Poverty of the Stimulus: Syntactic Islands

An induction problem by any other name...

One of the most controversial claims in linguistics is that children face an induction problem:

*Plato’s Problem* (Chomsky 1988, Dresher 2003)

Basic claim:
The data encountered are compatible with multiple hypotheses.

The data encountered are compatible with multiple hypotheses.

The data encountered are compatible with multiple hypotheses.

The data encountered are compatible with multiple hypotheses.

The data encountered are compatible with multiple hypotheses.

The data encountered are compatible with multiple hypotheses.

Extended claim:
Given this, the data are insufficient for identifying the correct hypothesis as quickly as children do (Legate & Yang 2002) – or at all.

Big question: How do children do it, then?

One answer: Children come prepared

- Children are not unbiased learners.
- But if children come equipped with helpful learning biases, then what is the nature of these necessary biases?
  - Are they innate or derived from the input somehow?
  - Are they domain-specific or domain-general?
  - Are they about the hypothesis space or about the learning mechanism?

The Universal Grammar (UG) hypothesis (Chomsky 1965, Chomsky 1973):
These biases are innate and domain-specific.
The Plan

(1) Look at syntactic islands: phenomena central to UG-based syntactic theories.

(2) Explicitly define the target knowledge state, based on adult acceptability judgments.

(3) Identify the kind of data children and adults have in their input, using realistic samples of child-directed and adult-directed input.

(4) Implement a computational learner that is able to reach the target knowledge state, given realistic data distributions, and see what kind of learning biases it requires. It turns out that none of these are necessarily innate and domain-specific, and so learning syntactic islands does not require UG-like biases.

Syntactic Islands

Dependencies can exist between two non-adjacent items, and these do not appear to be constrained by length (Chomsky 1965, Ross 1967).

What does Jack think _? What does Jack think that Lily said _? What does Jack think that Lily said that Sarah heard _? What does Jack think that Lily said that Sarah heard that Jareth stole _?

Syntactic Islands

However, if the gap position appears inside certain structures (called “syntactic islands” by Ross (1967)), the dependency seems to be ungrammatical.


Syntactic Islands

Predominant learning theory in generative syntactic theory: syntactic islands require innate, domain-specific learning biases.

Example: Subjacency

Bounding nodes: language-specific (CP, IP, and/or NP)

Learning biases:
(1) Innate, domain-specific knowledge of hypothesis space: Exclude hypotheses that allow dependencies crossing 2+ bounding nodes.

(2) Innate, domain-specific knowledge of hypothesis space: Hypothesis space consists of bounding nodes for all languages, and the child must identify the ones applicable to her language.
Sprouse et al. (2012) collected magnitude estimation judgments for four different islands:

**Complex NP islands**
*What did you make [the claim that Jack bought ___]?*

**Subject islands**
*What do you think [the joke about ___] offended Jack?*

**Whether islands**
*What do you wonder [whether Jack bought ___]?*

**Adjunct islands**
*What do you worry [if Jack buys ___]?*

---

**Complex NP islands**

Who __ claimed that Lily forgot the necklace?  short | non-island
What did the teacher claim that Lily forgot ___?  long | non-island
Who __ made the claim that Lily forgot the necklace?  short | island
*What did the teacher make the claim that Lily forgot ___?  long | island

---

**Subject islands**

Who __ thinks the necklace is expensive?  short | non-island
What does Jack think ___ is expensive?  long | non-island
Who __ thinks the necklace for Lily is expensive?  short | island
*Who does Jack think the necklace for ___ is expensive?  long | island
The target state:  
**Adult knowledge of syntactic islands**

Sprouse et al. (2012)’s factorial definition controls for two salient properties of island-crossing dependencies:
- length of dependency (short vs. long)
- presence of an island structure (non-island vs. island)

**Whether islands**

<table>
<thead>
<tr>
<th>Question</th>
<th>Short</th>
<th>Non-island</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who _ thinks that Jack stole the necklace?</td>
<td>short</td>
<td>non-island</td>
</tr>
<tr>
<td>What does the teacher think that Jack stole _?</td>
<td>long</td>
<td>non-island</td>
</tr>
<tr>
<td>Who _ wonders whether Jack stole the necklace?</td>
<td>short</td>
<td>island</td>
</tr>
<tr>
<td>*What does the teacher wonder whether Jack stole _?</td>
<td>long</td>
<td>island</td>
</tr>
</tbody>
</table>

**Adjunct islands**

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who _ thinks that Lily forgot the necklace?</td>
</tr>
<tr>
<td>What does the teacher think that Lily forgot _?</td>
</tr>
<tr>
<td>Who _ worries if Lily forgot the necklace?</td>
</tr>
<tr>
<td>*What does the teacher worry if Lily forgot _?</td>
</tr>
</tbody>
</table>

Superadditivity is visually salient

Superadditivity present for all islands tested

Knowledge that dependencies cannot cross these island structures is part of the adult knowledge state
The input: Induction problems

Data from three corpora of child-directed speech (Brown-Adam, Brown-Eve, Valian) from CHILDES (MacWhinney 2000): speech to 23 children between the ages of one and four years old.
Total words: 340,913
Utterances containing a wh-word and a verb: 14,260

Sprouse et al. (2012) stimuli types:

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>Long</th>
<th>Short</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex NP</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Subject</td>
<td>4</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Whose</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adjusted</td>
<td>4</td>
<td>177</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Total words: 340,913
Utterances containing a wh-word and a verb: 14,260

Sprouse et al. (2012) stimuli types:

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>Long</th>
<th>Short</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex NP</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Subject</td>
<td>4</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Whose</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adjusted</td>
<td>4</td>
<td>177</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

These kinds of utterances are fairly rare in general - the most frequent appears less than 0.01% of the time (177 of 14,260.)

Sprouse et al. (2012) stimuli types:

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>Long</th>
<th>Short</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex NP</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Subject</td>
<td>4</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Whose</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adjusted</td>
<td>4</td>
<td>177</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Being grammatical doesn’t necessarily mean an utterance will appear in the input at all.

Sprouse et al. (2012) stimuli types:

<table>
<thead>
<tr>
<th></th>
<th>Short</th>
<th>Long</th>
<th>Short</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex NP</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Subject</td>
<td>4</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Whose</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adjusted</td>
<td>4</td>
<td>177</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Unless the child is sensitive to very small frequencies, it’s difficult to tell the difference between grammatical and ungrammatical dependencies sometimes...
The input: Induction problems

Data from three corpora of child-directed speech (Brown-Adam, Brown-Eve, Valian) from CHILDES (MacWhinney 2000): speech to 23 children between the ages of one and four years old.
Total words: 3,409,133
Utterances containing a wh-word and a verb: 14,260

Sprouse et al. (2012) stimuli types:

<table>
<thead>
<tr>
<th>Complex NP</th>
<th>4</th>
<th>177</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setback</td>
<td>4</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Whether</td>
<td>4</td>
<td>177</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adjust</td>
<td>4</td>
<td>177</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

...and impossible to tell no matter what the rest of the time. This looks like an induction problem for the language learner.

Building a computational learner: Proposed learning biases

Learning Bias: Children track the occurrence of structures that can be derived from phrase structure trees - container nodes.

How to do this:
- Identifying container nodes
  - applies to language data: domain-specific
  - requires child to represent the hypothesis space a certain way
  - derived from ability to parse utterances
- Parsing utterances
  - requires chunking data into cohesive units: likely to be innate and domain-general
  - units being chunked are phrasal units: derived from distributional data and domain-specific

Building a computational learner: Proposed learning biases

Learning Bias: Children’s hypotheses are about what container node sequences are grammatical for dependencies in the language.
Learning Bias: Children’s hypotheses are about what container node sequences are grammatical for dependencies in the language.

How to do this:
- Identifying container nodes - applies to language data: domain-specific
  - requires child to represent the hypothesis space a certain way
  - derived from ability to parse utterances

Parsing utterances
- requires chunking data into cohesive units: likely to be innate and domain-general
  - units being chunked are phrasal units: derived from distributional data and domain-specific

What this does:
- longer dependencies are less probable than shorter dependencies, all other things being equal
- individual trigram frequency matters: short dependencies made of infrequent trigrams will be less probable than longer dependencies made of frequent trigrams

Effect: the frequencies observed in the input temper the detrimental effect of dependency length.
Building a computational learner: Proposed learning biases

Learning Bias: Implicitly assign a probability to a container node sequence by tracking trigrams of container nodes. A sequence’s probability is the smoothed product of its trigrams.

How to do this:
- have enough memory to hold the utterance and its dependency in mind: innate and domain-general
- track trigrams of units: innate, domain-general, learning mechanism

Building a computational learner:
Empirical grounding

Child-directed speech (Brown-Adam, Brown-Eve, Valian) from CHILDES: If we want to model child learners.

Adult-directed speech (Treebank-3-Switchboard corpus: Marcus et al. 1999) and text (Treebank-3-Brown corpus: Marcus et al. 1999): If we want to model adult learners, since we have adult data.

Note: Child-directed speech and adult-directed speech are qualitatively similar in being mostly IP-VP and IP dependencies, with many more IP-VP dependencies (child: 80% IP-VP/11% IP, adult: 73% IP-VP/17% IP). Adult-directed text is still mostly IP-VP and IP dependencies, but there are more IP dependencies compared to the speech samples (63% IP-VP/23% IP).
Building a computational learner: Empirical grounding

Hart & Risley (1995): Children hear approximately 1 million utterances in their first three years. Assumption: learning period for modeled learners is 3 years (e.g. between 2 and 5 years old for modeling children’s acquisition)

Estimating proportion of wh-dependencies in the input, based on child-directed speech sample: total learning period is 175,000 wh-dependency data points, drawn from distribution observed in speech and/or text samples.

Success metrics

Compare learned grammaticality preferences to Sprouse et al. (2012) judgment data.

To do this, we need to identify the container node sequences for each stimuli for each island type.

Complex NP islands

<table>
<thead>
<tr>
<th></th>
<th>short</th>
<th>long</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>non-island</td>
<td></td>
</tr>
<tr>
<td>IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP-VP</td>
<td>non-island</td>
<td></td>
</tr>
<tr>
<td>IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP-VP</td>
<td>island</td>
<td></td>
</tr>
<tr>
<td>*IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP-NP-PP</td>
<td>island</td>
<td></td>
</tr>
</tbody>
</table>

Subject islands

<table>
<thead>
<tr>
<th></th>
<th>short</th>
<th>long</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>non-island</td>
<td></td>
</tr>
<tr>
<td>IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP</td>
<td>non-island</td>
<td></td>
</tr>
<tr>
<td>IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP</td>
<td>island</td>
<td></td>
</tr>
<tr>
<td>*IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP-NP-PP</td>
<td>island</td>
<td></td>
</tr>
</tbody>
</table>

Whether islands

<table>
<thead>
<tr>
<th></th>
<th>short</th>
<th>long</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>non-island</td>
<td></td>
</tr>
<tr>
<td>IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP-VP</td>
<td>non-island</td>
<td></td>
</tr>
<tr>
<td>IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP-VP</td>
<td>island</td>
<td></td>
</tr>
<tr>
<td>*IP-VP-CP/CP&lt;sub&gt;opt&lt;/sub&gt;-IP-NP-PP</td>
<td>island</td>
<td></td>
</tr>
</tbody>
</table>
Success metrics

Compare learned grammaticality preferences to Sprouse et al. (2012) judgment data.

To do this, we need to identify the container node sequences for each stimuli for each island type.

Adjunct islands

| IP | short | non-island |
| IP-VP-CP/CP_{null}-IP-VP | long | non-island |

Then, for each island, we plot the predicted grammaticality preferences from the modeled learner on an interaction plot, using log probability of the dependency on the y-axis. Non-parallel lines indicate the presence of islands.

The non-UG learner

Using basic-level container nodes
- ex: only CP rather than CP_{null} CP_{null} etc.

Child-directed speech input

Complex NP and Subject islands have the correct superadditive behavior...
The non-UG learner
Using basic-level container nodes
- ex: only CP rather than $\text{CP}_{\text{null}}$, $\text{CP}_{\text{that}}$, etc.

Child-directed speech input

But Whether and Adjunct islands don't. In fact, the lines are overlapping - the learner thinks the grammatical long | non-island stimuli and ungrammatical long | island stimuli are equally good.

The same is true for adult-directed input: the learner has the correct preferences for Complex NP islands and Subject islands, but has the incorrect preferences for Whether and Adjunct islands.

The non-UG learner
Using basic-level container nodes
- ex: only CP rather than $\text{CP}_{\text{null}}$, $\text{CP}_{\text{that}}$, etc.

Adult-directed speech & text input

The learner does not distinguish between grammatical structures with the sequence IP-VP-$\text{CP}_{\text{null/that}}$ IP-VP

What did he think (that) she saw?

and structures with the ungrammatical sequence IP-VP-$\text{CP}_{\text{whether/if}}$ IP-VP

* What did he wonder whether/if she saw?

This means that Whether and Adjunct island violations, which contain specific types of CPs ($\text{CP}_{\text{whether}}$ and $\text{CP}_{\text{if}}$), are treated identically to grammatical utterances containing $\text{CP}_{\text{null}}$ or $\text{CP}_{\text{that}}$.
The non-UG learner
Using finer-grained container nodes: include CP specification
- ex: use CP_sub, CP_sub, etc.

Child-directed speech input

The same is true for the learner using adult-directed input: all four island plots show superadditivity for the ungrammatical island dependency.

The non-UG learner
Using finer-grained container nodes: include CP specification
- ex: use CP_sub, CP_sub, etc.

Problem solved! Superadditivity observed for all four island types.

Adult-directed speech & text input

Implications of this learner
Basic: A learner using no biases that would traditionally be considered part of UG (i.e., innate and domain-specific biases) was able to learn the correct grammaticality preferences for dependencies over four different island types. This suggests that adult knowledge of these syntactic islands does not implicate UG.

Though there appears to be an induction problem, it does not require UG to solve it.
Implications of this learner
Something useful for children to have: Complex learning biases that are made up of simpler biases. (So, perhaps a bias to combine existing biases.)

Ex: Tracking trigrams of container nodes
- basic unit is container node \(\text{derived, domain-specific, hypothesis space}\)
- tracking 3 unit sequences \(\text{intra, domain-general, learning mechanism}\)

What about the CP specification requirement? Is that UG?
Not necessarily:
- uncontroversial to assume that children learn to distinguish different types of CPs since the lexical content of CPs has substantial consequences for the semantics of a sentence (e.g., declaratives versus interrogatives)
- adult speakers are sensitive to the distribution of that versus null complementizers (Jaeger 2010)

Likely a \(\text{derived, domain-specific}\) learning bias about the representation of the hypothesis space.

A remaining issue
This learner can’t handle \text{parasitic gaps}\, which are dependencies that span an island (and so should be ungrammatical) but which are somehow rescued by another dependency in the utterance.

*Which book did you laugh \(\text{[before reading …]}\)?
  Which book did you judge \(\text{—-}\mathbf{\text{parasitic}}\) \(\text{[before reading —-parasitic]}\) ?

\(\text{Adjunct island}\)

*What did \(\text{the attempt to repair —-parasitic}\) ultimately damage the car?
  What did \(\text{the attempt to repair —-parasitic}\) ultimately damage \(\text{—-}\mathbf{\text{parasitic}}\) ?

\(\text{Complex NP island}\)

Why not? The current learner would judge the parasitic gap as \text{ungrammatical}\, since it is inside an island, irrespective of what other dependencies are in the utterance.

*Which book did you laugh \(\text{[before reading …]}\)?
  Which book did you judge \(\text{—-}\mathbf{\text{parasitic}}\) \(\text{[before reading —-parasitic]}\) ?

\(\text{Adjunct island}\)

*What did \(\text{the attempt to repair —-parasitic}\) ultimately damage the car?
  What did \(\text{the attempt to repair —-parasitic}\) ultimately damage \(\text{—-}\mathbf{\text{parasitic}}\) ?

\(\text{Complex NP island}\)

This may be able to be addressed in a learner that is able to combine information from multiple dependencies in an utterance (perhaps because the learner has observed multiple dependencies resolved in utterances in the input).
A developmental prediction

If children begin with only a basic specific of container nodes (CP instead of CP_{1,2}), we may expect a period of time when they recognize Complex NP and Subject islands but view dependencies spanning Whether and Adjunct islands as grammatical. Once they allow CP specification, they will recognize Whether and Adjunct islands as well.

Stage 1
* Complex NP island
* Subject island
* Whether island
* Adjunct island

Stage 2
* Complex NP island
* Subject island
* Whether island
* Adjunct island

de Villiers & Roeper (1995) suggest that children as young as 3 years old may view dependencies spanning wh-islands (such as whether islands) as ungrammatical. If they recognize whether islands as well, this suggests Stage 2 would be complete by this age.