
Lisa Pearl, University of California, Irvine
Jan 16, 2008
Psychobabble Talk Series, UCLA

Human Language Learning

Theoretical work:
- object of acquisition

Experimental work:
- time course of acquisition

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. Moreover, data are often ambiguous.

Syntactic System
- Observable form: word order
- Difficulty: interactive structural pieces

Metrical Phonology System
- Observable form: stress contour
- Difficulty: interactive structural pieces

One Aid: Constraints on Hypothesis Space

Premise: learner considers finite range of hypotheses
- (parameters: Halle & Vergnaud (1987), Chomsky (1981))
- or constraints: Tesar & Smolensky, (2000))

But this doesn’t solve the learning problem…

“Assuming that there are n binary parameters, there will be 2^n possible core grammars.” - Clark (1994)
How do learners choose among these hypotheses?

Size matters not.

Real learning seems to be gradual and somewhat robust to noise.

Probabilistic Learning: Naïve Parameter Learner


The Naïve Parameter (NPar) Learner

Probabilistic learning strategy explicitly compatible with parameterized grammars: learning is gradual & variable

“grammars that succeed in analyzing [a data point] are rewarded and those that fail are punished”

Probabilistic Learning: Bayesian Inference

Tenenbaum & Griffiths (2001)

Shifting the probability to the hypotheses most compatible with the observed data, using Bayes’ rule. Implementations can be gradual.

Hypotheses = opposing grammars (sets of parameter values) or - if adapted to parametric framework - opposing parameter values for a single parameter.

Complex System Woes

But this may not always work when we have complex systems with multiple parameters and noisy data.

Problem for learning probabilistically over grammars: What if the adult parametric system is actually represented by a minority of the available data? That is, data points consistent with all adult parameter values simultaneously are actually rare.

Extracting features from a data set: cat grammar

2 2 3 4 2

Extracting features from a data set: cat grammar

2 2 1 2 2
Most frequent type = most frequent “grammar” instantiation … but this doesn’t see quite right, compared against the rest.

Parameter values individually compatible with most other cats best-fit “grammar” of cat parameter values.

The point about data noise

In a system with multiple generalizations (e.g. a grammar containing multiple parameter values), data points signaling all the correct generalizations simultaneously may not be all that common.

Instead, the child must integrate information from multiple data points which individually may signal some incorrect generalizations (along with the correct ones).

This points to learning explicitly with parameters rather than over complete grammars.

Complex System Woes Revisited

Even if the child makes use of parameters when learning, there still may be trouble (foreshadowing)

Where might other constraints on the learning system come from?

Learning Framework: 3 Components

(1) Hypothesis space

(2) Data

(3) Update procedure
Investigating Data Intake Filtering
“Selective Learning”

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.

Unambiguous data: Fodor, 1998; Dresher, 1996; Lightfoot, 1998; Pearl & Weinberg, 2007

Ambiguous Data Woes: Feasibility of an Unambiguous Data Filter

“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)

Today’s Plan

Given a realistic complex system to learn and realistic data to learn from, we ask...

(1) Is something beyond probabilistic learning necessary? (Necessity)
(2) Is there a data sparseness problem for an unambiguous data filter? (Feasibility)
(3) Does learning from unambiguous data yield correct behavior? (Sufficiency)

Useful Tool: Modeling

Why? Can easily and ethically manipulate some part of the learning mechanism and observe the effect on learning.

Important: Empirically grounded in realistic data & psychologically plausible learning constraints

Road Map

I. The System
Parameterized Metrical Phonology

II. The Input

III. Learning Without Filters

IV. The Filter

V. Learning With Filters

VI. Good Ideas

Metrical Phonology

What tells you to put the Emphasis on a particular Syllable

Sample metrical phonology structure from parametric system (adapted from Dresher (1999))
Metrical Phonology Parameters

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

Quantity Sensitivity: QI
Quantity-Insensitive (QI): All syllables are treated the same (S)

\[
\begin{align*}
S & \quad S & \quad S \\
VV & \quad V & \quad VC & \text{rime only} \\
CVV & \quad CV & \quad CVC \\
\text{lu} & \quad \text{di} & \quad \text{crous}
\end{align*}
\]

Quantity Sensitivity: QS
Quantity-Sensitive (QS): Syllables are separated into Light and Heavy
- V are always L
- VV are always H

\[
\begin{align*}
H & \quad L & \quad L/H \\
VV & \quad V & \quad VC & \text{rime only} \\
CVV & \quad CV & \quad CVC \\
\text{lu} & \quad \text{di} & \quad \text{crous}
\end{align*}
\]

Quantity Sensitivity: Stress
Rule of Stress: If a syllable is Heavy, it should get stressed - unless some other parameter interacts with it

Metrical Phonology Parameters
Extrametricality, Metrical Feet, and Stress

Rule of Stress: If a syllable is extrametrical, it is not included in a metrical foot. If a syllable is not in a metrical foot, it cannot have stress.

Extrametricality: None

- Extrametricality: None (Em-None): All syllables are in metrical feet

Extrametricality: Some

- Extrametricality: Some (Em-Some): One edge syllable not in foot
  - Extrametricality-Left (Em-Left): Leftmost syllable not in foot - cannot have stress
  - Extrametricality-Right (Em-Right): Rightmost syllable not in foot - cannot have stress

Extrametricality: Left

- Metrical foot: (L L) (H)
  - VC VC VV
  - af ter noon

Extrametricality: Right

- Metrical foot: (H L) (H)
  - VV V VC
  - lu di crou s

Metrical Phonology Parameters

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

Feet Directionality

- Feet Direction Left: start from left edge
- (H L H)
- Feet Direction Right: start from right edge
- (H L H)
Feet Directionality

Feet Direction: What edge of the word metrical foot construction begins at

Feet Direction Left: start from left edge

(H L H)

Feet Direction Right: start from right edge

H L H

Metrical Phonology Parameters

Boundedness: Unbounded Feet

Unbounded: a metrical foot extends until a heavy syllable is encountered

(start from left) L L L H L

(start from left) (L L) H L
Boundedness: Unbounded Feet

Unbounded: a metrical foot extends until a heavy syllable is encountered.

\[
\text{start from left } \rightarrow (L \ L \ L)(H \ L) \\
\text{start from right } \rightarrow (L \ L \ L)(L)
\]
Boundedness: Unbounded Feet

Unbounded: a metrical foot extends until a heavy syllable is encountered

\[ (L \ L \ L)(H \ L) \]

Boundedness: Bounded Feet

Bounded: a metrical foot only extends a certain amount (cannot be longer)

\[ (L \ L \ L \ L) \]

\[ (L \ L \ L \ H)(L) \]

Bounded-2: a metrical foot only extends 2 units

\[ (L \ L \ L) \]

\[ (X \ X \ X \ X \ X) \]

Bounded-3: a metrical foot only extends 3 units

\[ (L \ L \ L \ L) \]

\[ (X \ X \ X \ X \ X) \]

\[ (L \ L \ L) \]

\[ (X \ X) (X \ X) (X) \]

Bounded-2: a metrical foot only extends 2 units

\[ (L \ L \ L \ L \ L) \]

\[ (X \ X \ X \ X \ X) \]

Bounded-3: a metrical foot only extends 3 units

\[ (L \ L \ L \ L \ L) \]

\[ (X \ X \ X \ X \ X) \]
Boundedness: Bounded Feet

**Bounded-Syllabic:** counting unit is **syllable**

**Bounded-Moraic:** counting unit is **mora**

- $H = 2$ moras, $L = 1$ mora

---

Boundedness: Bounded Feet

**Bounded-Syllabic:** counting unit is **syllable**

- Start from left
- **L H L L H**

**Bounded-Moraic:** counting unit is **mora**

- $H = 2$ moras, $L = 1$ mora

---

Boundedness: Bounded Feet

**Bounded-Syllabic:** counting unit is **syllable**

- Start from left
- (L H)(L L)(H)

**Bounded-Moraic:** counting unit is **mora**

- $H = 2$ moras, $L = 1$ mora

---

Boundedness: Bounded Feet

**Bounded-Syllabic:** counting unit is **syllable**

- Start from left
- (L H)(L L)(H)

**Bounded-Moraic:** counting unit is **mora**

- $H = 2$ moras, $L = 1$ mora

---

Boundedness: Bounded Feet

**Bounded-Syllabic:** counting unit is **syllable**

- Start from left
- (H H)(L L)(H)

**Bounded-Moraic:** counting unit is **mora**

- $H = 2$ moras, $L = 1$ mora
Boundedness: Bounded Feet

**Bounded-Syllabic:** counting unit is syllable

- start from left: \((L \ H)(L \ L)(H)\)
- bounded-2: \((H \ H)(L \ L)(H)\)

**Bounded-Moraic:** counting unit is mora

- \(H = 2\) moras, \(L = 1\) mora

- start from left: \(\mu \ \mu \ \mu \ \mu \ \mu \ \mu \ \mu \ \mu \ \mu \ \mu \)
- bounded-2: \(H \ L \ L \ L \ H\)

---

Metrical Phonology Parameters

- **Quantity Sensitivity**
- **Extrametricality**
- **Feet Directionality**
- **Feet Boundedness**

---

Feet Headedness

**Feet Headedness:** which syllable of metrical foot gets stress

- **Feet Head Left:** leftmost syllable in foot gets stress
  \((H)(L \ H)\)

- **Feet Head Right:** rightmost syllable in foot gets stress
  \((H)(L \ H)\)

---

Metrical Feet and Stress

- **Rule of Stress:** Exactly one syllable per metrical foot must have stress.
Feet Headedness

Feet Headedness: which syllable of metrical foot gets stress

Feet Head Left: leftmost syllable in foot gets stress

(\text{H}) (\text{L} \text{H})

Feet Head Right: rightmost syllable in foot gets stress

(\text{H}) (\text{L} \text{H})

Metrical Phonology Parameters

Road Map

I. The System
II. The Input
   English child-directed speech
   III. Learning Without Filters
   IV. The Filter
   V. Learning With Filters
   VI. Good Ideas

English Metrical Phonology

Non-trivial language: English (full of exceptions)
Input: data unambiguous for the incorrect value in the adult system
   - 27\% incompatible with correct grammar on at least one value
   - None are unambiguous for all correct values simultaneously

Adult English system values:

\text{QS}, \text{QSVCH}, \text{Em-Some}, \text{Em-Right}, \text{Ft Dir Right},
\text{Bounded}, \text{B-2}, \text{B-Syllabic}, \text{Ft Hd Left}

Exceptions:

\text{QI}, \text{QSVCL}, \text{Em-None}, \text{Ft Dir Left}, \text{Unbounded},
\text{B-3}, \text{B-Moraic}, \text{Ft Hd Right}
Empirical Grounding in Realistic Data: Estimating English Data Distributions

Caregiver speech to children between the ages of 6 months and 2 years (CHILDES [Brent & Bernstein-Ratner corpora]; MacWhinney, 2000)

Total Words: 540505
Mean Length of Utterance: 3.5

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

Road Map

I. The System
II. The Input
III. Learning Without Filters
   The Naïve Parameter Learner
   Parametric Bayesian Learner
IV. The Filter
V. Learning With Filters
VI. Good Ideas

Probabilistic Learning with Parameters

Naïve Parameter (NPar) Learner
Incremental learning: Learn from a single data point at a time (psychological plausibility)

For each parameter, the learner associates a probability with each of the competing parameter values

\[
\begin{align*}
Q_I &= 0.7 \\
Q_S &= 0.3 \\
E_m-\text{Some} &= 0.4 \\
F_t \text{ Dir Left} &= 0.8 \\
\text{Bounded} &= 0.5 \\
F_t \text{ Dir Right} &= 0.5 \\
\end{align*}
\]

\[
\begin{align*}
Q_I &= 0.7 \\
Q_S &= 0.3 \\
E_m-\text{None} &= 0.6 \\
F_t \text{ Dir Right} &= 0.2 \\
\text{Unbounded} &= 0.4 \\
F_t \text{ Dir Left} &= 0.5 \\
\end{align*}
\]

If the data point can be parsed, then all participating parameter values are rewarded (and opposing values are punished).

\[
\text{\textbf{Example:}} \quad \text{\textbf{parse}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]

\[
\text{\textbf{Probability}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]

\[
\text{\textbf{Probability}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]

\[
\text{\textbf{Probability}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]

\[
\text{\textbf{Probability}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]

\[
\text{\textbf{Probability}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]

\[
\text{\textbf{Probability}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]

\[
\text{\textbf{Probability}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]

\[
\text{\textbf{Probability}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}} \quad \text{\textbf{ð}}
\]
Probabilistic Learning with Parameters

**Naive Parameter (NPar) Learner**

If the data point cannot be parsed, then all participating parameter values are punished (and opposing values are rewarded).

---

**The NPar Learner on English Metrical Phonology**

Learning Period Length: 1,160,000 words (based on estimates of words heard in a 6 month period, using Aghtar et al. (2004)).

Learning rate: (0.01 ≤ γ ≤ 0.05)

---

Probabilistic Learning with Parameters

**Parametric Bayesian Learner**

Incremental learning: Learn from a single data point at a time (psychological plausibility)

Each parameter has two potential values (e.g. QI/QS, QSVCL/QSVCH, Em-Some/Em-None, etc.). View child as trying to decide what probability a binomial distribution should be centered at to maximize likelihood of observed data for each parameter.

Can use beta distribution function to estimate a posteriori probability.

\[
\text{Prob} = \frac{\alpha \cdot (1-\beta)}{\alpha + \beta \cdot (1-\beta) + \gamma}
\]

when α = β, bias is symmetric about p = 0.5 (each value equally likely)

when α, β < 1, bias is towards p = 0.0 and p = 1.0 (bias to pick one value or the other)
For each parameter, the learner associates a probability with each of the competing parameter values. (Initially, all are 0.5.)

Let \( p = \frac{x}{x+n} \), \( x \) = successes, \( n \) = total seen.

If the data point cannot be parsed, then all participating parameter values have their successes \( x \) left alone and the posterior probabilities are updated.

\[
\begin{align*}
\text{Qi} & = 0.5 \\
\text{Em-Some} & = 0.5 \\
\text{Fl Dir Left} & = 0.5 \\
\text{Bounded} & = 0.5 \\
\text{Fl Hd Left} & = 0.5 \\
\text{Unbounded} & = 0.5 \\
\text{Fl Hd Rl} & = 0.5 \\
\text{...}
\end{align*}
\]

For any data point encountered, a parameter value combination (grammar) is generated using the current probabilities. The learner attempts to parse the current data point with this combination.

\[
\begin{align*}
\text{Qi} & = 0.5 \\
\text{Em-Some} & = 0.5 \\
\text{Fl Dir Right} & = 0.5 \\
\text{B-Syl} & = 0.5 \\
\text{Ft Dir Right} & = 0.5 \\
\text{Ft Dir Left} & = 0.5 \\
\text{Unbounded} & = 0.5 \\
\text{Fl Hd Rl} & = 0.5 \\
\text{...}
\end{align*}
\]

The probability of a parameter value combination (grammar) is generated is

\[
\text{Parametric Bayesian Learner} = \frac{\text{prior probability}}{\text{prior probability} + \text{likelihood}}.
\]

\[
\begin{align*}
\text{Qi} & = 0.5 \\
\text{Em-Some} & = 0.5 \\
\text{Fl Dir Right} & = 0.5 \\
\text{B-Syl} & = 0.5 \\
\text{Ft Dir Right} & = 0.5 \\
\text{Ft Dir Left} & = 0.5 \\
\text{Unbounded} & = 0.5 \\
\text{Fl Hd Rl} & = 0.5 \\
\text{...}
\end{align*}
\]

\[
\begin{align*}
\text{Qi} & = 0.5 \\
\text{Em-Some} & = 0.5 \\
\text{Fl Dir Right} & = 0.5 \\
\text{B-Syl} & = 0.5 \\
\text{Ft Dir Right} & = 0.5 \\
\text{Ft Dir Left} & = 0.5 \\
\text{Unbounded} & = 0.5 \\
\text{Fl Hd Rl} & = 0.5 \\
\text{...}
\end{align*}
\]

\[
\begin{align*}
\text{Qi} & = 0.5 \\
\text{Em-Some} & = 0.5 \\
\text{Fl Dir Right} & = 0.5 \\
\text{B-Syl} & = 0.5 \\
\text{Ft Dir Right} & = 0.5 \\
\text{Ft Dir Left} & = 0.5 \\
\text{Unbounded} & = 0.5 \\
\text{Fl Hd Rl} & = 0.5 \\
\text{...}
\end{align*}
\]

\[
\begin{align*}
\text{Qi} & = 0.5 \\
\text{Em-Some} & = 0.5 \\
\text{Fl Dir Right} & = 0.5 \\
\text{B-Syl} & = 0.5 \\
\text{Ft Dir Right} & = 0.5 \\
\text{Ft Dir Left} & = 0.5 \\
\text{Unbounded} & = 0.5 \\
\text{Fl Hd Rl} & = 0.5 \\
\text{...}
\end{align*}
\]

\[
\begin{align*}
\text{Qi} & = 0.5 \\
\text{Em-Some} & = 0.5 \\
\text{Fl Dir Right} & = 0.5 \\
\text{B-Syl} & = 0.5 \\
\text{Ft Dir Right} & = 0.5 \\
\text{Ft Dir Left} & = 0.5 \\
\text{Unbounded} & = 0.5 \\
\text{Fl Hd Rl} & = 0.5 \\
\text{...}
\end{align*}
\]

\[
\begin{align*}
\text{Qi} & = 0.5 \\
\text{Em-Some} & = 0.5 \\
\text{Fl Dir Right} & = 0.5 \\
\text{B-Syl} & = 0.5 \\
\text{Ft Dir Right} & = 0.5 \\
\text{Ft Dir Left} & = 0.5 \\
\text{Unbounded} & = 0.5 \\
\text{Fl Hd Rl} & = 0.5 \\
\text{...}
\end{align*}
\]
A More Conservative Learner: NPar Learner + Batch

Naive Parameter Learner with Batch Learning (NPar + B Learner): More conservative about rewarding and punishing parameters. Meant for more complex systems with interactive parameters.

For any data point encountered, a parameter value combination (grammar) is generated using the current probabilities. The learner attempts to parse the current data point with this combination.
A More Conservative Learner: NPar Learner + Batch

If the data point can be parsed, then all participating parameter values have their batch counter incremented by 1.

\[
\begin{align*}
\text{Counter}_{\text{QS}} &= \text{Counter}_{\text{QS}} + 1 \\
\text{Counter}_{\text{QI/QS}} &= \text{Counter}_{\text{QI/QS}} + 1 \\
\text{Counter}_{\text{VC Left}} &= \text{Counter}_{\text{VC Left}} + 1 \\
\end{align*}
\]

If the data point cannot be parsed, then all participating parameter values have their batch counters decremented by 1.

\[
\begin{align*}
\text{Counter}_{\text{QS}} &= \text{Counter}_{\text{QS}} - 1 \\
\text{Counter}_{\text{QI/QS}} &= \text{Counter}_{\text{QI/QS}} - 1 \\
\text{Counter}_{\text{VC Left}} &= \text{Counter}_{\text{VC Left}} - 1 \\
\end{align*}
\]

A More Conservative Learner: NPar Learner + Batch

If the batch counter for a value reaches the upper limit \( b \), that parameter value is rewarded. If the batch counter reaches the lower limit \( -b \), that parameter value is punished. The counters are then reset to 0.

\[
\begin{align*}
\text{Counter}_{\text{QS}} &= 0 \\
\text{Counter}_{\text{QI/QS}} &= 0 \\
\text{Counter}_{\text{VC Left}} &= 0 \\
\end{align*}
\]

The NPar + B Learner on English Metrical Phonology

Learning Period Length: 1,160,000 words (based on estimates of words heard in a 6 month period, using Akhtar et al. (2004)).

0.01 ≤ learning rate ≤ 0.05 
2 ≤ batch size \( b \) ≤ 10
The Bayesian Learner + Batch on English Metrical Phonology

Learning Period Length: 1,160,000 words (based on estimates of words heard in a 6 month period, using Akhtar et al. (2004)).

Results using distributions in English child-directed speech:

- **Ft**
  - Ft Dir Left
  - Ft Dir Right
- **Hd**
  - Left, QS,
  - Right, QS,
- **Unb**
  - QSVCH, Ft Dir Left
  - Unbounded, Ft Dir Left

If learners ignore monosyllabic words (since such words don’t have a stress contour per se), less than 0.8% converge on English.

Examples of incorrect target languages Bayesian learners converged on:

- Unb, Ft Dir Left, QI, Em-None, Ft Dir Right
- Ft Dir Left, QI, Em-None, Ft Dir Left
- Ft Dir Right, Ft Dir Left, QI, Em-None
- Ft Dir Right, Ft Dir Left, Em-None, Ft Dir Right

The learner may already have knowledge of Ft Dir Left and QI.

Such words don’t have a stress.

The difference between long vowels and short vowels in syllables is some of the rhythmical properties of their language.

Infant research has shown that infants are sensitive to the difference between long vowels and short vowels in syllables (1995): English infants are sensitive to the difference between long vowels and short vowels in syllables.

The learner may already have knowledge of Ft Dir Left and QI.

Perhaps this knowledge of the system will allow a probabilistic learner to converge on the rest of the English values more reliably.
The Bayesian Batch Learner with Prior Knowledge on English Metrical Phonology

Learning Period Length: 1,160,000 words (based on estimates of words heard in a 6 month period, using Akhtar et al. (2004)).

batch size \( b = 5 \)

Results using distributions in English child-directed speech:

- Learners never converge on English.
- If learners ignore monosyllabic words (since such words don’t have a stress contour per se), 1% of learners converge on English.

Examples of incorrect target languages Bayesian + B learners with prior knowledge converged on:

- Ft Hd Left, QS, Bounded, Em-Some, QSVCH, Ft Dir Left, B-3, B-Mor
- Ft Hd Left, QS, Em-Some, QSVCH, Ft Dir Right, Unbounded
- Ft Hd Left, QS, Em-Some, QSVCH, Ft Dir Right, Bounded, Ft Dir Left, B-Mor
- Ft Hd Left, QS, Em-Some, QSVCH, Ft Dir Right, Bounded, Ft Dir Right, B-Mor

Big picture

For the complex linguistic system under consideration, knowledge of the parameters, learning explicitly with parameter values, and a probabilistic learning strategy doesn’t seem to yield the correct behavior. These are insufficient for learning by themselves.

Something else seems necessary for successful acquisition. Can selective learning help?

Filter Feasibility

How feasible is an unambiguous data filter for a complex system?

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

How could a child identify such data?

Road Map

I. The System
II. The Input
III. Learning Without Filters
IV. The Filter
    Selectively Learning From Unambiguous Data
V. Learning With Filters
VI. Good Ideas

Changing Knowledge States: Unambiguous Data is a Moving Target

Current knowledge of system influences perception of unambiguous data. The informativity of a data point changes over time.

Data initially ambiguous may later be perceived as unambiguous. Data initially unambiguous may later be perceived as exceptional.

Point: The order in which parameters are set may determine if they are set correctly (Dresher, 1989).
Identifying Unambiguous Data

Identifying unambiguous data:

**Cues** (Dresher, 1996; Lightfoot, 1999)

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

Important: psychological plausibility

Both cues and parsing operate over a single data point at a time and are thus compatible with incremental learning (that doesn’t require the child to see the whole data set at once)

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
  - ... H
- **Unb**: 3+ unstressed S/L syllables in a row
  - ... S S S...
  - ... L L L L
- **Ft Hd Left**: Leftmost foot has stress on leftmost syllable
  - S S S...
  - H L L ...

Parsing with Metrical Phonology Parameters

Sample Datum: VC VC VV ("afternoon")

<table>
<thead>
<tr>
<th>QS, QSVCL, Em-None, Ft Dir Right, B, B-2, B-Syl, Ft Hd Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) (x)</td>
</tr>
<tr>
<td>L L H</td>
</tr>
<tr>
<td>VC VC VV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>QS, QSVCL, Em-None, Ft Dir Left, B, B-2, B-Syl, Ft Hd Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x) (x)</td>
</tr>
<tr>
<td>L L H</td>
</tr>
<tr>
<td>VC VC VV</td>
</tr>
</tbody>
</table>

Values leading to successful parses of data point *afternoon*:

- (QI, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
- (QI, Em-None, Ft Dir Right, Ft Hd Left, B, B-2, B-Syl)
- (QI, Em-None, Ft Dir Left, Ft Hd Right, B, B-2, B-Syl)
- (QI, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)
- (QI, Em-None, Ft Dir Left, Ft Hd Left, Unb)

Data point is unambiguous for Em-None.
Parsing with Metrical Phonology Parameters

Values leading to successful parses of data point \textit{afternoon}:
\begin{itemize}
  \item \textit{QI}, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl
  \item \textit{QI}, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl
  \item \textit{QI}, QSVC, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl
  \item \textit{QI}, QSVC, Em-None, Ft Dir Left, Ft Hd Left, Unb
\end{itemize}

Data point is unambiguous for Em-None.

Perception of unambiguous data changes over time:
  \textit{If Bounded} already set, data point is unambiguous for \textit{Em-None}, B, B-2, and B-Syl.

Road Map

I. The System
II. The Input
III. Learning Without Filters
IV. The Filter
V. Learning With Filters
  \textit{Simulating What Children Do}
VI. Good Ideas

The Learning Process

Selective Learning: The learner encounters a data point and decides if it's unambiguous for any parameter values.

Updating Hypotheses: If so, the learner shifts probability to those parameter values.

Probabilistic Learning Intuition: The parameter value whose unambiguous data have a higher probability of being encountered by the learner will win when the learner is setting that parameter value.

Initial State of English Child-Directed Speech: Probability of Encountering Unambiguous Data

<table>
<thead>
<tr>
<th>Quantity Sensitivity</th>
<th>Extrametricality</th>
</tr>
</thead>
<tbody>
<tr>
<td>QI: 0.00398</td>
<td>QS: 0.0205</td>
</tr>
<tr>
<td>None: 0.0294</td>
<td>Some: 0.000259</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feet Directionality</th>
<th>Boundedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left: 0.000</td>
<td>Right: 0.0009925</td>
</tr>
<tr>
<td>Unbounded: 0.000000370</td>
<td>Bounded: 0.000435</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feet Headedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left: 0.00148</td>
</tr>
<tr>
<td>Right: 0.000</td>
</tr>
</tbody>
</table>

Initial State of English Child-Directed Speech: Probability of Encountering Unambiguous Data

- QI
- Em-None
- QS
- QSVC
- FT Dir Right
- FT Dir Left
- FT Head Left
- FT Head Right
- Bound
- Unb
- QSVC
- QS
Getting to English

The child must set all the parameter values in order to converge on a language system.

Current knowledge of the system influences the perception of unambiguous data. So, the order in which parameters are set influences the probability of encountering unambiguous data for unset parameters.

To get to English, the child must converge on QS, QSVCH, Em-Some, Em-Right, FtDirRt, Bound, Bounded, Bounded-Syl, FtHdLeft.

Will any parameter-setting orders lead the learner to English?

Getting to English: Exhaustive Search of All 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.
Parameter-Setting Orders: Knowledge Necessary for Acquisition Success

"Viable parameter-setting order" means...

If the learner manages to set the parameters in this order, the learner will converge on English.

But wouldn’t it be better if the viable orders could be captured more compactly, instead of being explicitly listed in the learner’s mind?

Order #23 looks good!

Unambiguous Data: 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.

 Parsing
Group 1:
  QS, Fl Hid 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.

Feasibility & Sufficiency of the Unambiguous Data Filter

Either method of identifying unambiguous data (cues or parsing) is successful. Given the non-trivial parametric 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)

Feasibility & Sufficiency of the Unambiguous Data Filter

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

(1) Unambiguous data exist and can be identified in sufficient relative quantities to learn a complex parametric system.

(2) The data intake filtering strategy is robust across a realistic (highly ambiguous, exception-filled) data set. It's feasible to identify such data, and the strategy yields sufficient learning behavior.

Predictions: Links to the Experimental Side

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

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

Are predicted parameter setting orders observed in real-time learning? E.g., whether cues or parsing is used. Quantity Sensitivity (QS, QSVC) is predicted to be set before Extrametricality (Em-Some, Em-Right).

And in fact, there is evidence that quantity sensitivity may be known quite early (Turk, Jusczyk, & Gerken, 1995)

Future Directions in Modeling

(1) Is the unambiguous data filter successful for other languages besides English? Other instantiations of metrical phonology? Other complex linguistic domains like syntax?

24
Take Home Message

(1) Modeling results for a realistic system and realistic data sets suggest the necessity of something beyond a simple probabilistic learning strategy, even if the hypothesis space of learners is already constrained and learners utilize its parametric nature.

(2) They also demonstrate the viability of the unambiguous data filter as an implementation of the selective learning strategy.

(3) Computational modeling is a very useful tool:
   
   (a) empirically test learning strategies that would be difficult to investigate with standard techniques
   
   (b) generate experimentally testable predictions about learning

Thank You

Amy Weinberg
Bill Idsardi

The audiences at
University of Southern California Linguistics Department
BUCLD 32
UC Irvine Department of Cognitive Science
CUNY Psycholinguistics Supper Club
UDelaware Linguistics Department
Yale Linguistics Department
UMaryland Cognitive Neuroscience of Language Lab

Jeff Lidz
Charles Yang
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 that change altogether.

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.

Cues vs. Parsing: Comparison

Cues:
- Easy identification of unambiguous data
  
- Can find information in sub-part of data point
  
- Can tolerate exceptions

Parsing:
- Fractal analysis
  
- Holistic interpretation
Cues vs. Parsing: Comparison

**Cues:**
- Are heuristic
- Require additional knowledge
- May rely on default values

**Parsing:**
- Resource-intensive identification of unambiguous data
- Needs complete parse of data point to get any information:
  - Cannot find information in sub-part of data point
  - Cannot tolerate exceptions

**Comparison Table:**

<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>

**Psychological plausibility:**
- does not require entire data set of cues to learn from
Calculating Unambiguous Data Probability: 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>

Calculating Unambiguous Data Probability: 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 (e.g. 2 syllables if cue is 2 syllable word with both syllables stressed)

<table>
<thead>
<tr>
<th>Relativized Probability</th>
<th>QI</th>
<th>QS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cues</td>
<td>0.132</td>
<td>0.777</td>
</tr>
<tr>
<td>Parsing: able to be parsed</td>
<td>$2140/p$</td>
<td>$11213/p$</td>
</tr>
</tbody>
</table>

Order Constraints

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, or they’re already known through other means.
Knowing Through Other Means

Infant research has shown that infants are sensitive to some of the rhythmic properties of their language.


Turk, Jusczyk, & Gerken (1995): English infants are sensitive to the difference between long vowels and short vowels in syllables.

The learner may already have knowledge of FT, Ft, HD, and QS, so these are set early.

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 may 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
Deriving Constraints: Cues

(a) QS-VC-Heavy before Em-Right
- Em-Right: absence of stress is less salient (data saliency); prior knowledge
- Bounded-Syl as default (default values)
(b) Em-Right before Bounded-Syl
- Em-Right: more unambiguous data than Bounded-Syl (data quantity)
- Bounded-Syl as default (default values)
(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, Hd, Left, Bounded

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

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

Group 1:
- QS, Ft, Hd, Left, Bounded

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

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

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

Other groupings cannot be derived from data quantity, however...

- Em-Some, Em-Right: absence of stress is less salient (data saliency)
Non-derivable Constraints: Predictions Across Languages?

Parsing Constraints

Group 1: QS, Ft Hld Left, Bounded

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

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

Do we find these same groupings if we look at other languages?

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 and unstressed...H

If a syllable is Heavy, it should be stressed.
If an edge syllable is Heavy and unstressed, an immediate solution (given the available parametric 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) retain the strengths that cues & parsing have separately?
(c) be a more psychologically plausible implementation of the unambiguous data filter?