Testing the Universal Grammar hypothesis: The contribution of computational modeling

Lisa S. Pearl
Assistant Professor
Department of Cognitive Sciences
SBSG 2314
University of California
Irvine, CA 92697

lpearl@uci.edu

Lisa Pearl

Computation of Language Laboratory
UC Irvine

California State University, Fullerton
Linguistics Symposium
April 9, 2012

An induction problem by any other name…

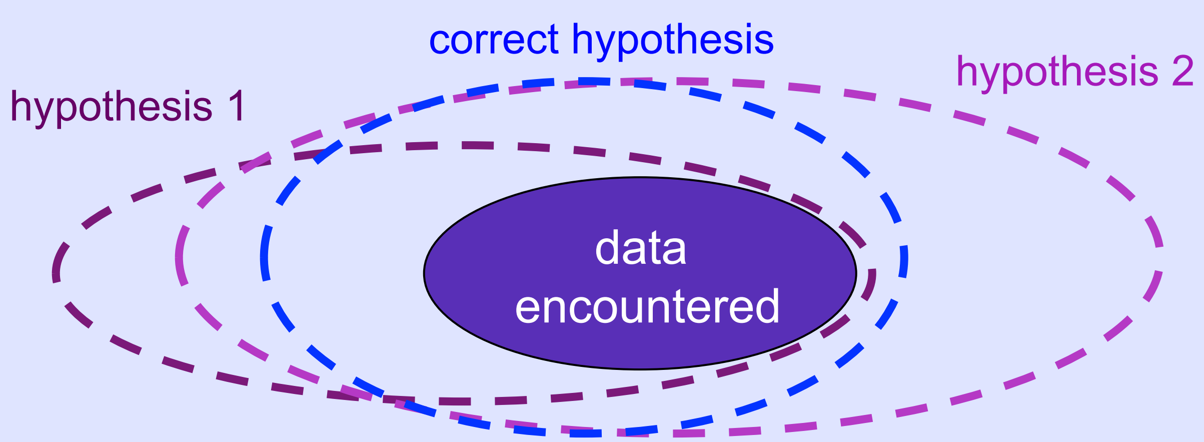
Children learning their native language face an induction problem:

"Poverty of the Stimulus" (Chomsky 1980, Crain 1991, Lightfoot 1989, Valian 2009)

"Logical Problem of Language Acquisition" (Baker 1981, Hornstein & Lightfoot 1981)

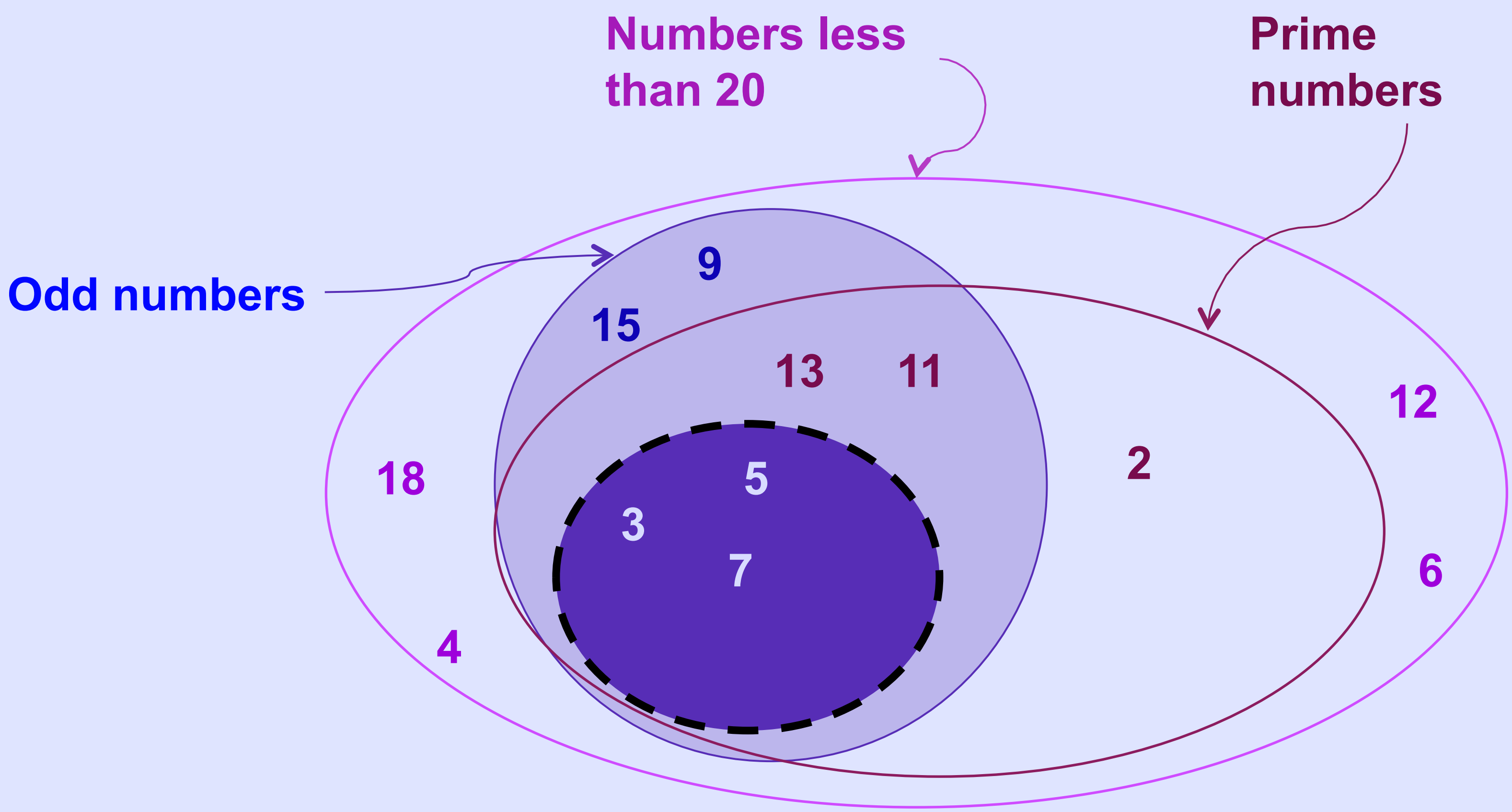
"Plato's Problem" (Chomsky 1988, Dresher 2003)

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



Poverty of the stimulus

A numerical analogy. Suppose you encounter the numbers 3, 5, and 7. What set are these numbers drawn from? That is, what is the right "number rule" for this language that will allow you to predict what numbers will appear in the future?



Poverty of the stimulus

Extended claim:

Given this, the data are insufficient for identifying the correct hypothesis.

Big question: How do children do it? (because we know they do)

One answer: Children come prepared

- Children are not unbiased learners. They come equipped with helpful learning biases.

- Big question: what is the nature of these necessary biases?

The nature of the necessary biases

- Bias kinds (at least three dimensions to consider):

The nature of the necessary biases

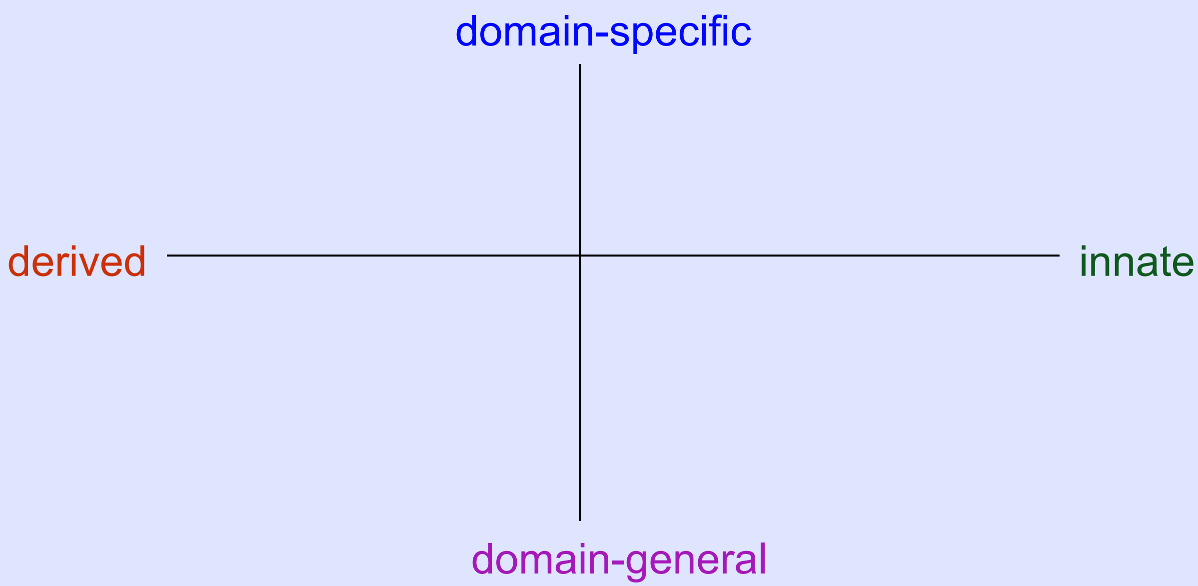
- Bias kinds (at least three dimensions to consider):
    - Domain-general or domain-specific?

domain-specific

domain-general

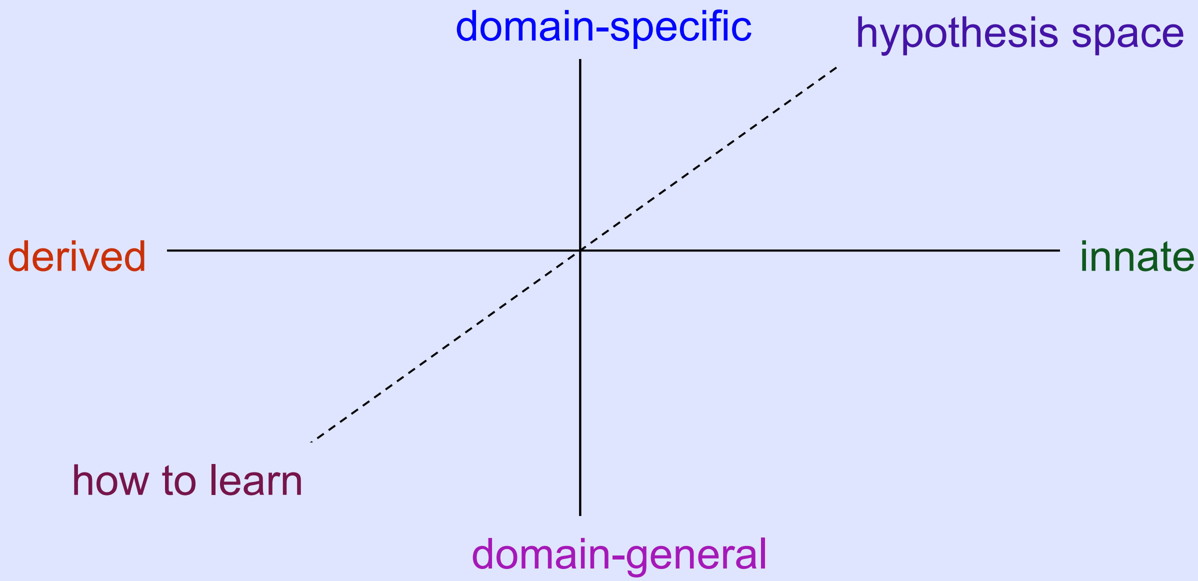
The nature of the necessary biases

- Bias kinds (at least three dimensions to consider):
    - Domain-general or domain-specific?
    - Innate or derived from prior linguistic experience?



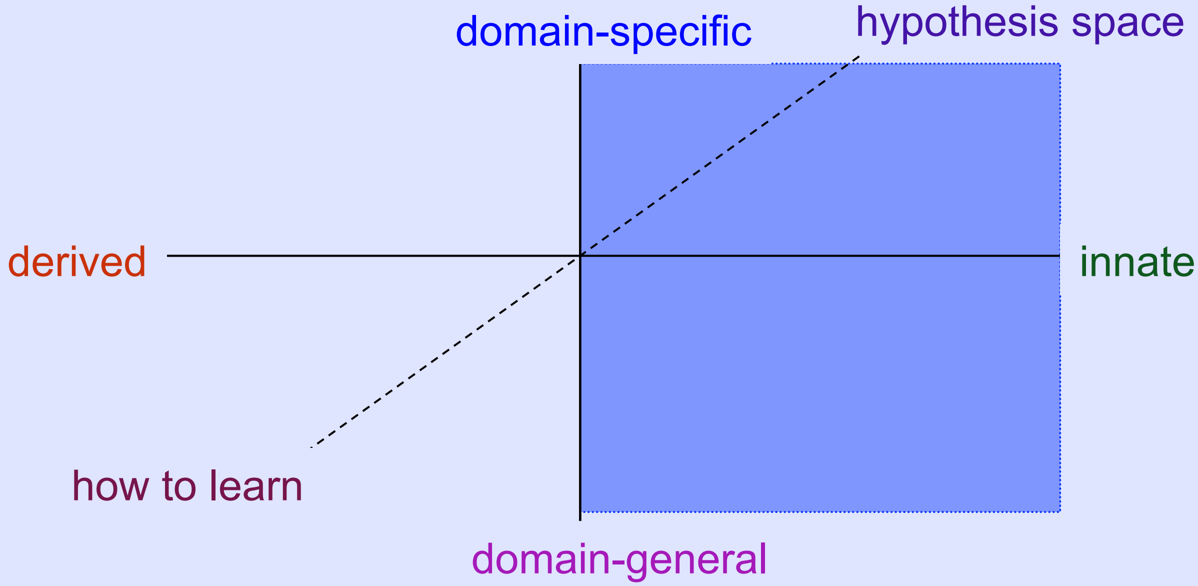
The nature of the necessary biases

- Bias kinds (at least three dimensions to consider):
  - Domain-general or domain-specific?
  - Innate or derived from prior linguistic experience?
  - Knowledge about the hypothesis space or knowledge about how to learn?



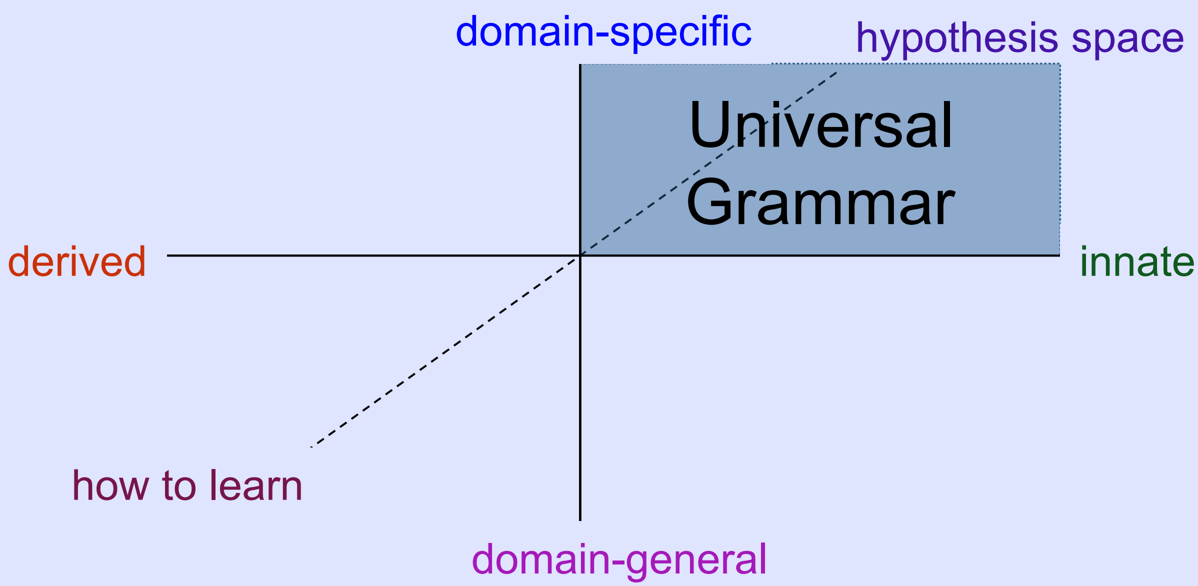
The nature of the necessary biases

- Nativists believe that the necessary knowledge is innate, but may be either domain-specific or domain-general.



The nature of the necessary biases

- Linguistic nativists believe that at least some necessary knowledge is both innate and domain-specific. This is sometimes called the Universal Grammar (UG) hypothesis (Chomsky 1965, Chomsky 1975). Because children have UG, they can solve the language acquisition problem.



The nature of the necessary biases

- How can we test different ideas about what the necessary knowledge might be?
  - Computational modeling studies can help us identify the necessary knowledge.

In a computational model, we can implement a specific learning strategy - which incorporates particular learning biases - and see how well a learner using this strategy is able to take realistic input and reach the desired target knowledge state.

What a "digital" child can tell us

We can construct a model where we have precise control over these:

- The hypotheses the child is considering at any given point [hypothesis space]

"I love my daxes."

Dax = that specific toy, teddy bear, stuffed animal, toy, object, …?

What a “digital” child can tell us

We can construct a model where we have precise control over these:

- The hypotheses the child is considering at any given point [hypothesis space]
- How the child represents the data & which data the child uses [data intake]

“I love my daxes.”

Dax = that specific toy, teddy bear, stuffed animal, toy, object, …?

What a "digital" child can tell us

We can construct a model where we have precise control over these:

- The hypotheses the child is considering at any given point
  [hypothesis space]

- How the child represents the data & which data the child uses
  [data intake]

- How the child changes belief based on those data
  [update procedure]

dax = that specific toy more probable

dax = any object less probable

A note on update procedures

Many current models rely on probabilistic learning as the update procedure. One common type of probabilistic learning that is used is Bayesian inference.

In Bayesian inference, the belief in a particular hypothesis (H) (or the probability of that hypothesis), given the data observed (D), can be calculated the following way:

$$P(H \mid D) \propto P(D \mid H) * P(H)$$

“The posterior probability of the hypothesis, given the data, is proportional to the likelihood of the data given the hypothesis multiplied by the prior probability of the hypothesis.”

A note on update procedures

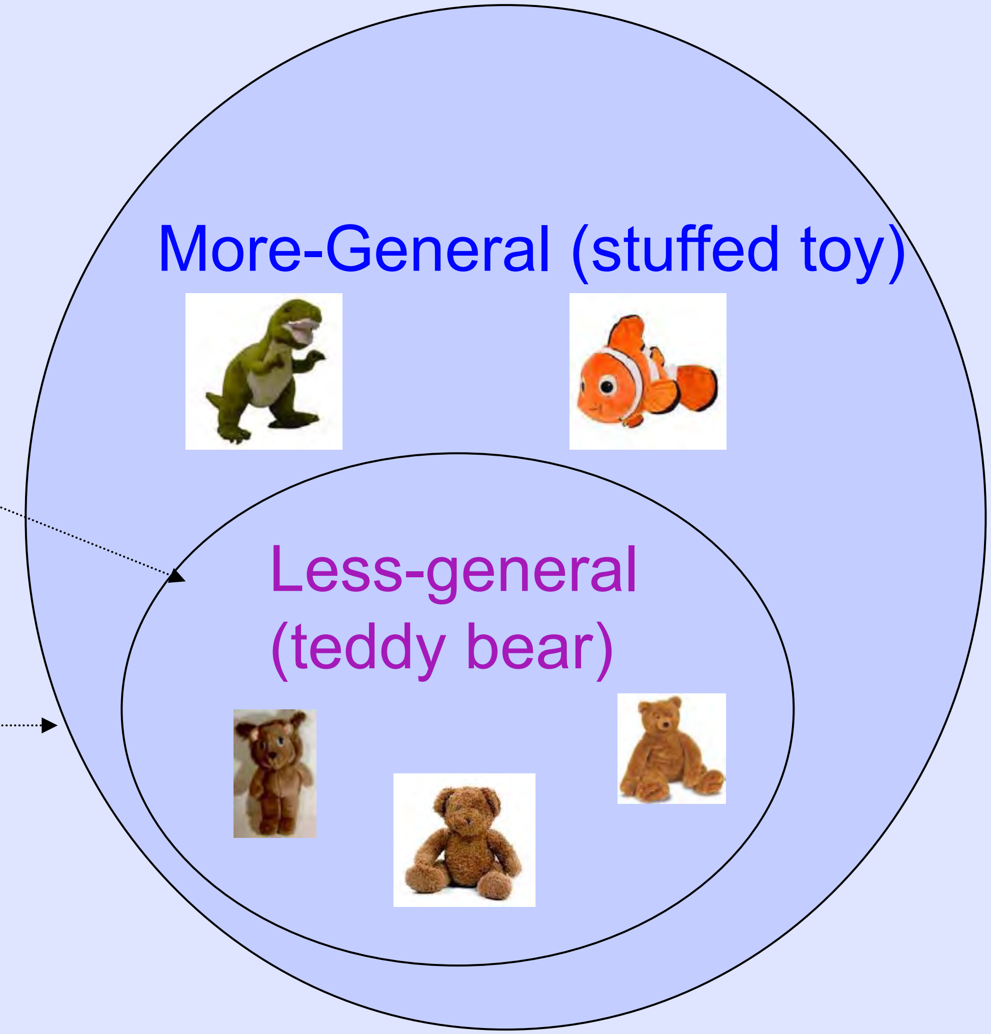
Bayesian inference is very useful when the hypotheses are in a subset relationship.

What does "dax" mean?

Suppose there are only 5 stuffed toys in the world that the child knows about, as shown in this diagram.

Hypothesis 1 (H1): The less-general hypothesis is true, and *dax* means teddy bear.

Hypothesis 2 (H2): The more-general hypothesis is true, and *dax* means stuffed toy.

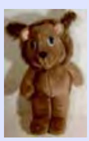


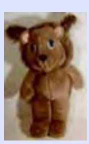
A note on update procedures

Bayesian inference is very useful when the hypotheses are in a subset relationship.

What does “dax” mean?

What’s the likelihood of selecting this toy for each hypothesis?

p( | H1) = 1/3
(since only three toys are possible)

p( | H2) = 1/5
(since all five toys are possible)

A note on update procedures

Bayesian inference is very useful when the hypotheses are in a subset relationship.

What does "dax" mean?

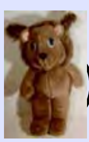
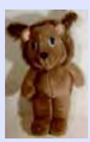
This means the likelihood for the less-general hypothesis is always going to be larger than the likelihood of the more-general hypothesis for data points that both hypotheses can account for.

A note on update procedures

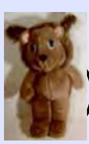
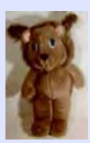
Bayesian inference is very useful when the hypotheses are in a subset relationship.

What does "dax" mean?

If the prior is equal (ex: before any data, both hypotheses are equally likely), then the posterior probability will be greater for the less-general hypothesis.

p(H1 | ) ∝ p( | H1) * p(H1)

∝ 1/3 * p(H1)

p(H2 | ) ∝ p( | H2) * p(H2)

∝ 1/5 * p(H2)

A note on update procedures

Bayesian inference is very useful when the hypotheses are in a subset relationship.

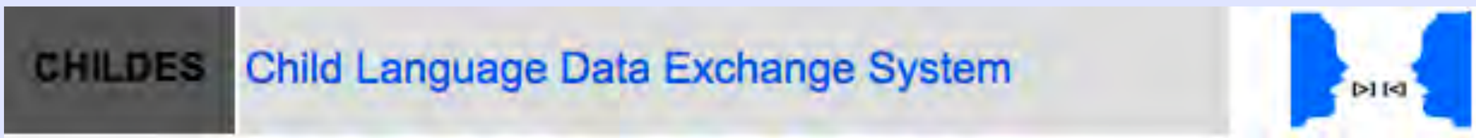
What does "dax" mean?

Upshot: Bayesian learners can learn something from ambiguous data that multiple hypotheses are compatible with. This can be useful for induction problems.

What a “digital” child can tell us

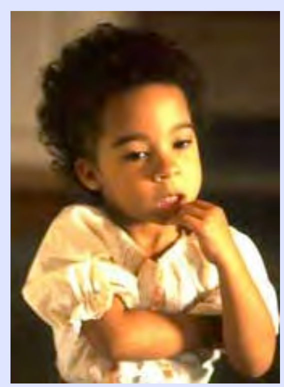
Models are most informative when they’re grounded empirically.

This is why most models make use of the child-directed speech data available through databases like CHILDES.



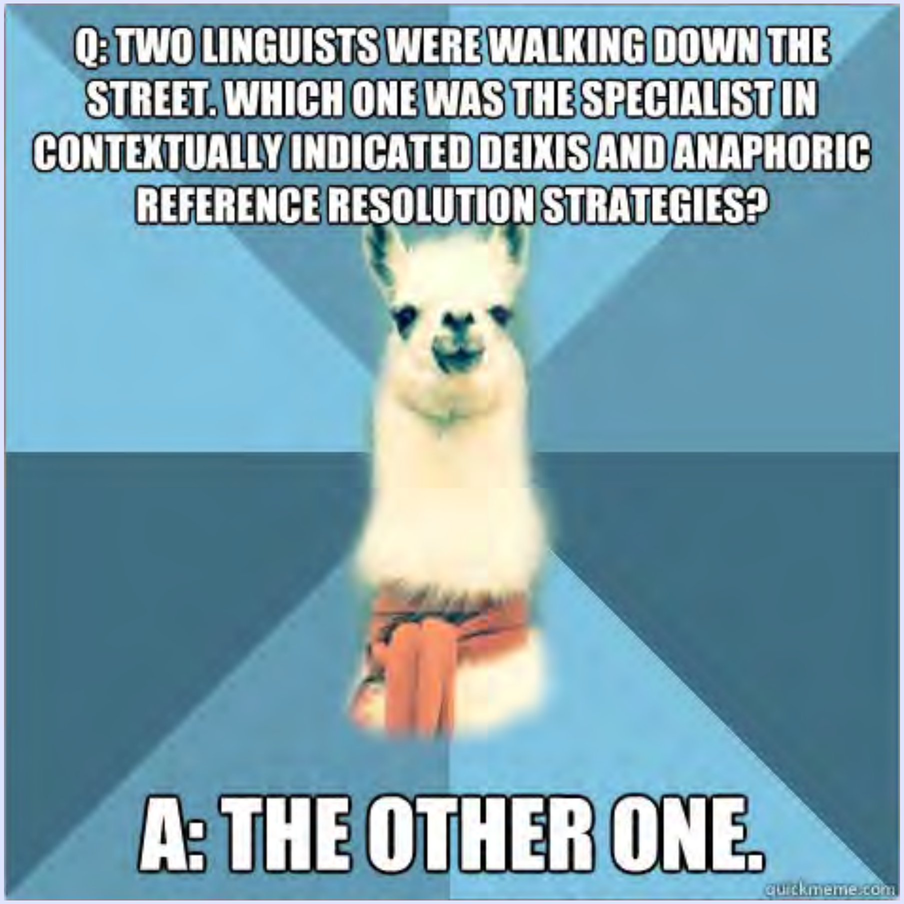
Many models will try to make cognitively plausible assumptions about how the child is representing and processing input data:

- Processing data points as they are encountered
- Assuming children have memory limitations (ex: memory of data points may decay over time)



Reasonable questions

- What are some examples of linguistic knowledge that seem to present a poverty of the stimulus problem?
  - Anaphoric *one* in English



Anaphoric *One*

Look - a red bottle!

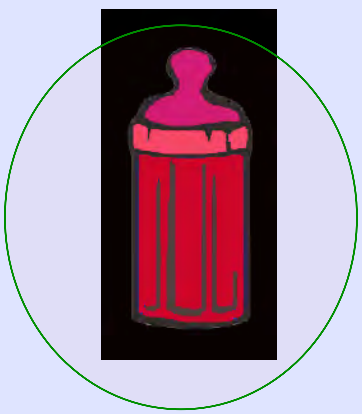
Do you see another *one*?

Anaphoric *One*

Look - a red bottle!

red bottle

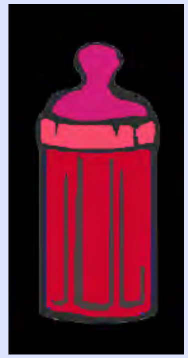
Do you see another *one*?



Process: First determine the linguistic antecedent of *one* (what words *one* is referring to). "red bottle"

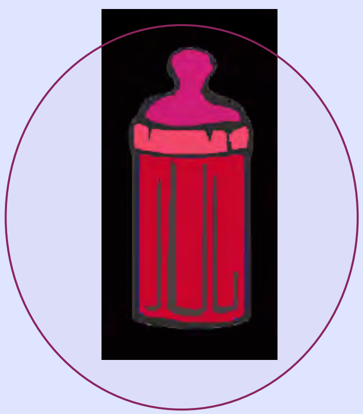
Anaphoric *One*

Look - a red bottle!



red bottle

Do you see another *one*?



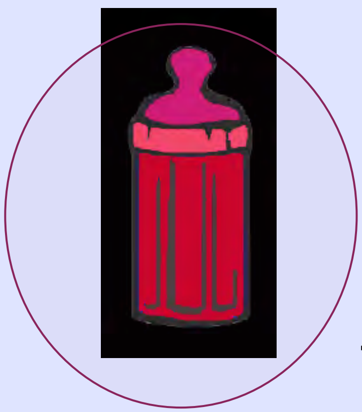
Process: Because the antecedent (“red bottle”) includes the modifier “red”, the property RED is important for the referent of *one* to have.

referent of *one* = RED BOTTLE

Anaphoric *One*

Look - a red bottle!

Do you see another *one*?



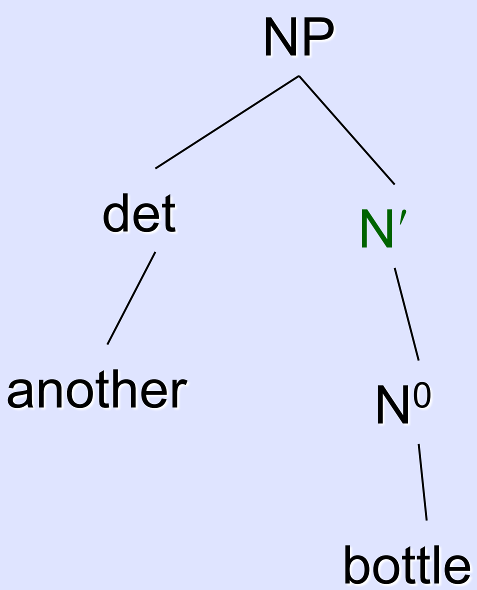
Two steps:

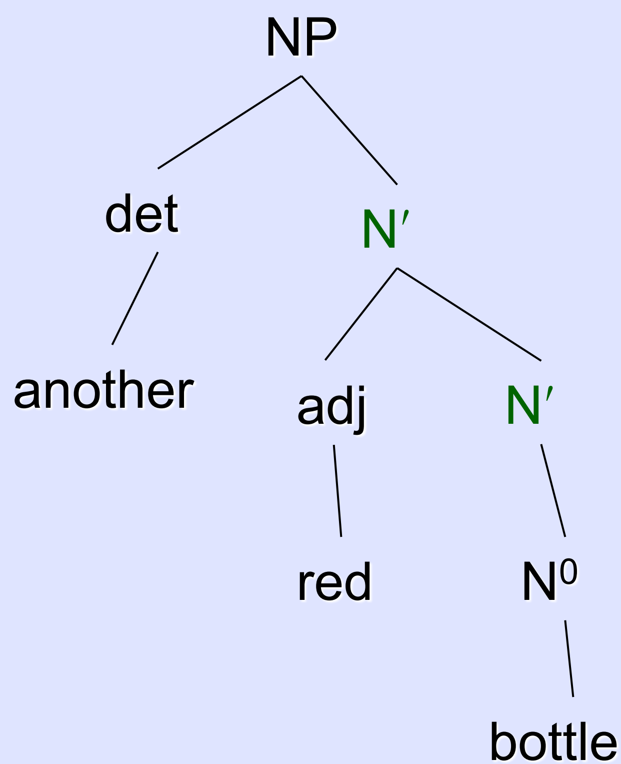
(1)Identify syntactic antecedent (based on syntactic category of *one*)

(2)Identify semantic referent (based on syntactic antecedent)

Anaphoric *One*: Syntactic Category

Standard linguistic theory says that *one* in these kind of utterances is a syntactic category smaller than an entire noun phrase, but larger than just a noun (N^0). This category is sometimes called N′. This category includes sequences like "bottle" and "red bottle".



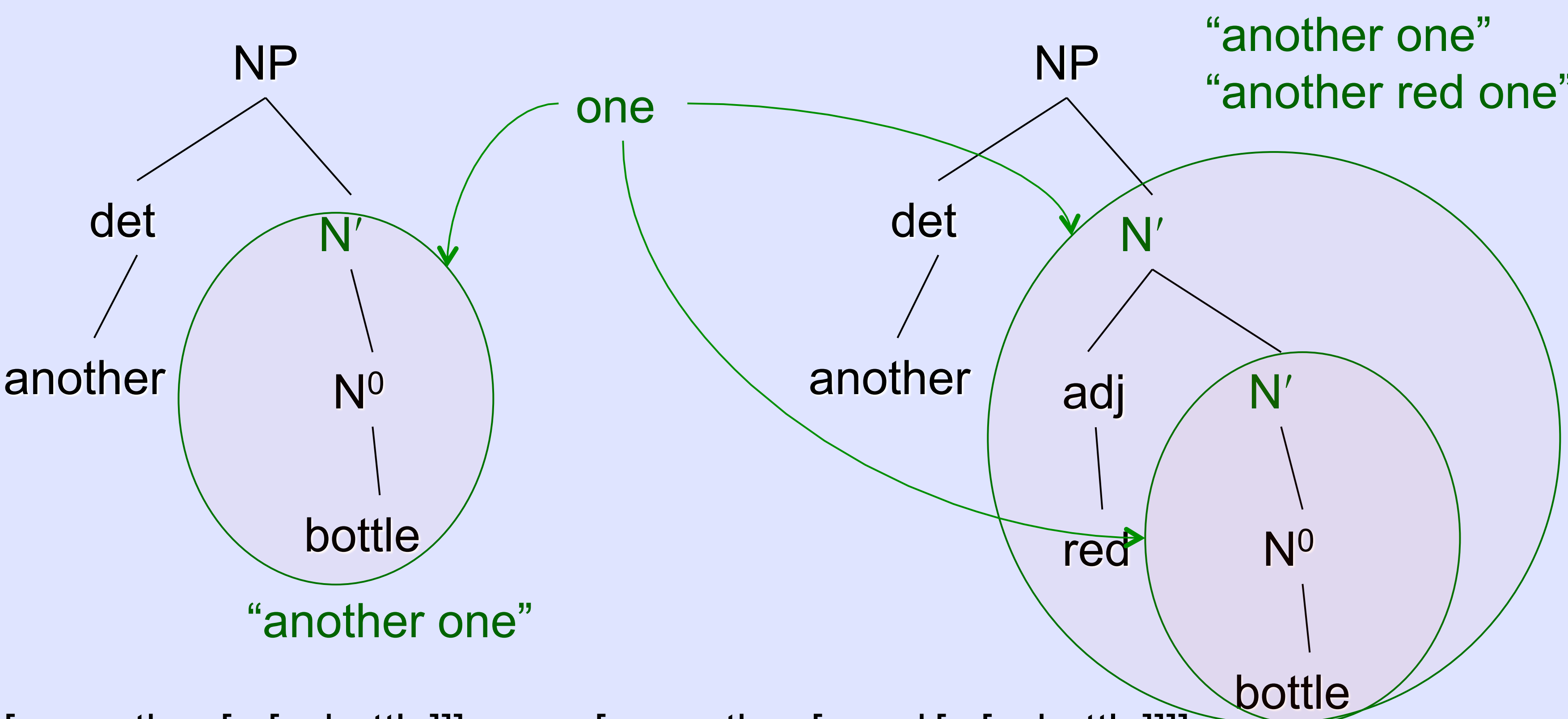


[$_{NP}$ another [$_{N'}$ [$_{N0}$ bottle]]]

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

Anaphoric *One*: Syntactic Category

Standard linguistic theory says that *one* in these kind of utterances is a syntactic category smaller than an entire noun phrase, but larger than just a noun (N^0). This category is sometimes called N'. This category includes sequences like "bottle" and "red bottle".

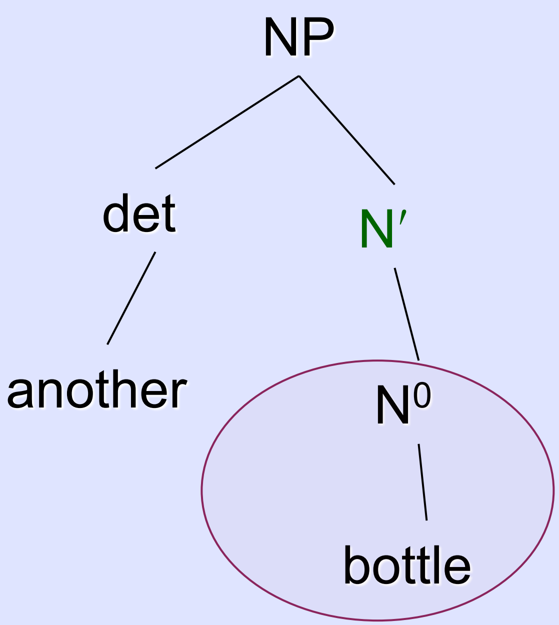


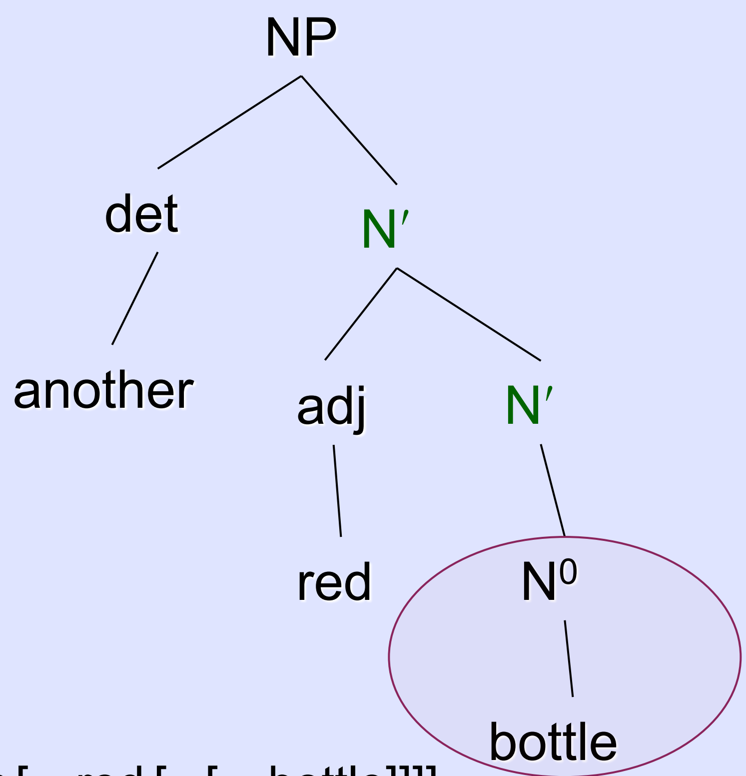
[$_{NP}$ another [$_{N'}$ [$_{N0}$ bottle]]]

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

Anaphoric *One*: Syntactic Category

Importantly, *one* is not N^0. If it was, it could only have strings like "bottle" as its antecedent, and could never have strings like "red bottle" as its antecedent.



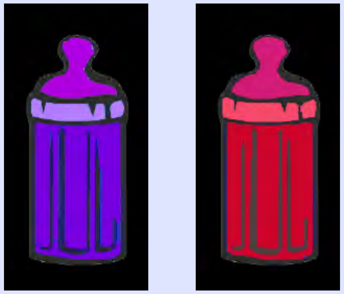


$[_{NP}\ \text{another}\ [_{N'}\ [_{N0}\ \text{bottle}]]]$

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

Anaphoric *One*: Interpretations based on Syntactic Category

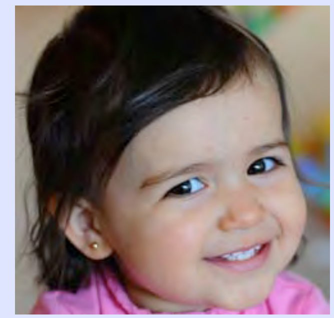
If *one* was N^0, we would have a different interpretation of

"Look – a red bottle!  Do you see another one?" 

Because *one*'s antecedent could only be "bottle", we would have to interpret the second part as "Do you see another *bottle*?" and the purple bottle would be a fine referent for *one*.

Since *one*'s antecedent is "red bottle", and "red bottle" cannot be N^0, *one* must not be N^0.

Anaphoric *One*: Children's Knowledge

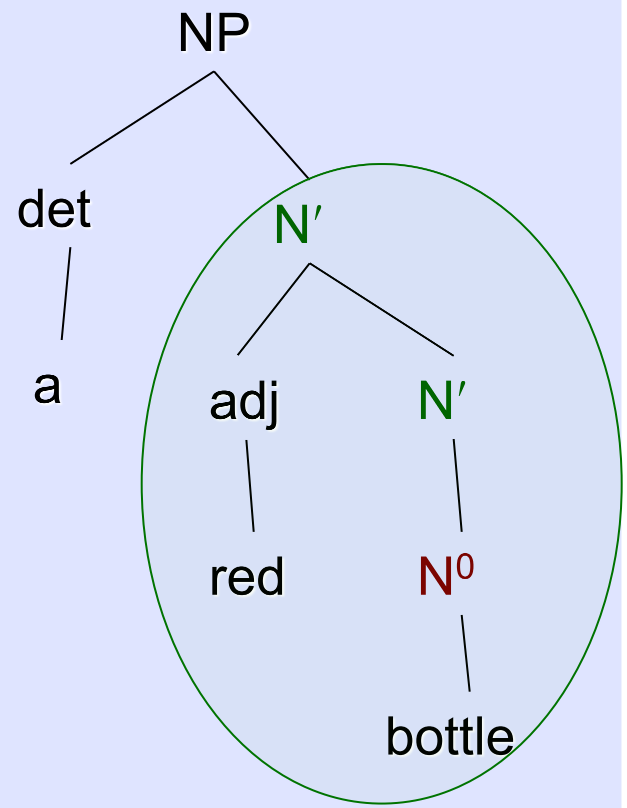


Lidz, Waxman, & Freedman (2003) found that 18-month-olds have a preference for the red bottle in the same situation.

"Look – a red bottle! Do you see another one?"

Lidz et al. interpretation & conclusion:

Preference for the RED BOTTLE means the preferred syntactic antecedent is "red bottle".



Lidz et al. conclude that 18-month-old knowledge =

syntactic category of *one* = N′

when modifier (like "red") is present, syntactic antecedent includes modifier (e.g., red) = referent must have modifier property

Anaphoric *One*: The induction problem

Acquisition: Children must learn the right syntactic category for *one*, and the right interpretation preference for *one* in situations with more than one option.

Anaphoric *One*: The induction problem

Acquisition: Children must learn the right syntactic category for *one*, and the right interpretation preference for *one* in situations with more than one option.

Problem: Most data children encounter are ambiguous.

Syntactically (SYN) ambiguous data:

"Look – a bottle! Oh, look – another one."

one's referent = BOTTLE

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

Anaphoric *One*: The induction problem

Acquisition: Children must learn the right syntactic category for *one*, and the right interpretation preference for *one* in situations with more than one option.

Problem: Most data children encounter are ambiguous.

Semantically and syntactically (SEM-SYN) ambiguous:

"Look – a red bottle! Oh, look – another one."

one's referent = RED BOTTLE or BOTTLE?

one's antecedent = $[_{N'}\ \text{red}[_{N'}[_{N0}\ \text{bottle}]]]$ or $[_{N'}[_{N0}\ \text{bottle}]]$ or $[_{N0}\ \text{bottle}]$?

Anaphoric *One*: The induction problem

Acquisition: Children must learn the right syntactic category for *one*, and the right interpretation preference for *one* in situations with more than one option.

Problem: Unambiguous data are extremely rare

Unambiguous (UNAMB) data:

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

one's referent = BOTTLE? If so, *one*'s antecedent = "bottle".
But it's strange to claim there's not another *bottle* here.
So, *one*'s referent must be RED BOTTLE, and *one*'s antecedent =

$[_{N'}$ red$[_{N'}[_{N0}$ bottle$]]]$.

Anaphoric *One*: The induction problem

Acquisition: Children must learn the right syntactic category for *one*, and the right interpretation preference for *one* in situations with more than one option.

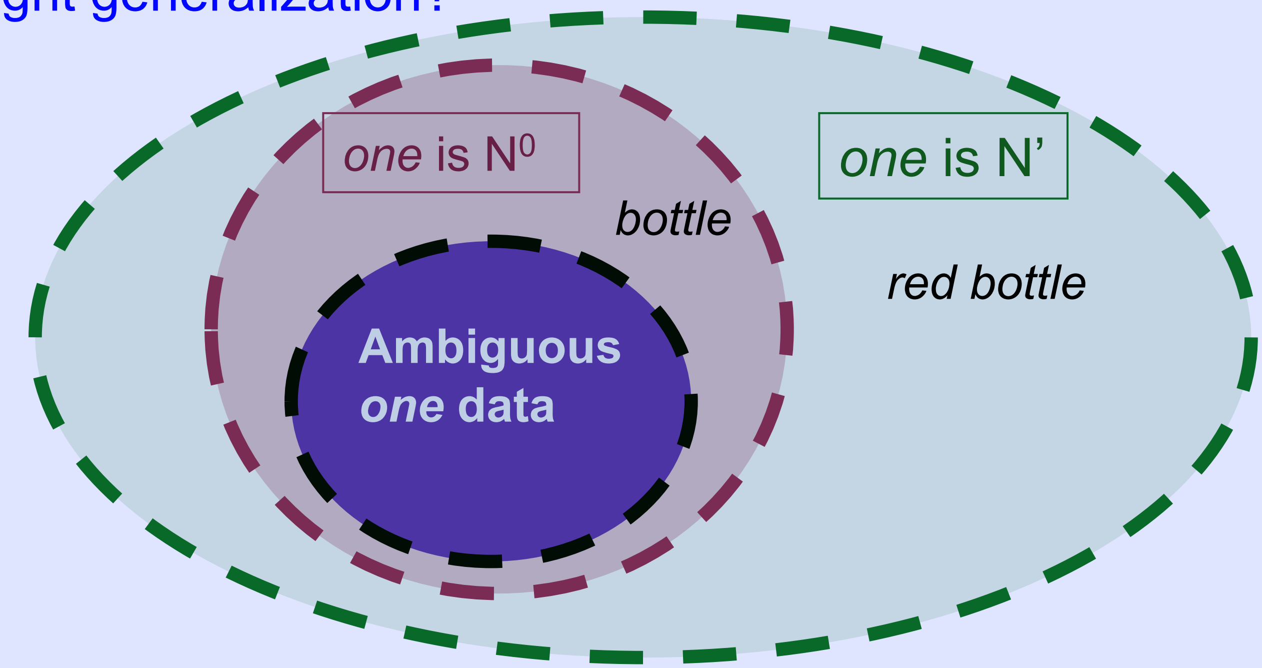
Problem: Unambiguous data are extremely rare

Pearl & Mis (2011) looked at ~17,500 child-directed speech utterances (from CHILDES), and discovered that *none* of them were unambiguous for anaphoric *one*.

Anaphoric *One*: The induction problem

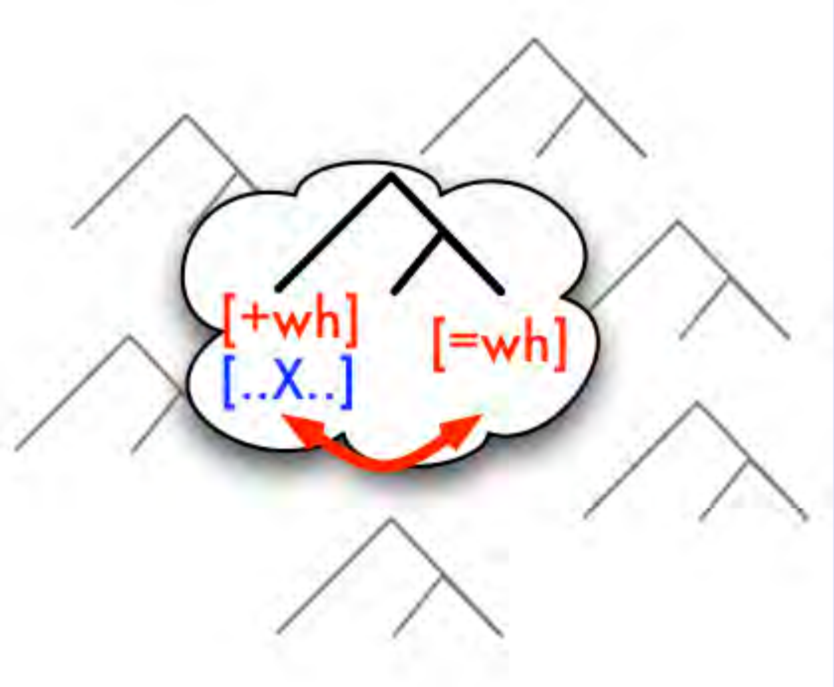
Acquisition: Children must learn the right syntactic category for *one*, so they end up with the right interpretation for *one*.

Problem: If children don't encounter unambiguous data often enough to notice them, they are left with data that are compatible with both hypotheses – that *one* is N^0 and that *one* is N′. How do children know which is the right generalization?



Reasonable questions

- What are some examples of linguistic knowledge that seem to present a poverty of the stimulus problem?
  - Anaphoric *one* in English
  - Syntactic islands



Syntactic Islands

Dependencies between a wh-word and where it's understood (its gap) can exist when these two items are not adjacent, and these dependencies do not appear to be constrained by length (Chomsky 1965, Ross 1967).

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

Syntactic Islands

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

*What did you make [the claim that Jack bought __]?
*What do you think [the joke about __] offended Jack?
*What do you wonder [whether Jack bought __]?
*What do you worry [if Jack buys __]?
*What did you meet [the scientist who invented __]?
*What did [that Jack wrote __] offend the editor?
*What did Jack buy [a book and __]?
*Which did Jack borrow [__ book]?

The input: Induction problems

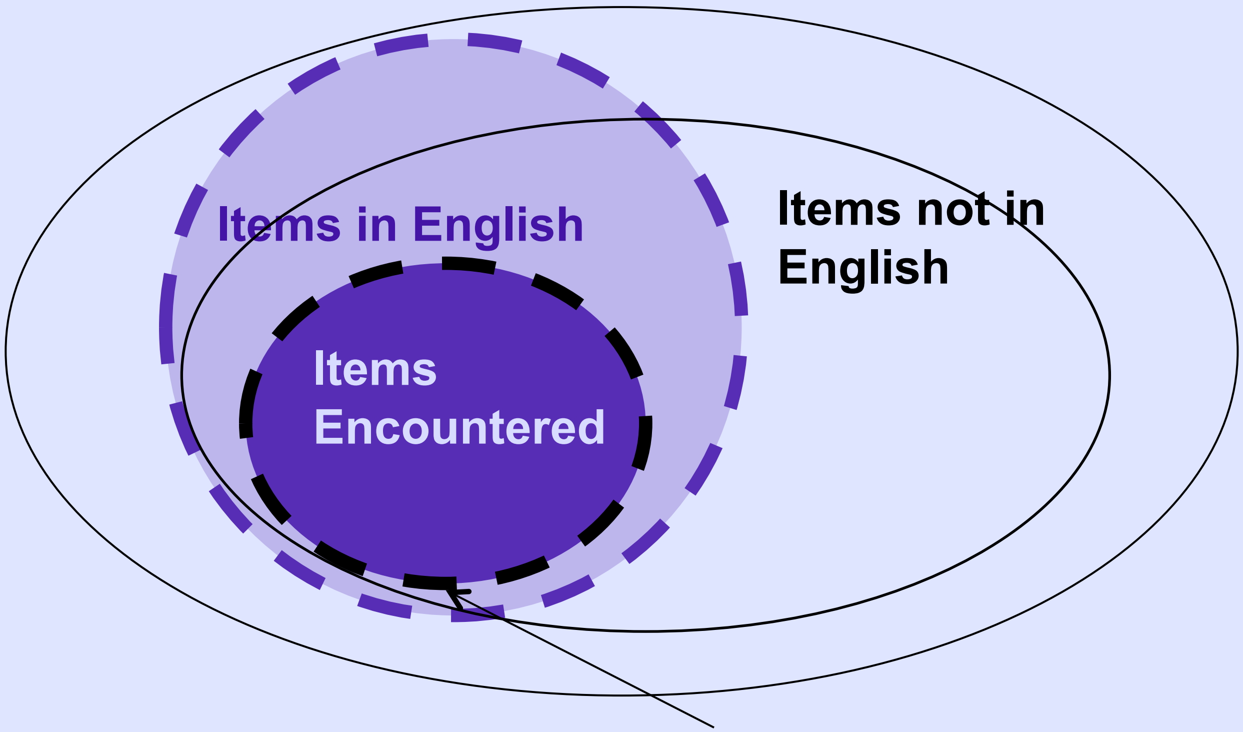
Data from five corpora of child-directed speech from CHILDES: speech to 25 children between the ages of one and four years old.

Utterances containing a *wh*-word and a verb: ~31,000

Pearl & Sprouse (2011, submitted) discovered that more complex dependencies were fairly rare in general (<0.01% of the input).

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.

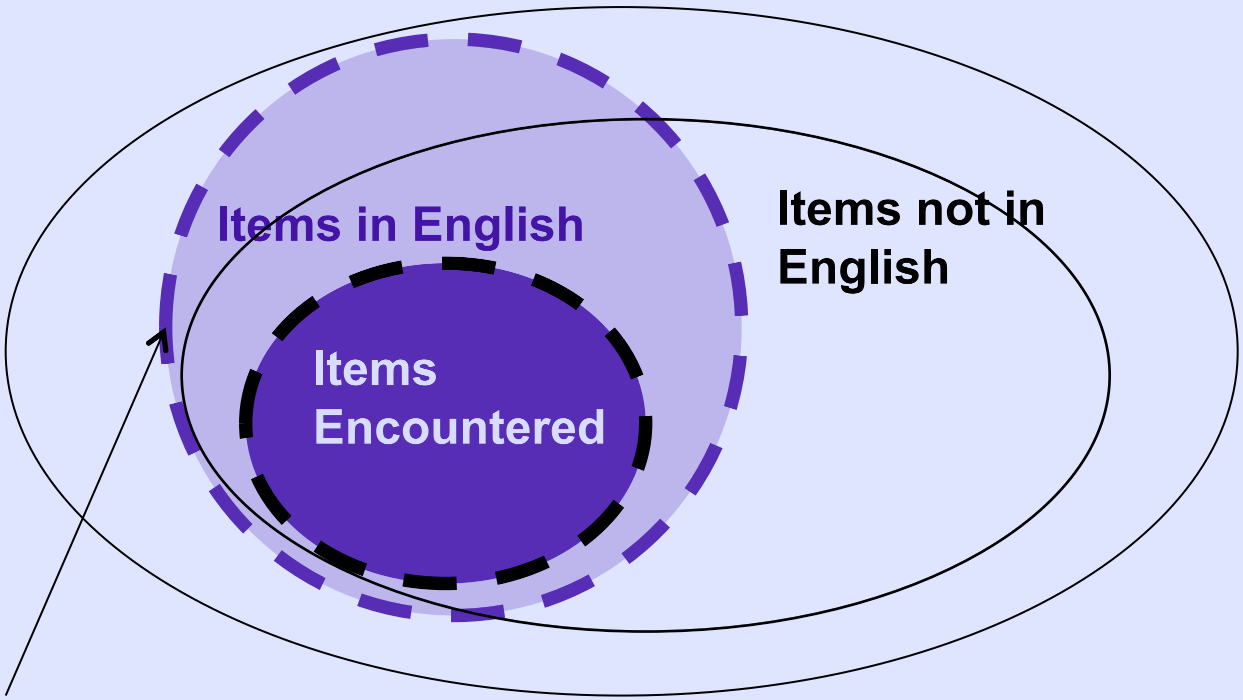
Syntactic Islands: Induction Problem



Wh-questions in input (usually fairly simple)

What did you see?
What happened?
…

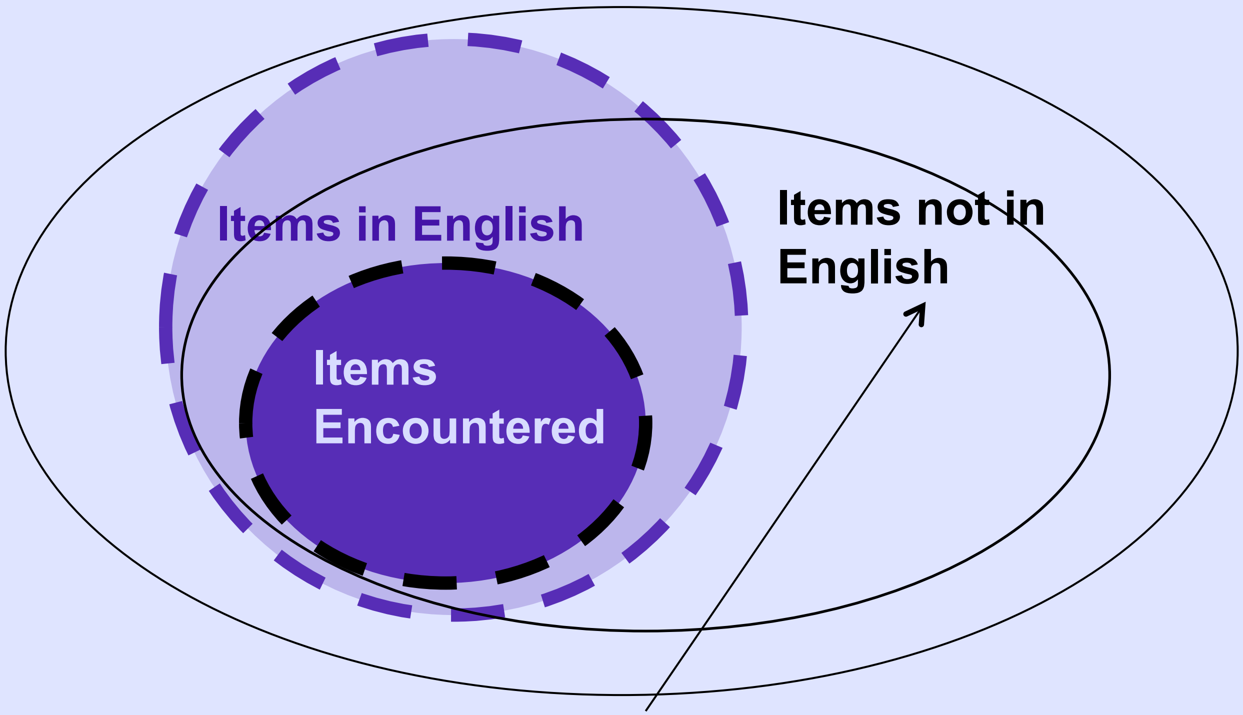
Syntactic Islands: Induction Problem



Grammatical wh-questions

What did you see?
What happened?
Who did Jack think that Lily saw?
What did Jack think happened?

Syntactic Islands: Induction Problem



Ungrammatical wh-questions: Syntactic islands

*What did you make [the claim that Jack bought __]?
*What do you think [the joke about __] offended Jack?
*What do you wonder [whether Jack bought __]?
*What do you worry [if Jack buys __]?

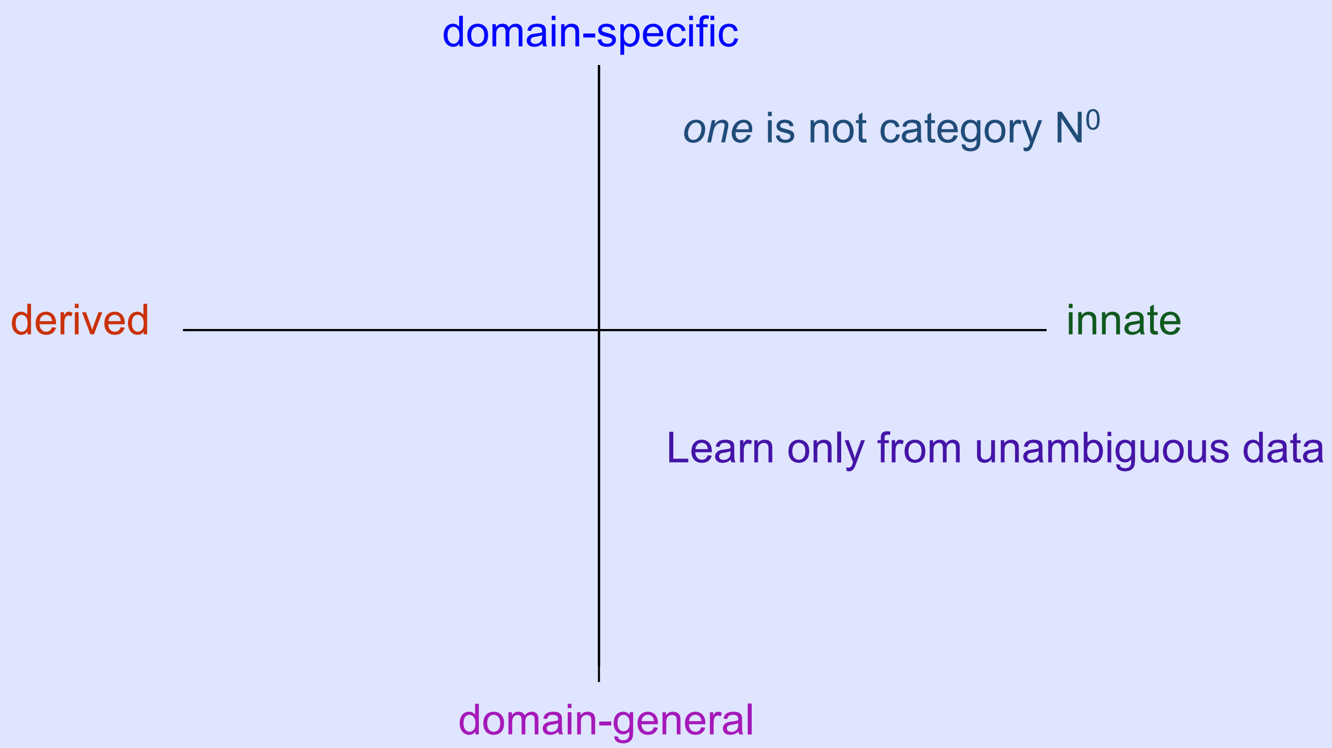
Computational modeling studies

Several recent computational models have attempted to address poverty of the stimulus questions, and rely on probabilistic learning (often Bayesian inference) as the main method of learning. By modeling the acquisition process for these linguistic phenomena, these models hope to pinpoint the kind of knowledge required for language acquisition.

- Anaphoric *one*: Regier & Gahl (2004), Foraker et al. (2009), Pearl & Lidz (2009), Pearl & Mis (2011, submitted)
- Syntactic islands: Pearl & Sprouse (2011, submitted)

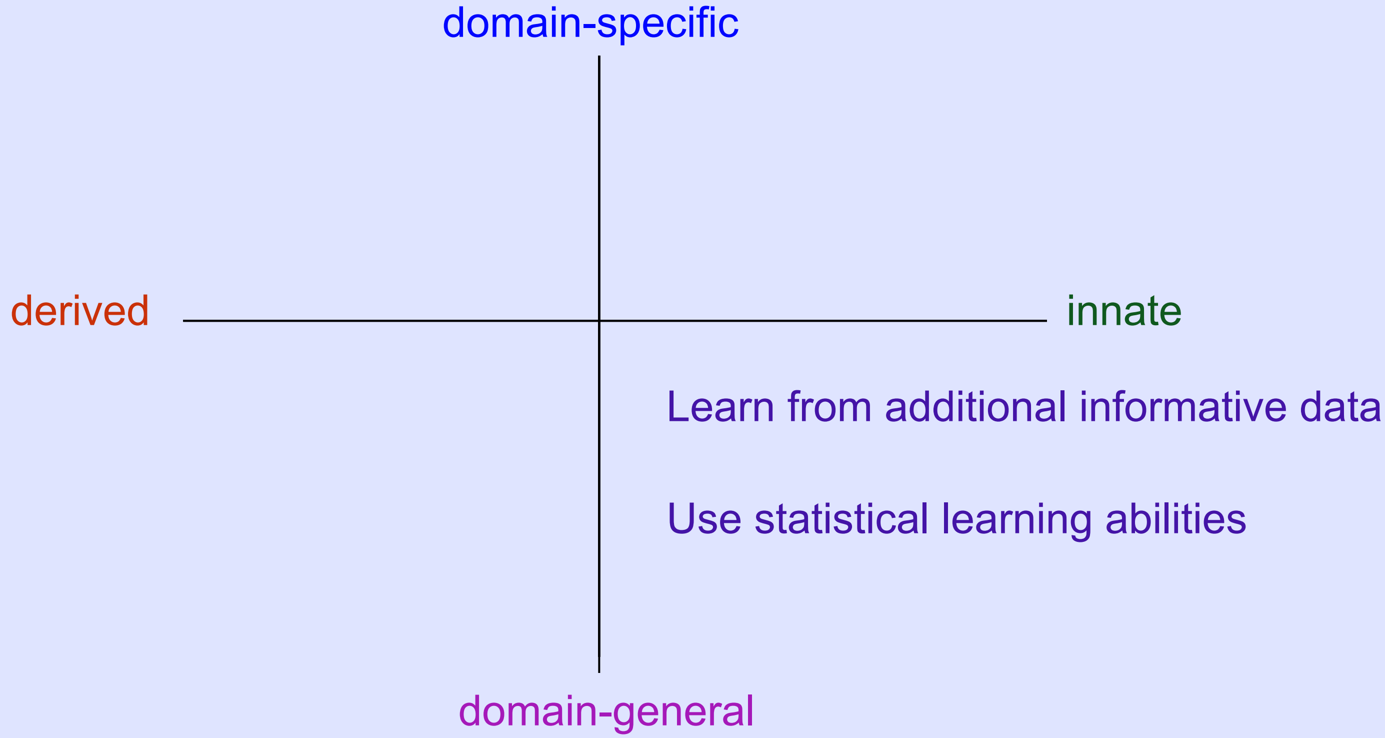
English anaphoric *one*

Baker (1978) assumed only unambiguous data are informative, and these data are rare. So, he proposed that children needed to know that *one* could not be syntactic category N^0.



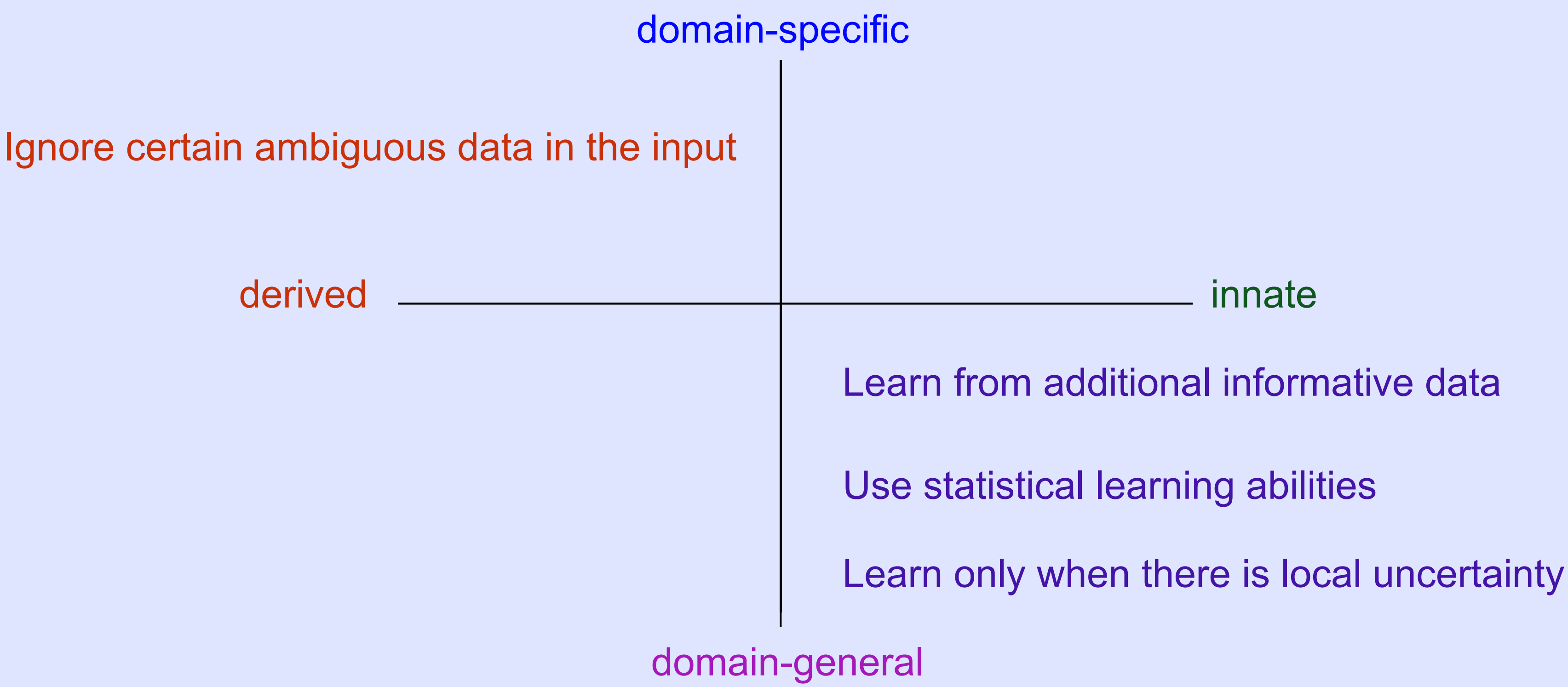
English anaphoric *one*

Regier & Gahl (2004) used a Bayesian learner computational model to show that children could learn *one* is category N′ if they learned from some of the available ambiguous data and used their statistical learning abilities to track suspicious coincidences in the input.



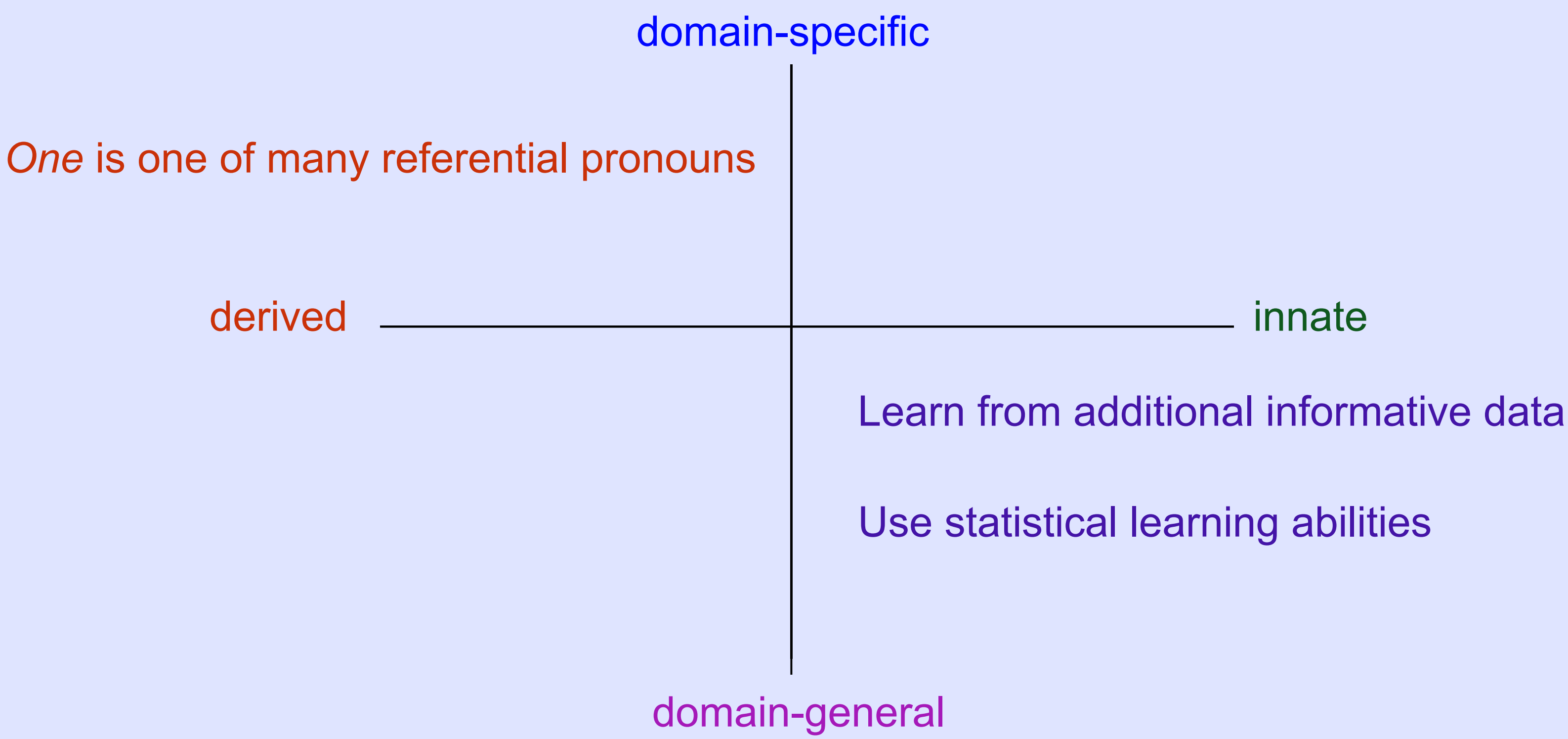
English anaphoric *one*

Pearl & Lidz (2009) discovered that a Bayesian learner must ignore certain ambiguous data (even if they're informative) in order to learn that *one* is category N′. This can be derived from an innate, domain-general preference for learning when there is uncertainty in the utterance heard.



English anaphoric *one*

Pearl & Mis (2011, submitted) discovered that a Bayesian learner can learn from all ambiguous *one* data and still learn to interpret *one* appropriately in experiments like Lidz, Waxman, & Freedman (2003), if the learner also learns from data containing other pronouns like *it.*



English anaphoric *one*

However, Pearl & Mis (2011, submitted) also discovered that the full adult representation of *one* (not just the one in the Lidz et al. 2003 experiment) still requires some innate, domain-specific knowledge about how to map meaning to structure.

domain-specific

One is one of many referential pronouns

specific knowledge mapping meaning to structure

derived

innate

Use statistical learning abilities

domain-general

A snapshot of the ideas about necessary learning biases over time

Anaphoric *one*

Baker 1978

UG: *one* is not N^0

Other: learn only from unambiguous *one* data

Regier & Gahl 2004

Other: learn from ambiguous *one* data, too

Other: use Bayesian inference

Pearl & Lidz 2009

Other: ignore some ambiguous *one* data

Other: use Bayesian inference

Pearl & Mis 2011

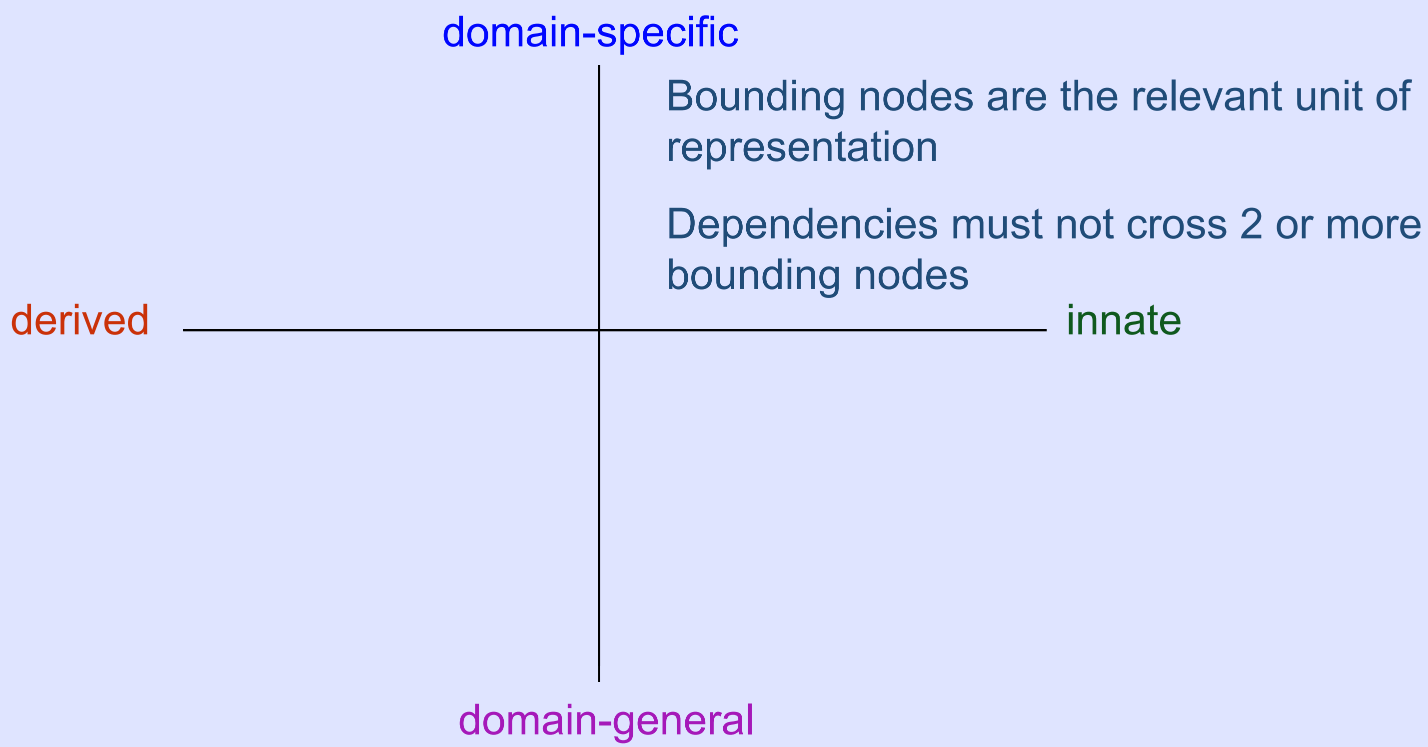
Other: learn from other pronoun data, too

Other: use Bayesian inference

UG: rules for mapping meaning to structure

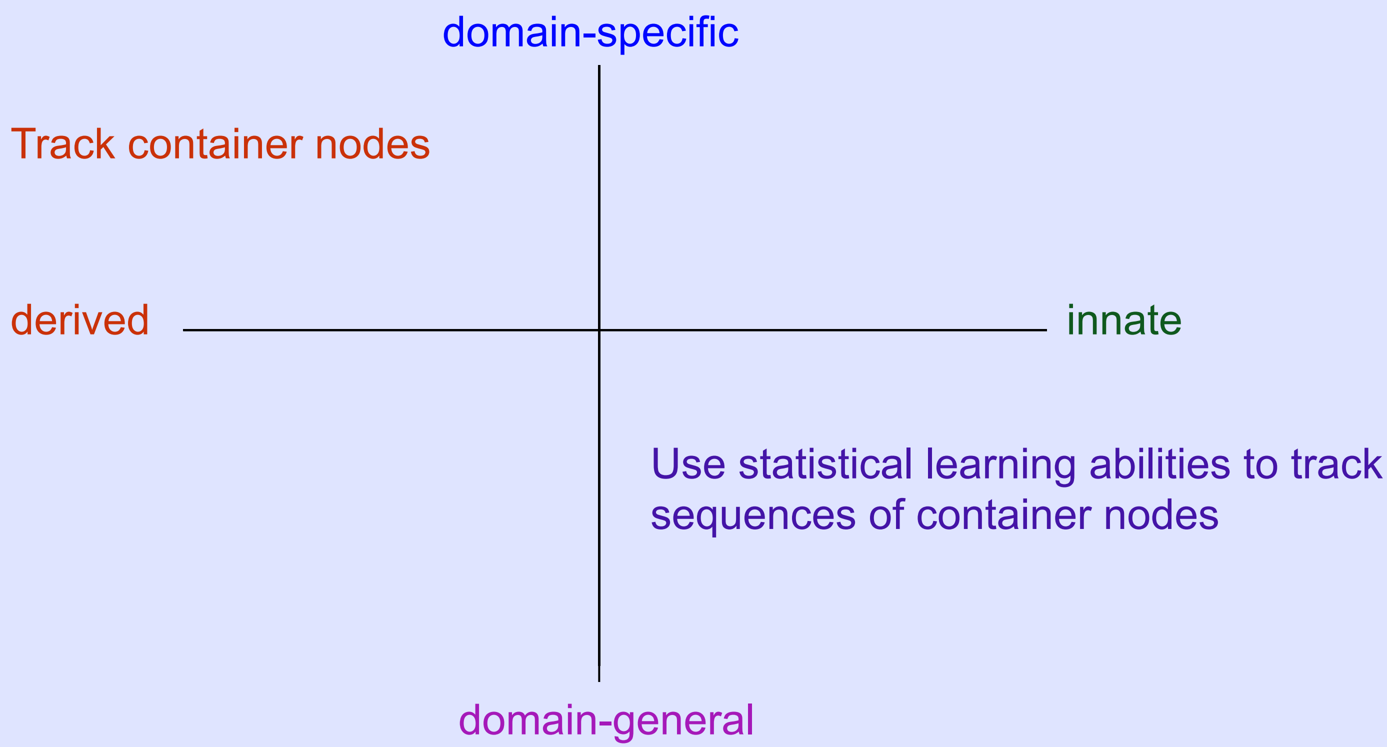
Syntactic islands

Chomsky (1973), Huang (1982), and Lasnik & Saito (1984) proposed that children must know that dependencies cannot cross 2 or more bounding nodes (a domain-specific representation).



Syntactic islands

Pearl & Sprouse (2011, submitted) discovered that a probabilistic learner that tracks sequences of container nodes (a derivable linguistic representation) can learn at least some of the syntactic islands.



A snapshot of the ideas about necessary learning biases over time

Syntactic islands

Chomsky 1973, Huang 1982, Lasnik & Saito 1984

UG: know that bounding nodes are relevant

UG: know dependencies crossing 2+ bounding nodes are bad

Pearl & Sprouse 2011

Other: know that container nodes are relevant

Other: use statistical learning

Big picture

- Universal Grammar has been proposed as one way to solve the induction problems faced by children learning their native language.
- While it's clear that children require some learning biases, there may be different kinds of learning biases that will work, especially when these biases are combined.
- Using computational modeling, we can examine specific learning biases and determine how well they do (or don't) work.
- For English anaphoric *one* and syntactic islands, some Universal Grammar biases may be less specific than previously thought, and some may be unnecessary after all.

Thank You!

Jon Sprouse

Benjamin Mis

Members of the Computation of Language Laboratory

Lisa S. Pearl
Assistant Professor
Department of Cognitive Sciences
SBSG 2314
University of California
Irvine, CA 92697

lpearl@uci.edu

Computation of Language Laboratory

UC Irvine