The Computation of Language:
Information processing

One way to think about the computation of language is from an information processing standpoint.

Natural language processing: How do people and machines extract information about the world from the language data they encounter?

Input

“I’m not that a nice kitty?”
“That...is not a dog.”

Internal representation

Output

persuasion

surprise
One way to think about the computation of language is from an information
processing standpoint.

Natural language processing:
Recent work on mindprints and writeprints:
Linguistic feature-based “fingerprints” in text indicating mental states and identity.

Language acquisition:
How do children extract information about language from the language
data they encounter?

Lidz & Gagliardi 2015
Sophisticated framework that makes explicit the
different components of the acquisition process.
The Computation of Language: Information processing

One way to think about the computation of language is from an information processing standpoint.

Language acquisition:
How do children extract information about language from the language data they encounter?

Input
lʊkətðəkɪɾi
“What’s that?”
“Do you see it?”

Internal representation
{l(look, at, the, kitty)}

Output
“Where’s the kitty?”

The Computation of Language: Information processing

One way to think about the computation of language is from an information processing standpoint.

Language acquisition:
How do children extract information about language from the language data they encounter?

Input
lʊkətðəkɪɾi
“What’s that?”
“Do you see it?”

Internal representation
{l(look, at, the, kitty)}

Output
“Where’s the kitty?”

Language acquisition: Methods of investigation

Theoretical methods:
What knowledge of language is (and what children have to learn)

LOOK at the KItty

lʊkətðəkɪɾi
Language acquisition: Methods of investigation

Experimental methods:
When knowledge is acquired, what the input looks like, & plausible capabilities underlying how acquisition works

Computational methods:
Strategies for how children acquire knowledge, sophisticated quantitative analysis of children’s input & output

Language acquisition: Representation & Development

Language acquisition involves complex knowledge that builds on itself over the course of linguistic development, embedded in a developing cognitive system.

This means there’s a natural dependence between theories of knowledge representation and theories of knowledge development.

Language acquisition: Foundational knowledge

Language acquisition involves complex knowledge that builds on itself over the course of linguistic development, embedded in a developing cognitive system.

Examples of “foundational” processes that children use for building more sophisticated knowledge:
speech segmentation
syntactic categorization

phonology
look at the kitty

Nouns = X

syntax
LOOK at the Kitty

look (me, the kitty)

semantics
Language acquisition: Foundational knowledge

Language acquisition involves complex knowledge that builds on itself over the course of linguistic development, embedded in a developing cognitive system.

Examples of “foundational” processes that children use for building more sophisticated knowledge:
- speech segmentation
- syntactic categorization

A recent finding: When the underlying representation (i.e., assumptions about language structure) is immature, immature processing capabilities may be helpful rather than harmful.


Language acquisition: More sophisticated knowledge

Language acquisition involves complex knowledge that builds on itself over the course of linguistic development, embedded in a developing cognitive system.

Examples of more sophisticated knowledge that depends on the foundational knowledge:
- metrical stress

A current finding: Some linguistic representations may be less acquirable from cognitively plausible child-directed input than previously assumed unless certain learning biases are in place.

Language acquisition: More sophisticated knowledge

Language acquisition involves complex knowledge that builds on itself over the course of linguistic development, embedded in a developing cognitive system.

Examples of more sophisticated knowledge that depends on the foundational knowledge:
- syntactic islands
- English anaphoric one
  where arguments appear syntactically

A current finding: The knowledge needed to create the right acquisitional intake may not necessarily look like we thought it did (e.g., what’s in Universal Grammar).

Examples of more sophisticated knowledge that depends on the foundational knowledge:
- syntactic islands
- English anaphoric one
  where arguments appear syntactically

Today’s Plan

Investigating Universal Grammar (UG)

Characterizing learning problems precisely enough to informatively model them

UG modeling forays
Motivating Universal Grammar

The argument from acquisition: one explicit motivation that highlights the natural link between linguistic representation and language acquisition.


What’s so hard about acquiring language?
There seem to be induction problems, given the available data.
(Poverty of the Stimulus, Logical Problem of Language Acquisition, Plato’s Problem)

So if the data themselves don’t pick out the right answer (and children all seem to), something internal to children must be guiding them.

If that something is both innate and domain-specific, we consider it part of Universal Grammar (UG) (Chomsky 1965, Chomsky 1975, Pearl & Sprouse 2013).
Motivating the contents of UG

Proposals have traditionally come from characterizing a specific acquisition problem for a particular linguistic phenomenon, and describing the (UG) solution to that specific characterization.

**Structure-dependent rules**

(Chomsky 1980, Anderson & Lightfoot 2000; Fodor & Crowther 2002; Berwick et al. 2011; Anderson 2013)

Pirates who can dance can often fight well.

Can pirates who can dance __ often fight well?

Motivating the contents of UG

Syntactic islands: Constraints on long-distance dependencies


Where did Jack think Lily bought the necklace from __?

*Where did Jack think the necklace from __ was too expensive?

Motivating the contents of UG

English anaphoric one representation

(Baker 1978, Pearl & Mis 2011, 2016)

Look – a red bottle! Do you see another one?
How to generate a learning theory proposal:
Characterize the learning problem precisely and identify a potential solution.

How to evaluate a learning theory proposal:
See if it’s successful when embedded in a model of the acquisition process for that learning problem.

Benefit of computational modeling:
We can make sure the learning problem is characterized precisely enough to implement. It’s not always obvious what pieces are missing until you try to build a model of the learning process.
(Pearl 2014, Pearl & Sprouse 2015)

Recently, in computational modeling, we’ve seen the integration of rich hypothesis spaces with probabilistic/statistical learning mechanisms (Sakas & Fodor 2001, Yang 2004, Pearl 2011, Dillon et al. 2013, Pearl & Sprouse 2013, Pearl et al. 2014, Pearl & Mis 2016, among many others).
**UG proposals: Generation & evaluation**

*How to generate a learning theory proposal:*
Characterize the learning problem precisely and identify a potential solution.

*How to evaluate a learning theory proposal:*
See if it’s successful when embedded in a model of the acquisition process for that learning problem.

We’ve also seen the development of more sophisticated acquisition frameworks that highlight the precise role of UG (Lidz & Gagliardi 2015).

Example: UG determines what data from the perceived input are relevant (acquisitional intake).

---

**UG proposals: Generation & evaluation**

*How to generate a learning theory proposal:*
Characterize the learning problem precisely and identify a potential solution.

*How to evaluate a learning theory proposal:*
See if it’s successful when embedded in a model of the acquisition process for that learning problem.

This computational modeling feedback helps us refine our theories about both the knowledge representation the learning theory relies on and the acquisition process that uses that representation.

---

**UG proposals: Generation & evaluation**

*How to generate a learning theory proposal:*
Characterize the learning problem precisely and identify a potential solution.

*How to evaluate a learning theory proposal:*
See if it’s successful when embedded in a model of the acquisition process for that learning problem.

*How to decide if any components of the proposal are UG:*
Examine the components of the successful learning solution.

Are they necessarily both domain-specific and innate?

*Note: We may use “innate” as a placeholder until we can determine if it’s impossible to derive the relevant component (Pearl 2014, Pearl & Mis 2016).*
UG proposal refinement: Recent successful forays

Syntactic islands (constraints on wh-dependencies):
Pearl & Sprouse 2013a, 2013b, 2015

English anaphoric one:
Pearl & Mis 2011, 2016

Recurring themes:
(1) Broadening the set of relevant data in the acquisitional intake
(2) Evaluating output by how useful it is

UG proposal refinement: Recent successful forays

Syntactic islands (constraints on wh-dependencies):
Pearl & Sprouse 2013a, 2013b, 2015

English anaphoric one:
Pearl & Mis 2011, 2016

Recurring themes:
(1) Broadening the set of relevant data in the acquisitional intake
(2) Evaluating output by how useful it is
(3) Not necessarily needing the prior knowledge we thought we did
Today's Plan

Investigating Universal Grammar (UG)

Characterizing learning problems precisely enough to informatively model them

UG modeling forays

Characterizing learning problems

Initial state:

- initial knowledge state
  - syntactic categories exist and can be identified
  - phrase structure exists and can be identified
  - participant roles can be identified

Agent, Patient, Goal, ...

- learning biases & capabilities
  - frequency information can be tracked
  - distributional information can be leveraged

h1, h2 more likely
Characterizing learning problems

Initial state: initial knowledge state + learning biases & capabilities

Data intake:

- encoding + acquisitional intake = data perceived as relevant for learning
  (Fodor 1998, Lidz & Gagliardi 2015)
  ex: all wh-utterances for learning about wh-dependencies
  ex: all pronoun data when learning about anaphoric one
  ex: syntactic and conceptual data for learning syntactic knowledge that links with conceptual knowledge
  [defined by knowledge & biases/capabilities in the initial state]

Learning period:

- how long children have to reach the target knowledge state (when inference & iteration happen)
  ex: 3 years, ~1,000,000 data points
  ex: 4 months, ~36,500 data points

Pearl & Sprouse 2015, Pearl & Mis 2016
Characterizing learning problems

Initial state: initial knowledge state + learning biases & capabilities

Data intake: data perceived as relevant for learning

Learning period: how long children have to learn

Target state:

- the knowledge children are trying to attain (as indicated by their behavior)

ex: *Where did Jack think the necklace from __ was too expensive?*

ex: one is category N' when it is not NP

ex: [image]

z-score rating

Target state: the knowledge children must attain

Once we have all these pieces specified, we should be able to implement an informative model of the learning process.

Informing UG (+ acquisition theory)

When we identify a successful learning strategy via modeling, this is an existence proof that children could solve that learning problem using the learning biases, knowledge, and capabilities comprising that strategy.

This identifies useful learning strategy components, which we can then examine to see where they might come from.
Today’s Plan

Investigating Universal Grammar (UG)

Characterizing learning problems precisely enough to informatively model them

UG modeling forays

Syntactic islands

• **Why?** Central to UG-based syntactic theories.

• **What?** Dependencies can exist between two non-adjacent items. They do not appear to be constrained by length (Chomsky 1965, Ross 1967), but rather by whether the dependency crosses certain structures (called “syntactic islands”).

```
What does Jack think __?
What does Jack think that Lily said that Sarah heard that Jareth believed __?
```

Syntactic islands: Acquisition target

Adult knowledge as measured by acceptability judgment behavior

```
Complex NP island:
*What did you make [the claim that Jack bought __]?

Subject island:
*What do you think [the joke about __] offended Jack?

Whether island:
*What do you wonder [whether Jack bought __]?

Adjunct island:
*What do you worry [if Jack buys __]?
```

Pearl & Sprouse 2013a, 2013b, 2015
Syntactic islands: Acquisition target

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)

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.

Importance for acquisition: This is one kind of target behavior that we’d like a learner to produce.
Syntactic islands: Representations

(1) A dependency cannot cross two or more bounding nodes.

(Wh ... [BN1 ... [BN2 ... ___] [CP, IP, NP]?)

Bounding nodes are language-specific
(CP, IP, and/or NP – must learn which ones are relevant for language)

---

**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)
(2) A dependency cannot cross a very low probability region of structure
(represented as a sequence of container nodes).

Container node: phrase structure node that contains dependency

{[Wh do you like ___ [NP in this picture?]]}

---

**In common:** Both rely on local structure anomalies (at some level)

---

Syntactic islands: Representations

(1) A dependency cannot cross two or more bounding nodes.

(Wh ... [BN1 ... [BN2 ... ___] [CP, IP, NP]?)

---

**Subjacency-ish** (Pearl & Sprouse 2013a, 2013b, 2015)
(2) A dependency cannot cross a very low probability region of structure
(represented as a sequence of container nodes).

---

Low probability regions are language-specific
(defined by sequences of container nodes that must be learned)
Syntactic islands: Representations

(1) A dependency cannot cross two or more bounding nodes.

\[
\text{Wh} \rightarrow \ldots [\text{BN}_1 \ldots [\text{BN}_2 \ldots \ldots \ldots]]
\]
(i) Dependencies defined over bounding nodes — track those
(ii) Bounding node = ?
(iii) 2+ bounding nodes = 

Subjacency-ish (Pearl & Sprouse 2013a, 2013b, 2015)
(2) A dependency cannot cross a very low probability region of structure
(represented as a sequence of container nodes).

\[
\text{Wh} \rightarrow [\text{CN}_1 \ldots [\text{CN}_2 \ldots [\text{CN}_3 \ldots [\text{CN}_4 \ldots \ldots \ldots]]]]
\]
(i) Dependencies defined over container node structure — track that already
(ii) Container node = ?
(iii) low probability = 

Different: Amount of language-specific knowledge built in just for islands
Pearl & Sprouse 2013a, 2013b, 2015

Syntactic islands: Subjacency-ish

Subjacency-ish implementation:
A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

\[
\text{Wh} \rightarrow [\text{CN}_1 \ldots [\text{CN}_2 \ldots [\text{CN}_3 \ldots [\text{CN}_4 \ldots \ldots \ldots]]]]
\]

Initial state:
(i) Dependencies defined over container node structure
(ii) Container nodes recognized
(iii) Track probability of short container node sequences (trigrams)
Pearl & Sprouse 2013a, 2013b, 2015

Subjacency-ish: Initial state implementation

Because \textit{wh}-dependencies are perceived as sequences of container nodes, local pieces of dependency structure can be characterized by container node trigrams.

\[
[\text{IP Who did} [\text{NP she} [\text{VP think} [\text{IP [the gift} [\text{VP was} [\text{VP from} \ldots]]]]]]]]
\]

\begin{align*}
\text{begin-IP-VP-CP} & \rightarrow \text{IP-VP-PP-end} \\
\text{begin-IP-VP} & \rightarrow [\text{IP-VP-CP}]
\end{align*}

Because wh-dependencies are perceived as sequences of container nodes, local pieces of dependency structure can be characterized by container node trigrams.

\[
[\text{IP Who did} [\text{NP she} [\text{VP think} [\text{IP [the gift} [\text{VP was} [\text{VP from} \ldots]]]]]]]
\]

\begin{align*}
\text{begin-IP-VP-CP} & \rightarrow \text{IP-VP-PP-end} \\
\text{begin-IP-VP} & \rightarrow [\text{IP-VP-CP}]
\end{align*}
A child learns about the frequency of container node trigrams...

begin-IP-VP

IP-VP null

...and at the end of the learning period has a sense of the probability of any given container node trigram, based on its relative frequency.

Subjacency-ish: Developing knowledge

Any wh-dependency can then have a probability, based on the product of the smoothed probabilities of its trigrams.

This allows the modeled learner to generate judgments about the grammaticality of any dependency.

Higher probability dependencies are more grammatical while lower probability dependencies are less grammatical.
Syntactic islands: Subjacency-ish

Subjacency-ish input & intake:
A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

Data intake: defined by initial state =
wh-dependencies in child-directed speech, as characterized by container nodes

But which wh-dependencies? Just the ones being evaluated in the target state?

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 __?

Learning period: defined by empirical estimates from Hart & Risley (1995) (~3 years of data)
>= 200,000 wh-dependency data points

Syntactic islands: Subjacency-ish

Subjacency-ish input & intake:
A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

Data intake: defined by initial state =
wh-dependencies in child-directed speech, as characterized by container nodes

But which wh-dependencies? Just the ones being evaluated in the target state?

No! Any wh-dependency has relevant information about container node trigrams used to determine the grammaticality of wh-dependencies in general.

Syntactic islands: Subjacency-ish

Subjacency-ish input & intake:
A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

Data intake: defined by initial state =

Syntactic islands: Subjacency-ish

Subjacency-ish input & intake:
A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

Data intake: defined by initial state =

Learning period: defined by empirical estimates from Hart & Risley (1995) (~3 years of data)
>= 200,000 wh-dependency data points
Syntactic islands: Subjacency-ish

Subjacency-ish input & intake:
A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

\[ \text{Wh} \quad [\text{CN1} \quad \ldots \quad [\text{CN2} \ldots [\text{CN3} \ldots [\text{CN4} \ldots [\text{CN5} \ldots \ldots ]]]] \]

Target state: Behavioral evidence of syntactic islands knowledge

Non-parallel lines indicate superadditivity, which indicates knowledge of islands.

But how do we get acceptability judgment equivalents?

\[ \text{Who} \quad _{-1} \quad _{-0.5} \quad _{0} \quad _{0.5} \quad _{1} \quad _{1.5} \quad _{2} \quad z \quad \text{score rating} \]

\[ \text{matrix embedded} \]

\[ \text{island structure} \quad \text{non-island structure} \quad \text{island effect} \]

\[ \text{embedded} \]

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.

Syntactic islands: Subjacency-ish

Subjacency-ish input & intake:
A dependency cannot cross a very low probability region of structure (represented as a sequence of container nodes).

\[ \text{Wh} \quad [\text{CN1} \quad \ldots \quad [\text{CN2} \ldots [\text{CN3} \ldots [\text{CN4} \ldots [\text{CN5} \ldots \ldots ]]]] \]

Target state: Behavioral evidence of syntactic islands knowledge

Non-parallel lines indicate superadditivity, which indicates knowledge of islands.

But how do we get acceptability judgment equivalents?

\[ \text{Who} \quad _{-1} \quad _{-0.5} \quad _{0} \quad _{0.5} \quad _{1} \quad _{1.5} \quad _{2} \quad z \quad \text{score rating} \]

\[ \text{matrix embedded} \]

\[ \text{island structure} \quad \text{non-island structure} \quad \text{island effect} \]

\[ \text{embedded} \]

The Subjacency-ish representation that relies on container node trigram probabilities can solve this learning problem.
**Subjacency-ish: Take away**

If dependencies are represented as container node sequences, acquisition works well for these four syntactic islands.

**Subjacency-ish vs. Subjacency: What’s in UG?**

**UG = innate + domain-specific**

<table>
<thead>
<tr>
<th>Innate</th>
<th>Derived</th>
<th>Domain-specific</th>
<th>Domain-general</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Innate

Attend to container nodes of a particular kind

Low probability items are dispreferred

Derived

Attend to bounding nodes (BNs)

Dependencies crossing 2+ BNs are not allowed

**Recurring themes: Syntactic islands**

Informing theories of representation & acquisition

Recurring themes, as seen in syntactic island acquisition:

1. Broadening the set of relevant data in the acquisitional intake to include all wh-dependencies

2. Evaluating output by how useful it is for generating acceptability judgment behavior

*"Wh ... [CN1 -> CN2 -> CN3 -> CN4 -> VN5 -> ...]"*
Recurring themes: Syntactic islands

Informing theories of representation & acquisition

Recurring themes, as seen in syntactic island acquisition:

1. Broadening the set of relevant data in the acquisitional intake to include all wh-dependencies
2. Evaluating output by how useful it is for generating acceptability judgment behavior
3. Not necessarily needing the prior knowledge we thought we did in UG: container nodes rather than bounding nodes, no domain-specific constraint on length

Open questions

This learning strategy relying on the Subjacency-ish representation for wh-dependencies makes some developmental predictions – can we verify these experimentally?

“that-trace” effect prediction:
Children initially disprefer all dependencies containing that, even ones adults allow, due to the infrequency of container node trigrams with CP that in child-directed speech

Subject extraction
*Who do you think that ___ read the book?
Who do you think ___ read the book?

Object extraction
What do you think that he read ___?
What do you think he read ___?

Open questions

This learning strategy relying on the Subjacency-ish representation for wh-dependencies makes some developmental predictions – can we verify these experimentally?

“that-trace” effect prediction:
Children initially disprefer all dependencies containing that, even ones adults allow, due to the infrequency of container node trigrams with CP that in child-directed speech

Subject extraction
*Who do you think that ___ read the book?
Who do you think ___ read the book?

Object extraction
What do you think that he read ___?
What do you think he read ___?
Open questions

How does this learning strategy for wh-dependencies measure up cross-linguistically?

Island effects vary.
Ex: Italian does not have a subject island effect when the wh-dependency is part of a relative clause, though it does when the wh-dependency is part of a question. (Sprouse et al. in press)

Would the input naturally lead the Subjacency-ish learner to this distinction?

Open questions

Can we extend this learning strategy to create an integrated theory of syntactic acquisition?

Related phenomena: The distribution of gaps

Parasitic gaps: 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 [before reading ___]?  Adjunct island
Which book did you judge [true] [before reading ___parasitic___]?

Open questions

Can we extend this learning strategy to create an integrated theory of syntactic acquisition?

Related phenomena: The distribution of gaps

Across-the-board (ATB) extraction: Similar situation.

Which book did you [(read ___) and (then review ___)]?
dependency for both gaps: IP-VP-VP

*Which book did you [(read the paper) and (then review ___)]?
dependency for gap: IP-VP-VP

*Which book did you [(read ___) and (then review the paper)]?
dependency for gap: IP-VP-VP

Open questions

Can we extend this learning strategy to create an integrated theory of syntactic acquisition?

Semi-related phenomena: Binding dependencies

There don’t appear to be the same restrictions on binding dependencies that there are on wh-dependencies.

The boy thought the joke about himself was really funny.

*Who did the boy think [the joke about ___] was really funny?  Subject island
Today’s Plan

- Investigating Universal Grammar (UG)
- Characterizing learning problems precisely enough to informatively model them
- UG modeling forays

English anaphoric one

- Why? A traditional poverty-of-the-stimulus problem used to motivate specific proposals for the contents of UG.
- What? Look - a red bottle!
- Do you see another one?
- Why? A traditional poverty-of-the-stimulus problem used to motivate specific proposals for the contents of UG.
- What? Look - a red bottle!
- Do you see another one?

Process of interpretation: First determine the linguistic antecedent of one (what expression one is referring to) based on its syntactic category.
→ antecedent of one = “red bottle”
English anaphoric *one*

- **Why?** A traditional poverty-of-the-stimulus problem used to motivate specific proposals for the contents of UG.

- **What?** Look - a red bottle! Do you see another *one*?

Process of interpretation: Because the antecedent ("red bottle") includes the modifier "red", the property RED is important for the referent of *one* to have.

→ referent of *one* = RED BOTTLE

English anaphoric *one*: Acquisition target

Two steps:
1. Identify linguistic antecedent (based on *one*'s syntactic category)
2. Identify referent (based on linguistic antecedent)

Adult Knowledge

Standard linguistic theory (Chomsky 1970, Jackendoff 1977) has posited that *one* in these kinds of utterances is a syntactic category smaller than an entire noun phrase (NP), but larger than just a noun (N°). This category has been called N°, and includes strings like "bottle" and "red bottle".
Because one is thought to be this same category (N'), available adult interpretations for one include both “Do you see another bottle?” and “Do you see another red bottle?” Additional preferences allow adults to choose the appropriate interpretation from these options in context.
English anaphoric one: Acquisition target

Child knowledge as measured by looking time behavior

Look - a red bottle!

Now look...

Child behavior at 18 months: Lidz et al. 2003

Control/Noun:
“Do you see another bottle?”
“Do you see another red bottle?”
Prefer to look at novel bottle.
(0.459 to same color)

Anaphoric/Adjective-Noun:
“Do you see another one?”
“Do you see another red bottle?”
Prefer to look at same color bottle.
(0.587 to same color)

Preference

Adjusted familiarity preference
Average probability of looking to same color bottle: 0.587

Prefer to look at same color bottle.

Developed knowledge according to Lidz et al. 2003: 18-month-olds interpret one’s antecedent as “red bottle” (an N’) and its referent as the RED BOTTLE.
English anaphoric one: Acquisition target

Target state for acquisition: knowledge and behavior

Look - a red bottle!
Now look...

Child behavior at 18 months: Lidz et al. 2003

Control/Noun:
“What do you see now?”
“Do you see another bottle?”
Prefer to look at novel bottle
(0.459 to same color)

Anaphoric/Adjective-Noun:
“Do you see another one?”
“Do you see another red bottle?”
Prefer to look at same color bottle
(0.587 to same color)

Developed knowledge according to Lidz et al. 2003:
18-month-olds interpret one’s antecedent as “red bottle” (an N) and its referent as the RED BOTTLE.

English anaphoric one: Representations

Proposed solutions for necessary knowledge & learning biases

Things in common:

- **Syntactic categories** exist (particularly NP, N’, and N°), and can be recognized.
- Anaphoric elements like **one** take linguistic antecedents of the same category.

Things that differ:

- Which input is considered relevant from the perceptual intake = acquisitional intake

Proposed solutions for necessary knowledge & learning biases

Things in common:

- **Syntactic categories** exist (particularly NP, N’, and N°), and can be recognized.
**English anaphoric one: Representations**

**Proposed solutions for necessary knowledge & learning biases**

Things that differ:
- Which input is considered relevant from the perceptual intake = acquisitional intake

**Baker (1978): One that won’t work = DirUnamb**

Only utterances directly using one are relevant for learning about anaphoric one.

Only utterances where one's antecedent is unambiguous are relevant.

**DirUnamb**: specific combination of utterance and situation

"Look – a red bottle! Hmmm - there doesn’t seem to be another one here, though."

- Children already know that one can’t be N°, so it must be N'.

This solves the problem of one’s syntactic category.

---

**English anaphoric one: Representations**

**Proposed solutions for necessary knowledge & learning biases**

Things that differ:
- Which input is considered relevant from the perceptual intake = acquisitional intake

**Baker (1978): One that won’t work = DirUnamb**

Only utterances directly using one are relevant for learning about anaphoric one.

Only utterances where one's antecedent is unambiguous are relevant.

Why won’t it work? The direct unambiguous data are too sparse. There's nothing to learn from.

Pearl & Mis 2011, 2016 affirmation:

0 examples in the 17,521 utterances in the Brown-Eve corpus (Brown 1973) from CHILDES.

---

**English anaphoric one: Representations**

**Proposed solutions for necessary knowledge & learning biases**

Things that differ:
- Which input is considered relevant from the perceptual intake = acquisitional intake

**Baker (1978): One that could work = DirUnamb + N’**

Only utterances directly using one are relevant for learning about anaphoric one.

Only utterances where one's antecedent is unambiguous are relevant.

Children already know that one can’t be N°, so it must be N’.

This solves the problem of one’s syntactic category.

---

**English anaphoric one: Representations**

**Proposed solutions for necessary knowledge & learning biases**

Things that differ:
- Which input is considered relevant from the perceptual intake = acquisitional intake

**Pearl & Lidz 2009: One that doesn’t work = DirEO**

Only utterances directly using one are relevant for learning about anaphoric one.

Use probabilistic inference to leverage ambiguous information about one.

All ambiguous data are relevant (Equal Opportunity).
English anaphoric one: Representations

Proposed solutions for necessary knowledge & learning biases

Things that differ:

- Which input is considered relevant from the perceptual intake = acquisitional intake

Pearl & Lidz 2009: One that doesn’t work = DirEO
Only utterances directly using one are relevant for learning about anaphoric one.
Use probabilistic inference to leverage ambiguous information about one.
All ambiguous data are relevant (Equal Opportunity).

DirRefSynAmb: Ambiguous about whether antecedent is “bottle” (N₀, N') or “red bottle” (N').
“Look – a red bottle! Oh, look – another one!”
0.66% of utterances containing a pronoun in Brown-Eve corpus

Pearl & Mis 2011, Pearl & Mis 2016

Pearl & Lidz 2009, Regier & Gahl 2004: One that does work for target knowledge = DirFiltered
Only utterances directly using one are relevant for learning about anaphoric one.
Use probabilistic inference to leverage ambiguous information about one.
Filter out the harmful DirSynAmb data.

DirSynAmb: Ambiguous about antecedent category (bottle = N₀, N').
“Look – a bottle! Oh, look – another one!”
7.52% of utterances containing a pronoun in Brown-Eve corpus

Turn out to be harmful to learning - they cause the learner to think one's category should be N₀.

Pearl & Mis 2011, Pearl & Mis 2016

Turn out to be harmful to learning - they cause the learner to think one's category should be N₀.

Pearl & Mis 2011, Pearl & Mis 2016
English anaphoric *one*: Representations

Proposed solutions for necessary knowledge & learning biases

Things that differ:

- Which input is considered relevant from the perceptual intake = acquisitional intake

Pearl & Mis 2011, 2016: One that could work = IndirPro

Utterances directly using *one* are relevant for learning about anaphoric *one*.

Use probabilistic inference to leverage ambiguous information about *one*.

Utterances using other pronouns anaphorically are relevant for learning about anaphoric *one*.

This is indirect evidence coming from other pronouns.

IndirUnamb: Relevant because indicates whether antecedent includes the mentioned property (it always does here), which is helpful when choosing between different interpretation options in other contexts.

“Look – a red bottle! I want *one/*it.”

8.42% of utterances containing a pronoun in Brown-Eve corpus

Learning proposal comparisons

<table>
<thead>
<tr>
<th>DirUnamb</th>
<th>one</th>
<th>ProInf</th>
<th>-DirSynAmb</th>
<th>+OtherPro</th>
<th>Unamb</th>
<th>one</th>
<th>ProInf</th>
<th>-DirSynAmb</th>
<th>+OtherPro</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Successful?

□ □

□ □

□ □

□ □

Pearl & Mis 2011, Pearl & Mis 2016
**English anaphoric one: Representations**

**Proposed solutions for necessary knowledge & learning biases**

Things that differ:
- Which input is considered relevant from the perceptual intake = *acquisitional intake*

Learning proposal comparisons

<table>
<thead>
<tr>
<th>Unamb</th>
<th><em>one</em>/=N?</th>
<th>ProblInf</th>
<th>~DirSynAmb</th>
<th>+OtherPro</th>
<th>Successful?</th>
<th>Representations</th>
<th>Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>DirUnamb</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DirUnamb + N</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DirFiltered</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IndirPro</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Pearl & Mis 2011, Pearl & Mis 2016*

---

**English anaphoric one: Data intake**

**Data intake: The data relevant for learning**

Data potentially in the acquisitional intake

<table>
<thead>
<tr>
<th>Data type</th>
<th>Example</th>
<th>Learning strategies using these data</th>
</tr>
</thead>
<tbody>
<tr>
<td>DirUnamb</td>
<td>Look - a red bottle! There isn’t another one here, though.</td>
<td>DirUnamb, DirUnamb + N, DirFiltered, DirEO, IndirPro</td>
</tr>
<tr>
<td>DirRefSynAmb</td>
<td>Oh, look – another one!</td>
<td>DirFiltered, DirEO, IndirPro</td>
</tr>
<tr>
<td>DirSynAmb</td>
<td>Look - a bottle!</td>
<td>DirEO, IndirPro</td>
</tr>
<tr>
<td>IndirUnamb</td>
<td>Look a red bottle! I want it, please.</td>
<td>IndirPro</td>
</tr>
</tbody>
</table>
English anaphoric one: Learning period

Learning period: How long children have to learn = how much data

Before this learning process can begin, children need to know something about syntactic categories. Experimental data from Booth & Waxman (2003) suggests they recognize linguistic markers of categories like Noun and Adjective around 14 months.

Beginning: 14 months
End: 18 months

Using empirical estimates from Hart & Risley (1995), we can estimate this as approximately 36,500 data points containing an anaphoric pronoun.

Beginning: 14 months
End: 18 months = 4 months’ worth of data

English anaphoric one: Target state

Target state: knowledge and behavior

Child behavior at 18 months: Lidz et al. 2003

Control/Noun: “What do you see now?” “Do you see another bottle?” Prefer to look at novel bottle (0.459 to same color)

Anaphoric/Adjective-Noun: “Do you see another one?” “Do you see another red bottle?” Prefer to look at same color bottle (0.587 to same color)

Developed knowledge according to Lidz et al. 2003: 18-month-olds interpret one’s antecedent as “red bottle” (an N’) and its referent as the RED BOTTLE.
Model of understanding a referential expression involving an anaphoric pronoun, which includes both syntactic information and referential information when determining the antecedent which then picks out the referent.

Developed knowledge according to Lidz et al. 2003: 18-month-olds interpret one’s antecedent as "red bottle" (an N) and its referent as the RED BOTTLE.

Control/Noun: “What do you see now?” “Do you see another bottle?” Prefer to look at novel bottle. (0.459 to same color)

Anaphoric/Adjective-Noun: “Do you see another one?” “Do you see another red bottle?” Prefer to look at same color bottle. (0.587 to same color)

$p_{beh}$ = probability of producing target behavior (looking to same color bottle)

$p_{beh} = \frac{\alpha + \beta}{1 + \frac{1}{\beta}}$

$d_0 = 1$ for every data point encountered

$p_{beh} = \frac{\alpha + \beta}{1 + \frac{1}{\beta}}$

$p_{beh} = \frac{\alpha + \beta}{1 + \frac{1}{\beta}}$

$p_{beh} = \frac{\alpha + \beta}{1 + \frac{1}{\beta}}$
English anaphoric *one*: Learning results

Averages over 1000 simulations, standard deviations in parentheses.

Note: Target $p_{\text{beh}} = 0.587$, all other target $p = 1.000$

<table>
<thead>
<tr>
<th></th>
<th>DirUnamb</th>
<th>DirUnamb + N'</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_N$</td>
<td>0.500 (&lt;0.01)</td>
<td>0.500 (&lt;0.01)</td>
</tr>
<tr>
<td>$p_{\text{incl}}$</td>
<td>0.500 (&lt;0.01)</td>
<td>0.500 (&lt;0.01)</td>
</tr>
<tr>
<td>$p_{\text{beh}}$</td>
<td>0.475 (&lt;0.01)</td>
<td>0.475 (&lt;0.01)</td>
</tr>
<tr>
<td>$p_{\text{rep}/\text{beh}}$</td>
<td>0.158 (&lt;0.01)</td>
<td>0.158 (&lt;0.01)</td>
</tr>
</tbody>
</table>

Note: Target $p_{\text{beh}} = 0.587$, all other target $p = 1.000$

A learner who only looks at direct unambiguous data has no data to learn from, so it learns nothing. (Poverty of the stimulus.)

It’s at chance for having the target *syntactic* and *referential* knowledge necessary to choose the correct antecedent.

It doesn’t generate the observed toddler looking preference, and it’s unlikely to have the target representation if it looks at the familiar bottle.

Implication: Something else is needed.  
(Baker (1978)’s original observation)

What if the learner also knows that *one* is category N'? (Baker 1978)
**English anaphoric one: Learning results**

Averages over 1000 simulations, standard deviations in parentheses.

Note: Target $p_{\text{beh}} = 0.587$, all other target $p = 1.000$

<table>
<thead>
<tr>
<th></th>
<th>DirUnamb</th>
<th>DirUnamb + $N'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{N'}}$</td>
<td>0.500</td>
<td>0.991 (&lt;0.01)</td>
</tr>
<tr>
<td>$p_{\text{incl}}$</td>
<td>0.500</td>
<td>0.963 (&lt;0.01)</td>
</tr>
<tr>
<td>$p_{\text{beh}}$</td>
<td>0.475</td>
<td>0.574 (&lt;0.01)</td>
</tr>
<tr>
<td>$p_{\text{rep/beh}}$</td>
<td>0.158</td>
<td>0.918 (&lt;0.01)</td>
</tr>
</tbody>
</table>

This learner still has no data to learn from, so it learns nothing about the correct referential knowledge necessary to choose the correct antecedent.

This lack of referential knowledge causes it not to generate the observed toddler looking preference in context, and even if it happens to look at the familiar bottle, to be unlikely to have the target representation when doing so.

**Implication:** Knowing one is category $N'$ isn’t sufficient to generate target behavior if only direct unambiguous data are relevant.

---

The DirFiltered learner (Regier & Gahl 2004, Pearl & Lidz 2009) believes *one* is *N’* when it is smaller than NP and a mentioned property should be included in the antecedent, as found previously.

It’s also close to generating the observed toddler looking preference, and is likely to have the target representation when looking at the familiar bottle.

**Implication:** This new finding suggests this is a pretty successful learning strategy for matching the available behavioral data.
English anaphoric one: Learning results

Averages over 1000 simulations, standard deviations in parentheses.
Note: Target $p_{beh} = 0.587$, all other target $p = 1.000$

<table>
<thead>
<tr>
<th></th>
<th>DirUnamb</th>
<th>DirUnamb + N'</th>
<th>DirFiltered</th>
<th>DirEO</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{N'}$</td>
<td>0.500 (±0.01)</td>
<td>1.000</td>
<td>0.991 (±0.01)</td>
<td>0.246 (±0.01)</td>
</tr>
<tr>
<td>$p_{incl}$</td>
<td>0.500 (±0.01)</td>
<td>0.500 (±0.01)</td>
<td>0.963 (±0.01)</td>
<td>0.379 (±0.05)</td>
</tr>
<tr>
<td>$p_{beh}$</td>
<td>0.475 (±0.01)</td>
<td>0.492 (±0.01)</td>
<td>0.574 (±0.01)</td>
<td>0.464 (±0.01)</td>
</tr>
<tr>
<td>$p_{rep</td>
<td>beh}$</td>
<td>0.158 (±0.01)</td>
<td>0.306 (±0.01)</td>
<td>0.918 (±0.01)</td>
</tr>
</tbody>
</table>

The DirEO learner (explored by Pearl & Lidz 2009) prefers one to be N0 when it is smaller than NP, and does not believe the mentioned property should be included in the antecedent. Neither of these is the target knowledge.

This causes the learner not to generate the observed toddler looking preference, and not to have the target representation if it looks at the familiar bottle.

**Implication:** This new finding suggests this isn’t a good learning strategy for matching the available behavioral data.

The IndirPro learner robustly decides the antecedent should include the mentioned property. However, this learner has a moderate dispreference for believing one is N' when it is smaller than NP. This isn’t the target representation, w.r.t syntactic category.

**Why?** The learner believes very strongly that the mentioned property must be included in the antecedent.

Only one antecedent allows this: $[\text{i.e., red}, \text{in}, \text{the bottle}]$
**English anaphoric one:** Learning results

Averages over 1000 simulations, standard deviations in parentheses.
Note: Target $p_{beh} = 0.587$, all other target $p = 1.000$

<table>
<thead>
<tr>
<th></th>
<th>DirUnamb</th>
<th>DirUnamb + $N'$</th>
<th>DirFiltered</th>
<th>DirEO</th>
<th>IndirPro</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_N'$</td>
<td>0.500</td>
<td>1.000</td>
<td>0.991</td>
<td>0.246</td>
<td>0.368</td>
</tr>
<tr>
<td>$p_{incl}$</td>
<td>0.500</td>
<td>0.500</td>
<td>0.963</td>
<td>0.379</td>
<td>1.000</td>
</tr>
<tr>
<td>$p_{beh}$</td>
<td>0.475</td>
<td>0.492</td>
<td>0.574</td>
<td>0.587</td>
<td>(0.01)</td>
</tr>
<tr>
<td>$p_{rep/beh}$</td>
<td>0.158</td>
<td>0.306</td>
<td>0.918</td>
<td>0.050</td>
<td>0.998</td>
</tr>
</tbody>
</table>

So, because the antecedent includes the mentioned property, it and the pronoun referring to it (one) must be $N'$ in this context - even if the learner believes one is not $N'$ in general.

Only one antecedent allows this: $[n_r \text{ red} [n_t \text{bottle}]]$

Let’s look at the strategies that worked and see what the implications are for Universal Grammar, as compared to the original UG proposal by Baker that didn’t work.

**Implication:** A learner viewing other pronoun data as relevant can generate target behavior without necessarily reaching the target knowledge state – instead, this learner has a context-sensitive representation (depending on whether a property was mentioned).
English anaphoric *one*: Strategy components

**Things in common:**
It may be possible to *derive* the domain-specific knowledge of the specific syntactic categories needed using distributional clustering techniques over words...but that remains to be shown.

Some innate knowledge may be necessary (UG).

Pearl & Mis 2011, Pearl & Mis 2016

Similarly, the preference to use probabilistic inference to leverage the information in ambiguous data seems likely to be innate and domain-general.

While this is a new strategy component, it’s unlikely to be part of UG.

Pearl & Mis 2011, Pearl & Mis 2016
English anaphoric one: Strategy components

- **DirFiltered**
  - Filter out
  - DirSynAmb

- **IndirPro**
  - + Indirect evidence = pronouns
  - Probabilistic inference

**Syntactic categories**
- Antecedent = Same Category
- + Direct positive evidence

**Utterances**
- "Look – a bottle!
Oh, look – another one!"

**Old unsuccessful proposal**

The domain-specific knowledge that one is not category N₀ was thought to be innate and so part of UG.

**Successful DirFiltered proposal**

The domain-specific preference to filter out data where only the syntactic category is uncertain (while the referent is clear) may be innate and so part of UG, or it may be derived from an innate, domain-general preference to learn in cases of uncertainty (Pearl & Lids 2009).

**DirSynAmb:** Ambiguous about antecedent category (bottle = N₀, N').

*Look – a bottle! Oh, look – another one!*

**Domain-specific**
- Universal Grammar
- Innate
- Derived
- Domain-specific

**Domain-general**
- Innate
- Derived
- Universal Grammar
English anaphoric *one*: Strategy components

**Successful IndirPro proposal**

The **domain-specific** knowledge to consider other pronouns relevant may be **innate** and so part of **UG** or it may **derive** from an overhypothesis (Kemp et al 2007) the learner forms about the similarity of *one* with other anaphoric pronouns in terms of their distribution.

**Both successful proposals**

The new components required may not necessarily need to be built into **UG**. However, if they are, they are **less-specific** knowledge than the previous proposal supposed (which didn’t actually capture children’s behavior anyway).

**Some open questions**

For each component that may be **derivable** from the input, can we create a learner than can actually derive that component from the available linguistic information? And if so, what are the learning components required to do so?

How general-purpose are these learning components? Are the components we find useful for making syntactic generalizations about anaphoric one useful for making other syntactic generalizations? What about other linguistic generalizations? Or other non-linguistic generalizations?
Recurring themes: English anaphoric one

Informing theories of representation & acquisition

Recurring themes:
1. Broadening the set of relevant data in the acquisitional intake to include all pronouns
2. Evaluating output by how useful it is for generating toddler looking time behavior
3. Not necessarily needing the prior knowledge we thought we did in UG: “good enough” derived data filter or derived overhypothesis about pronouns rather than specific knowledge about syntactic category

Big picture:
Understanding how children acquire syntactic knowledge

If we precisely define the components of any acquisition task by drawing on the insights from different methodologies, we can make progress on how children solve that acquisition task.

In particular, we can understand the nature of children’s language acquisition toolkit — what fundamental building blocks they use are, and what is (or is not) part of Universal Grammar.

Theoretical methods
Computational methods
Experimental methods
Big picture: Understanding how children acquire syntactic knowledge

If we precisely define the components of any acquisition task by drawing on the insights from different methodologies, we can make progress on how children solve that acquisition task.

In particular, we can understand the nature of children's language acquisition toolkit — what fundamental building blocks they use are, and what is (or is not) part of Universal Grammar.

This technique is a useful tool — so let's use it to inform our theories of representation and acquisition!

Thank you!

Jon Sprouse  Benjamin Mis
Greg Carlson  LouAnn Gerken  Jeff Lidz
Computational Models of Language Learning seminar, UC Irvine 2010


This work was supported in part by NSF grants BCS-0843896 and BCS-1147028.