Psych 215L: Language Acquisition

Lecture 15
Poverty of the Stimulus IV:
Structure Dependence

Reminder: Poverty of the Stimulus

The Logic of Poverty of the Stimulus (The Logical Problem of Language Acquisition)

1) Suppose there are some data.

2) Suppose there is an incorrect hypothesis compatible with the data.

3) Suppose children behave as if they never entertain the incorrect hypothesis.

   Addendum (interpretation): Or children converge on the correct hypothesis much earlier than expected (Legate & Yang 2002).

Conclusion: Children possess innate knowledge ruling out the incorrect hypothesis from the hypothesis space considered.

   Addendum (Interpretation): The initial hypothesis space does not include all hypotheses. Specifically, the incorrect ones of a particular kind are not in the child's hypothesis space.

Legate & Yang (2002): Poverty of the Stimulus Lives

Child Input

Very frequent
Is Hoggle t_{is} running away from Jareth?

Very infrequent, if ever
Can someone who can solve the Labyrinth t_{can} show someone who can't how?

Perfors, Tenenbaum, & Regier (2006): Or does it?

Two Issues

(1) Unclear how much evidence is "enough". Forms do occur, even if they do so rarely.

(2) Previous statistical models using a distributional approach did not really engage with the notion of linguistic structure that is central to the auxiliary-fronting phenomenon.

"Many linguists and cognitive scientists tend to discount these results because they ignore a principal feature of linguistic knowledge, namely that it is based on structured symbolic representations. Secondly, connectionist networks and n-gram models tend to be difficult to understand analytically. For instance, the models used by Reali and Christiansen (2004) and Lewis and Elman (2001) measure success by whether they predict the next word in a sequence, rather than based on examination of an explicit grammar. Though the models perform above chance, it is difficult to tell why and what precisely they have learned."

Perfors, Tenenbaum, & Regier (2006): Or does it?

Important point about their Bayesian learning approach

"This is an ideal learnability analysis: our question is not whether a learner without innate language-specific biases *must* be able infer that linguistic structure is hierarchical, but rather whether it is *possible* to make that inference. It thus addresses the exact challenge posed by the PoS argument, which holds that such an inference is not possible."

Note: It might be worth modifying this to "possible by a child with limited processing and memory capabilities". (Difference between computational and algorithmic approaches to language acquisition modeling.)

Perfors, Tenenbaum, & Regier (2006): Or does it?

Another important point

"PoS arguments are sensible only when phenomena are considered as part of a linguistic system, rather than taken in isolation"

Worth noting if children can make use of indirect (and ambiguous evidence), which they seem able to. It's not necessarily enough to show that unambiguous data are sparse.

Perfors, Tenenbaum, & Regier (2010 Manuscript): Or does it?

A note about innateness vs. domain-specificity

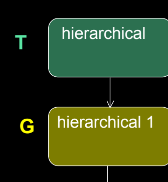
Figure 1: A schematic view of the theoretical landscape for language acquisition in cognitive science. The vertical axis reflects the nature of the representation. The horizontal axis reflects the source of inductive bias: "innate" and "learned" are in parentheses because they are often conflated with "language-specific" and "domain-general", which we suggest is closer to the real issue. The two most prominent approaches are represented by the two opposite shaded quadrants. We explore a different part of the landscape, represented by the shaded oval: assuming that mental representations of language are based on structured symbolic grammars (the upper half plane of the picture), we attempt to assess whether their form could be inferred based on domain-general learning mechanisms (the upper-right quadrant) or instead must be constrained by language-specific innate knowledge (the upper-left quadrant).

Perfors, Tenenbaum, & Regier (2006): Or does it?

Bayesian Model Selection

First, pick a type of grammar T (ex: linear, regular, hierarchical).

Then, pick an instance of T, G, from which the data D are generated.

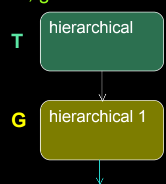


"Is the dwarf who is being teased grumpy?"

Perfors, Tenenbaum, & Regier (2006): Or does it?

Posterior probability of G and T, given D

$p(G,T|D) \propto p(D|G,T)p(G|T)p(T)$

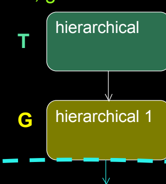


"Is the dwarf who is being teased grumpy?"

Perfors, Tenenbaum, & Regier (2006): Or does it?

Posterior probability of G and T, given D

$p(G,T|D) \propto p(D|G,T)p(G|T)p(T)$



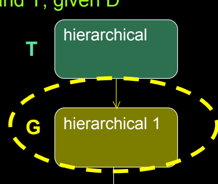
"Is the dwarf who is being teased grumpy?"

is proportional to the probability of generating the data from G and T [p(D | G,T)]

Perfors, Tenenbaum, & Regier (2006): Or does it?

Posterior probability of G and T, given D

$p(G,T|D) \propto p(D|G,T)p(G|T)p(T)$



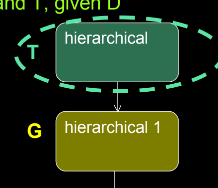
"Is the dwarf who is being teased grumpy?"

is proportional to the probability of generating the data from G and T [p(D | G,T)], multiplied by the probability of picking G from all grammars in T [p(G | T)],

Perfors, Tenenbaum, & Regier (2006): Or does it?

Posterior probability of G and T, given D

$p(G,T|D) \propto p(D|G,T)p(G|T)p(T)$



"Is the dwarf who is being teased grumpy?"

is proportional to the probability of generating the data from G and T [p(D | G,T)], multiplied by the probability of picking G from all grammars in T [p(G | T)], multiplied by the prior probability of picking T period [p(T)].

Perfors, Tenenbaum, & Regier (2006): Or does it?

The Corpus, slightly simplified

Adam corpus (American English), each word (mostly) replaced with its syntactic category:

determiners (det) [ex: *the, a, an*]
nouns (n) [ex: *cat, penguin, dream*]
adjectives (adj) [ex: *adorable, stinky*]
comments (c) [ex: *mmhm*]
prepositions (prep) [ex: *to, from, of*]
pronouns (pro) [ex: *he, she, it, one*]
proper nouns (prop) [ex: *Jareth, Sarah, Hoggle*]
infinitives (to) [ex: *to* in *I want to go*]
participles (part) [ex: She would have *gone*, I'm *going*]
infinitive verbs (vinf) [ex: I want to *go*]
conjugated verbs (v) [ex: he *went*]
auxiliary verbs (aux) [ex: he *can* go]
complementizers (comp) [ex: I thought *that* I should go.]
wh-question words (wh) [ex: *what* are you doing]

Adverbs (ex: *too, very*) and negations (ex: *not*) were removed from all sentences.

Perfors, Tenenbaum, & Regier (2006): Or does it?

The Corpus, slightly simplified

Ungrammatical and the most complex grammatical sentences were also removed: (available at http://www.psychology.adelaide.edu.au/personalpages/staff/amyperfors/research/cognitionpos/index.html)

topicalized sentences
ex: "Here he is."

(some) sentences with subordinate clauses
ex: "if you want to."

(some) sentential complements
ex: "He thought that she ought to watch the movie."

conjunctions (ex: *and, or, but*)

serial verb constructions
ex: "You should go play outside."

Perfors, Tenenbaum, & Regier (2006): Or does it?

Test corpora

Separate by frequency (idea: less complex sentences occur more frequently)

Level 1 (500+ times) = 8 unique types
Level 2 (300+ times) = 13 types
Level 3 (100+ times) = 37 types
Level 4 (50+ times) = 67 types
Level 5 (10+ times) = 268 types
Level 6 (complete corpus) = 2338 unique types, including interrogatives, wh-questions, relative clauses, prepositional and adverbial phrases, command forms, and auxiliary as well as non-auxiliary verbs.

Perfors, Tenenbaum, & Regier (2006): Or does it?

The grammars

Structure-dependent, hierarchical grammar: represented with context-free phrase structure rules

14 terminals, 14 non-terminals, 69 productions

Context-free grammar
NP → NP PP | NP CP | NP C | N | det N | adj N
pro | prop
N → n | adj N

Structure-independent grammar 1 = flat grammar: represented as simply a list of the sentences in the corpus
(2338 rules of the form Sentence → "det n")

Structure-independent (?) grammar 2 = regular grammar: represented with regular rules of the form A → a or A → aB

14 terminals, 85 non-terminals, 390 productions

Regular grammar	
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

Perfors, Tenenbaum, & Regier (2006): Or does it?

Priors for the grammars

Probability of grammar, given all other grammars of that type:

$$p(G|T) = p(P)p(n)\prod_{i=1}^{P} p(N_i)\prod_{j=1}^{N_i}\frac{1}{V}$$

p(P) = probability of P productions
p(n) = probability of n nonterminals
$p(N_i)$ = probability of non-terminal symbol N_i for production under consideration
V = vocabulary items used in production under consideration

Perfors, Tenenbaum, & Regier (2006): Or does it?

Likelihoods for the grammars

Two component model of Goldwater et al. (2005)
(1) Assign probability distribution over syntactic forms accepted in the language
(2) Generate finite observed corpus from that probability distribution (use power-law generation, so a few syntactic types are very frequent while most are infrequent)

Focus on first part (assignment of probability distribution) since concerned with the acceptability of sentence types (syntactic forms).

Perfors, Tenenbaum, & Regier (2006): Or does it?

Likelihoods for the grammars

(Log) likelihood of the data D, given the grammar G and grammar type T:

$$\log(p(D|G,T)) = \sum_{i=1}^{k}\log(p(S_i|G,T))$$

Assuming k unique sentence types observed.
The likelihood of generating sentence S_i with that syntactic form
$p(S_i \mid G, T)$

is the sum of all the probabilities of all the parses (rules & production combinations) that lead to that observed sentence as output, given that grammar. The probability of any specific parse is the product of all the productions used to derive that output form.

Perfors, Tenenbaum, & Regier (2006): Or does it?

Likelihoods for the grammars

(Log) likelihood of the data D, given the grammar G and grammar type T:

$$\log(p(D|G,T)) = \sum_{i=1}^{k}\log(p(S_i|G,T))$$

$p(S_i \mid G, T)$ for S_i = "That's an idea for him" = "pro aux det n prep pro"

Grammar G under consideration:
(1) Sentence → NP VP

(.5) NP → pro
(.2) NP → NP PP
(.3) NP → det n

(.3) VP → aux NP PP
(.7) VP → aux NP

(1) PP → prep NP

Production 1:
Sentence → NP VP → pro VP → pro aux NP
→ pro aux NP PP → pro aux det n PP
→ pro aux det n prep NP
→ pro aux det n prep pro

Parse 1:
$(_S\ (_{NP}\ \text{pro})\ (_{VP}\ \text{aux}\ (_{NP}\ (_{NP}\ \text{det n})\ (_{PP}\ \text{prep}\ (_{NP}\ \text{pro})))))$

Prob parse 1: 1 * .5 * .7 *.2 * .3 * 1 * .5 = .0105

Perfors, Tenenbaum, & Regier (2006): Or does it?

Likelihoods for the grammars

(Log) likelihood of the data D, given the grammar G and grammar type T:

$$\log(p(D|G,T)) = \sum_{i=1}^{k} \log(p(S_i|G,T))$$

$p(S_i$ | G, T) for S_i = “That’s an idea for him” = “pro aux det n prep pro”

Grammar G under consideration:

(1) Sentence → NP VP

(.5) NP → pro
(.2) NP → NP PP
(.3) NP → det n

(.3) VP → aux NP PP
(.7) VP → aux NP

(1) PP → prep NP

Production 2:
Sentence → NP VP → pro VP → pro aux NP PP
→ pro aux det n PP
→ pro aux det n prep NP
→ pro aux det n prep pro

Parse 2:
$(_S\ (_{NP}$ pro) $(_{VP}$ aux $(_{NP}\ (_{NP}$ det n)) $(_{PP}$ prep $(_{NP}$ pro))))

Prob parse 2: 1 * .5 * .3 * .3 * 1 * .5 = .0225

Perfors, Tenenbaum, & Regier (2006): Or does it?

Likelihoods for the grammars

(Log) likelihood of the data D, given the grammar G and grammar type T:

$$\log(p(D|G,T)) = \sum_{i=1}^{k} \log(p(S_i|G,T))$$

$p(S_i$ | G, T) for S_i = “That’s an idea for him” = “pro aux det n prep pro”

Grammar G under consideration:

(1) Sentence → NP VP

(.5) NP → pro
(.2) NP → NP PP
(.3) NP → det n

(.3) VP → aux NP PP
(.7) VP → aux NP

(1) PP → prep NP

Prob parse 1: 1 * .5 * .7 *.2 * .3 * 1 * .5 = .0105
Prob parse 2: 1 * .5 * .3 * .3 * 1 * .5 = .0225

$p(S_i$ | G, T) = .0105 + .0225 = .0330

Perfors, Tenenbaum, & Regier (2006): Or does it?

Likelihoods for the grammars

(Log) likelihood of the data D, given the grammar G and grammar type T:

$$\log(p(D|G,T)) = \sum_{i=1}^{k} \log(p(S_i|G,T))$$

$p(S_i$ | G, T) for S_i = “That’s an idea for him” = “pro aux det n prep pro”

Grammar G under consideration:

(1) Sentence → NP VP

(.3) NP → pro
(.3) NP → NP PP
(.3) NP → det n

(.5) VP → aux NP PP
(.5) VP → aux NP

(1) PP → prep NP

Simplification: “all productions with the same left-hand side have the same probability, in order to avoid giving grammars with more productions more free parameters to adjust in fitting the data.”

Perfors, Tenenbaum, & Regier (2006): Or does it?

Priors, likelihoods, and posteriors
(negative log probability = smaller numbers are better)

	Prior			Likelihood			Posterior		
Corpus	Flat	PRG	PCFG	Flat	PRG	PCFG	Flat	PRG	PCFG
Level 1	**-68**	-116	-133	**-17**	-19	-29	**-85**	-135	-162
Level 2	**-112**	-165	-180	**-33**	-36	-56	**-145**	-201	-236
Level 3	-405	-394	**-313**	**-134**	-179	-243	**-539**	-573	-556
Level 4	-783	-560	**-384**	**-281**	-398	-522	-1064	-958	**-906**
Level 5	-4087	-1343	**-541**	**-1499**	-2379	-2891	-5586	-3722	**-3432**
Level 6	-51505	-5097	**-681**	**-18128**	-36392	-38421	-69633	-41489	**-39102**

Perfors, Tenenbaum, & Regier (2006): Or does it?

Priors, likelihoods, and posteriors
(negative log probability = smaller numbers are better)

	Prior			Likelihood			Posterior		
Corpus	Flat	PRG	PCFG	Flat	PRG	PCFG	Flat	PRG	PCFG
Level 1	**-68**	-116	-133	**-17**	-19	-29	**-85**	-135	-162
Level 2	**-112**	-165	-180	**-33**	-36	-56	**-145**	-201	-236
Level 3	-405	-394	**-313**	**-134**	-179	-243	**-539**	-573	-556
Level 4	-783	-560	**-384**	**-281**	-398	-522	-1064	-958	**-906**
Level 5	-4087	-1343	**-541**	**-1499**	-2379	-2891	-5586	-3722	**-3432**
Level 6	-51505	-5097	**-681**	**-18128**	-36392	-38421	-69633	-41489	**-39102**

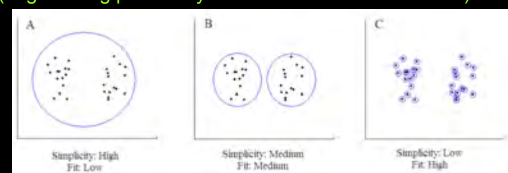
Flat grammar is simpler/more compact when the sentences are simpler

Flat grammar always has a better fit.

Combined, flat grammar is only better when the sentences are simpler

Perfors, Tenenbaum, & Regier (2006): Or does it?

Priors, likelihoods, and posteriors
(negative log probability = smaller numbers are better)



Flat grammar is most like generalization C – it has excellent fit and is extremely simple, at the cost of not being able to generalize well to new data points. It overfits.

The hierarchical grammar is more like generalization B – not a perfect fit, and not perfectly simple, but better at generalizing.

Perfors, Tenenbaum, & Regier (2006): Or does it?

Generalizability of hierarchical grammars is better

	% types			% tokens		
Grammar	Flat	RG	CFG	Flat	RG	CFG
Level 1	0.3%	0.7%	2.4%	9.8%	31%	40%
Level 2	0.5%	0.8%	4.3%	13%	38%	47%
Level 3	1.4%	4.5%	13%	20%	62%	76%
Level 4	2.6%	13%	32%	25%	74%	88%
Level 5	11%	53%	87%	34%	93%	98%

Table 3: Proportion of sentences in the full corpus that are parsed by smaller grammars of each type. The *Level 1* grammar is the smallest grammar of that type that can parse the *Level 1* corpus. All *Level 6* grammars parse the full corpus.

Flat grammar generalized poorly, especially by types

Hierarchical grammars generalize well much earlier on.

Perfors, Tenenbaum, & Regier (2006): Or does it?

Specific generalizability: Aux-inversion in complex yes/no questions – only hierarchical grammar has productions allowing it to parse/generate this structure

				Can parse?		
Type	Subject NP	in input?	Example	Flat	RG	CFG
Decl	Simple	Y	He is happy. (pro aux adj)	Y	Y	Y
Int	Simple	Y	Is he happy? (aux pro adj)	Y	Y	Y
Decl	Complex	Y	The boy who is reading is happy. (det n comp aux part aux adj)	Y	Y	Y
Int	Complex	N	Is the boy who is reading happy? (aux det n comp aux part adj	N	N	Y

Table 4: Ability of each grammar to parse specific sentences. Only the PCFG can parse the complex interrogative sentence.

Question: Does it have productions allowing it to parse the mistaken formation – "Is the boy who reading is happy?"

No → see Perfors, Tenenbaum, & Regier (2010 manuscript) for details

Perfors, Tenenbaum, & Regier (2006): Or does it?

Implications about useful data

"Our findings also make a general point that has sometimes been overlooked in considering stimulus poverty arguments, namely that children learn grammatical rules as a part of a *system* of knowledge. As with auxiliary fronting, most PoS arguments consider some isolated linguistic phenomenon and conclude that because there is not enough evidence for that phenomenon in isolation, it must be innate. We have shown here that while there might not be direct evidence for an individual phenomenon, there may be enough evidence about the system of which it is a part to explain the phenomenon itself."

Perfors, Tenenbaum, & Regier (2006): Or does it?

Important point

"Are we trying to argue that the knowledge that language is structure-dependent *is not* innate? No. All we have shown is that, contra the PoS argument, structure dependence need not be a part of innate linguistic knowledge. It is true that the ability to represent PCFGs is "given" to our model, but this is a relatively weak form of innateness: few would argue that children are born without the capacity to represent the thoughts they later grow to have, since if they were no learning would occur. Furthermore, everything that is built into the model – the capacity to represent each grammar as well as the details of the Bayesian inference procedure – is domain- general, not language-specific as the original PoS claim suggests."

More specifically: Bias for structure dependence need not be there a priori

Perfors, Tenenbaum, & Regier (2010 Manuscript): Or does it?

Another point about Bayesian learner's ability to learn more abstract knowledge before more specific knowledge – useful to think about since domain-specific knowledge is often described as abstract knowledge acquired very early

"While there are infinitely many possible specific grammars *G*, there are only a small number of possible grammar types *T*. It may thus require less evidence to identify the correct *T* than to identify the correct *G*. More deeply, because the higher level of *T* affects the grammar of the language as a whole while any component of *G* affects only a small subset of the language produced, there is in a sense much more data available about *T* than there is about any particular component of *G*…every sentence offers at least some evidence about the grammar type *T* – about whether language has hierarchical or linear phrase structure – based on whether rules generated from a hierarchical or linear grammar tend to provide a better account of that sentence. Higher-order generalizations may thus be learned faster simply because there is much more evidence relevant to them."