How Far Can Indirect Evidence Take Us? Anaphoric *One* Revisited

Lisa Pearl & Benjamin Mis, University of California, Irvine

Overview

One of the most controversial claims in theoretical and developmental linguistics: children learning their native language face an **induction problem** ("poverty of the stimulus"), where the data in children's input are insufficient to identify the correct language knowledge as quickly as children do (Baker 1981, Chomsky 1980, Chomsky 1988, Crain 1991, Dresher 2003, Hornstein & Lightfoot 1981, Legate & Yang 2002, Lightfoot 1982, Lightfoot 1989).

Implication: **Children have helpful learning biases.**

Open question: The nature of those biases
Some dimensions of variation:
- Innate vs.. Derived
- Domain-specific vs. Domain-general
- What to learn vs. How to learn

Universal Grammar (Chomsky 1965) = learning biases that are innate and domain-specific.

Main question

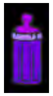
When induction problems exist, what does it take to solve them?
- What indirect evidence is available?
- Are the necessary biases innate and domain-specific?

Anaphoric *one*

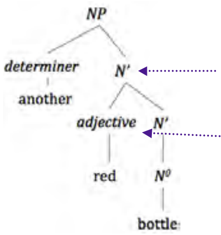
One linguistic phenomenon argued to present an induction problem.

"Look - a red bottle!"

"Do you see another one?"



Adult response



$[_{NP}$ **another** $[_{N'}$ **red** $[_{N'}\ [_{N0}$ **bottle**$]]]]$

Adult representation:
Syntactic antecedent = "red bottle"
(*one* = N')
Semantic referent =
RED BOTTLE
(modifier property is important)

Child behavior:
Same at 18 months as adults
(Lidz, Waxman, & Freedman 2003)

Assumption: 18-month-old representation is same as adult representation.

Question: How do children learn to produce this behavior and have this representation, given the data they're exposed to?

Why is this a potential induction problem? The direct evidence available

Most data children encounter are ambiguous

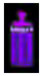
Syntactically ambiguous (SYN)
"Look - a bottle!"
"Oh look - another one!"

one's referent = BOTTLE
one's antecedent = $[_{N'}[_{N0}$ **bottle**$]]$ or $[_{N0}$ **bottle**$]$?

Semantically + syntactically ambiguous (SEM-SYN)
"Look - a red bottle!"
"Oh look - another one!"

one's referent = RED BOTTLE or BOTTLE?
one's antecedent = $[_{N'}$ **red**$[_{N'}[_{N0}$ **bottle**$]]]$ or $[_{N'}[_{N0}$ **bottle**$]]$ or $[_{N0}$ **bottle**$]$?

Unambiguous data are rare (requiring a specific coincidence of utterance and situation)



Unambiguous (UNAMB)
"Look - a red bottle! Hmmm - there doesn't seem to be another one around, though."

one's referent = BOTTLE? If so, *one*'s antecedent = "bottle". But it's strange to claim there's not another *bottle* here, since there clearly is another bottle.
So, *one*'s referent must be RED BOTTLE, and *one*'s antecedent = $[_{N'}$ **red**$[_{N'}[_{N0}$ **bottle**$]]]$.

Previous proposals for using the direct evidence: Input restrictions

Baker 1978 (**Baker**):
Only UNAMB data are informative. Children must have innate, domain-specific knowledge that *one* cannot be category N⁰ because UNAMB data rarely occur.

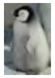
Regier & Gahl 2004 (**R&G**):
Leverage SEM-SYN ambiguous data in addition to UNAMB data. Children use innate, domain-general statistical learning abilities to track suspicious coincidences in the properties that *one*'s referents have.

Pearl & Lidz 2009 (**P&L**):
Filter out SYN ambiguous data, even if using SEM-SYN ambiguous data - otherwise, children will learn *one* is category N^0. Children employ a domain-specific bias to ignore these data, which can be derived from an innate domain-general preference for learning in cases of local uncertainty.

A potential source of indirect evidence

Pearl & Mis (P&M) Observation: Other pronouns can also be used anaphorically.

"Look! A cute penguin.
I want to hug it."



$[_{NP}$ **a** $[_{N'}$ **cute** $[_{N'}\ [_{N0}$ **penguin**$]]]]$ ⟶ $[_{NP}$ **it**$]$

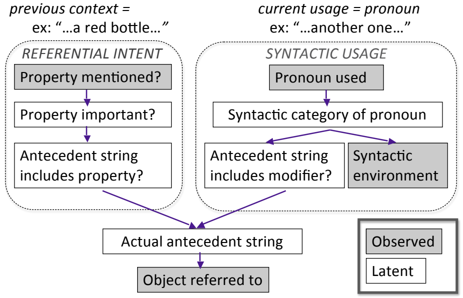
They are always category NP (UNAMB NP), as evidenced by their syntactic environment and antecedent. The referent is unambiguous w.r.t. to having the mentioned property - i.e., the referent must have the property mentioned (e.g., *cute*). Note that *one* can also be category NP sometimes.

"Look! A cute penguin. I want one."
≈ Look! A cute penguin. I want *a cute penguin*."

Why are these data useful?
They can help a child decide in general if the referent of an anaphoric pronoun should have a property mentioned in the potential syntactic antecedent.

P&M proposal:
Use UNAMB NP data in addition to all the other data. To do this: The child must recognize *one* is similar to other anaphoric pronouns (look at syntactic and semantic distribution of *one* and other pronouns, using innate domain-general statistical learning abilities).

Modeling information in the data



Sample values: SEM-SYN ambiguous data (*...a red bottle...another one...*)

Property mentioned? Yes	Pronoun used: *one*
Property important? Yes or No	Syntactic category: N' or N⁰
Antecedent includes prop? Yes or No	Antecedent includes modifier? Yes or No
	Syntactic environment: smaller than NP

Actual antecedent: "red bottle" or "bottle"
Object referred to: has property mentioned (RED BOTTLE)

Important attributes for correct representation

Syntactic category = N', if not NP
p(syntactic category of pronoun = N' | syntactic environment indicates category is smaller than NP)
= $p_{N'}$

Referent should have property mentioned in potential antecedent
p(property important? = yes | property mentioned = yes)
= p_I

Online probabilistic framework

General form of update equations (Chew 1971)

data seen suggesting x is true

$$p_x = \frac{\alpha + data_x}{\alpha + \beta + totaldata_x}, \alpha = \beta = 1$$

A very weak prior

total informative data seen w.r.t x

After every informative data point encountered:

$$data_x = data_x + \phi_x$$

Incremented by probability that data point suggests x is true

$$totaldata_x = totaldata_x + 1$$

One informative data point seen

$$\phi_{N'} = p(N'|\pi, \sigma =< NP, \mu, \omega) = \frac{p(\pi, \mu, \omega | N', \sigma =< NP) * p_{N'}}{p(\pi, \mu, \omega | \sigma =< NP)}$$

$$\phi_I = p(I|\pi, \sigma, \mu = yes, \omega) = \frac{p(\pi, \sigma, \omega | I, \mu = yes) * p_I}{p(\pi, \sigma, \omega | \mu = yes)}$$

I: property important=yes N': syntactic category=N'
π: what pronoun was mentioned; σ: what the syntactic environment is; μ: whether the previous context mentioned a property; ω: whether the object has the mentioned property

Learner input

Derived from frequencies in Brown-Eve corpus (Brown 1973) and the number of utterances children hear (Akhtar et al. 2004), assuming children learn anaphoric *one* between 14 and 18 months (Pearl & Lidz 2009).

	Baker	R&G, P&L	P&L - no filter	P&M
UNAMB	0	0	0	0
SEM-SYN	0	242	242	242
SYN	0	0	2743	2743
UNAMB NP	0	0	0	3073
Uninformative	36500	36258	33515	30442

Success metrics

Want $p_{N'}$ near 1, p_I near 1, and reproducing infant learner behavior (p_{beh}) to be near 1 $p_{beh} = p(\omega = hasproperty | \pi = \text{one}, \sigma =< NP, \mu = yes)$

Results & Implications

Averages over 1000 simulations, standard deviations in parentheses.

	Baker	R&G, P&L	P&L - no filter	P&M
$p_{N'}$	.50 (<.01)	.97 (<.01)	.17 (.02)	.37 (.04)
p_I	.50 (<.01)	.95 (<.01)	.02 (<.01)	>.99 (<.01)
p_{beh}	.56 (<.01)	.93 (<.01)	.50 (<.01)	>.99 (<.01)

P&M learner: Correct behavior, even if the representation is incorrect ($p_{N'}$ is low). This is due to the additional indirect evidence data, since other learners produce the same qualitative results found previously.

Why does this happen? If the property is important (p_I), the antecedent must contain the modifier (e.g., *red bottle*) and so the referent must have that property (RED BOTTLE). This produces the correct behavior in this context, even if $p_{N'}$ is low for the general case.

Main Answers
- Child anaphoric *one* behavior can be reproduced without requiring innate domain-specific learning biases, provided the child learns from indirect evidence. No input filtering is required.
- However, this does not lead to the adult representation. That representation may be learned during a second stage of acquisition, and may require innate domain-specific learning biases to do so.