Necessary Bias in Natural Language Learning
Lisa Pearl
University of Maryland
May 3, 2007

Natural Language Learning

Theoretical work: object of acquisition
Experimental work: time course of acquisition

worthwhile: mechanism of acquisition
given the boundary conditions provided by
(a) linguistic representation
(b) the trajectory of learning

The Learning Problem

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

The Mechanism of Language Learning:
Some Bias = Parameters

Premise: learner considers finite range of hypotheses
(parameters)

"Assuming that there are n binary parameters, there will
be 2^n possible core grammars." - Clark (1994)
The Mechanism of Language Learning: Extracting Systematicity Is Hard

"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?
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?

Computational Modeling of Data Intake Filtering

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

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

Road Map

Learning Framework Overview

Computational Case Studies:
Brief Highlights: Old English OV/VO word order
Details: English Metrical Phonology
Highlights: English Anaphoric 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

Important Feature: Case studies grounded in empirical data searching realistic data space for evidence of underlying system

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, 2008)
- 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

Hypothesis Space: theoretical and experimental work on what hypotheses children entertain (e.g. Lidz, Waxman, & Freedman, 2003; Thornton & Crain, 1999; Hamburger & Crain, 1984)

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.

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 Studies of Data Intake Filters

Case One: Old English Syntax
Hypothesis Space: parameters (OV/VO word order)
Proposed Filtering: Degree-0 unambiguous data only
Update Procedure: Bayesian updating
Interesting Feature: target state is a probability distribution

Case Two: English Metrical Phonology
Hypothesis Space: parameters
Proposed Filtering: unambiguous data only
Update Procedure: Bayesian updating
Interesting Feature: multiple interactive parameters; noisy data
Modeling Case Studies of Data Intake Filters

Case Three: 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

Learning Framework Overview

Computational Case Studies:

Brief Highlights: Old English OV/VO word order
- unambiguous degree-0 data filtering
  - feasibility
  - sufficiency & necessity
Details: English Metrical Phonology
Highlights: English Anaphoric One

Old English Filters

Filter 1: Use data perceived as unambiguous (Dresher, 1999; Lightfoot, 1999; Fodor, 1998)

Filter 2: Use structurally "simple" data - matrix clause or "degree-0" data (Lightfoot, 1991)

Jack told his mother that he stole the golden goose.
[----Degree-0-------]
[-------------Degree-1----------]
Problems: Feasibility

Potential feasibility problem: data sparseness
degree-0 unambiguous data set is significantly smaller than entire input set

How could a learner find unambiguous data for OV/VO word order?

Perceived Unambiguous Data: Making “Unambiguous” Feasible

Definitions of data perceived as unambiguous are heuristic and/or involve only partial knowledge of the adult linguistic system (Lightfoot 1999, Dresher 1999, Fodor 1998)

OV: 
[...]_p ... Object TensedVerb ...
... Object Verb-Marker ...

VO: 
[...]_p [...]_p ... TensedVerb Object ...
... Verb-Marker Object ...

This allows the learner to identify some data points as unambiguous (even if they’re actually not for someone with full knowledge of the adult linguistic system)
Sufficiency of Filters: Correct Behavior (Population-Level)

- Avg. \( p_{ij} \) in Population Over

Necessity of Filters: Removal = Incorrect Behavior

Using ambiguous data

Using ambiguous & degree-1 data

Big Questions for Filtering: Old English Syntax

(1) **Feasibility**
No data sparseness problem.

(2) **Sufficiency**
Filtering yields the correct behavior.

(3) **Necessity**
Removing the filters yields incorrect behavior.
Road Map

Learning Framework Overview

Computational Case Studies:
- Brief Highlights: Old English OV/VO word order
- Details: English Metrical Phonology
  - unambiguous data feasibility in a complex system: cues vs. parsing
  - metrical phonology overview: interacting parameters
  - cues vs. parsing in metrical phonology
  - English metrical phonology
  - sufficiency: logical problem of language acquisition
- Highlights: English Anaphoric One

Feasibility

Unambiguous data filter feasibility in a complex system

Data sparseness: are there unambiguous data? (Clark 1992)

How could a learner identify such data?

Metrical phonology (9 interacting parameters)

Interactive Parameters

The order in which parameters are set may determine if they are set correctly (Dresher, 1999): parameter-setting influences what data are identified as "unambiguous".

Identifying unambiguous data:
- **Cues** (Dresher, 1999; Lightfoot, 1999)

Parsing (Fodor, 1998; Sakas & Fodor, 2001)

Cues vs. Parsing: Overview

A cue is a local "specific configuration in the input" that corresponds to a specific parameter value. A cue matches an unambiguous data point. (Dresher, 1999)
Cues vs. Parsing: Overview

A cue is a local "specific configuration in the input" that corresponds to a specific parameter value. A cue matches an unambiguous data point. (Brether, 1999)

 Parsing tries to analyze a data point with "all possible parameter value combinations", conducting an "exhaustive search of all parametric possibilities." (Fodor, 1998)

Cues vs. Parsing: Comparison

<table>
<thead>
<tr>
<th></th>
<th>Cues</th>
<th>Parsing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy identification of unambiguous data</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Can find information in datum sub-part</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Can tolerate exceptions</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Is not heuristic</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Does not require additional knowledge</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Does not use default values</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

Cues vs. Parsing in a Probabilistic Framework

"Both models ... cannot capture the variation in and the gradualness of language development...when a parameter is set, it is set in an all-or-none fashion." - Yang (2002)

Benefit of using learning framework to sidestep this problem - separable components used in combination:

1. **cues/parsing** to identify unambiguous data
2. probabilistic framework of **gradual updating** based on unambiguous data

Road Map

Computational Case Studies:
- Brief Highlights: Old English OV/VO word order
- Details: English Metrical Phonology
  - unambiguous data feasibility in a complex system: cues vs. parsing
  - metrical phonology overview: interacting parameters
  - cues vs. parsing in metrical phonology
  - English metrical phonology
  - sufficiency: logical problem of language acquisition

Highlights: English Anaphoric One
Metrical Phonology

What tells you to put the emphasis on a particular syllable

Sample metrical phonology structure

\[
(\text{em} \quad \text{pha} \quad \text{sis})
\]

Metrical Phonology Parameters

- Quantity Sensitivity
- Extrametricality
- Feet Directionality
- Feet Boundedness
- Feet Headedness

Road Map

Learning Framework Overview

Computational Case Studies:
- Brief Highlights: Old English OV/VO word order
- Details: English Metrical Phonology
  - unambiguous data feasibility in a complex system:
    - cues vs. parsing
  - metrical phonology overview: interacting parameters
  - cues vs. parsing in metrical phonology
  - English metrical phonology
  - sufficiency: logical problem of language acquisition
- Highlights: English Anaphoric One

Cues for Metrical Phonology Parameters

Recall: Cues match local surface structure (sample cues below)

- **QS:** 2 syllable word with 2 stresses
  - VV VV

- **Em-Right:** Rightmost syllable is Heavy and unstressed
  - LH H

- **Unb:** 3+ unstressed S/L syllables in a row
  - ...SSS...
  - ...LLLLL

- **Ft Hd Left:** Leftmost foot has stress on leftmost syllable
  - SS S...
  - HLLL...
If all successful parses of a data point share one value of a parameter (e.g., "Extrametrical None"), that data point is considered unambiguous for that parameter value.
If (Values leading to successful parses of data point):
(QI, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QI, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)

Data point is unambiguous for Em-None.

If QI already set, data point is unambiguous for Em-None, B, B-2, and B-Syl.
Finding Unambiguous Data: English Metrical Phonology

Non-trivial system: metrical phonology

Non-trivial language: English (full of exceptions)

data unambiguous for the incorrect value in the adult system

Adult English system values:

- QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, B-2, B-Syllabic, Ft Hd Left

Exceptions:

- Qi, QSVCL, Em-None, Ft Dir Left, Unbounded, B-3, B-Moraic, Ft Hd Right

Empirical Grounding in Realistic Data: Estimating English Data Distributions

Caretaker speech to children between the ages of 6 months and 2 years (CHILDES: MacWhinney, 2000)

Total Words: 540505
Mean Length of Utterance: 3.5

Words parsed into syllables and assigned stress using the American English CALLHOME database of telephone conversation (Canavan et al., 1997) & the MRC Psycholinguistic database (Wilson, 1988)

Road Map

Learning Framework Overview

Computational Case Studies:

- Brief Highlights: Old English OV/VO word order
- Details: English Metrical Phonology
  - unambiguous data feasibility in a complex system:
    - cues vs. parsing
    - metrical phonology overview: interacting parameters
    - cues vs. parsing in metrical phonology
    - English metrical phonology
    - sufficiency: logical problem of language acquisition
- Highlights: English Anaphoric One

Sufficient Filters: Viable Parameter-Setting Orders

Can learners using unambiguous data (identified by either cues or parsing) learn the English system? What parameter-setting orders are viable?

Viable orders are derived for each method via an exhaustive walk through all possible parameter-setting orders.
Viable Parameter-Setting Orders: Encapsulating the Knowledge for Acquisition Success

**Worst Case:** learning with filters produces *insufficient* behavior. No orders lead to correct system.

**Better Cases:** learning with filters produces *sufficient* behavior.

- Slightly Better Case: Viable orders available, but fairly random.
- Better Case: Viable orders available, can be captured by small number of *order constraints*.
- Best Case: All orders lead to correct system.

---

**Identifying Viable Parameter-Setting Orders**

(a) For all currently unset parameters, determine the unambiguous data distribution in the corpus.

(b) Choose a currently unset parameter to set. The value chosen for this parameter is the value that has a higher probability in the data the learner perceives as unambiguous.

(c) Repeat steps (a-b) until all parameters are set.
Identifying Viable Parameter-Setting Orders

(a) For all currently unset parameters, determine the unambiguous data distribution in the corpus.

(b) Choose a currently unset parameter to set. The value chosen for this parameter is the value that has a higher probability in the data the learner perceives as unambiguous.

(c) Repeat steps (a-b) until all parameters are set.

(d) Compare final set of values to English set of values. If they match, this is a viable parameter-setting order.

(e) Repeat (a-d) for all parameter-setting orders.

Cues: Parameter-Setting Orders

Cues: Sample viable orders
(a) QS, QS-VC-Heavy, Bound, Bound-2, Feet HD Left, Feet Dir Right, Em-Some, Em-Right, Bound-Syl
(b) Feet Dir Right, QS, Feet HD Left, Bound, QS-VC-Heavy, Bound-2, Em-Some, Em-Right, Bound-Syl

Cues: Sample failed orders
(a) QS, Bound, Feet HD Left, Feet Dir Right, QS-VC-Heavy, Em-Some, Em-Right, Bound-Syl, Bound-2
(b) Feet HD Left, Feet Dir Right, Bound, Bound-Syl, Bound-2, QS, QS-VC-Heavy, Em-Some, Em-Right

…but only for certain assumptions about probability relativization.

Sufficiency of an Unambiguous Filter

Are there any viable parameter-setting orders for a learner using either method (cues or parsing)? What constraints are there?
Parsing: Parameter-Setting Orders

Parsing: Sample viable orders
(a) Bounded, QS, Feet Hd Left, Feet Dir Right, QS-VC-Heavy, Bounded-Syl, Em-Some, Em-Right, Bounded-2
(b) Feet Hd Left, QS, QS-VC-Heavy, Bounded, Feet Dir Right, Em-Some, Em-Right, Bounded-Syl, Bounded-2

Parsing: Sample failed orders
(a) Feet Dir Right, QS, Feet Hd Left, Bounded, QS-VC-Heavy, Bounded-2, Em-Some, Em-Right, Bounded-Syl
(b) Em-Some, Em-Right, QS, Bounded, Feet Hd Left, Feet Dir Right, QS-VC-Heavy, Bounded-Syl, Bounded-2

...irrespective of what probability relativization assumptions are made.

Cues vs. Parsing: Order Constraints

Cues
(a) QS-VC-Heavy before Em-Right
(b) Em-Right before Bounded-Syl
(c) Bounded-2 before Bounded-Syl

The rest of the parameters are freely ordered w.r.t. each other.
Note: Constraints are derivable from properties of the learning system.

Parsing
Group 1:
QS, Ft Head Left, Bounded
Group 2:
Ft Dir Right, QS-VC-Heavy
Group 3:
Em-Some, Em-Right, Bounded-2, Bounded-Syl

The parameters are freely ordered w.r.t. each other within each group.
Note: Most constraints are not derivable from properties of the learning system.

Feasibility & Sufficiency of the Unambiguous Data Filter

Either method of identifying unambiguous data (cues or parsing) is successful. Given the non-trivial system (9 interactive parameters) and the non-trivial data set (English is full of exceptions), this is no small feat.

"It is unlikely that any example ... would show the effect of only a single parameter value" - Clark (1994)

(1) Feasibility & Sufficiency:
- Unambiguous data identified in sufficient quantities
- Correct systematicity can be extracted

(2) This filter is robust across a realistic (highly ambiguous, exception-filled) data set.
Big Questions for Filtering: English Metrical Phonology

(1) Feasibility
No data sparseness problem, even in complex system with multiple interactive parameters.

(2) Sufficiency
Filtering yields the correct behavior.

(3) Necessity
Future investigation

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

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 = red ball

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

one = ball
Anaphoric One: Adult Knowledge

Syntax: \( \text{one} = N' \)

Preference when two \( N' \) constituents = pick larger one

"Jack likes this [red [ball] \( \_\_\_\_\_\_\_\_\_\_\_\_\) and Lily likes that one."

Semantic consequences: more restrictive set of referents (red balls vs. all balls)

Anaphoric One: Infant Behavior (LWF 2003)

"Look! A red bottle."

18-month old baby
"Do you see another one?"
(Same results as "Do you see another red bottle?")

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]₀. 18-month olds, like adults, don’t think one can have an N₀ antecedent.

Available Anaphoric One Data
By 18 months, estimated 4017 anaphoric one data points. But…only 10 of these are unambiguous.

"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.)

Feasibility problem: data sparseness
Potential Solution: Utilize ambiguous data somehow
Using Ambiguous Data

Type I: 183 data points
“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 N0 structure.

Type II: 3805 data points
“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.

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

Additional Information Source: Exploiting the Hypothesis Space Layout

Subset-superset hypothesis space

Size principle (Tenenbaum & Griffiths, 2001):
favor the subset hypothesis when encountering an ambiguous data point

Size principle logic:
- Likelihood of ambiguous data point d
- Learner expectation of set of data points d₁, d₂, … dₙ
Anaphoric One: Hypothesis Space Layout

(Towards the right hypothesis) Type I Ambiguous: “…red ball…one…”

(Towards the wrong hypothesis) Type II Ambiguous: “…ball…one…”

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
Data Intake Filtering

Filter: Use only Unambiguous & Type I Ambiguous data
- less data sparseness (feasibility)
- data will bias learner in the correct direction (Regier & Gahl (2004) insight)
- Note: Use both syntactic & semantic information

Metric of Success: Does learner steadily increase probability of interpreting anaphoric one as real 18-month olds do? (sufficiency)

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

Data Intake Filtering: Sufficiency

Feasible: can find sufficient data

Sufficient: produces behavior qualitatively similar to human learners

Necessary?
What happens if we remove the filter and learn from all available data (specifically type II ambiguous, which biases the learner in the wrong direction)?

Equal-Opportunity Learner

Probability of adult interpretation of anaphoric one for different quantities of data encountered
Data Intake Filtering
Filter: Use only Unambiguous & Type I Ambiguous data
   **Feasible:** can find sufficient data
   **Sufficient:** produces behavior qualitatively similar to human learners
   **Necessary:** incorrect behavior results when we remove the filtering

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 (perhaps contrary to intuition)
(4) Computational modeling: tool for exploring questions of the learning mechanism & generating testable predictions

Thank You

My Fabulous Thesis Committee:
Amy Weinberg, Jeff Lidz, Bill Idsardi, Charles Yang, Jim Reggia

My Awesome Intellectual/Moral Support:
Norbert Hornstein, Philip Resnik, Colin Phillips, David Poeppel, Peggy Antonisse, Andrea Zukowski, Howard Lasnik, Michelle Hugue, Heather Taylor, Brian Dillon, Yuval Marton, Rachel Shorey, Annie Gagliardi, Raven Alder, Elizabeth Royston, Robert Snyder, Bill Sakas, Cedric Boeckx, Ivano Caponigro, the CNL Lab at UMaryland

Causes of Language Change

Old Norse influence before 1000 A.D.: VO-biased
If sole cause of change, requires exponential influx of Old Norse speakers.

Old French at 1066 A.D.: embedded clauses predominantly OV-biased (Kibler, 1984)
Matrix clauses often SVO (ambiguous)
OV-bias would have hindered Old English change to VO-biased system.

Evidence of individual probabilistic usage in Old English
Historical records likely not the result of subpopulations of speakers who use only one order.
Scandinavian Influence, Perfect Learning

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

\[
\text{Max}(\text{Prob}(p_{VO} | u)) = \text{Max}\left(\frac{\text{Prob}(u | p_{VO}) \cdot \text{Prob}(p_{VO})}{\text{Prob}(u)}\right)
\]

Bayes’ Rule, find maximum of a posteriori (MAP) probability
Manning & Schütze (1999)

\[
\text{Prob}(u | p_{VO}) = \text{probability of seeing unambiguous data point } u, \text{ given } p_{VO} = p_{VO}
\]

\[
\text{Prob}(p_{VO}) = \text{probability of seeing } r \text{ out of } n \text{ data points that are unambiguous for VO, for } 0 \leq r \leq n
\]

\[
= \binom{n}{r} \cdot p_{VO}^r \cdot (1 - p_{VO})^{n-r}
\]

Scandinavian Influence, Perfect Learning
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

\[
\max(\text{Prob}(p_{VO} | u)) = \max(p_{VO} \cdot r^n \cdot (1 - p_{VO})^{n-r} \cdot \text{Prob}(u)) \quad \text{(for each point } r, 0 \leq r \leq n) \\
\frac{d}{dp_{VO}} \left( \frac{p_{VO}^r \cdot (1 - p_{VO})^{n-r}}{\text{Prob}(u)} \right) = 0 \\
p_{VO} = \frac{r + 1}{n + 1}
\]

Estimating Historical \(p_{VO}\)

Known quantities:
Unambiguous and ambiguous data in \(d_0\) and \(d_1\)

\(\text{OV Unamb} \quad \text{Amb} \quad \text{VO Unamb}\)

\(D_0\)

\(\text{OV Unamb} \quad \text{Amb} \quad \text{VO Unamb}\)

\(D_1\)
Estimating Historical $p_{VO}$

Known quantities: Unambiguous and ambiguous data in $d_0$ and $d_1$

1. Normalize $d_1$ to $d_0$ distribution: estimate how much $d_1$ unambiguous data was "lost" in $d_0$

2. Calculate $OV$ to $VO$ "loss ratio"
Estimating Historical $p_{VO}$

Known quantities: Unambiguous and ambiguous data in $d_0$ and $d_1$

Normalize $d_1$ to $d_0$ distribution: estimate how much $d_1$ unambiguous data was "lost" in $d_0$

Calculate OV to VO "loss ratio"

Assume $d_1$-to-$d_0$ "loss ratio" is same as underlying-to-$d_1$ "loss" ratio

Use "loss ratio" to estimate how much underlying unambiguous data was "lost" in $d_1$
Estimating Historical $p_{VO}$

Known quantities:
Unambiguous and ambiguous data in $d_0$ and $d_1$

Normalize $d_1$ to $d_0$ distribution: estimate how much $d_1$ unambiguous data was "lost" in $d_0$

Calculate $p_{VO}$ from estimated underlying unambiguous data distribution

Use "loss ratio" to estimate how much underlying unambiguous data was "lost" in $d_1$

Assume $d_1$-to-$d_0$ "loss ratio" is same as underlying-to-$d_1$ "loss ratio"
Why Parameters?
Why posit parameters instead of just associating stress contours with words?

Arguments from stress change over time (Dresher & Lahiri, 2003):
(1) If word-by-word association, expect piece-meal change over time at the individual word level. Instead, historical linguists posit changes to underlying systems to best explain the observed data: many words changing at once.
(2) If stress contours are not composed of pieces (parameters), expect start and end states of change to be near each other. However, examples exist where start & end states are not closely linked from perspective of observable stress contours.
Relativizing Probabilities

Relativize-against-all:
- probability conditioned against entire input set
- relativizing set is constant across methods

<table>
<thead>
<tr>
<th>Cues or Parsing</th>
<th>Qi</th>
<th>Qs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous Data Points</td>
<td>2140</td>
<td>11213</td>
</tr>
<tr>
<td>Relativizing Set</td>
<td>540505</td>
<td>540505</td>
</tr>
<tr>
<td>Relativized Probability</td>
<td>0.00396</td>
<td>0.0207</td>
</tr>
</tbody>
</table>

Relativizing Probabilities

Relativize-against-potential:
- probability conditioned against set of data points that meet preconditions of being an unambiguous data point
- relativizing set is not constant across methods

Cues: have correct syllable structure

<table>
<thead>
<tr>
<th>Cues: have correct syllable structure</th>
<th>Qi</th>
<th>Qs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous Data Points</td>
<td>2140</td>
<td>11213</td>
</tr>
<tr>
<td>Relativizing Set</td>
<td>2755</td>
<td>85288</td>
</tr>
<tr>
<td>Relativized Probability</td>
<td>0.777</td>
<td>0.132</td>
</tr>
</tbody>
</table>

Relativizing Probabilities

Relativize-against-potential:
- probability conditioned against set of data points that meet preconditions of being an unambiguous data point
- relativizing set is not constant across methods

Parsing: able to be parsed

<table>
<thead>
<tr>
<th>Parsing: able to be parsed</th>
<th>Qi</th>
<th>Qs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous Data Points</td>
<td>2140</td>
<td>11213</td>
</tr>
<tr>
<td>Relativizing Set</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>Relativized Probability</td>
<td>2140/p</td>
<td>11213/p</td>
</tr>
</tbody>
</table>

Cues vs. Parsing Again

Is there any (additional) reason to prefer one method of identifying unambiguous data over the other?

<table>
<thead>
<tr>
<th>Cues vs. Parsing Again</th>
<th>Cues</th>
<th>Parsing</th>
</tr>
</thead>
<tbody>
<tr>
<td>W W L H H ... L L L L</td>
<td>QI, Em-None, F1, F1-Hdi-L, S, T-S, S-Syl</td>
<td>P1, Dir-Left, F1-Hdi-L, S, T-S, S-Syl</td>
</tr>
<tr>
<td>H L L ... S S S S S S</td>
<td>QI, Em-None, F1, F1-Hdi-L, S, T-S, S-Syl</td>
<td>P1, Dir-Left, F1-Hdi-L, S, T-S, S-Syl</td>
</tr>
<tr>
<td>QI, Em-None, F1, F1-Hdi-L, S, T-S, S-Syl</td>
<td>QI, Em-None, F1, F1-Hdi-L, S, T-S, S-Syl</td>
<td>P1, Dir-Left, F1-Hdi-L, S, T-S, S-Syl</td>
</tr>
</tbody>
</table>
Cues vs. Parsing: Success Across Relativization Methods

<table>
<thead>
<tr>
<th></th>
<th>Cues</th>
<th>Parsing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative-Against-All</td>
<td>Successful</td>
<td>Successful</td>
</tr>
<tr>
<td>Relative-Against-Potential</td>
<td>Unsuccessful</td>
<td>Successful</td>
</tr>
</tbody>
</table>

...so parsing seems more robust across relativization methods.

Another Consideration: Constraint Derivability

Good: Order constraints exist that will allow the learner to converge on the adult system, provided the learner knows these constraints.

Better: These order constraints can be derived from properties of the learning system, rather than being stipulated.

Deriving Constraints from Properties of the Learning System

Data saliency: presence of stress is more easily noticed than absence of stress, and indicates a likely parametric cause

Data quantity: more unambiguous data available

Default values (cues only): if a value is set by default, order constraints involving it disappear

Note: data quantity and default values would be applicable to any system. Data saliency is more system-dependent.

Deriving Constraints: Cues

(a) QS-VC-Heavy before Em-Right

(b) Em-Right before Bounded-Syl

(c) Bounded-2 before Bounded-Syl
(a) QS-VC-Heavy before Em-Right

**Em-Right**: absence of stress is less salient (data saliency)

(b) Em-Right before Bounded-Syl

**Bounded-Syl** as default (default values)

(c) Bounded-2 before Bounded-Syl

**Bounded-2** has more unambiguous data once Em-Right is set; Em-Right has much more than Bounded-2 or Bounded-Syl (data quantity)
Deriving Constraints: Cues

(a) QS-VC-Heavy before Em-Right

Em-Right: absence of stress is less salient (data saliency)

(b) Em-Right before Bounded-Syl

Bounded-Syl as default (default values)

Em-Right: more unambiguous data than Bounded-Syl (data quantity)

(c) Bounded-2 before Bounded-Syl

Bounded-Syl as default (default values)

Bounded-2 has more unambiguous data once Em-Right is set; Em-Right has much more than Bounded-2 or Bounded-Syl (data quantity)

Deriving Constraints: Parsing

Group 1:
QS, Ft Head Left, Bounded

Group 2:
Ft Dir Right, QS-VS-Heavy

Group 3:
Em-Some, Em-Right, Bounded-2, Bounded-Syl

Other groupings cannot be derived from data quantity, however…

Em-Some, Em-Right: absence of stress is less salient (data saliency)
Cues vs. Parsing for Unambiguous Data

The order constraints a learner would need to succeed can be derived in a principled manner for cues but must be mostly stipulated for parsing.

Open Questions

1. Can we combine the strengths of cues and parsing?

Combining Cues and Parsing

Cues and parsing have a complementary array of strengths and weaknesses

Problem with cues: require prior knowledge
Problem with parsing: requires parse of entire data point

Viable combination of cues & parsing:

  parsing of data point subpart = derivation of cues?

Combining Cues and Parsing

Em-Right: Rightmost syllable is Heavy …H[H]
and unstressed

If a syllable is Heavy, it should be stressed.
If an edge syllable is Heavy and unstressed, an immediate solution (given the available parameteric system) is that the syllable is extrametrical.
Combining Cues and Parsing

Viable combination of cues & parsing:

*Parsing of data point subpart = derivation of cues?*

Would partial parsing

(a) derive cues that lead to successful acquisition?
(b) be successful across relativization methods?
(c) have derivable order constraints?
(d) be a more realistic representation of the learning mechanism?

Open Questions

(1) Can we combine the strengths of cues and parsing?
(2) Are order constraints not derivable from the learning system consistent cross-linguistically?

Non-derivable Constraints

 Parsing Constraints

<table>
<thead>
<tr>
<th>Group 1:</th>
<th>QS, Ft Head Left, Bounded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 2:</td>
<td>Ft Dir Right, QS-VS-Heavy</td>
</tr>
<tr>
<td>Group 3:</td>
<td>Em-Some, Em-Right, Bounded-2, Bounded-Syl</td>
</tr>
</tbody>
</table>

Open Questions

(1) Can we combine the strengths of cues and parsing?
(2) Are order constraints not derivable from the learning system consistent cross-linguistically?
(3) Are predicted parameter-setting orders observed in real-time learning?
Experimental Predictions for English

Cues
(a) QS-VC-Heavy before Em-Right
(b) Em-Right before Bounded-Syl
(c) Bounded-2 before Bounded-Syl

Parsing
Group 1: QS, Ft Head Left, Bounded
Group 2: Ft Dir Right, QS-VS-Heavy
Group 3: Em-Some, Em-Right, Bounded-2, Bounded-Syl

Open Questions
(1) Can we combine the strengths of cues and parsing?
(2) Are order constraints not derivable from the learning system consistent cross-linguistically?
(3) Are predicted parameter-setting orders observed in real-time learning?
(4) Is the unambiguous data filter successful for other languages besides English? Other complex linguistic domains?

Additional Information Source: Exploiting the Hypothesis Space Layout

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 \( a + b \).

So, \( A \) has a higher probability of having produced \( d \). Thus, \( A \) is favored when encountering ambiguous data.
How does a learner know to use the no-type-II-ambiguous 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 consequences (N vs. N'). Then, only updating domains from semantic information, not semantic & syntactic. Result: lower probability of correct interpretation.

Semantic-referents-only filter

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
The Simple Variational Model: Subset/Superset

Suppose two grammars, $G_1$ and $G_2$.

For whichever grammar is chosen, if $G_1$ can parse the sentence (reward):
$$\text{prob}(G_1) = \text{old\_prob}(G_1) + \gamma(1-\text{old\_prob}(G_1))$$

if $G_1$ can’t parse the sentence (punish):
$$\text{prob}(G_1) = (1-\gamma)\text{old\_prob}(G_1)$$

where $\gamma$ is the learning rate

$$\text{prob}(G_2) = 1 - \text{prob}(G_1)$$ since there are only 2 grammars in this world

---

The Simple Variational Model: Subset/Superset

Subset-Superset: English vs. French wh-questions

- English: wh-fronting
- French: wh-fronting & in-situ

G1= English
G2 = French

What if only subset data points are encountered (learning English)?

English: wh-fronting
French: wh-fronting & in-situ

Convergence to either grammar

Final Value for prob(English) after 100,000 sentences. 50 start, all English (subset) data on input

Subset (prob = 1) doesn’t win.
Also, learner doesn’t stay at 50-50, especially as gamma increases. (Tendency to converge on one grammar)