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 running away from Jareth?

Very infrequent, if ever
Can someone who can solve the Labyrinth 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 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

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.

$T$ hierarchical

$G$ hierarchical 1

“Is the dwarf who is being teased grumpy?”
Perfors, Tenenbaum, & Regier (2006): Or does it?

Posterior probability of G and T, given D

\[ p(G, T | D) \propto p(D | G, T) p(G | T) \]

"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)].
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]
- prepositions (prep) [ex: to, from, of]
- proper nouns (pro) [ex: Jareth, Sarah, Hoggle]
- infinitives (to) [ex: to in / want to go]
- participles (part) [ex: She would have gone, I'm going]
- 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.

The Corpus, slightly simplified

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/cognitiontop/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.”

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.

The grammars

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

14 terminals, 14 non-terminals, 69 productions

Structure-independent grammar 1 = flat grammar: represented as simply a list of the sentences in the corpus (2338 rules of the form Sentence $\rightarrow$ “det n”)

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

14 terminals, 85 non-terminals, 390 productions
Perfors, Tenenbaum, & Regier (2006): Or does it?

Priors for the grammars
Probability of grammar, given all other grammars of that type:

$$p(G | T) = \frac{1}{\sum_{G'} p(G')}$$

- $p(P)$ = 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

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).

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

$$p(S_i | G, T) = \log(p(S_i | G, T))$$

Grammar $G$ under consideration:

1. Sentence $\rightarrow$ NP VP
2. NP $\rightarrow$ pro
3. NP $\rightarrow$ NP PP
4. NP $\rightarrow$ NP det n
5. VP $\rightarrow$ aux NP
6. VP $\rightarrow$ aux NP PP
7. PP $\rightarrow$ prep NP

Production 1:

Sentence $\rightarrow$ NP VP $\rightarrow$ pro VP $\rightarrow$ pro aux NP
$\rightarrow$ pro aux NP PP $\rightarrow$ pro aux det n PP $\rightarrow$ pro aux det n prep NP $\rightarrow$ pro aux det n prep pro

Parse 1:

Sentence $\rightarrow$ NP pro aux (det n) (prep (prep pro))

Prob parse 1: $1 * .5 * .7 * .3 * 1 * .5 = .0105$
Perfors, Tenenbaum, & Regier (2006): Or does it?

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

$$\log p(S_i | G, T) = \sum \log p(S_i | G, T)$$

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

Grammar $G$ under consideration:

1. Sentence $\rightarrow$ NP VP
   - Production 2:
     - Sentence $\rightarrow$ NP VP $\rightarrow$ pro VP $\rightarrow$ pro aux NP PP
     - $p(S_i | G, T) = 1 \cdot .5 = .5$
     - Prob parse 2: $1 \cdot .5 \cdot .3 \cdot .3 \cdot 1 = .0225$

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?

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

$$\log p(S_i | G, T) = \sum \log p(S_i | G, T)$$

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

Grammar $G$ under consideration:

1. Sentence $\rightarrow$ NP VP
   - Prob parse 1: $1 \cdot .5 \cdot .7 \cdot 1 \cdot .3 \cdot .5 = .0105$
   - Prob parse 2: $1 \cdot .5 \cdot .3 \cdot .3 \cdot 1 \cdot .5 = .0225$
   - $p(S_i | G, T) = .0105 + .0225 = .0330$

Priors, likelihoods, and posteriors
(negative log probability = smaller numbers are better)
Perfors, Tenenbaum, & Regier (2006): Or does it?

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Prior</th>
<th>Likelihood</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>-112</td>
<td>-353</td>
<td>-353</td>
</tr>
<tr>
<td>Level 3</td>
<td>-98</td>
<td>-364</td>
<td>-364</td>
</tr>
<tr>
<td>Level 4</td>
<td>-783</td>
<td>-590</td>
<td>-590</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Prior</th>
<th>Likelihood</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>-135</td>
<td>-353</td>
<td>-353</td>
</tr>
<tr>
<td>Level 3</td>
<td>-135</td>
<td>-364</td>
<td>-364</td>
</tr>
<tr>
<td>Level 4</td>
<td>-590</td>
<td>-590</td>
<td>-590</td>
</tr>
</tbody>
</table>

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?

Generalizability of hierarchical grammars is better

<table>
<thead>
<tr>
<th>% types</th>
<th>% tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>0.3% 0.7% 2.4% 8.8% 31% 61%</td>
</tr>
<tr>
<td>Level 2</td>
<td>0.5% 0.8% 4.3% 13% 30% 47%</td>
</tr>
<tr>
<td>Level 3</td>
<td>1.4% 4.5% 13% 20% 62% 76%</td>
</tr>
<tr>
<td>Level 4</td>
<td>2.6% 17% 32% 25% 74% 87%</td>
</tr>
<tr>
<td>Level 5</td>
<td>5.5% 33% 87% 30% 93% 96%</td>
</tr>
</tbody>
</table>

Table 3: Proportion of sentences in the full corpus that are parsed by smaller grammars of each type. The first row gives the % types, and the second row gives the % tokens.

Flat grammar generalized poorly, especially by types

Hierarchical grammars generalize 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

<table>
<thead>
<tr>
<th>Type</th>
<th>Suggestion</th>
<th>Error</th>
<th>Case creator</th>
<th>Context</th>
<th>Can compete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dead</td>
<td>Simple</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Dead</td>
<td>Complex</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Not</td>
<td>Complex</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Table 4: Ability of each grammar to parse specific structures. Only the PCFG can parse the complex presupposition structures.

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
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.

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.

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."