Jack only learns from this data point, but Lily learns from that one, too.

Lisa Pearl, University of Maryland
(in collaboration with Jeff Lidz)
University of Rochester: Center for Language Sciences
May 14, 2007
Overview of the Plan

Human language learning: mechanism
   investigating one component: data filtering
   interests: feasibility, sufficiency, necessity

Case Study: English Anaphoric One
   tool: computational modeling
   empirical grounding: experimental results, child-directed speech data
   conclusion: data filtering is feasible, sufficient, & necessary
Road Map

**Language Learning Mechanism**

- Learning language and why it’s hard
- Potentially helpful bias
- Computational modeling utility

**Learning Framework**

**Case Study: English Anaphoric One**
Human Language Learning: The How

worthwhile quest: understanding the mechanism of acquisition
given the boundary conditions provided by

(a) linguistic representation
   from theoretical work

(b) the trajectory of learning
   from experimental work

\[
\text{NP} \quad \frac{}{\text{det} \quad \text{N'}}
\quad \frac{}{\text{this} \quad \text{N}^0}
\quad \frac{}{\text{data point}}
\]
Why is learning tricky?

The linguistic system is made up of many different pieces... and there is often a non-transparent relationship between the observable form of the data and the underlying system that produced it.

Syntactic System

Observable form: word order
Interference: movement rules

Subject  Verb  \( t_{\text{Subject}} \)  Object  \( t_{\text{Verb}} \)
Why is learning tricky?

The linguistic system is made up of many different pieces... and they may be linked across different levels of representation, corresponding to different information sources.

*linguistic structure*  
*referent in the world*

```
NP
   \_______________/
  det  N'
   \___________/
this  N^0
   \________/
   ball

this ball
```
Road Map

Language Learning Mechanism
- Learning language and why it’s hard
- Potentially helpful bias
- Computational modeling utility

Learning Framework

Case Study: English Anaphoric One
Some Potentially Helpful Bias = Parameters

Premise: learner considers finite range of hypotheses (parameters) for the linguistic system

“Assuming that there are $n$ binary parameters, there will be $2^n$ possible core grammars.” - Clark (1994)
Not Completely Helpful Bias = Parameters

“It is unlikely that any example … would show the effect of only a single parameter value; rather, each example is the result of the interaction of several different principles and parameters” - Clark (1994)
Not Completely Helpful Bias = Parameters

“It is unlikely that any example ... would show the effect of only a single parameter value; rather, each example is the result of the interaction of several different principles and parameters” - Clark (1994)

Potential solution: the learner focuses in on a subset of the data perceived as “informative”.

Additional Bias = Filter on data intake
Big Questions for Filtering

1. Feasibility
   - Is there a data sparseness problem?

2. Sufficiency
   - Can we filter and get correct behavior?

3. Necessity
   - Must we filter to get correct behavior?
Big Questions for Filtering

(1) **Feasibility**
Is there a data sparseness problem?
Big Questions for Filtering

(1) **Feasibility**
Is there a data sparseness problem?

(2) **Sufficiency**
Can we filter and get correct behavior?
Big Questions for Filtering

(1) **Feasibility**
Is there a data sparseness problem?

(2) **Sufficiency**
Can we filter and get correct behavior?

(3) **Necessity**
Must we filter to get correct behavior?
Road Map

Language Learning Mechanism
- Learning language and why it’s hard
- Potentially helpful bias
- Computational modeling utility

Learning Framework

Case Study: English Anaphoric One
Computational Modeling of Data Intake Filtering

Why?

(1) Can easily (and ethically) restrict data intake to simulated learners and observe the effect on learning.

\[ P_A = .1 \]
\[ P_B = .9 \]

(2) Can empirically ground with data from experimental work & corpora: learners searching through realistic data space for evidence of the underlying system.

Recent computational modeling surge: Yang, 2000; Sakas & Fodor, 2001; Yang, 2002; Pearl, 2005; Pearl & Weinberg, 2007
Road Map

Language Learning Mechanism

Learning Framework
- Separable Components
- Investigating Data Filtering

Case Study: English Anaphoric One
Learning Framework: 3 Separable Components

(1) Hypothesis space

(2) Data intake

(3) Update procedure
Benefits of Learning Framework

Components:
(1) hypothesis space (2) data intake (3) update procedure

Application to a wide range of learning problems, provided these three components are defined

Ex: hypothesis space defined in terms of parameter values (Yang, 2002) or in terms of how much structure is posited for the language (Perfors, Tenenbaum, & Regier, 2006)

Can combine discrete representations (hypothesis space) with probabilistic components (update procedure) to get gradualness and variation found in human language learning
The Hypothesis Space & The Update Procedure


**Update Procedure**: recent experimental work on probabilistic learning as feasible in adults (Tenenbaum, 2000; Thompson & Newport, 2007) and infants (Newport & Aslin, 2004; Gerken, 2006).
The Hypothesis Space &
The Update Procedure


**Update Procedure**: recent experimental work on probabilistic learning as feasible in adults (Tenenbaum, 2000; Thompson & Newport, 2007) and infants (Newport & Aslin, 2004; Gerken, 2006).

**Bayesian updating**
- Infers likelihood of given hypothesis, given data. Amount of probability shifted depends on layout of hypothesis space.
Road Map

Language Learning Mechanism

Learning Framework
- Separable Components
- Investigating Data Filtering

Case Study: English Anaphoric One
Investigating Data Intake Filtering

Intuition 1: Use all available data to uncover a full range of systematicity, and allow probabilistic model enough data to converge.

Intuition 2: Use more “informative” data or more “accessible” data only.
Modeling Case Study of Data Intake Filters

Case Study: English Anaphoric *One*

Hypothesis Space: *structures & associated referents in world*

Proposed Filtering: *ignore some (pervasive) ambiguous data*

Update Procedure: *Bayesian updating + hypothesis space layout information*

Interesting Feature: *multiple sources of information across domains*
Big Questions for Filtering

(1) **Feasibility**
Is there a data sparseness problem?

(2) **Sufficiency**
Can we filter and get correct behavior?

(3) **Necessity**
Must we filter to get correct behavior?
Road Map

**Language Learning Mechanism**

**Learning Framework**

**Case Study: English Anaphoric One**

- Interesting problems, adult knowledge, & infant behavior
- Linked hypothesis spaces & additional sources of information
- No filters: available data & equal-opportunity learners
- Filters: feasibility considerations
- Data intake filters: sufficiency & necessity
Anaphoric *One*: Why Is It Interesting?

“Look, a red bottle! Do you see another *one*?”

Representations that are linked across domains (syntactic structure & semantic reference)

Available information: linguistic antecedent (*red bottle*) + referent in world
Anaphoric One: Adult Knowledge

“Jack likes this red ball, and Lily likes that one.

\( one = \text{red ball} \)

“Jack likes this ball, and Lily likes that one.

\( one = \text{ball} \)
One = N' (not N^0)
Anaphoric One: Adult Knowledge

Syntax: $one = N'$

Preference when two $N'$ constituents = pick larger one

“Jack likes this [red [ball]$_{N'}$]$_{N'}$, and Lily likes that $one$.”

Semantic consequences: more restrictive set of referents (red balls vs. all balls)
Anaphoric One: Infant Behavior
(Lidz, Waxman, & Freedman 2003)

“Look! A red bottle.”

18-month old baby

TV

camera

0

"Look! A red bottle."

18-month old baby

TV

camera

"Do you see another one?"

(Same results as "Do you see another red bottle?")

18-month old baby

TV

camera
18-month olds have looking preference for red bottle.

LWF (2003) interpretation & conclusion:
Red bottle preference = semantic consequence of syntactic knowledge that 
*one* = [*red bottle*]_{N'}. 18-month olds, like adults, believe *one* has an N’ antecedent (since *red bottle* can’t be N^0).
Road Map

Language Learning Mechanism

Learning Framework

Case Study: English Anaphoric One

- Interesting problems, adult knowledge, & infant behavior
- Linked hypothesis spaces & additional sources of information
- No filters: available data & equal-opportunity learners
- Filters: feasibility considerations
- Data intake filters: sufficiency & necessity
Syntactic Hypothesis Space: Structure
“What is the antecedent of one?”

All elements in the sets described by the hypotheses are possible antecedents of one.

All elements in the $N^0$ set (ex: ball, bottle) are also elements of the $N'$ set. In addition, there are elements in the $N'$ set (ex: red ball, ball behind his back) that are not elements of $N^0$.

Subset-superset relationship
All elements in the sets described by the hypotheses are possible referents of *one*.

All elements in the N’-property set (ex: red balls) are also elements of the any-property set. In addition, there are elements in the any-property set (ex: non-red balls) that are not elements of the N’-property set.

**Subset-superset relationship**

“Jack wants a red ball, and Lily has another *one*.”
Size principle (Tenenbaum & Griffiths, 2001): favor the subset hypothesis when encountering an ambiguous data point.

Specific application to learning anaphoric *one* (Regier & Gahl, 2004).

Size principle logic:
- Likelihood of ambiguous data point $d$
- Learner expectation of set of data points $d_1, d_2, \ldots, d_n$
Exploiting the Hypothesis Space Layout

Subset-superset hypothesis space

Likelihood of $d$ Logic:

Suppose the learner encounters an ambiguous data point $d$.

Let the number of examples covered by subset $A$ be $a$. Let the number of examples covered by superset $B$ be $a + b$. 
The likelihood that $d$ was produced from $A$ is $1/a$. The likelihood that $d$ was produced from $B$ is $1/(a+b)$.

$$1/a > 1/(a+b)$$

So, $A$ has a higher probability of having produced $d$. Thus, $A$ is favored when encountering ambiguous data.
Learner Expectation Logic:

If B were correct, learner should encounter some **unambiguous data points for B**.
Additional Information Source: Exploiting the Hypothesis Space Layout

Subset-superset hypothesis space

Learner Expectation Logic:

If **only subset data points** are encountered, a restriction to the subset A becomes more and more likely.

The more subset data points encountered (while not encountering superset B data points), the more the learner is **biased towards A**.
“Jack wants a ball, and Lily has another one”
"Jack wants a red ball, and Lily has another one"
Road Map

Language Learning Mechanism

Learning Framework

Case Study: English Anaphoric One
- Interesting problems, adult knowledge, & infant behavior
- Linked hypothesis spaces & additional sources of information
- No filters: available data & equal-opportunity learners
- Filters: feasibility considerations
- Data intake filters: sufficiency & necessity
Available Anaphoric One Data

By 18 months, estimated 4017 anaphoric one data points. (CHILDES database)
Note: data points are pairing of utterance and situation

Unambiguous data points: only 10

“Jack wants a red ball, but Lily doesn’t have another one.”
Situation: Lily doesn’t have another red ball. She has a red and a purple one, and wants to keep a red ball herself.
“Jack wants a red ball, but Lily doesn’t have another one”
Available Anaphoric One Data

**Type I Ambiguous** data points: 183
(potential antecedents with modifiers)

“Jack wants a red ball, and Lily has another one for him.”
(Situation: Lily has another red ball. She has two - one for herself, and one for Jack.)

Why ambiguous: She has another ball, as well. One could refer to ball, which is compatible with the N^0 structure.
“Jack wants a red ball, and Lily has another one for him”
Available Anaphoric One Data

Type II Ambiguous data points: 3805
(potential antecedents without modifiers)

“Jack wants a ball, and Lily has another one for him.”
(Situation: Lily has another ball. She has two - one for herself, and one for Jack.)

Why ambiguous: One refers to ball, which is compatible with the N0 structure.
“Jack wants a ball, and Lily has another one for him”
Modeling Anaphoric *One* Learning

Initial State for learner:

Both hypotheses are equiprobable in each hypothesis space

Syntax: $p_{N0} = 0.5$, $p_{N'} = 0.5$

Semantic referents: $p_{N'-property} = 0.5$, $p_{any-property} = 0.5$

Updating, based on data points encountered:

(1) Update probabilities *within* each domain
(2) Update probabilities *across* domains
   (linked hypothesis spaces)
(3) Update for each source of information
   (syntactic & semantic)
Updating Within Domains: Syntax

Two hypotheses: *one* has an antecedent that is $N^0$ or $N'$

Track $p_{N'}$ ($p_{N0} = 1 - p_{N'}$)

Max($\text{Prob}(p_{N'} \mid u)$) = Max($\frac{p_{N'} \cdot \binom{t}{r} \cdot p_{N'}^r \cdot (1 - p_{N'})^{t-r}}{\text{Prob}(u)}$) (for each point $r$, $0 \leq r \leq t$)

$$\frac{d}{dp_{N'}} \left( \frac{p_{N'} \cdot \binom{t}{r} \cdot p_{N'}^r \cdot (1 - p_{N'})^{t-r}}{\text{Prob}(u)} \right) = 0$$

$$\frac{d}{dp_{N'}} \left( \frac{p_{N'} \cdot \binom{t}{r} \cdot p_{N'}^r \cdot (1 - p_{N'})^{t-r}}{\text{Prob}(u)} \right) = 0 \quad (P(u) \text{ is constant with respect to } p_{N'})$$

$$p_{N'} = \frac{r + 1}{t + 1}, \quad r = p_{N' \text{ old}} \cdot t$$

$$p_{N'} = \frac{p_{N' \text{ old}} \cdot t + 1}{t + 1}$$
Two hypotheses: \textit{one} has an antecedent that is \(N^0\) or \(N'\)
Track \(p_{N'} (p_{N0} = 1 - p_{N'})\)

Update: \textbf{Unambiguous Data Point} (10 of 4017)

\[p_{N'} = \frac{p_{N'_\text{old}} \times t + 1}{t + 1}\]

\(t = \# \text{ of data points expected (amount of change allowed)} = 4017\)

“Jack wants a red ball, but Lily doesn’t have another \textit{one}”
Two hypotheses: *one* has an antecedent that is $N^0$ or $N'$

Track $p_{N'}$ ($p_{N0} = 1 - p_{N'}$)

Update: **Unambiguous Data Point** (10 of 4017)

\[
p_{N'} = \frac{p_{N'} \text{ old} \cdot t + 1}{t + 1}
\]

Intuition: 1 added to numerator since learner is fully confident that unambiguous data point signals $N'$ hypothesis

“Jack wants a red ball, but Lily doesn’t have another *one*”
Two hypotheses: *one* has an antecedent that is $N^0$ or $N'$

Track $p_{N'}$ ($p_{N0} = 1 - p_{N'}$)

Update: **Unambiguous Data Point** (10 of 4017)

$$p_{N'} = \frac{p_{N'}^{old} t + 1}{t + 1}$$

Intuition: 1 added to denominator since 1 data point seen

“Jack wants a red ball, but Lily doesn’t have another *one*”
Updating Within Domains: Syntax

Update: **Unambiguous Data Point** (10 of 4017)
Two hypotheses: *one* has an antecedent that is $N^0$ or $N'$

Track $p_{N'}$ ($p_{N0} = 1 - p_{N'}$)

Update: **Type II Ambiguous Data Point** (3805 of 4017)

$p_{N'} = \frac{p_{N'} \cdot t + p_{N'}|a}}{t + 1}$

Intuition: number added should be less than 1, since learner is not certain that type II ambiguous data point signals $N'$ hypothesis

“Jack wants a ball, and Lily has another *one* for him”
Updating Within Domains: Syntax

Two hypotheses: *one* has an antecedent that is \( N^0 \) or \( N' \)

Track \( p_{N'} (p_{N0} = 1 - p_{N'}) \)

Update: **Type II Ambiguous Data Point** (3805 of 4017)

\[
p_{N'} = p_{N' \text{ old}} \frac{t}{t + 1} + p_{N'|a}
\]

Value added is partial confidence value, \( p_{N'|a} \), which will be < 1. Using size principle, where the relative sizes of the hypotheses influence how much bias there is for the subset \( (N^0) \)

“Jack wants a ball, and Lily has another *one* for him”
Type II Ambiguous: $N^0$ Subset Bias

If hypotheses are defined by what **word strings** they cover, the $N^0$ set is much smaller than the $N'$ set (based on vocabulary).

The bias towards the subset $N^0$ is stronger = more bias towards the incorrect hypothesis.

MacArthur CDI (Dale & Fenson, 1996) estimates: subset-to-superset ratio $\approx 1/50$
Type II Ambiguous: \(N^0\) Subset Bias

If hypotheses are defined by what category strings they cover, the \(N^0\) set is more comparable to the \(N'\) set.

The bias towards the subset \(N^0\) is weaker = less bias towards the incorrect hypothesis.

For generous estimates of learner performance: use category instantiation.

subset-to-superset ratio = 1/4
Updating Within Domains: Syntax

Two hypotheses: one has an antecedent that is $N^0$ or $N'$

Track $p_{N'}$ ($p_{N0} = 1 - p_{N'}$)

Update: **Type II Ambiguous Data Point** (3805 of 4017)

$$p_{N'} = p_{N'}^{old} \cdot t + \frac{p_{N'} | a}{t + 1}$$

Example Update for **Type II Ambiguous**

$$p_{N'} = 0.5, \ t = 4017, \ \text{subset-to-superset ratio} = 0.25$$

$$p_{N'} = 0.5 \cdot 4017 + 0.2 = \frac{.499925}{4017 + 1}$$ (slight bias for $N^0$)
Updating Within Domains: Syntax

Two hypotheses: one has an antecedent that is $N^0$ or $N'$
Track $p_{N'}$ ($p_{N0} = 1 - p_{N'}$)

Update: **Type I Ambiguous Data Point** (183 of 4017)

\[
p_{N'} = \frac{p_{N',\text{old}}^t + ???}{t + 1}
\]

Intuition: value should be < 1 (learner not fully confident).

“Jack wants a red ball, and Lily has another one for him”
Two hypotheses: *one* has an antecedent that is $\text{N}^0$ or $\text{N}'$

Track $p_{\text{N}'}$ ($p_{\text{N0}} = 1 - p_{\text{N}'}$)

Update: **Type I Ambiguous Data Point** (183 of 4017)

\[
p_{\text{N}'} = \frac{p_{\text{N}'}_{\text{old}} * t + 1}{t + 1}
\]

However, we’ll be generous and allow full confidence. This gives an overestimation of the learner’s probability of converging on the $\text{N}'$ hypothesis.

“Jack wants a red ball, and Lily has another *one* for him”
Two hypotheses: one has referent with \textbf{any-prop} or \textbf{N'-prop}

Track $p_{\text{N'-prop}} (p_{\text{any-prop}} = 1 - p_{\text{N'-prop}})$

Update: \textbf{Unambiguous} + Type I Ambiguous (193 of 4017)

\[
p_{\text{N'}} = p_{\text{N'}_{\text{old}}} \frac{t + ???}{t + 1}
\]

“…red ball…”
Updating Within Domains: Semantics

Two hypotheses: one has referent with \textit{any-prop} or \textit{N'-prop}

Track $p_{N'-\text{prop}}$ ($p_{\text{any-prop}} = 1 - p_{N'-\text{prop}}$)

Update: \textbf{Unambiguous} + \textbf{Type I Ambiguous} (193 of 4017)

\[ p_{N'} = \frac{p_{N' \text{ old}}^t + p_{N'-\text{prop} | s}}{t + 1} \]

“…red ball…”

Value added is partial confidence value, $p_{N'-\text{prop}|s}$, which will be < 1. Using size principle, where the relative sizes of the hypotheses influence how much bias there is for the subset (N’-prop)
If the learner is aware of many types of balls in the world (so that red balls are a small subset), the bias for the subset is greater. This is the \textit{correct bias}.

Generous: Assume number of ball types corresponds to number of adjectives known at 18 months (MacArthur CDI $\approx 49$) even though all won’t necessarily apply to the balls in the situation.

“…red ball…”

\textit{Lily}
Two hypotheses: one has referent with any-prop or N’-prop

Track $p_{N’\text{-prop}} (p_{\text{any-prop}} = 1 - p_{N’\text{-prop}})$

Update: Unambiguous + Type I Ambiguous (193 of 4017)

$$p_{N’} = \frac{p_{N’ \text{ old}} \cdot t + p_{N’\text{-prop} \mid s}}{t + 1}$$

“…red ball…”
Two hypotheses: *one* has referent with **any-prop** or **N’-prop**

Track \( p_{N’-prop} = 1 - p_{any-prop} \)

Update: **Type II Ambiguous** (3805 of 4017)

No update function invoked for semantic referents, because no subset is defined. (No N’-property.)
Updating Across Domains & From Multiple Data Sources

syntax

\[ N' \]
Prob = 0.5

\[ N^0 \]
Prob = 0.5

semantics

any-property
Prob = 0.5

N'-property
Prob = 0.5
Updating Across Domains & From Multiple Data Sources

- **syntax**
  - N'
    - Prob = 0.5
  - N^0
    - Prob = 0.5

- **semantics**
  - any-property
    - Prob = 0.5
  - N'-property
    - Prob = 0.5
Encounter data point: Unambiguous/Type I Ambiguous

Unambiguous/Type I Ambiguous Data point

syntax: “…red ball…one…” (N’)

semantics: N’-property
Choose one domain to update (Syntax hypotheses)

Unambiguous/Type I Ambiguous Data point

**Syntax:** “…red ball…one…” (N’)

**Semantics:** N’-property

\[ N' \]
\[ \text{Prob} = 0.5 \]

\[ N^0 \]
\[ \text{Prob} = 0.5 \]

\[ \text{any-property} \]
\[ \text{Prob} = 0.5 \]

Choose one domain to update (Syntax hypotheses)
Choose one domain to update (Syntax hypotheses)

Unambiguous/Type I Ambiguous Data point

Syntax: “…red ball…one…” (N’)

Semantics: N’-property
Update linked hypotheses (Semantic consequences)

syntax

N'
Prob = 0.6

N^0
Prob = 0.4

semantics

any-property
Prob = 0.4

N’-property
Prob = 0.6

Unambiguous/Type I Ambiguous Data point

syntax: “…red ball…one…” (N’)

semantics: N’-property
Unambiguous/Type I Ambiguous Data point

**syntax:** “…red ball…one…” (N’)

**semantics:** N’-property

Prob = 0.4

Prob = 0.6

Prob = 0.4
Update the other domain (Semantic hypotheses)

Unambiguous/Type I Ambiguous Data point

syntax: “…red ball…one…” (N’)

semantics: N’-property
Unambiguous/Type I Ambiguous Data point

syntax: “…red ball…one…” (N’)

semantics: N’-property
Unambiguous/Type I Ambiguous Data point

syntax: “...red ball...one...” (N')

semantics: N’-property
Prob = 0.63
Prob = 0.37

Unambiguous/Type I Ambiguous Data point

syntax: “…red ball…one…” (N’)

semantics: N’-property
Encounter data point: Type II Ambiguous

Syntax:

- N' with Prob = 0.63
- N^0 with Prob = 0.37

Semantics:

- Any-property with Prob = 0.37
- N'-property with Prob = 0.63

Syntax: “…ball…one…” (N^0 bias)

Semantics: N/A
Type II Ambiguous Data point

syntax: “…ball…one…” (N^0 bias)

semantics: N/A
Type II Ambiguous Data point

syntax: “…ball…one…” (N⁰ bias)

semantics: N/A
Type II Ambiguous Data point
syntax: “…ball…one…” (N⁰ bias)

syntax: N’
Prob = 0.58

N⁰
Prob = 0.42

semantics:

any-property
Prob = 0.37

N’-property
Prob = 0.63

semantics: N/A
Metric of Success: Does an equal-opportunity learner (no data filters) steadily increase the probability of interpreting anaphoric one correctly? (sufficiency)

\[ one = N' \ (p_{N'}) \]

semantic referent = set corresponding to larger \( N' \) (\( p_{N'\text{-prop}} \))

"Look! A red bottle. Do you see another one?"

\[ \text{Prob(correct interpretation)} = p_{N'} \ast p_{N'\text{-prop}} \]

initial = 0.5*0.5 = 0.25
Learning Without Filters: The Equal-Opportunity Learner

The equal-opportunity learner has incorrect behavior: learning without filters is **insufficient** even with generous estimates of variables involved.
Road Map

**Language Learning Mechanism**

**Learning Framework**

**Case Study: English Anaphoric One**

- Interesting problems, adult knowledge, & infant behavior
- Linked hypothesis spaces & additional sources of information
- No filters: available data & equal-opportunity learners
- Filters: feasibility considerations
- Data intake filters: sufficiency & necessity
Data Intake Filtering

Possible Filter: Use only **Unambiguous** data (Pearl & Weinberg, 2007; Dresher, 1999; Lightfoot, 1999; Fodor, 1998)

problem: **feasibility**

Estimate from CHILDES: Only 10 data points are unambiguous for the correct interpretation of anaphoric *one* - out of months and months of available data

**Data sparseness!**
Data Intake Filtering

Possible Filter: Use Unambiguous & Type I Ambiguous data
- less data sparseness (feasibility): 193 total
- data will bias learner in the correct direction
- Note: Still use both syntactic & semantic information (different from Regier & Gahl, 2004)

Metric of Success: Does learner steadily increase probability of interpreting anaphoric one correctly (sufficiency)

“Look! A red bottle. Do you see another one?”
Road Map

Learning Framework Overview

Computational Case Studies:

Brief Highlights: Old English OV/VO word order
Details: English Metrical Phonology
Highlights: English Anaphoric *One*
  - interesting problems, adult knowledge, & infant behavior
  - available data & filter feasibility considerations
  - additional sources of information: hypothesis space layout
  - data intake filters: sufficiency & necessity
The learner that uses data intake filtering has correct behavior: learning without filters is *sufficient*.
Data Intake Filtering: Big Questions

Filter: Use only Unambiguous & Type I Ambiguous data

**Feasible:** can find sufficient data

**Sufficient:** produces behavior qualitatively similar to human learners

**Necessary:** removing the filter and learning from all available data (specifically type II ambiguous) produces behavior unlike human learners
How does a learner know to use this filter?

Want: Filter to ignore type II ambiguous data to result from some principled strategy for learning

Principled strategy: Learn only in cases of uncertainty (Shannon 1948; Gallistel 2001) - that’s where information is gained
How does a learner know to use this filter?

Want: Filter to ignore type II ambiguous data to result from some principled strategy for learning

Principled strategy: Learn only in cases of uncertainty (Shannon 1948; Gallistel 2001) - that's where information is gained

Need to ignore: data points where potential antecedent has no modifier

Jack wants a ball and Lily has another one for him.
How does a learner know to use this filter?

Want: Filter to ignore type II ambiguous data to result from some principled strategy for learning

Possibility 1: Look for situations where there is uncertainty in the semantic referent set (e.g. balls vs. red balls) only. This will occur when the utterance has a modifier on the potential antecedent (e.g. red ball).

Jack wants a red ball and Lily has/doesn’t have another one for him.
Problem: Learner must only care about semantic referents and not about syntactic structure ($N'$ vs. $N^0$). (~Regier & Gahl, 2004) Then, only updating hypotheses from semantic information, not semantic & syntactic. Result: lower probability of correct interpretation.
How does a learner know to use this filter?

Want: Filter to ignore type II ambiguous data to result from some principled strategy for learning

Possibility 2: Syntactocentric approach, and solving the problem of *which N’ antecedent* is correct when there is more than one. Only relevant data are those with multiple potential N’ antecedents (e.g. nouns with modifiers like *red ball*).

Jack wants a red ball and Lily has/doesn’t have another one for him.
Syntactocentric Approach

Requirement: Prior knowledge that the antecedent of one is N’. 
Methods:
- Innate constraints (Hornstein & Lightfoot, 1981)
- Syntactocentric filter over distribution of one vs. distribution of other nouns w.r.t complements (Foraker et al. in press)

Benefit: learner uses syntactic data to update as well since this is a question of which syntactic antecedent (larger or smaller N’) is correct

Jack wants a red ball and Lily has/doesn’t have another one for him.

\[\text{one} = \text{N’}\]
Syntactocentric Approach
Anaphoric One: Filters (Recap)

Feasible:

Jack only learns from this unambiguous data point, but Lily learns from that ambiguous one, too.

Jack has a data sparseness problem. Lily doesn’t.

Data filters can be made feasible for this case study.
Anaphoric One: Filters (Recap)

**Feasible:** Data filters can be made feasible for this case study.

**Sufficient:**

*Jack used this semantocentric filter, and Lily used that syntactocentric one.*

Filter used: Ignore type II ambiguous data.

**Learner instantiation:**

- **Good:** semantocentric approach, views only semantic data as relevant
- **Better:** syntactocentric approach, still allowing multiple sources of information (syntactic & semantic referents)

Filtering produced qualitatively **correct behavior**.
Anaphoric One: Filters (Recap)

**Feasible:** Data filters can be made feasible for this case study.

**Sufficient:** Filtering produced qualitatively correct behavior.

**Necessary:**

*Jack only learns from this ambiguous data point, but Lily learns from that one, too.*

Lily fails if she’s using type II ambiguous data (i.e. no filter).

Filtering was **necessary** for correct behavior.
Anaphoric One: Filters (Recap)

**Feasible:** Data filters can be made feasible for this case study.

**Sufficient:** Filtering produced qualitatively correct behavior.

**Necessary:** Filtering was necessary for correct behavior.
Big Picture
Big Picture

(1) Explaining language learning: theory of the mechanism
Big Picture

(1) Explaining language learning: theory of the mechanism

(2) Learning framework: separable components that can be explored individually
Big Picture

(1) Explaining language learning: theory of the mechanism

(2) Learning framework: separable components that can be explored individually

(3) Data intake filtering: feasibility, sufficiency, necessity
Big Picture

(1) Explaining language learning: theory of the mechanism

(2) Learning framework: separable components that can be explored individually

(3) Data intake filtering: *feasibility, sufficiency, necessity*

(4) Computational modeling: tool for exploring questions of the learning mechanism
Thank You

Jeff Lidz
Colin Phillips
Amy Weinberg
Norbert Hornstein
Howard Lasnik
Bill Idsardi
Paul Pietroski

the Psycho-Acquisition Lab Group
at the University of Maryland

the Cognitive Neuroscience of Language Lab
at the University of Maryland
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

\[
\text{Max}(\text{Prob}(p_{N'}|u)) = \text{Max}\left(\frac{\text{Prob}(u|p_{N'}) \times \text{Prob}(p_{N'})}{\text{Prob}(u)}\right)
\]

Bayes’ Rule, find maximum of a posteriori (MAP) probability
Manning & Schütze (1999)
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

\[
\text{Max}(\text{Prob}(p_{N'}|u)) = \text{Max}\left(\frac{\text{Prob}(u|p_{N'}) \times \text{Prob}(p_{N'})}{\text{Prob}(u)}\right)
\]

\text{Prob}(u | p_{N'}) = \text{probability of seeing unambiguous data point } u, \text{ given } p_{N'} = p_{N'}

\text{Prob}(p_{N'}) = \text{probability of seeing } r \text{ out of } t \text{ data points that are unambiguous for } N', \text{ for } 0 \leq r \leq t

= \binom{t}{r} \times p_{N'}^r \times (1 - p_{N'})^{t-r}
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

\[ \text{Max}(\text{Prob}(p_{N'} | u)) = \text{Max}(\frac{p_{\text{vo}} * \binom{t}{r} * p_{N'}^r * (1 - p_{N'})^{t-r}}{\text{Prob}(u)}) \]  
(for each point \( r, 0 \leq r \leq t \))

\[ \frac{d}{dp_{N'}} \left( \frac{p_{N'}^r * \binom{t}{r} * (1 - p_{N'})^{t-r}}{\text{Prob}(u)} \right) = 0 \]

\[ \frac{d}{dp_{N'}} \left( \frac{p_{N'}^r * \binom{t}{r} * (1 - p_{N'})^{t-r}}{\text{Prob}(u)} \right) = 0 \]  
(P\( (u) \) is constant with respect to \( p_{N'} \))

\[ p_{N'} = \frac{r + 1}{t + 1} \]
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

\[ p_{N'} = \frac{r + 1}{t + 1}, \quad t = p_{N' \text{ old}} \times t \]

\[ p_{N'} = \frac{p_{N' \text{ prev}} \times t + 1}{t + 1} \]
Ambiguous Data Points: Type II (Syntactic)

\[ p_{N'} = \frac{p_{N'} \text{old} \ast t + p_{N'} \ast a}{t + 1}, \text{ ambiguous} = "...ball..." \]

\[ p_{N'} \ast (\frac{n}{n + o}) + (1 - p_{N'}) \ast 1 \]

\[ p_{N'} \ast \sum_{\text{hypothesis}} p(a \mid \text{hypothesis}) \ast p(a \mid p_{N'}) + p_{N0} \ast (a \mid p_{N0}) \]

\[ p_{N'} \ast \frac{n}{n + o} \]

\[ p_{N'} \ast \frac{n}{n + o} \]
Ambiguous Data Points: Type II (Semantic)

\[ p_{N'}^{\text{-prop}} = \frac{p_{N'}^{\text{-prop old}} \ast t + p_{N'}^{\text{-prop}} \mid a}{t + 1}, \text{ ambiguous } = \text{ ball of } N'\text{-property} \]

\[ p_{N'}^{\text{-prop}} \mid a = \frac{\text{Prob}(a \mid N'\text{-prop}) \ast \text{Prob}(N'\text{-prop})}{\text{Prob}(a)} = \frac{1 \ast p_{N'}^{\text{-prop}}}{p_{N'}^{\text{-prop}} \ast 1 + (1 - p_{N'}^{\text{-prop}}) \ast \frac{1}{c}} \]

\[ \sum_{\text{hypotheses}} p_{\text{hypothesis}} \ast p(a \mid p_{\text{hypothesis}}) \\
\sum_{\text{hypotheses}} p_{N'}^{\text{-prop}} \ast p(a \mid p_{N'}^{\text{-prop}}) + p_{\text{any} - \text{prop}} \ast p(a \mid p_{\text{any} - \text{prop}}) \\
p_{N'}^{\text{-prop}} \ast 1 + (1 - p_{N'}^{\text{-prop}}) \ast \frac{1}{c} \]