Computational models of syntactic acquisition

Lisa Pearl
University of California, Irvine

August 4, 2017:
Norwegian Summer Institute on Language & Mind
University of Oslo
Today’s Plan:
Computational models of syntactic acquisition

I. Some non-parametric examples

II. About linguistic parameters

III. Learning with parameters
Today’s Plan:
Computational models of syntactic acquisition

I. Some non-parametric examples

- Who does... Is pretty?
- another one

syntax, semantics
Some non-parametric examples

This kitty was bought as a present for someone.

Lily thinks this kitty is pretty.

What’s going on here?

Who does Lily think the kitty for is pretty? 😞

What does Lily think is pretty, and who does she think it’s for? 😊
Some non-parametric examples

**syntax**

Who does Lily think the kitty for is pretty?

*What’s going on here?*

There’s a **dependency** between the *wh*-word *who* and where it’s understood (the gap)

![Diagram](image)

This dependency is **not allowed** in English.

*One explanation:* The dependency crosses a “syntactic island” (Ross 1967)
Some non-parametric examples

What’s going on here? syntactic island (Ross 1967)

Who does Lily think the kitty for ___ is pretty?

Jack is somewhat tricksy.

He claimed he bought something.

What did Jack make the claim that he bought ___?
Some non-parametric examples

What’s going on here? X syntactic island (Ross 1967)

Who does Lily think the kitty for ____ is pretty?
What did Jack make the claim that he bought ____?

Jack is somewhat tricksy.
He claimed he bought something.
Elizabeth wondered if he actually did and what it was.

What did Elizabeth wonder whether Jack bought ____?
Some non-parametric examples

What’s going on here? syntactic island (Ross 1967)

Who does Lily think the kitty for ___ is pretty?
What did Jack make the claim that he bought___ ?
What did Elizabeth wonder whether Jack bought ___ ?

Jack is somewhat tricksy.
He claimed he bought something.
Elizabeth worried it was something dangerous.

What did Elizabeth worry if Jack bought ___ ?
Some non-parametric examples

What’s going on here? syntactic island (Ross 1967)

Who does Lily think the kitty for ____ is pretty?
What did Jack make the claim that he bought ____?
What did Elizabeth wonder whether Jack bought ____?
What did Elizabeth worry if Jack bought ____?

Jack bought something.

Elizabeth met him afterwards.

Lily asks Elizabeth about it.
Some non-parametric examples  

What’s going on here? syntactic island

Who does Lily think the kitty for ___ is pretty?
What did Jack make the claim that he bought ___ ?
What did Elizabeth wonder whether Jack bought ___ ?
What did Elizabeth worry if Jack bought ___ ?
What did you meet the pirate who bought ___ ?

Jack bought something.
Elizabeth was surprised by it.

What did that Jack bought ____ surprise you ?

Lily asks Elizabeth about it.
Some non-parametric examples

What’s going on here? syntactic island

Who does Lily think the kitty for ___ is pretty?
What did Jack make the claim that he bought ___?
What did Elizabeth wonder whether Jack bought ___?
What did Elizabeth worry if Jack bought ___?
What did you meet the pirate who bought ___?
What did that Jack bought ___ surprise you?

Jack bought two things - a kitty and something else.

Elizabeth wants to know about the other thing.
Some non-parametric examples

What’s going on here? syntactic island

Who does Lily think the kitty for ___ is pretty?
What did Jack make the claim that he bought ___ ?
What did Elizabeth wonder whether Jack bought ___ ?
What did Elizabeth worry if Jack bought ___ ?
What did you meet the pirate who bought ___ ?
What did that Jack bought ___ surprise you?
What did you buy a kitty and ___ ?

Which did you buy ___ kitty?

Jack bought a specific kind of kitty.

Elizabeth wants to know about the kind.
Some non-parametric examples

What’s going on here? syntactic island

Who does Lily think the kitty for ___ is pretty?
What did Jack make the claim that he bought ___?
What did Elizabeth wonder whether Jack bought ___?
What did Elizabeth worry if Jack bought ___?
What did you meet the pirate who bought ___?
What did that Jack bought ___ surprise you?
What did you buy a kitty and ___?
Which did you buy ___ kitty?

Important: It’s not about the length of the dependency.

(Chomsky 1965, Ross 1967)
Some non-parametric examples

What’s going on here? **≠** syntactic island

Who does Lily think the kitty for ___ is pretty?
What did Jack make the claim that he bought
What did Elizabeth wonder whether Jack bought
What did Elizabeth worry if Jack bought ___
What did you meet the pirate who bought
What did that Jack bought ___ surprise you
What did you buy a kitty and ___?
Which did you buy ___ kitty?

What did Elizabeth think ___?

✔

It’s not about the length of the dependency.
Some non-parametric examples

What’s going on here? syntactic island

Who does Lily think the kitty for is pretty?
What did Jack make the claim that he bought?
What did Elizabeth wonder whether Jack bought?
What did Elizabeth worry if Jack bought?
What did you meet the pirate who bought?
What did that Jack bought surprise you?
What did you buy a kitty and?
Which did you buy kitty?

What did Elizabeth think Jack said?

It’s not about the length of the dependency.
Some non-parametric examples

What’s going on here? × syntactic island

Who does Lily think the kitty for ___ is pretty?
What did Jack make the claim that he bought ___?
What did Elizabeth wonder whether Jack bought ___?
What did Elizabeth worry if Jack bought ___?
What did you meet the pirate who bought ___?
What did that Jack bought ___ surprise you?
What did you buy a kitty and ___?
Which did you buy ___ kitty?

What did Elizabeth think Jack said Lily saw ___?

✔

It’s not about the length of the dependency.
Some non-parametric examples

Who does Lily think the kitty for is pretty?

Adults judge these dependencies to be far worse than many others, including others that are very similar except that they don’t cross syntactic islands (Sprouse et al. 2012).
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

Sprouse et al. (2012) collected magnitude estimation judgments for four different islands, using a factorial definition that controlled for two salient properties of island-crossing dependencies:
- length of dependency (matrix vs. embedded)
- presence of an island structure (non-island vs. island)
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

**Sprouse et al. (2012)**

- **Length** of dependency
  - (matrix vs. embedded)
- Presence of an island structure
  - (non-island vs. island)

Complex NP island stimuli

Who __ claimed that Lily forgot the necklace?
What did the teacher claim that Lily forgot __?
Who __ made the claim that Lily forgot the necklace?
*What did the teacher make the claim that Lily forgot __?*
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

Sprouse et al. (2012)
- length of dependency (matrix vs. embedded)
- presence of an island structure (non-island vs. island)

Subject island stimuli

Who __ thinks the necklace is expensive?  
What does Jack think __ is expensive?  
Who __ thinks the necklace for Lily is expensive?  
*Who does Jack think the necklace for__ is expensive?  

Lidz & Gagliardi 2015
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

*Sprouse et al. (2012)*
- **length** of dependency
- presence of an **island** structure
- (non-island vs. island)

**Whether island stimuli**

Who ___ thinks that Jack stole the necklace?

What does the teacher think that Jack stole ___?

Who ___ wonders whether Jack stole the necklace?

*What does the teacher wonder whether Jack stole ___?*
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

*Sprouse et al.* (2012)

- **length** of dependency
- (matrix vs. embedded)
- presence of an island structure
- (non-island vs. island)

Adjunct island stimuli

Who __ thinks that Lily forgot the necklace?
What does the teacher think that Lily forgot __?
Who __ worries if Lily forgot the necklace?
*What does the teacher worry if Lily forgot__?*
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

Syntactic island = superadditive interaction of the two factors (additional unacceptability that arises when the two factors — length & presence of an island structure — are combined, above and beyond the independent contribution of each factor).

Lidz & Gagliardi 2015
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

Sprouse et al. (2012): acceptability judgments from 173 adult subjects

Superadditivity present for all islands tested = Knowledge that dependencies cannot cross these island structures is part of adult knowledge about syntactic islands.

Pearl & Sprouse 2013a, 2013b, 2015
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

*Sprouse et al. (2012): acceptability judgments from 173 adult subjects*

Importance for acquisition: This is one kind of target behavior that we’d like a modeled child to produce.

*Pearl & Sprouse 2013a, 2013b, 2015*
Adult judgments: Target behavior

Adult knowledge as measured by acceptability judgment behavior

*Sprouse et al. (2012)*: acceptability judgments from 173 adult subjects

So if we’re focusing on these *wh*-dependencies and that specific target state, what does children’s input look like?
Children’s input

Children’s input really doesn’t look so helpful

Data from five corpora of child-directed speech (Brown-Adam, Brown-Eve, Brown-Sarah, Suppes, Valian) from CHILDES (MacWhinney 2000): speech to 25 children between the ages of one and five years old.

= 813,036 words

= 31,247 utterances containing a wh-dependency
Children’s input

Children’s input really doesn’t look so helpful

Data from five corpora of child-directed speech = \textbf{31,247} utterances containing a \textit{wh}-dependency

<table>
<thead>
<tr>
<th>grammatical stimuli</th>
<th>syntactic island</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATRIX + NON-ISLAND</td>
<td>EMBEDDED + NON-ISLAND</td>
</tr>
<tr>
<td>Complex NP</td>
<td>7</td>
</tr>
<tr>
<td>Subject</td>
<td>7</td>
</tr>
<tr>
<td>Whether</td>
<td>7</td>
</tr>
<tr>
<td>Adjunct</td>
<td>7</td>
</tr>
</tbody>
</table>

These kinds of utterances are fairly rare in general - the most frequent appears about 0.9% of the time (295 of 31,247.)
Children’s input

Children’s input really doesn’t look so helpful

Data from five corpora of child-directed speech = 31,247 utterances containing a *wh*-dependency

<table>
<thead>
<tr>
<th></th>
<th>MATRIX + NON-ISLAND</th>
<th>EMBEDDED + NON-ISLAND</th>
<th>MATRIX + ISLAND</th>
<th>EMBEDDED + ISLAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex NP</td>
<td>7</td>
<td>295</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Subject</td>
<td>7</td>
<td>29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Whether</td>
<td>7</td>
<td>295</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Adjunct</td>
<td>7</td>
<td>295</td>
<td>15</td>
<td>0</td>
</tr>
</tbody>
</table>

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

Pearl & Sprouse 2013a, 2013b, 2015
Children’s input

Children’s input really doesn’t look so helpful

Data from five corpora of child-directed speech = \textbf{31,247} utterances containing a \textit{wh}-dependency

<table>
<thead>
<tr>
<th>grammatical stimuli</th>
<th>syntactic island</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATRIX + NON-ISLAND</td>
<td>EMBEDDED + NON-ISLAND</td>
</tr>
<tr>
<td>Complex NP</td>
<td>7</td>
</tr>
<tr>
<td>Subject</td>
<td>7</td>
</tr>
<tr>
<td>Whether</td>
<td>7</td>
</tr>
<tr>
<td>Adjunct</td>
<td>7</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...

Pearl & Sprouse 2013a, 2013b, 2015
Children’s input

Children’s input really doesn’t look so helpful

Data from five corpora of child-directed speech = 31,247 utterances containing a \textit{wh}-dependency

<table>
<thead>
<tr>
<th></th>
<th>grammatical stimuli</th>
<th>syntactic island</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MATRIX + NON-ISLAND</td>
<td>EMBEDDED + NON-ISLAND</td>
</tr>
<tr>
<td>Complex NP</td>
<td>7</td>
<td>295</td>
</tr>
<tr>
<td>Subject</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>Whether</td>
<td>7</td>
<td>295</td>
</tr>
<tr>
<td>Adjunct</td>
<td>7</td>
<td>295</td>
</tr>
</tbody>
</table>

...and impossible to tell no matter what the rest of the time. This looks like an \textbf{induction problem} for the language learner if we’re looking for direct evidence in the input. 

\textit{Pearl & Sprouse 2013a, 2013b, 2015}
Children’s input really doesn’t look so helpful

Data from five corpora of child-directed speech = 31,247 utterances containing a *wh*-dependency

Important: Some grammatical utterances never appeared at all. This means that **only a subset of grammatical utterances appeared**, and the child has to **generalize appropriately from this subset**.
Data from five corpora of child-directed speech = 31,247 utterances containing a *wh*-dependency

So what kinds of dependencies *are* in the input?
Children’s input

So what kinds of dependencies *are* in the input?

Data from five corpora of child-directed speech = **31,247** utterances containing a *wh*-dependency

A lot of simpler ones!

76.7%  *What did you see* __?  
12.8%  *What* __ happened?  
5.6%   *What did she want to do* __?  
2.5%   *What did she read from* __?  
1.1%   *What did she think he said* __?  
...

*Pearl & Sprouse 2013a, 2013b, 2015*
Children’s input

The induction problem

*wh*-questions in input (usually fairly simple)

What did you see __?
What __ happened?
...

Lidz & Gagliardi 2015

Pearl & Sprouse 2013a, 2013b, 2015
Children’s input

The induction problem

Grammatical *wh*-questions

- What did you see __?
- What __ happened?
- Who did Jack think that Lily saw __?
- What did Jack think __ happened?
Ungrammatical *wh*-questions: Syntactic islands

*Who does Lily think the kitty for ___ is pretty?*

*What did Jack make the claim that he bought ___?*

*What did Elizabeth wonder whether Jack bought ___?*

*What did Elizabeth worry if Jack bought ___?*
Learning strategies

Previous learning theories suggested children need syntactic-island-specific innate knowledge.
Learning strategies

A dependency cannot cross two or more bounding nodes.

Wh ... [BN1 ... [BN2 ... __]]

Bounding nodes come from a fixed set (CP, IP, and/or NP). The ones that act as a bounding nodes for a given language must be learned.

\[
Wh \quad \ldots \quad [BN_1 \quad \ldots \quad [BN_2 \quad \ldots \quad \_]]
\]

\(\{CP, IP, NP\}\)?
Learning strategies


can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)
Learning strategies


Can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)

An alternative learning strategy proposes children need less-specific linguistic prior knowledge along with probabilistic learning.

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)
Learning strategies


can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability region of structure
can’t cross 2+ bounding nodes
from a fixed set (CP, IP, and/or NP)

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)
A dependency can’t cross a very low probability region of structure
Dependencies represented as a sequence of container nodes
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very **low probability region of structure**

Dependencies represented as a sequence of **container nodes**

How to describe this dependency:
What phrases is the gap inside but the *wh*-word isn’t inside?
Container nodes

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability region of structure

Dependencies represented as a sequence of container nodes

How to describe this dependency:
What phrases is the gap inside but the wh-word isn’t inside?

What did you see ___?
= What did [IP you [VP see ___]]?
= IP-VP
Container nodes

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability region of structure

Dependencies represented as a sequence of container nodes

What did you see __?
= What did [IP you [VP see __]]?
= IP-VP

What __ happened?
= What [IP __ happened]?
= IP

... [CN1 ... CN2 ... CN3 ... CN4 ... CN5 ... __]]
Container nodes

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)
A dependency can’t cross a very low probability region of structure
Dependencies represented as a sequence of container nodes

What did you see __?
= What did [IP you [VP see __]]?
= IP-VP

What __ happened?
= What [IP __ happened]?
= IP

What did she want to do __?
= What did [IP she [VP want [IP to [VP do __]]]]?
= IP-VP-IP-VP
Container nodes

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability region of structure

Dependencies represented as a sequence of container nodes

What did you see __?
= What did [IP you [VP see __]]?
= IP-VP

What __ happened?
= What [IP __ happened]?
= IP

What did she want to do __?
= What did [IP she [VP want [IP to [VP do __]]]]?
= IP-VP-IP-VP

What did she read from __?
= What did [IP she [VP read [PP from __]]]]?
= IP-VP-PP
Learning strategies


can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)

**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability region of structure

Dependencies represented as a sequence of container nodes

Container node: phrase structure node that contains dependency

\[
\begin{align*}
\text{[CP } & \text{What } \quad \text{do } \quad [\text{IP } & \text{you } \quad [\text{VP } & \text{like } \quad \_ ] ] ] ]
\end{align*}
\]
Learning strategies


.can’t cross 2+ bounding nodes
from a fixed set (CP, IP, and/or NP)

**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability region of structure
Dependencies represented as a sequence of container nodes

Sequence of container nodes characterizes dependencies

[start-IP-VP-end]
Learning strategies


- can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

- A dependency can’t cross a very low probability region of structure
- Dependencies represented as a sequence of container nodes

Ungrammatical dependencies have low probability segments

- [CP Who did [IP Lily [VP think [CP [IP [NP the kitty [PP for] ___ ]] was pretty ?]]]]

start-IP-VP-CP-IP-NP-PP-end
Learning strategies


can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)

**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability region of structure

Dependencies represented as a sequence of *container nodes*

Low probability container node sequences have to be learned for the language
**Learning strategies**


can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)

![Diagram](image1)

**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability sequence of container nodes

![Diagram](image2)

In common: Local structural anomaly is the problem
**Subjacency-ish** *(Pearl & Sprouse 2013a, 2013b, 2015)*

A dependency can’t cross a very low probability sequence of container nodes

\[ Wh \ldots [CN_1 \ldots [CN_2 \ldots [CN_3 \ldots [CN_4 \ldots [CN_5 \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ ld...
A dependency can’t cross a very low probability sequence of container nodes.

Intuition: Learn what you can from the dependencies you do actually observe in the data and apply it to make a judgment about the dependencies you haven’t seen before, like these syntactic islands.
A dependency can’t cross a very low probability sequence of container nodes

Intuition: Learn what you can from the dependencies you do actually observe in the data and apply it to make a judgment about the dependencies you haven’t seen before, like these syntactic islands.

That is, leverage a broader set of data to make syntactic generalizations.
**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)

What information is there to leverage exactly?
What information is there to leverage exactly?

This relates to the strategy children use for learning and then generating predictions about the grammaticality of dependencies.
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

What information is there to leverage exactly?

Strategy

(1) Pay attention to the structure of dependencies.

What did she want to do ___?

= What did [IP she [VP want [IP to [VP do ___]]]]?

= IP-VP-IP-VP
Strategy
(1) Pay attention to dependency structure.

(2) Break these dependency structures into smaller pieces made up of three units (trigrams) that you can track the frequency of in the input you encounter.

\[
\begin{align*}
\text{IP-VP} &= \text{begin-IP-VP} \\
&\quad \text{IP-VP-end} \\
\text{IP-VP-IP-VP} &= \text{begin-IP-VP} \\
\text{IP-VP-IP} &= \text{begin-IP-VP} \\
\text{VP-IP-VP} &= \text{begin-IP-VP} \\
\text{IP-VP-end} &= \text{begin-IP-VP} \\
\text{IP-VP-PP} &= \text{VP-PP-end}
\end{align*}
\]
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

Strategy
(1) Pay attention to dependency structure.

(2) Break these dependency structures into smaller pieces made up of three units (trigrams) that you can track the frequency of in the input you encounter.

\[
\begin{align*}
\text{IP-VP} & = \text{begin-IP-VP} & \text{IP} & = \text{begin-IP-end} & \text{begin-IP-VP} & = 86/225 \\
& & \text{IP-VP-end} & & \text{IP-VP-end} & = 83/225 \\
\text{IP-VP-IP-VP} & = \text{begin-IP-VP} & \text{IP-VP-PP} & = \text{begin-IP-VP} & \text{begin-IP-VP} & = 13/225 \\
& \text{IP-VP-IP} & & \text{IP-VP-PP} & \text{IP-VP-IP} & = 6/225 \\
& \text{VP-IP-VP} & & \text{VP-PP-end} & \text{VP-IP-VP} & = 6/225 \\
& \text{IP-VP-end} & & \text{VP-PP-end} & \text{IP-VP-PP} & = 3/225 \\
& & & \text{VP-PP-end} & \text{VP-PP-end} & = 3/225 \\
& & & & \text{...} & 
\end{align*}
\]
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

What information is there to leverage exactly?

Strategy
(1) Pay attention to dependency structure.

(2) Break these dependency structures into smaller pieces made up of three units (trigrams) that you can track the frequency of in the input you encounter.

IP-VP = begin-IP-VP
IP-VP-end

IP-VP-VP
= begin-IP-VP
IP-VP-IP
VP-IP-VP
IP-VP-end

IP = begin-IP-end

begin-IP-VP = 86/225
IP-VP-end = 83/225
begin-IP-end = 13/225
IP-VP-IP = 6/225
VP-IP-VP = 6/225
IP-VP-PP = 3/225
VP-PP-end = 3/225
...

Note that some of these trigrams appear in multiple dependencies that commonly occur in children’s input. This will be helpful!
Strategy

1. Pay attention to dependency structure.
2. Break dependency structures into trigrams that you can track the frequency of.
3. Use trigram frequency to calculate the probability of that trigram occurring in a dependency.

\[
\begin{align*}
\text{begin-IP-VP} &= \frac{86}{225} & p(\text{begin-IP-VP}) &= 0.38 \\
\text{IP-VP-end} &= \frac{83}{225} & p(\text{IP-VP-end}) &= 0.37 \\
\text{begin-IP-end} &= \frac{13}{225} & p(\text{begin-IP-end}) &= 0.06 \\
\text{IP-VP-IP} &= \frac{6}{225} & p(\text{IP-VP-IP}) &= 0.03 \\
\text{VP-IP-VP} &= \frac{6}{225} & p(\text{VP-IP-VP}) &= 0.03 \\
\text{IP-VP-PP} &= \frac{3}{225} & p(\text{IP-VP-PP}) &= 0.01 \\
\text{VP-PP-end} &= \frac{3}{225} & p(\text{VP-PP-end}) &= 0.01 \\
\end{align*}
\]
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

Strategy

(1) Pay attention to dependency structure.

(2) Break dependency structures into trigrams that you can track the frequency of.

(3) Calculate the trigram probability in a dependency.

(4) When you see a new dependency, break it down into its trigrams and then calculate its probability, based on the trigram probabilities.

What does Jack want __?
= What does $[_{IP} \text{Jack} \ [_{VP} \text{want} \ __]]$?
= IP-VP
= $begin$-IP-VP
  
  $IP-VP-end$

$p(IP-VP) = p(begin-IP-VP) * p(IP-VP-end)$

$= 0.38 \times 0.37 = 0.14$
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

Strategy
(1) Pay attention to dependency structure.
(2) Break dependency structures into trigrams that you can track the frequency of.
(3) Calculate the trigram probability in a dependency.
(4) When you see a new dependency, break it down into its trigrams and then calculate its probability, based on the trigram probabilities.

What does Jack want to do that for __?
= What does \([_{IP} \text{ Jack } [_{VP} \text{ want } [_{IP} \text{ to } [_{VP} \text{ do that } [_{PP} \text{ for ___]]]}}\]?
= IP-VP-IP-VP-PP
= \text{begin-IP-VP}
  \text{IP-VP-IP}
  \text{VP-IP-VP}
  \text{IP-VP-PP}
  \text{VP-PP-end}

\[ p(\text{IP-VP-IP-VP-PP}) = p(\text{begin-IP-VP})*p(\text{IP-VP-IP})*p(\text{VP-IP-VP})*p(\text{IP-VP-PP})*p(\text{VP-PP-end}) = 0.38*0.03*0.03*0.01*0.01 = 0.000000034 \]
Subjects island dependency

What do you think that the joke about ___ offended Jack?

= What do [ip you [vp think [cp that [ip [np the joke [pp about ___]]]]]] offended Jack?

= IP-VP-CP-NP-PP
= begin-IP-VP

\[
p(IP-VP-CP-IP-NP-PP) = p(begin-IP-VP) \times p(IP-VP-CP) \times p(VP-CP-S) \times p(CP-IP-NP) \times p(IP-NP-PP) \times p(NP-PP-end) \\
= 0.86 \times 0.01 \times 0.001 \times 0.00 \times 0.00 \times 0.02 = 0.00
\]
**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)

What information is there to leverage exactly?

**Strategy**

1. Pay attention to dependency structure.
2. Break dependency structures into trigrams that you can track the frequency of.
3. Calculate the trigram probability in a dependency.
4. Break a new dependency into its trigrams and calculate its probability.
5. Use calculated dependency probabilities as the basis for grammaticality judgments. Lower probability dependencies are dispreferred, compared to higher probability dependencies.

\[
p(\text{IP-VP}) = 0.14 \\
p(\text{IP-VP-IP-VP-PP}) = 0.000000034 \\
p(\text{IP-VP-CP-IP-NP-PP}) = 0.00
\]
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

Wh ... [CN1 ... [CN2 ... [CN3 ... [CN4 ... [CN5 ... , __]]

Use calculated dependency probabilities as the basis for grammaticality judgments. Lower probability dependencies are dispreferred, compared to higher probability dependencies.

For each set of island stimuli from Sprouse et al. (2012), we generate grammaticality preferences for the modeled learner based on the dependency’s perceived probability and use this as a stand-in for acceptability.

Looking for superadditivity as a sign of syntactic island knowledge
Subjacency-lish (Pearl & Sprouse 2013a, 2013b, 2015)

Use calculated dependency probabilities as the **basis for grammaticality judgments**. Lower probability dependencies are **dispreferred**, compared to higher probability dependencies.

Looking for **superadditivity** as a sign of syntactic island knowledge
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

Use calculated dependency probabilities as the basis for grammaticality judgments. Lower probability dependencies are dispreferred, compared to higher probability dependencies.

Each dependency is characterized by a container node sequence, whose probability can be calculated and then plotted.
Superadditivity observed for all four islands — the qualitative behavior suggests that this learner has knowledge of these syntactic islands.

The Subjacency-ish representation that relies on container node trigram probabilities can solve this learning problem using this learning strategy.
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

Note: We’re careful to say “qualitative” behavior fit because there are lots of other factors that impact acceptability judgment behavior, and we’ve only modeled one (presumably) large part of them, which is the grammaticality of the dependency.
But is this all we can say?

No! One useful aspect of models is that we can look inside the modeled child to see why it’s behaving the way that it is. (This is something that’s harder to do with real children — that is, opening up their minds and seeing how they work.)
Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)

What’s going on?
Why are the island-spanning dependencies so much worse than the grammatical ones?
What’s going on?
Why are the island-spanning dependencies so much worse than the grammatical ones?

Let’s look inside them and see!
Let’s look inside them and see!

It turns out that each island-spanning dependency contains at least one very low probability container node trigram. So these are the relevant “island” representations.

a. Complex NP
   (i) * What did [IP the teacher [VP make [NP the claim CP that that [IP Lily VP forgot __ ]]]]]?
   (ii) \textit{start-IP-VP-NP-CP_{that}-IP-VP-end}
   (iii) Low probability.
      
   \textbf{VP-NP-CP_{that}}
   \textbf{NP-CP_{that}-IP
Let’s look inside them and see!

It turns out that each island-spanning dependency contains at least one very low probability container node trigram. So these are the relevant “island” representations.

b. Subject
   (i) * Who does [IP Jack [VP think [CP<null> [IP [NP the necklace [PP for __ ]] is expensive]]]]?
   (ii) start-IP-VP-CP<null>-IP-NP-PP-end
   (iii) Low probability: CP<null>-IP-NP
Let’s look inside them and see!

It turns out that each island-spanning dependency contains at least one very low probability container node trigram. So these are the relevant “island” representations.

c. Whether
   (i) * What does $[IP$ the teacher $[VP$ wonder $[CP_{whether}$ whether $[IP$ Jack $[VP$ stole __ $]]]]$?
   (ii) start-IP-VP-CP$_{whether}$-IP-VP-end
   (iii) Low probability:
         IP-VP-CP$_{whether}$
         VP-CP$_{whether}$-IP
         CP$_{whether}$-IP-VP
Let’s look inside them and see!

It turns out that each island-spanning dependency contains at least one very low probability container node trigram. So these are the relevant “island” representations.
Learning strategies


can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)


can’t cross 2+ bounding nodes from a fixed set (CP, IP, and/or NP)

![Diagram showing subjacency constraints](image)

**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)

A dependency can’t cross a very low probability sequence of container nodes

![Diagram showing subjacency-ish constraints](image)

**In common: Local structural anomaly is the problem**

The way Subjacency-ish implements this local structural anomaly can allow the development of syntactic island knowledge without relying on prior knowledge about bounding nodes and how many a dependency is limited to crossing.

Less reliance on island-specific prior knowledge
Learning strategies

**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)

Less reliance on island-specific prior knowledge

---

**Diagram Description**

- **Perceptual encoding**
  - Developing grammar
  - Parsing procedures
  - Extralinguistic systems (vision, pattern recognition, memory, theory of mind, etc.)

- **Perceptual intake** (linguistic representations)

- **Behavior**
  - Production systems

- **Inference engine**
  - Acquisitional intake
  - Universal grammar

- **Input**

---

**Symbols**

- Wh
- \[CN1 \ldots CN2 \ldots CN3 \ldots CN4 \ldots CN5 \ldots \]
Today’s Plan:
Computational models of syntactic acquisition

I. Some non-parametric examples
Pronoun interpretation

“Oh look — a pretty kitty!”

“Look — there’s another one!”
Pronoun interpretation

“Oh look — a pretty kitty!”

“Look — there’s another one!”

Interpretation: another pretty kitty

same

syntactic category

as antecedent

???
Pronoun interpretation

antecedent

“Oh look — a pretty kitty!”

Interpretation: another

same

syntactic category

as antecedent

???

bigger than a plain Noun

Noun

| pretty kitty
Pronoun interpretation

antecedent

“Oh look — a pretty kitty!”

“Look — there’s another one!”

Interpretation: another the pretty kitty

same syntactic category as antecedent

???

smaller than a full Noun Phrase

Noun Phrase

Noun

| pretty kitty
Pronoun interpretation

“Look — there’s another one!”

Interpretation: another

same syntactic category as antecedent

In-between category Noun’ that includes strings with nouns and modifiers+nouns

antecedent

“Oh look — a pretty kitty!”
Pronoun interpretation

antecedent

“Oh look — a pretty kitty!”

“Look — there’s another one!”

Interpretation: another

same syntactic category as antecedent

This is why we can also interpret one as just kitty.
Pronoun interpretation

“Oh look — a pretty kitty!”

“Do you see another one?”

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
"Oh look — a pretty kitty!"

"Do you see another one?"

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“Do you see another one?”

pretty kitty

Noun’

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation

syntax, semantics

another one

“Oh look — a pretty kitty!”

“What do you see now?”

another one

pretty kitty

Noun’

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“What do you see now?”

another one

pretty kitty

Noun’

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“What do you see now?”

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations

Shows baseline looking preference

Pronoun interpretation

“Oh look — a pretty kitty!”

“What do you see now?”

Shows baseline looking preference which is counteracted with “Do you see another one?”

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“Do you see another kitty?”

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“Do you see another kitty?”

another one

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation  
syntax, semantics  

another one

“Oh look — a pretty kitty!”

“Do you see another kitty?”

Shows baseline looking preference


another one  

pretty kitty

Noun’

Lidz, Waxman, & Freedman 2003:  
18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“Do you see another pretty kitty?”

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“Do you see another pretty kitty?”

Lidz, Waxman, & Freedman 2003: 18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“Do you see another pretty kitty?”

Same looking pattern as “another one”

Lidz, Waxman, & Freedman 2003:
18-month-old interpretations
Pronoun interpretation

“Oh look — a pretty kitty!”

“Do you see another one?”

Several learning strategies implemented with algorithmic-level modeled learners, given realistic samples of English child-directed speech.

Pearl & Mis 2016
**Pronoun interpretation**

**syntax, semantics**

*another one*

---

**English child-directed speech**

**Problem:** Most direct evidence children encounter is ambiguous.

**Syntactically (SYN) ambiguous data**

(92% according to corpus study by Pearl & Mis 2011, 2016):

“Look – a kitty! Oh, look – another one.”
English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Syntactically (SYN) ambiguous data

(92% according to corpus study by Pearl & Mis 2011, 2016):

“Look – a **kitty**! Oh, look – another **one**.”
English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Syntactically (SYN) ambiguous data
(92% according to corpus study by Pearl & Mis 2011, 2016):
“Look – a kitty! Oh, look – another one.”

Antecedent = “kitty”
Referent

Syntactic category?

Noun’

Pretty kitty
English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Referentially and syntactically (REF-SYN) ambiguous

(8% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty kitty! Oh, look – another one.”
Pronoun interpretation

**syntax, semantics**

English child-directed speech

**Problem:** Most direct evidence children encounter is ambiguous.

Referentially and syntactically (REF-SYN) ambiguous

(8% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty **kitty**! Oh, look – another **one**.”

92% SYN ambiguous

Noun’

**pretty kitty**
English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Referentially and syntactically (REF-SYN) ambiguous

(8% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty kitty! Oh, look – another one.”

Antecedent = “pretty kitty”
OR
Antecedent = “kitty”
Referent

92% SYN ambiguous

Pronoun interpretation

syntax, semantics
Pronoun interpretation

92% SYN ambiguous

English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Referentially and syntactically (REF-SYN) ambiguous

(8% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty kitty! Oh, look – another one.”

Antecedent = “pretty kitty”

Antecedent = “kitty”

Referent

Syntactic category?

Noun’

???

Noun

kitty
English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Referentially and syntactically (REF-SYN) ambiguous

(8% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty kitty! Oh, look – another one.”

Antecedent = “pretty kitty”
Antecedent = “kitty”
Referent

Syntax, semantics

92% SYN ambiguous

Pronoun interpretation

Noun’
Noun’
Noun

Syntactic category?

preuy kiZy Noun
Noun’
Noun’
Noun’
Noun

pretty kitty

kitty
Pronoun interpretation

92% SYN ambiguous
8% REF-SYN ambiguous

English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Unambiguous (UNAMB) data
What we wish were there but isn’t
(0% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty kitty!
Hmmm - there doesn’t seem to be another one here, though.”
English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Unambiguous (UNAMB) data

What we wish were there but isn’t

(0% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty kitty!  

Hmmm - there doesn’t seem to be another one here, though.”

Can’t have “kitty” as its antecedent, because there is another kitty here. This would be a false thing to say.
Pronoun interpretation

English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Unambiguous (UNAMB) data
What we wish were there but isn’t
(0% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty kitty!
Hmmm - there doesn’t seem to be another one here, though.”
Pronoun interpretation

92% SYN ambiguous
8% REF-SYN ambiguous

English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

Unambiguous (UNAMB) data

What we wish were there but isn’t

(0% according to corpus study by Pearl & Mis 2011, 2016)

“Look – a pretty kitty!

Hmmm - there doesn’t seem to be another one here, though.”

Must have “pretty kitty” as its antecedent.

and be a Noun’ category.
Pronoun interpretation

English child-directed speech
Problem: Most direct evidence children encounter is ambiguous.

92% SYN ambiguous
8% REF-SYN ambiguous

How do children learn the right generalizations for interpreting *one*?

**syntactic category**

- One is Noun
- One is Noun’
- Kitty
- Pretty kitty

**referent in context**

- Ambiguous one
- Pretty kitty
- KITTY

*Noun’ pretty kitty*
Pronoun interpretation

English child-directed speech
Problem: Most direct evidence children encounter is ambiguous.

92% SYN ambiguous
8% REF-SYN ambiguous

How do children learn the right generalizations for interpreting one?

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence (being more selective about what you learn from) & learning from it in more sophisticated ways

Pearl & Mis (2016): Leveraging a broader set of data to learn from & learning from it in more sophisticated ways
English child-directed speech
Problem: Most direct evidence children encounter is ambiguous.

92% SYN ambiguous
8% REF-SYN ambiguous

How do children learn the right generalizations for interpreting one?

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence

Pearl & Mis (2016):
Leveraging a broader set of data

Learning from it in more sophisticated ways
Pronoun interpretation

English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

92% SYN ambiguous
8% REF-SYN ambiguous

How do children learn the right generalizations for interpreting one?

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence

Pearl & Mis (2016):
Leveraging a broader set of data

Learning from it in more sophisticated ways

Probabilistic reasoning about input:
Bayesian inference
Pronoun interpretation

English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

92% SYN ambiguous
8% REF-SYN ambiguous

How do children learn the right generalizations for interpreting one?

Pearl & Mis (2016):
Leveraging a broader set of data

Learning from it in more sophisticated ways

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence

Diagram:

- **Input:**
  - Perceptual encoding
    - Developing grammar
    - Parsing procedures
    - Extralinguistic systems (audition, pattern recognition, memory, theory of mind, etc.)

- **Behavior:**
  - Production systems
  - Perception intake (linguistic representations)

- **Inference engine:**
  - Acquisitional intake
  - Universal grammar

- **Developing grammar**
How do children learn the right generalizations for interpreting *one*?

Pearl & Mis (2016):
Leveraging a broader set of data

Learning from it in more sophisticated ways

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence

Ignore these data 92% SYN ambiguous

“Look – a *kitty*!
Oh, look – another *one*.”
Pronoun interpretation

English child-directed speech

Problem: Most direct evidence children encounter is ambiguous.

How do children learn the right generalizations for interpreting one?

Pearl & Mis (2016): Leveraging a broader set of data

Learning from it in more sophisticated ways

Regier & Gahl (2004), Pearl & Lidz (2009): Filtering the direct evidence

Ignore these data 92% SYN ambiguous

and learn from these data using Bayesian inference 8% REF-SYN ambiguous

“Look – a pretty kitty!
Oh, look – another one.”
Pronoun interpretation
English child-directed speech
Problem: Most direct evidence children encounter is ambiguous.

92% SYN ambiguous
8% REF-SYN ambiguous

How do children learn the right generalizations for interpreting one?

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence

Learning from it in more sophisticated ways

Pearl & Mis (2016):
Leveraging a broader set of data
**Pronoun interpretation**

*English child-directed speech*

Problem: Most direct evidence children encounter is ambiguous.

92% SYN ambiguous
8% REF-SYN ambiguous

How do children learn the right generalizations for interpreting *one*?

Regier & Gahl (2004), Pearl & Lidz (2009):
*Filtering the direct evidence*

Learning from it in more sophisticated ways

Pearl & Mis (2016):
*Leveraging a broader set of data*

Learn from data like these that involve other pronouns

“Look – *a pretty kitty*!
I want to pet *it*."

```
syntax, semantics
another one
```

```
Noun'
pretty kitty
```
Pronoun interpretation

English child-directed speech
Problem: Most direct evidence children encounter is ambiguous.

92% SYN ambiguous
8% REF-SYN ambiguous

How do children learn the right generalizations for interpreting one?

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence

Learning from it in more sophisticated ways

Pearl & Mis (2016):
Leveraging a broader set of data

Learn from data like these that involve other pronouns

“Look – a pretty kitty!
I want to pet it.”

Key: modifier is included in antecedent.
Implication: May want to include the modifier whenever it’s an option.
Pronoun interpretation

Algorithmic-level implementation of these strategies
Evaluated on whether they matched 18-month-old looking preferences.

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence
Learning from it in more sophisticated ways

Pearl & Mis (2016):
Leveraging a broader set of data
Pronoun interpretation

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence
Learning from it in more sophisticated ways

Pearl & Mis (2016):
Leveraging a broader set of data

Algorithmic-level

Both were successful at generating the 18-month-old behavior. We can then look inside the modeled learners and see what the underlying representations were.
Pronoun interpretation

Learning from it in more sophisticated ways

Pearl & Mis (2016): Leveraging a broader set of data

Algorithmic-level

Regier & Gahl (2004), Pearl & Lidz (2009): Filtering the direct evidence

Adult representations

Noun’

But...required additional situational context to be present to succeed.
Pronoun interpretation

Learning from it in more sophisticated ways

Pearl & Mis (2016):
Leveraging a broader set of data

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence

Adult representations
Noun’
pretty kitty

But...required additional situational context to be present to succeed.

"Look – a pretty kitty!
Oh, look – another one."

Needed to have a lot of alternative options so it’s a suspicious coincidence that the referent is pretty if "pretty" wasn’t actually included in the antecedent.
Pronoun interpretation

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence ✓ Less robust
Learning from it in more sophisticated ways

Algorithmic-level

Leveraging a broader set of data

Immature representations
✓ Noun’ only in certain linguistic contexts
✓ pretty kitty

Pearl & Mis (2016):
“Look – a pretty kitty!
Oh, look – another one.”

Pronoun interpretation syntax, semantics another one
Pronoun interpretation

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence ✓ Less robust
Learning from it in more sophisticated ways

Pearl & Mis (2016):
Leveraging a broader set of data
Immature representations
✓ Noun’ only in certain linguistic contexts
✓ pretty kitty  × otherwise Noun

“Look – a kitty!
Oh, look – another one.”

But...does this for pretty much any situational context.
More robust
Pronoun interpretation

Regier & Gahl (2004), Pearl & Lidz (2009):

Filtering the direct evidence ✓ Less robust
Learning from it in more sophisticated ways

Pearl & Mis (2016):

✗ More robust
Leveraging a broader set of data

Algorithmic-level

By modeling, we have two concrete proposals for how children learn the knowledge they do by 18 months.

This also motivates future experimental work to distinguish these two possibilities.
Pronoun interpretation

Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence ✓ Less robust
Learning from it in more sophisticated ways
Pearl & Mis (2016): × ✓ More robust
Leveraging a broader set of data

Algorithmic-level

This also motivates future experimental work to distinguish these two possibilities.

“This kitty likes the cup of milk but not the one of water.”

Adults generally don’t like this because it forces one to be category Noun.
Pronoun interpretation

Regier & Gahl (2004), Pearl & Lidz (2009):
- Filtering the direct evidence ✓ Less robust
- Learning from it in more sophisticated ways

Pearl & Mis (2016):
- Less robust
- Leveraging a broader set of data

This also motivates future experimental work to distinguish these two possibilities.

“This kitty likes the cup of milk but not the one of water.”

When do children have this same judgment? Is it before 18 months?
Learning from it in more sophisticated ways

Pearl & Mis (2016): More robust
Leveraging a broader set of data

Algorithmic-level

By 18 months
Regier & Gahl (2004), Pearl & Lidz (2009):
Filtering the direct evidence

“This kitty likes the cup of milk but not the one of water.”

When do children have this same judgment? Is it before 18 months?
Pronoun interpretation

By 18 months
Regier & Gahl (2004),
Pearl & Lidz (2009):
Filtering the direct evidence
✓

Not by 18 months
Pearl & Mis (2016):
Leveraging a broader set of data
✗

“This kitty likes the cup of milk but not the one of water.”

When do children have this same judgment? Is it before 18 months?
Today’s Plan:
Computational models of syntactic acquisition

I. Some non-parametric examples

II. About linguistic parameters

III. Learning with parameters

 udpating parameters
Today’s Plan: 
Computational models of syntactic acquisition

II. About linguistic parameters
About linguistic parameters

What are linguistic parameters?
How do they work?
What exactly are they supposed to do?
About linguistic parameters

A parameter is meant to be something that can account for multiple observations in some domain.

Parameter for a statistical model: determines what the model predicts will be observed in the world in a variety of situations.

Parameter for our mental (and linguistic) model: determines what *we* predict will be observed in the world in a variety of situations.
The normal distribution is a statistical model that uses **two parameters**: 
- \( \mu \) for the mean 
- \( \sigma \) for the standard deviation

If we know the **values of these parameters**, we can make predictions about the probability of data we rarely or never see.
Statistical parameter

\( \mu \) for the mean

\( \sigma \) for the standard deviation

Suppose this is a model of how many minutes late I’ll be to class.

Let’s use the model with \( \mu = 0 \) and \( \sigma^2 = 0.2 \).
How probable is it that I’ll be 5 minutes late, given these parameter values?

Let’s use the model with $\mu = 0$ and $\sigma^2 = 0.2$.

How probable is it that I’ll be 5 minutes late, given these parameter values?

$\phi_{\mu, \sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$
About linguistic parameters

Statistical parameter

$\mu$ for the mean

$\sigma$ for the standard deviation

Let’s use the model with $\mu = 0$ and $\sigma^2 = 0.2$.

5 minutes late? ✗

What about right on time? ✓

$\phi_{\mu, \sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$

Much more probable!
About linguistic parameters

Statistical parameter

- $\mu$ for the mean
- $\sigma$ for the standard deviation

Let’s use the model with $\mu = 0$ and $\sigma^2 = 0.2$.

5 minutes late? $\times$

On time? $\checkmark$

What about 2 minutes early? $\times$

We can tell this just by knowing the values of the two statistical parameters. These parameter values allow us to infer the probability of the observable behavior.
About linguistic parameters

Statistical parameter

\( \mu \) for the mean

\( \sigma \) for the standard deviation

Let’s shift to the model with \( \mu = -2 \) and \( \sigma^2 = 0.5 \).

\[
\varphi_{\mu,\sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}
\]
About linguistic parameters

Statistical parameter

\( \mu \) for the mean
\( \sigma \) for the standard deviation

Let’s shift to the model with \( \mu = -2 \) and \( \sigma^2 = 0.5 \).

How probable is it that I’ll be 5 minutes late, given these parameter values?

Not very probable!
Statistical parameter
\( \mu \) for the mean
\( \sigma \) for the standard deviation

Let’s shift to the model with \( \mu = -2 \) and \( \sigma^2 = 0.5 \).

5 minutes late? ✗

What about right on time? ✗

Not very probable!
About linguistic parameters

Statistical parameter

$\mu$ for the mean

$\sigma$ for the standard deviation

Let’s shift to the model with $\mu = -2$ and $\sigma^2 = 0.5$.

5 minutes late? ✗

On time? ✗

What about 2 minutes early? ✅

Changing the parameter values changes the behavior we predict we’ll observe.

$$\varphi_{\mu,\sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$
About linguistic parameters

Statistical parameter

\( \mu \) for the mean

\( \sigma \) for the standard deviation

Observing different quantities of data with particular values can tell us which values of \( \mu \) and \( \sigma^2 \) are most likely, if we know we’re trying to determine the values of \( \mu \) and \( \sigma^2 \) in function \( \phi(X) \)

\[
\phi_{\mu,\sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}
\]

Observing data points distributed like the green curve tells us that \( \mu \) is likely to be around -2 and \( \sigma^2 \) is likely to be around 0.5.
About linguistic parameters

Statistical parameter

μ for the mean

σ for the standard deviation

Important similarity to linguistic parameters:

We don’t see the process that generates the data, but only the data themselves. This means that in order to form our expectations about X, we are, in effect, reverse engineering the observable data.

\[
\varphi_{\mu,\sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}
\]
About linguistic parameters

Statistical parameter

$\mu$ for the mean

$\sigma$ for the standard deviation

Our knowledge of the underlying function/principle that generates these data - $\phi(X)$ - as well as the associated parameters - $\mu$, and $\sigma^2$ - allows us to represent an infinite number of expectations about the behavior of variable $X$.

$$\varphi_{\mu,\sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$
Both linguistic principles and linguistic parameters are often thought of as innate domain-specific abstractions that connect to many structural properties about language.

Linguistic **principles** correspond to the properties that are invariant across all human languages.
Comparison: $\mu$ and $\sigma^2$ determine the exact form of the curve that represents the probability of observing certain data. While different values for these parameters can produce many different curves, these curves share their underlying form due to the common invariant function.

$$q_{\mu,\sigma^2}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

Both linguistic principles and linguistic parameters are often thought of as innate domain-specific abstractions that connect to many structural properties about language.

Linguistic parameters correspond to the properties that vary across human languages.
Parameters connecting to multiple structural properties is a very good thing from the perspective of someone trying to acquire language (like a child). This is because a child can learn about a parameter’s value by observing many different kinds of examples in the language.
About linguistic parameters for language acquisition

“The richer the deductive structure associated with a particular parameter, the greater the range of potential ‘triggering’ data which will be available to the child for the ‘fixing’ of the particular parameter” – Hyams (1987)
Parameters can be especially useful when a child is trying to learn the things about language structure that are otherwise hard to learn, perhaps because they are very complex properties themselves or because they appear very infrequently in the available data.
About linguistic parameters for language acquisition

An issue: The observable data are often the result of a combination of interacting parameters.

This can make it hard to figure out what parameter values might have produced the observable data - even if the child already knows what the parameters are.

Observable data can be ambiguous for which parameter values they signal.

Observable data
“"I love kitties.""
An issue: The observable data are often the result of a combination of interacting parameters.

Observable data can be ambiguous for which parameter values they signal.

"I love kitties."
Interacting parameters

Example Parameter 1: Head-directionality

Edo/English: Head-first

Basic word order:
Subject Verb Object [SVO]

Prepositions:
Preposition Noun Phrase
Interacting parameters

Example Parameter 1: Head-directionality
Edo/English: Head-first

Japanese/Navajo: Head-final

Basic word order:
Subject Object Verb [SOV]

Postpositions:
Noun Phrase Postposition
Interacting parameters

Example Parameter 1: Head-directionality
- Edo/English: Head-first
- Japanese/Navajo: Head-final

Example Parameter 2: Verb Second (V2)
- German: +V2
  - Verb moves to second phrasal position, some other phrase moves to the first position

Sarah das Buch liest
- *Sarah the book reads*

*Underlying form of the sentence*
Interacting parameters

Example Parameter 1: Head-directionality
- Edo/English: Head-first
- Japanese/Navajo: Head-final

Example Parameter 2: Verb Second (V2)
- German: +V2
  - Verb moves to second phrasal position, some other phrase moves to the first position

Observable form of the sentence

Sarah liest das Buch
Sarah reads the book
Interacting parameters

Example Parameter 1: Head-directionality
  Edo/English: Head-first
  Japanese/Navajo: Head-final

Example Parameter 2: Verb Second (V2)

German: +V2
Verb moves to second phrasal position, some other phrase moves to the first position

Sarah das Buch liest
Sarah the book reads

Underlying form of the sentence
Interacting parameters

Example Parameter 1: Head-directionality ☢
  Edo/English: Head-first ☢
  Japanese/Navajo: Head-final ☢

Example Parameter 2: Verb Second (V2) ☢

German: +V2 ☢
Verb moves to second phrasal position, some other phrase moves to the first position

Das Buch liest Sarah das Buch liest
The book reads Sarah

Observable form of the sentence
Interacting parameters

Example Parameter 1: Head-directionality
- Edo/English: Head-first
- Japanese/Navajo: Head-final

Example Parameter 2: Verb Second (V2)
- German: +V2
- English: -V2

Verb doesn’t move.

Sarah reads the book

- Underlying form of the sentence
- Observable form of the sentence
Interacting parameters

Head-directionality  Verb Second (V2)

Grammars available

- **G1**: Head-first \(+V2\)
- **G2**: Head-final \(+V2\)
- **G3**: Head-first \(-V2\)
- **G4**: Head-final \(-V2\)
Interacting parameters

Head-directionality  Verb Second (V2)

Data point

Subject  Verb  Object

“I love kittens.”

G1  Head-first +V2
G2  Head-final +V2
G3  Head-first -V2
G4  Head-final -V2
Interacting parameters

Head-directionality  Verb Second (V2)

“"I love kitties.""

Which grammars can analyze this data point?

Subject  Verb  Object

G1  Head-first  +V2  

G2  Head-final  +V2  

G3  Head-first  -V2  

G4  Head-final  -V2  

"I love kitties."
“I love kittens.”

Interacting parameters

Head-directionality  Verb Second (V2)

Subject  Verb  Verb  Object

G1

Head-first  +V2

G2

Head-final  +V2

G3

Head-first  -V2

G4

Head-final  -V2

✓ +head-first predicts SVO
✓ +V2 predicts Verb moved to second position
Interacting parameters

Head-directionality
Verb Second (V2)

Subject Verb Subject Object Verb

“I love kitties.”

G1
Head-first +V2

G2
Head-final +V2

G3
Head-first -V2

G4
Head-final -V2

✔ head-final predicts SOV
✔ +V2 predicts Verb moved to second position
Interacting parameters

Head-directionality  Verb Second (V2)

Subject  Verb  Object

“I love kitties.”

head-first predicts SVO
-V2 predicts Verb doesn’t move

G1  Head-first  +V2

G2  Head-final  +V2

G3  Head-first  -V2

G4  Head-final  -V2

Subject  Verb  Object

✔  ✔  ✔

✔  ✔  ✔
Interacting parameters

Head-directionality  Verb Second (V2)

“I love kitties.”

Subject  Verb  Object

- Head-final predicts SOV
- V2 predicts Verb doesn’t move

- Head-first +V2
- Head-final +V2
- Head-first -V2
Interacting parameters

- Head-directionality
- Verb Second (V2)

“I love kitties.”

What do the grammars that can analyze this data point have in common?

G1: Head-first +V2
G2: Head-final +V2
G3: Head-first -V2
G4: Head-final -V2

Subject  Verb  Object
Interacting parameters

Head-directionality  Verb Second (V2)

“"I love kitties.""

Subject  Verb  Object

G1  Head-first +V2

G2  Head-final +V2

G3  Head-first -V2

G4  Head-final -V2

We don’t know whether the true grammar is head-first or head-final since there’s a grammar of each kind.
We don’t know whether the true grammar is head-first or head-final since there’s a grammar of each kind.

(though there are more head-first)
We don’t know whether the true grammar is +V2 or -V2 since there’s a grammar of each kind.
Interacting parameters

Head-directionality

Verb Second (V2)

“\text{i love kitties.}”

\text{Subject} \quad \text{Verb} \quad \text{Object}

\begin{itemize}
  \item G1: Head-first +V2
  \item G2: Head-final +V2
  \item G3: Head-first -V2
  \item G4: Head-final -V2
\end{itemize}

We don’t know whether the true grammar is +V2 or -V2 since there’s a grammar of each kind.

\text{(though there are more +V2)}
Interacting parameters

Head-directionality  Verb Second (V2)

“I love kitties.”

Subject  Verb  Object

G1  Head-first  +V2
G2  Head-final  +V2
G3  Head-first  -V2
G4  Head-final  -V2

This data point isn’t unambiguous for any of the parameters we’re interested in because the parameters interact... even though we feel like it might be somewhat informative for head-first and +V2 because these occur in more grammars that are compatible.
Interacting parameters

Head-directionality
- Edo/English: Head-first
- Japanese/Navajo: Head-final

Example Parameter 3: Subject drop

Spanish: +subj-drop
Allows Subject to be overt or dropped

✔ Ellos beben
  they drink-3rd-pl

✔ Beben
  drink-3rd-pl

“They drink”
Interacting parameters

Head-directionality
Edo/English: Head-first
Japanese/Navajo: Head-final

Example Parameter 3: Subject drop
Spanish: +subj-drop

English: -subj-drop
Subject must be overt

✔ They drink

✗ Drink  “They drink”
Interacting parameters

Head-directionality  Subject drop (subj-drop)

Grammars available

- **G1**  Head-first  +subj-drop
- **G2**  Head-final  +subj-drop
- **G3**  Head-first  -subj-drop
- **G4**  Head-final  -subj-drop
Interacting parameters

Head-directionality  Subject drop (subj-drop)

“…dass ich Kätzchen liebe.”
...that I Kitties love

Subject  Object  Verb

Which grammars can analyze this data point?

G1  Head-first +subj-drop
G2  Head-final +subj-drop

G3  Head-first -subj-drop
G4  Head-final -subj-drop

"…dass ich Kätzchen liebe.”
...that I Kitties love
Interacting parameters

Head-directionality  Subject drop (subj-drop)

Subject  Object  Verb

“...dass ich Kätzchen liebe.”  ...
...that I Kitties love

✗ head-first predicts SVO
✓ +subj-drop allows subject to be overt

G1  Head-first  +subj-drop

G2  Head-final  +subj-drop

G3  Head-first  -subj-drop

G4  Head-final  -subj-drop

"...dass ich Kätzchen liebe."  ...
...that I Kitties love
Interacting parameters

Head-directionality  Subject drop (subj-drop)

Subject  Object  Verb

G1
Head-first +subj-drop

G2
Head-final  +subj-drop

G3
Head-first -subj-drop

G4
Head-final -subj-drop

“...dass ich Kätzchen liebe.”
...that I Kitties love

✔ head-final predicts SOV
✔ +subj-drop allows subject to be overt

“…dass ich Kätzchen liebe.”
…that I Kitties love
Interacting parameters

Head-directionality  Subject drop (subj-drop)

“...dass ich Kätzchen liebe.”
...that I Kitties love

Subject  Object  Verb

- subj-drop requires subject to be overt

G3  Head-first  -subj-drop

G2  Head-final  +subj-drop

G4  Head-final  -subj-drop

G1  Head-first  +subj-drop

“...dass ich Kätzchen liebe.”
...that I Kitties love
Interacting parameters

Head-directionality  Subject drop (subj-drop)

“…dass ich Kätzchen liebe.”
...that I Kitties love

 ✓ head-final predicts SOV
 ✓ -subj-drop requires subject to be overt

G1  Head-first  +subj-drop
G2  Head-final  +subj-drop
G3  Head-first  -subj-drop
G4  Head-final  -subj-drop
There’s more than one grammar compatible with this data point...even though we feel like it *should definitely* be informative for head-final (since that’s the only value in the compatible grammars).
But technically, this is still an ambiguous data point because more than one grammar will work.

---

“...dass ich Kätzchen liebe.”

...that *I* *Kitties* love
Interacting parameters

Head-directionality  Subject drop (subj-drop)

Subject  Object  Verb

G2  Head-final  +subj-drop

G4  Head-final  -subj-drop

So what can we do?

…dass ich Kätzchen liebe.
...that I Kitties love
Today’s Plan:
Computational models of syntactic acquisition

I. Some non-parametric examples

Who does ... is pretty?

syntax

II. About linguistic parameters

III. Learning with parameters

0.2 0.3 0.8 0.7 0.1
0.8 0.7 0.2 0.3 0.9
Today’s Plan:
Computational models of syntactic acquisition

III. Learning with parameters

0.2 0.3 0.8 0.7 0.1

0.8 0.7 0.2 0.3 0.9
Learning with parameters

A language’s grammar = combination of parameter values

G2: Head-final + subj-drop
G4: Head-final - subj-drop
Learning with parameters

A language’s grammar = combination of parameter values
Variational learning (Yang 2002, 2004, 2012): use reinforcement learning to learn which value (for each parameter) that the native language uses for its grammar. This is a combination of using linguistic knowledge & statistical learning.
Learning with parameters

Variational learning

Idea taken from evolutionary biology:
In a population, individuals compete against each other. The fittest individuals survive while the others die out.

How do we translate this to learning with parameters?
Learning with parameters

Variational learning

The fittest **individuals** survive while the others die out.

Individual = grammar (combination of parameter values that represents the structural properties of a language)
Learning with parameters

Variational learning

The **fittest** individuals survive while the others die out.

Fitness = how well a grammar can analyze the data the child encounters
A child’s mind consists of a population of grammars that are competing to analyze the data in the child’s native language.
Intuition: The most successful (fittest) grammar will be the native language grammar because it can analyze all the data the child encounters. This grammar will “win”, once the child encounters enough native language data. This is because none of the other competing grammars can analyze all the data.
If this is the native language grammar, this grammar can analyze all the intake while the others can’t.
At any point in time, a grammar in the population will have a **probability** associated with it. This represents the child’s belief that this grammar is the correct grammar for the native language.
Before the child has encountered any native language data, all grammars are *equally likely*. So, initially all grammars have the same probability, which is $1$ divided the number of grammars available.
Since there are 11 grammars here, each begins with probability $1/11$. 
As the child encounters data from the native language, some of the grammars will be more fit because they are better able to account for the syntactic properties of the intake.

Other grammars will be less fit because they cannot account for some of the data encountered.
Grammars that are more compatible with the native language data intake will have their probabilities increased while grammars that are less compatible will have their probabilities decreased over time.
After the child has encountered enough data from the native language, the native language grammar should have a probability near 1.0 while the other grammars have a probability near 0.0.
Learning with parameters
Variational learning

The power of unambiguous data:
Unambiguos data from the native language can only be analyzed by grammars that use the native language’s parameter value.
This makes unambiguous data very influential data for the child to encounter, since these data are only compatible with the parameter value that is correct for the native language.
Problem: Do unambiguous data exist for entire grammars? This requires data that are incompatible with every other possible parameter value of every other possible grammar....
This seems unlikely for real language data because linguistic parameters connect with different types of patterns, which may have nothing to do with each other, or parameters may interact with each other.
Learning with parameters

Variational learning

Key: Parameters are separable components of grammars
A variational learner can take advantage of the fact that grammars are really sets of parameter values.
Parameter values can be probabilistically accessed, depending on the level of belief (probability) the learner currently has in each one.
Learning with parameters

Variational learning

Parameter values can be probabilistically accessed, depending on the level of belief (probability) the learner currently has in each one.

\[ p = 0.2 \times 0.3 \times 0.8 \times 0.7 \times 0.1 \]

\[ p = 0.8 \times 0.7 \times 0.2 \times 0.3 \times 0.9 \]
Learning with parameters

Variational learning

Parameter values can be probabilistically accessed, depending on the level of belief (probability) the learner currently has in each one.
Parameter values can be probabilistically accessed, depending on the level of belief (probability) the learner currently has in each one.
Parameter values can be probabilistically accessed, depending on the level of belief (probability) the learner currently has in each one.
Parameter values can be probabilistically accessed, depending on the level of belief (probability) the learner currently has in each one.
Learning with parameters
The learning algorithm

Variational learning

For each data point encountered in the input...
For each data point encountered in the input...

(1) Choose a grammar to test out on a particular data point. Select a grammar by choosing a set of parameter values, based on the probabilities associated with each parameter value.

Denison, Bonawitz, Gopnik, & Griffiths 2013: Experimental evidence from 4 and 5-year-olds suggests that children are sensitive to the probabilities of complex representations (which parameters are), and so this kind of sampling is not unrealistic.
For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

If this grammar can analyze the data point, increase the probability of all participating parameter values slightly (reward each value).
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for reward:

\[ p_v = .2 \]
\[ p_o = .8 \]

\[ p = .8 \times .3 \times .8 \times .3 \times .9 \]

\( p_v \) = previous value of successful parameter value

\( p_o \) = previous value of opposing parameter value
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for reward:

$p_v = 0.8$
$p_o = 0.2$
Learning with parameters
The learning algorithm

Variational learning

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for reward:

\[ p_v = 0.8 \]
\[ p_o = 0.2 \]

\[ p_{v_{\text{updated}}} = p_v + \gamma(1-p_v) \]
\[ p_{o_{\text{updated}}} = (1-\gamma)p_o \]

\[ \gamma = \text{learning rate (ex: } \gamma = .125) \]
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for reward:

\[ p_v = 0.8 \]
\[ p_o = 0.2 \]

\[ p_{v_{\text{updated}}} = 0.8 + 0.125(1 - 0.8) \]
\[ p_{o_{\text{updated}}} = (1 - 0.125)0.2 \]

\[ \gamma = \text{learning rate (ex: } \gamma = 0.125) \]
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for reward:

\[
p_v = 0.8 \\
p_o = 0.2 \\
p_{v\_updated} = 0.825 \\
p_{o\_updated} = 0.175
\]
Learning with parameters
The learning algorithm

Variational learning

For each data point encountered in the input...

1. Choose a grammar.
2. Try to analyze the data point with this grammar.
3. Update parameter value probabilities.

Actual update equation for reward:

\[
p_v = 0.8 \\
p_o = 0.2 \\
p_{v_{\text{updated}}} = 0.825 \\
p_{o_{\text{updated}}} = 0.175
\]

Do this for all the other parameters, too.
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.
(2) Try to analyze the data point with this grammar.
(3) Update parameter value probabilities.

\[
p = 0.8 \times 0.3 \times 0.8 \times 0.3 \times 0.9
\]
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.
(2) Try to analyze the data point with this grammar.
(3) Update parameter value probabilities.

But what happens if the selected grammar can’t account for the data point?

Then all the participating parameter values are punished.
Learning with parameters
The learning algorithm

Variational learning

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for punishment:

\[ p_v = .2 \]
\[ p_o = .8 \]

\[ p = .8 \times .3 \times .8 \times .3 \times .9 \]

\( p_v \) = previous value of unsuccessful parameter value

\( p_o \) = previous value of opposing parameter value
For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for **punishment**: \( p_v = 0.8 \)
\( p_o = 0.2 \)

\[ p = 0.8 \times 0.3 \times 0.8 \times 0.3 \times 0.9 \]

**Variational learning**

Subject Object Verb
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for punishment:

\[ p_{v} = 0.8 \]
\[ p_{o} = 0.2 \]
\[ p_{v\_updated} = (1-\gamma)p_{v} \]
\[ p_{o\_updated} = \gamma + (1-\gamma)p_{o} \]

\[ \gamma = \text{learning rate (ex: } \gamma = .125) \]
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.
(2) Try to analyze the data point with this grammar.
(3) Update parameter value probabilities.

Actual update equation for punishment:

\[
\begin{align*}
\text{p}_v &= 0.8 \\
\text{p}_o &= 0.2 \\
\text{p}_{v\text{ updated}} &= (1-0.125)0.8 \\
\text{p}_{o\text{ updated}} &= 0.125 + (1-0.125)0.2
\end{align*}
\]
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for punishment:

\[ p_v = 0.8 \]
\[ p_o = 0.2 \]

\[ p_{v_{\text{updated}}} = 0.70 \]
\[ p_{o_{\text{updated}}} = 0.30 \]
Learning with parameters
The learning algorithm

Variational learning

For each data point encountered in the input...

(1) Choose a grammar.

(2) Try to analyze the data point with this grammar.

(3) Update parameter value probabilities.

Actual update equation for punishment:

\[ p_v = 0.8 \]
\[ p_o = 0.2 \]

\[ p_{v_{\text{updated}}} = 0.70 \]
\[ p_{o_{\text{updated}}} = 0.30 \]

Do this for all the other parameters, too.
Learning with parameters
The learning algorithm

For each data point encountered in the input...

(1) Choose a grammar.
(2) Try to analyze the data point with this grammar.
(3) Update parameter value probabilities.

Variational learning

Subject Object Verb

\[ p = 0.3 \times 0.26 \times 0.7 \times 0.74 \times 0.21 \]

\[ p = 0.7 \times 0.74 \times 0.3 \times 0.26 \times 0.79 \]
Learning with parameters
The learning algorithm

Variational learning

For each data point encountered in the input...

(1) Choose a grammar.
(2) Try to analyze the data point with this grammar.
(3) Update parameter value probabilities.

Problem ameliorated!

**Unambiguous data** are much more likely to exist for **individual parameter values** instead of entire grammars.
Unambiguous data are much more likely to exist for individual parameter values instead of entire grammars.

“...dass ich Kätzchen liebe.”

...that I Kitties love
In this case, if either $G_2$ or $G_4$ were selected, head-final would be rewarded (in addition to whichever subj-drop value was used).

Head-directionality  Subject drop (subj-drop)

“...dass ich Kätzchen liebe.”
...that I Kitties love
In this case, if either G1 or G3 were selected, head-first would be punished (in addition to whichever subj-drop value was used).

“…dass ich Kätzchen liebe.”

...that I Kitties love
Learning with parameters
The learning algorithm
Variational learning

Because this data point is unambiguous for head-final, grammars using that value would be rewarded and its probability as a parameter value would become 1.0 over time.

“...dass ich Kätzchen liebe.”
...that I Kitties love

Subject drop (subj-drop)

Subject  Object  Verb

Head-final
G2 ✔ +subj-drop

Head-first
G1 ✗ +subj-drop

G3 ✗ -subj-drop

G4 ✔ -subj-drop

Head-final
G4 ✔ -subj-drop
Meanwhile, grammars using **head-first** would be punished every time, and its probability as a parameter value would approach 0.0 over time.
Learning with parameters
The learning algorithm
Variational learning

Implication: The more unambiguous data there are, the faster the native language’s parameter value will “win” (reach a probability near 1.0). This means that the child will learn the associated structural pattern faster.
Learning with parameters
The learning algorithm
Variational learning

Example: the more unambiguous head-final data the child encounters, the faster a child should learn that the native language prefers objects before verbs as the basic order.

Subject  Object  Verb

"...dass ich Kätzchen liebe."
...that I Kitties love
Is it true that the amount of unambiguous data the child encounters for a particular parameter strongly impacts when the child learns that structural property of the language?
Learning with parameters
The learning algorithm
Variational learning

Striking evidence that this is true

Table 1: The qualitative fit Yang discovered between the unambiguous data advantage (Adv) perceived by a VarLearner in its acquisitional intake and the observed age of acquisition (AoA) in children for six parameter values across different languages.

<table>
<thead>
<tr>
<th>Param Value</th>
<th>Language</th>
<th>Unambiguous Form</th>
<th>Unambiguous Ex</th>
<th>Adv</th>
<th>AoA</th>
</tr>
</thead>
<tbody>
<tr>
<td>+wh-fronting</td>
<td>English</td>
<td>wh-fronting in questions</td>
<td>Who did you see?</td>
<td>25%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td>+topic-drop</td>
<td>Chinese</td>
<td>null objects</td>
<td>Wǒ méi kànjiàn I not see “I didn’t see (him)”</td>
<td>12%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td>+pro-drop</td>
<td>Italian</td>
<td>null subjects in questions</td>
<td>Chi hai visto who have seen “Who have you seen?”</td>
<td>10%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td>+verb-raising</td>
<td>French</td>
<td>Verb Adverb</td>
<td>Jean voit souvent Marie Jean sees often Marie “Jean often sees Marie”</td>
<td>7%</td>
<td>1;8</td>
</tr>
<tr>
<td>-pro-drop</td>
<td>English</td>
<td>expletive subjects</td>
<td>There’s a penguin on the ice.</td>
<td>1.2%</td>
<td>3;0</td>
</tr>
<tr>
<td>+verb-second</td>
<td>German</td>
<td>Object Verb Subject</td>
<td>Pinguine liebe ich. penguins like I “I like penguins”</td>
<td>1.2%</td>
<td>3;0-3;2</td>
</tr>
<tr>
<td>-scope-marking</td>
<td>English</td>
<td>long-distance wh questions without medial-wh</td>
<td>Who do you think is on the ice?</td>
<td>0.2%</td>
<td>&gt;4.0</td>
</tr>
</tbody>
</table>
Learning with parameters
The learning algorithm
Variational learning

Striking evidence that this is true

Table 1: The qualitative fit Yang discovered between the unambiguous data advantage (Adv) perceived by a VarLearner in its acquisitional intake and the observed age of acquisition (AoA) in children for six parameter values across different languages.

<table>
<thead>
<tr>
<th>Param Value</th>
<th>Language</th>
<th>Unambiguous Form</th>
<th>Unambiguous Ex</th>
<th>Adv</th>
<th>AoA</th>
</tr>
</thead>
<tbody>
<tr>
<td>+wh-fronting</td>
<td>English</td>
<td>wh-fronting in questions</td>
<td>Who did you see?</td>
<td>25%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td>+topic-drop</td>
<td>Chinese</td>
<td>null objects</td>
<td>Wǒ méi kàn jiàn</td>
<td>12%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>I not see</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“I didn’t see (him)”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+pro-drop</td>
<td>Italian</td>
<td>null subjects in questions</td>
<td>Chi hai visto</td>
<td>10%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>who have seen</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Who have you seen?”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+verb-raising</td>
<td>French</td>
<td>Verb Adverb</td>
<td>Jean voit souvent Marie</td>
<td>7%</td>
<td>1;8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Jean sees often Marie</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Jean often sees Marie”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-pro-drop</td>
<td>English</td>
<td>expletive subjects</td>
<td>There’s a penguin on the ice.</td>
<td>1.2%</td>
<td>3;0</td>
</tr>
<tr>
<td>+verb-second</td>
<td>German</td>
<td>Object Verb Subject</td>
<td>Pinguine liebe ich.</td>
<td>1.2%</td>
<td>3;0-3;2</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td></td>
<td>penguins like I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“I like penguins”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-scope-marking</td>
<td>English</td>
<td>long-distance wh questions</td>
<td>Who do you think is on the ice?</td>
<td>0.2%</td>
<td>&gt;4.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>without medial-wh</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The **more unambiguous data** there are for one value over another (its advantage)...
Learning with parameters
The learning algorithm

Variational learning

Striking evidence that this is true

Table 1: The qualitative fit Yang discovered between the unambiguous data advantage (Adv) perceived by a VarLearner in its acquisitional intake and the observed age of acquisition (AoA) in children for six parameter values across different languages.

<table>
<thead>
<tr>
<th>Param Value</th>
<th>Language</th>
<th>Unambiguous Form</th>
<th>Unambiguous Ex</th>
<th>Adv</th>
<th>AoA</th>
</tr>
</thead>
<tbody>
<tr>
<td>+wh-fronting</td>
<td>English</td>
<td>wh-fronting in questions</td>
<td>Who did you see?</td>
<td>25%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td>+topic-drop</td>
<td>Chinese</td>
<td>null objects</td>
<td>Wǒ méi kànjiàn I not see</td>
<td>12%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“I didn’t see (him)”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+pro-drop</td>
<td>Italian</td>
<td>null subjects in questions</td>
<td>Chi hai visto who have seen</td>
<td>10%</td>
<td>&lt;1;8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Who have you seen?”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+verb-raising</td>
<td>French</td>
<td>Verb Adverb</td>
<td>Jean voit souvent Marie</td>
<td>7%</td>
<td>1;8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Jean sees often Marie</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Jean often sees Marie”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-pro-drop</td>
<td>English</td>
<td>expletive subjects</td>
<td>There’s a penguin on the ice.</td>
<td>1.2%</td>
<td>3;0</td>
</tr>
<tr>
<td>+verb-second</td>
<td>German</td>
<td>Object Verb Subject</td>
<td>Pinguine liebe ich. penguins like</td>
<td>1.2%</td>
<td>3;0-3;2</td>
</tr>
<tr>
<td></td>
<td>Dutch</td>
<td></td>
<td>I “like penguins”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-scope-marking</td>
<td>English</td>
<td>long-distance wh questions</td>
<td>Who do you think is on the ice?</td>
<td>0.2%</td>
<td>&gt;4.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>without medial-wh</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The more unambiguous data there are for one value over another (its advantage), the earlier it seems to be learned.
Thank you!

This work was supported in part by NSF grants BCS-0843896 and BCS-1347028.