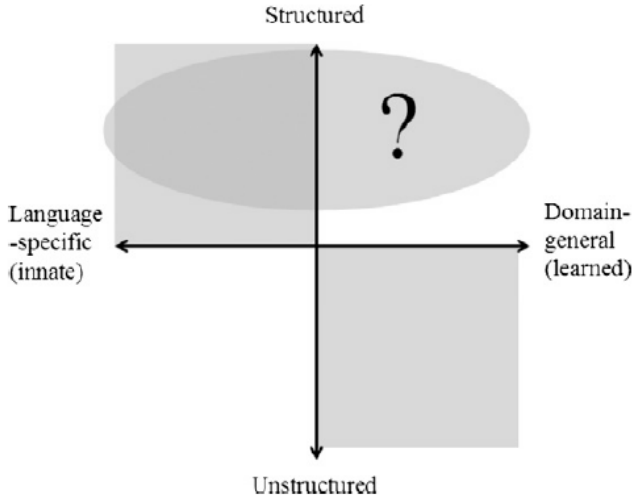
- Language learners make grammatical generalizations that go beyond what is immediately justified by the evidence in the input.
- For example: children appear to favor hierarchical rules of language over linear rules, even in the absence of evidence for this preference.
- This suggests that learners may innately know such things as an organizing principle of language.

Goal of the Paper

- Reevaluate the PoS argument for innate language-specific knowledge by formalizing the problem of language acquisition within a Bayesian framework for rational inductive inference.
- Consider an ideal learner with the following domain-general capacities:
  1. Can represent structured grammars of various forms
  2. Does Bayesian statistical inference
- Argue that certain core aspects of linguistic knowledge can be inferred without language-specific capabilities.

Highlights of the Approach

1. Consider the learnability of entire grammars, rather than the learnability of any specific linguistic rule.
2. The question of whether human learners have language-specific or domain-general knowledge is separable from the question of whether linguistic knowledge is structured or unstructured.



Auxiliary Fronting

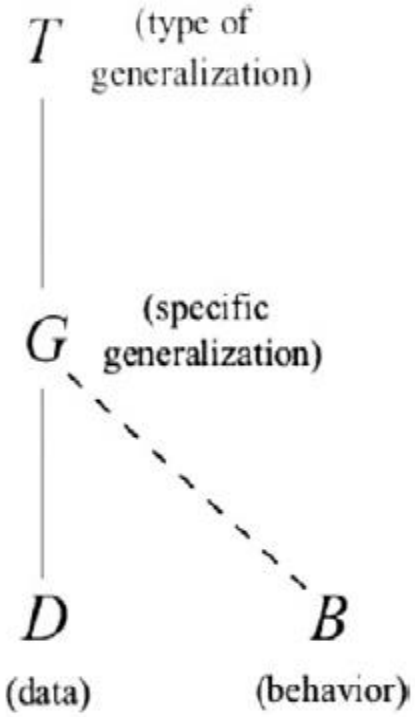
- Movement of auxiliary verb to the front of the sentence:

(1a) The man was hungry.
(1b) Was the man hungry?
(2a) The boy is smiling.
(2b) Is the boy smiling?

- With two identical auxiliary verbs:

(4a) The boy who is smiling is happy.
*(4b) Is the boy who smiling is happy?
(4c) Is the boy who is smiling happy?

Auxiliary Fronting



- There is not a unique way to characterize declarative and interrogative forms.
- Complex interrogative sentences are not common in child-directed speech.
- In spite of PoS, children can form correct complex interrogative sentences and do not produce incorrect forms.
- Children must have some innate knowledge of the hierarchical structure of phrases.
- What is the nature of this innate bias?

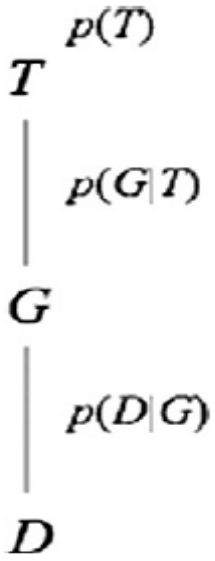
Main Results

1. An unbiased learner that can represent both linear and hierarchical grammars, can infer that the hierarchical grammar is a better fit to typical child-directed input.

   → It is possible to acquire domain-specific knowledge about the form of representations via domain-general learning mechanisms.

2. Hierarchical phrase-structure grammar succeeds in the auxiliary fronting task even when no direct evidence is available in the input data.

Method



Prior distribution of grammar T (equal for each type T)

Conditional probability of a specific grammar G (assigning higher probability to simpler grammars)

Conditional probability of the data (measure of fit of grammar to corpus data)

- Inferences from D to G and T captured by joint posterior probability:

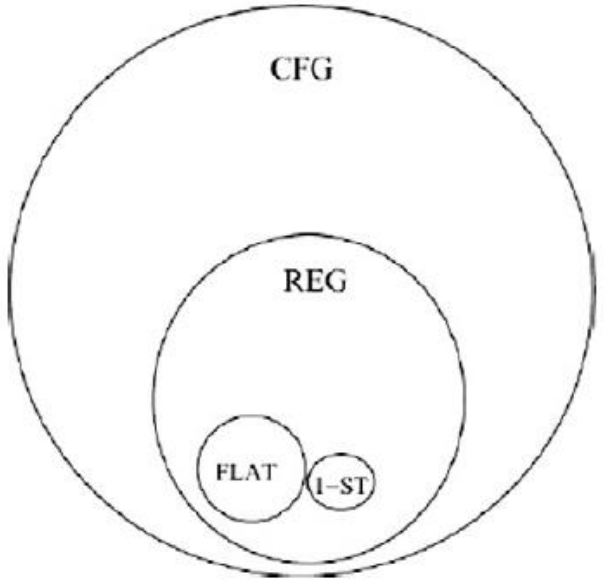
$$p(G,T|D) \propto_p (D|G)_p (G|T_p(T).$$

The Corpora

- Adam corpus of the CHILDES database.
- Each word replaced by its syntactic category.
- Ungrammatical and grammatically complex sentences removed.
- Six corpus levels: depending on token frequency.
- Five corpus epochs: depending on age of the child.

Hypothesis Space of Grammars

1. To represent grammars with hierarchical phrase-structure
   - Context-free grammars
2. To represent grammars without hierarchical phrase-structure
   - Regular grammars – finite-state grammars
   - FLAT grammar
   - One-state (1-ST) grammar



Results

Hand-designed grammars:

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**

Results

- To what extent are these results dependent on particular hand-designed grammars? Local search from hand-designed grammars:

Corpus	Probability	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Level 1	Prior	−99	−99	−99	−99	−94	−133	−148
	Likelihood	−17	−19	−20	−19	−36	−26	−25
	Posterior	**−116**	−118	−119	−118	−130	−159	−173
Level 2	Prior	−630	−385	−423	−384	−201	−355	−404
	Likelihood	−134	−151	−158	−155	−275	−189	−188
	Posterior	−764	−536	−581	−539	**−476**	−544	−592
Level 3	Prior	−1198	−653	−569	−529	−211	−433	−521
	Likelihood	−282	−320	−339	−346	−553	−402	−380
	Posterior	−1480	−973	−908	−875	**−764**	−835	−901
Level 4	Prior	−5839	−1514	−1099	−837	−234	−566	−798
	Likelihood	−1498	−1770	−1868	−2008	−3104	−2088	−1991
	Posterior	−7337	−3284	−2967	−2845	−3338	**−2654**	−2789
Level 5	Prior	−10,610	−1771	−1279	−956	−244	−615	−817
	Likelihood	−2856	−3514	−3618	−3816	−5790	−3931	−3781
	Posterior	−13,466	−5285	−4897	−4772	−6034	**−4546**	−4598
Level 6	Prior	−67,612	−5169	−2283	−1943	−257	−876	−1111
	Likelihood	−18,118	−24,299	−25,303	−25,368	−40,108	−27,032	−25,889
	Posterior	−85,730	−29,468	−27,586	−27,311	−40,365	−27,908	**−27,000**

Results

Regular grammar by automated search:

REG-AUTO				Other best grammars (posterior).						
Corpus	Prior	Likelihood	Posterior	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Level 1	−105	−18	−123	**−116**	−118	−119	−118	−130	−159	−173
Level 2	−302	−193	−495	−764	−536	−581	−539	**−476**	−544	−592
Level 3	−356	−505	−841	−1480	−973	−908	−875	**−764**	−835	−901
Level 4	−762	−2204	−2966	−7337	−3284	−2967	−2845	−3338	**−2654**	−2789
Level 5	−1165	−3886	−5051	−13,466	−5285	−4897	−4772	−6034	**−4546**	−4598
Level 6	−3162	−25,252	−28,414	−85,730	−29,468	−27,586	−27,311	−40,365	−27,908	**−27,000**

Results

- A context-free grammar always has the highest posterior probability on the largest type-based corpus, compared to plausible linear grammars.
- Though the ability of the hierarchical phrase-structure grammars to generate a higher variety of sentences from fewer productions typically results in a lower likelihood, this compression helps in the prior.
- This type of grammar thus consistently maximizes the tradeoff between data fit and complexity.

Results

Sentence tokens vs. sentence types

- Evaluate a grammar based on how well they account for which sentences occur, rather than their frequency distribution.
- Linear grammars were preferred to context-free grammars.
- The context-free grammars overgeneralize – the data has more tokens, but not more variety.
- This suggests that if the hierarchical phrase structure of syntax is to be inferred from observed data, the learner may need to have a disposition to evaluate grammars with respect to type-based rather than token-based data.

Results

Age-based stratification:

Corpus	Probability	FLAT	REG-N	REG-M	REG-B	1-ST	CFG-S	CFG-L
Epoch 0 (2;3)	Prior	−3968	−1915	−1349	−1166	−244	−698	−864
	Likelihood	−881	−1265	−1321	−1322	−2199	−1489	−1448
	Posterior	−4849	−3180	−2670	−2488	−2433	**−2187**	−2312
Epoch 1 (2;3–2;8)	Prior	−22,832	−3791	−1974	−1728	−257	−838	−1055
	Likelihood	−5945	−7811	−8223	−8164	−13,123	−8834	−8467
	Posterior	−28,777	−11,602	−10,197	−9892	−13,380	−9672	**−9522**
Epoch 2 (2;3–3;1)	Prior	−34,908	−4193	−2162	−1836	−257	−865	−1096
	Likelihood	−9250	−12,164	−12,815	−12,724	−20,334	−13,675	−13,099
	Posterior	−44,158	−16,357	−14,977	−14,560	−20,591	−14,540	**−14,195**
Epoch 3 (2;3–3;5)	Prior	−48,459	−4621	−2202	−1862	−257	−876	−1111
	Likelihood	−12,909	−17,153	−17,975	−17,918	−28,487	−19,232	−18,417
	Posterior	−61,368	−21,774	−20,177	−19,780	−28,744	−20,108	**−19,528**
Epoch 4 (2;3–4;2)	Prior	−59,625	−4881	−2242	−1903	−257	−876	−1111
	Likelihood	−15,945	−21,317	−22,273	−22,293	−35,284	−23,830	−22,793
	Posterior	−75,570	−26,198	−24,515	−24,196	−35,541	−24,706	**−23,904**
Epoch 5 (2;3–5;2)	Prior	−67,612	−5169	−2283	−1943	−257	−876	−1111
	Likelihood	−18,118	−24,299	−25,303	−25,368	−40,108	−27,032	−25,889
	Posterior	−85,730	−29,468	−27,586	−27,311	−40,365	−27,908	**−27,000**

Results

Generalizability:

- The percentage of the full (Level 6) corpus that can be parsed by the best grammars learned for subsets (Level 1–5) of the full corpus.

Grammar	FLAT (%)	REG-N (%)	REG-M (%)	REG-B (%)	1-ST (%)	CFG-S (%)	CFG-L (%)
% types							
Level 1	0.3	0.7	0.7	0.7	100	2.4	2.4
Level 2	1.4	3.7	5.1	5.5	100	31.5	16.4
Level 3	2.6	9.1	9.1	32.2	100	53.1	46.8
Level 4	10.9	50.7	61.2	75.2	100	87.6	82.7
Level 5	18.7	68.8	80.3	88.0	100	91.8	88.7
% tokens							
Level 1	9.9	32.6	32.6	32.6	100	40.2	40.2
Level 2	21.4	58.8	61.7	60.7	100	76.4	69.7
Level 3	25.4	72.5	70.9	79.6	100	87.8	85.8
Level 4	34.2	92.5	94.3	96.4	100	98.3	97.5
Level 5	36.9	95.9	97.6	98.5	100	99.0	98.6

Results

Generalizability:

- Do context-free grammars simply generalize more than the regular grammars, or do they generalize in the right way?

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

Results

Linguistic adequacy:

- Compare accuracy of grammars on a gold standard parsed corpus.

Automatically parsed				Hand-parsed			
Grammar	Precision	Recall	*F*-score	Grammar	Precision	Recall	*F*-score
CFG-L	51.8	94.1	66.8	CFG-L	89.6	90.4	90.0
CFG-S	50.8	92.3	65.6	CFG-S	88.3	89.1	88.7
REG-B	50.1	90.8	64.6	REG-B	85.1	85.6	85.3
REG-M	50.1	90.8	64.6	REG-M	85.1	85.6	85.3
REG-N	49.4	89.6	63.7	REG-N	84.1	84.5	84.3
RB	50.7	92.4	65.5	RB	86.9	87.3	87.1
LB	32.0	62.9	42.4	LB	33.1	33.6	33.4

Appendix A

Table A1
Sample merges for context-free and regular grammars. Identical merges for right-hand side items were also used.

CFG merge example		REG merge example	
Old	New	Old	New
$A \rightarrow B\ C$	$A \rightarrow B\ F$	$A \rightarrow b\ C$	$A \rightarrow b\ F$
$A \rightarrow B\ D$	$F \rightarrow C$	$A \rightarrow b\ D$	$F \rightarrow d$
$A \rightarrow B\ E$	$F \rightarrow D$	$A \rightarrow b\ E$	$F \rightarrow g\ E$
	$F \rightarrow E$	$C \rightarrow g\ E$	$F \rightarrow e\ D$
		$D \rightarrow d$	
		$E \rightarrow e\ D$	

Appendix B

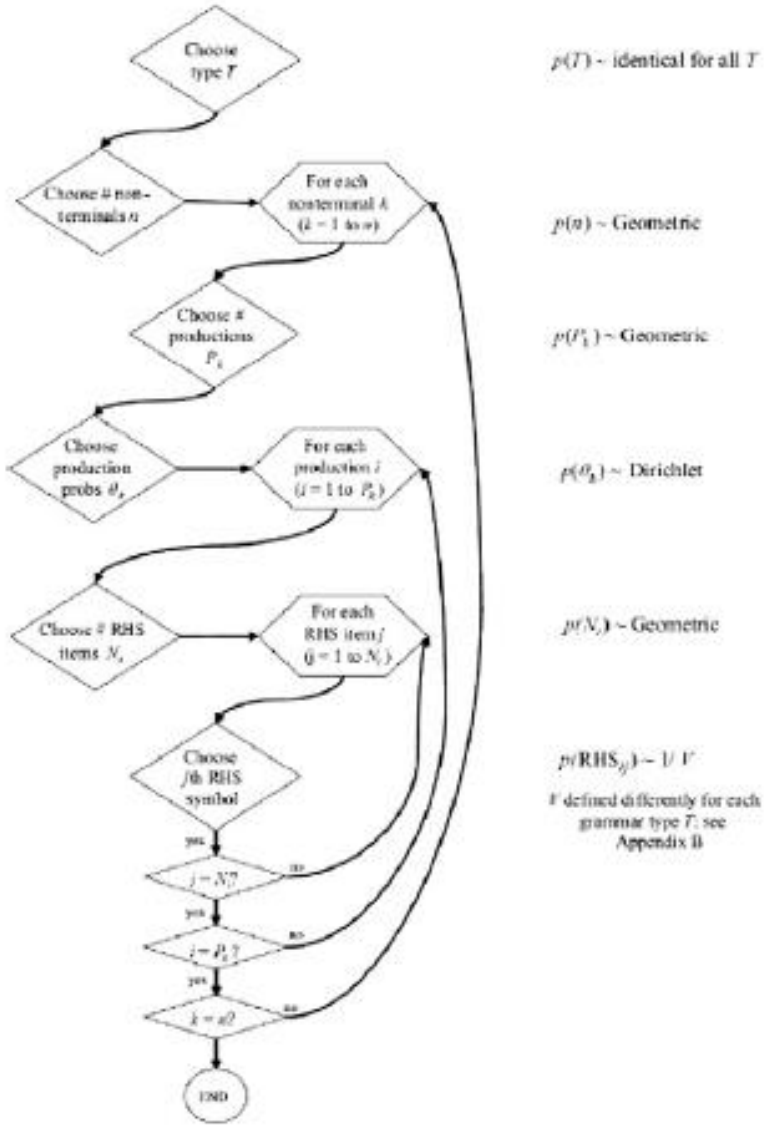


Fig. B1. Flowchart depicting the series of choices required to generate a grammar. More subtle differences between grammar types are discussed in the text.

Appendix B

$$p(G|T) = p(n)\prod_{k=1}^{n} p(P_k)p(\theta_k)\prod_{i=1}^{P_k} p(N_i)\prod_{j=1}^{N_i}\frac{1}{V}. \tag{2}$$

$$p(1-p)^{n-1} \tag{3}$$

$$p(D|G) = \prod_{i=1}^{M} p(S_i|G). \tag{4}$$

$$P(w|\theta) = \sum_{z,\ell}\left(\prod_{k=1}^{K(z)}\theta_{t_k}\right)\cdot\frac{\Gamma(K(z))}{\Gamma(N)}\cdot a^{K(z)-1} \cdot\left(\prod_{k=1}^{K(z)}\frac{\Gamma(n_k^{(z)}-a)}{\Gamma(1-a)}\right) \tag{5}$$

Appendix B

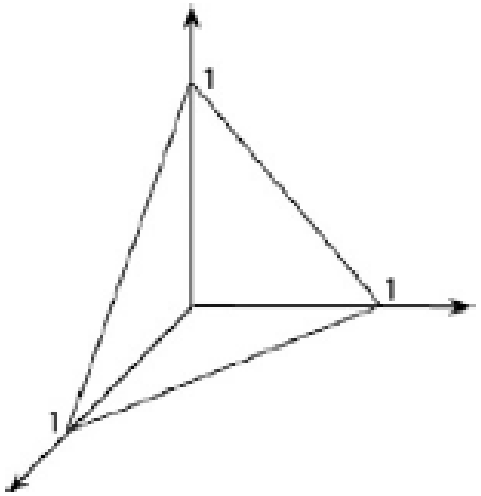


Fig. B2. The unit simplex for $m_k = 3$ (a triangle), corresponding to the Dirichlet distribution with $\alpha = 1$ on a θ_k vector of production-probability parameters with three productions.

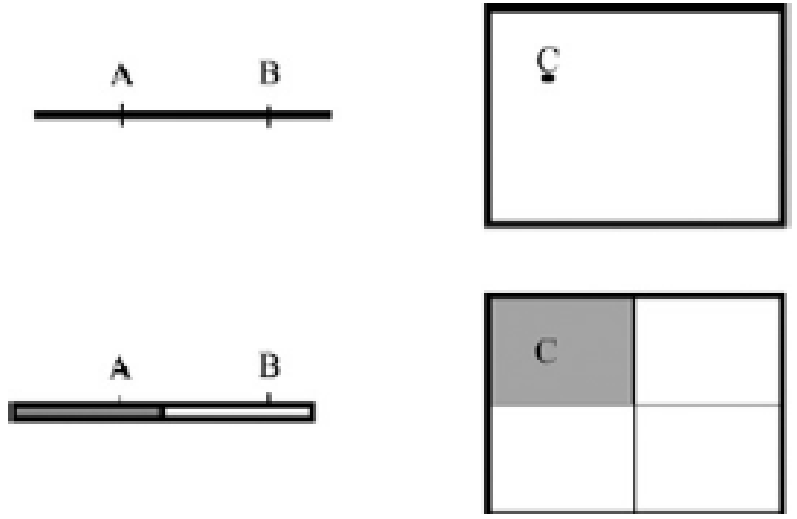


Fig. B3. Top: one cannot compare the probability of continuous points A and C with different dimensionality. Bottom: when A and C correspond to discrete points, comparison is possible.