Learning English Metrical Phonology: Beyond Simple Probability

Lisa Pearl
University of California, Irvine
Sept 4, 2008
GALANA 3

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
Complex Linguistic Systems

What is the generative system that creates the observed (structured) data of language (ex: syntax, metrical phonology)?

 Observable data: word order  Subject  Verb  Object

Complex Linguistic Systems

What is the generative system that creates the observed (structured) data of language (ex: syntax, metrical phonology)?

 Observable data: word order  Subject  Verb  Object

Kannada

Subject  Verb  Object

German

Subject  Verb  Object

English

Subject  Verb  Object
Complex linguistic systems
General problems
Parametric systems
Parametric metrical phonology
Case study: English metrical phonology

Learnability of complex linguistic systems
General learnability framework
Previous learning successes: biased learners
Unbiased probabilistic learning
Where the problem lies: tricky data

Where next? Implications & Extensions
General Problems with Learning Complex Linguistic Systems

What children encounter: the output of the generative linguistic system

What children must learn: the components of the system that combine to generate this observable output

Why this is tricky:
There is often a non-transparent relationship between the observable form of the data and the underlying system that produced it. Hard to know what parameters of variation to consider.

The Hypothesis Space

Hypothesis for a language consists of a combination of generalizations about that language (grammar). But this leads to a theoretically infinite hypothesis space.

Observation:
Languages only differ in constrained ways from each other. Not all generalizations are possible.
Hypothesis for a language consists of a combination of generalizations about that language (grammar). But this leads to a theoretically infinite hypothesis space.

Observation: Languages only differ in constrained ways from each other. Not all generalizations are possible.

Idea: Bias on hypothesis space - children’s hypotheses are constrained so they only consider generalizations that are possible in the world’s languages.

Linguistic parameters give the benefit of a finite hypothesis space. Still, the hypothesis space can be quite large.

For example, assuming there are $n$ binary parameters, there are $2^n$ core grammars to choose from.

Exponentially growing hypothesis space


Linguistic parameters = finite (if large) hypothesis space of possible grammars

Learning Parametric Linguistic Systems

Also, data are often ambiguous between competing hypotheses, since multiple grammars can account for the same data point.

Parametric Metrical Phonology

Metrical phonology: What tells you to put the emphasis on a particular syllable

Process speakers use:

Basic input unit: syllables

Larger units formed: metrical feet

Stress assigned within metrical feet

Observable Data: stress contour of word
Metrical phonology: What tells you to put the emphasis on a particular syllable?

Process speakers use:
Basic input unit: syllables
Larger units formed: metrical feet
The way these are formed varies from language to language.
Stress assigned within metrical feet
The way this is done also varies from language to language.
Observable Data: stress contour of word

Parametric Metrical Phonology system here: 5 main parameters, 4 sub-parameters (adapted from Dresher 1999 and Hayes 1995) - 156 viable grammars

Most parameters involve metrical foot formation

Sub-parameters: options that become available if main parameter value is a certain one

All combine to generate stress contour output

Quantity Sensitivity
Extrametricality
Feet Directionality
Roundedness
Feet Headedness

System parameters of variation to be determined by learner from available data

All combine to generate stress contour output

 Observable Data: stress contour of word

Parametric Metrical Phonology

Metrical phonology: What tells you to put the emphasis on a particular syllable?

Process speakers use:
Basic input unit: syllables
Larger units formed: metrical feet
The way these are formed varies from language to language.
Stress assigned within metrical feet
The way this is done also varies from language to language.
Observable Data: stress contour of word

Parametric Metrical Phonology system here: 5 main parameters, 4 sub-parameters (adapted from Dresher 1999 and Hayes 1995) - 156 viable grammars

Most parameters involve metrical foot formation

Sub-parameters: options that become available if main parameter value is a certain one

All combine to generate stress contour output

Quantity Sensitivity
Extrametricality
Feet Directionality
Roundedness
Feet Headedness

System parameters of variation to be determined by learner from available data

All combine to generate stress contour output
Generating a Stress Contour

Are syllables differentiated?
Yes - by rhyme.
VC & VV syllables are Heavy, V syllables are Light.

Quantity Sensitivity

Are any syllables extrametrical?
Yes.
Rightmost syllable is not included in metrical foot.

Extrametricality

Which direction are feet constructed from?
From the right.

Feet Directionality

Process speaker uses to generate stress contour

em pha sis
Generating a Stress Contour

Process speaker uses to generate stress contour

Are feet unrestricted in size?

No.

2 syllables per foot.

(\text{H} \ \text{L}) \ \text{H} \\
\text{VC} \ \text{CV} \ \text{CVC} \\
\text{em} \ \text{pha} \ \text{sis}

Generating a Stress Contour

Process speaker uses to generate stress contour

Which syllable of the foot is stressed?

Leftmost.

(\text{H} \ \text{L}) \ \text{H} \\
\text{VC} \ \text{CV} \ \text{CVC} \\
\text{em} \ \text{pha} \ \text{sis}

Generating a Stress Contour

Process speaker uses to generate stress contour

Learner’s task: Figure out which parameter values were used to generate this contour.

(\text{H} \ \text{L}) \ \text{H} \\
\text{VC} \ \text{CV} \ \text{CVC} \\
\text{EM} \ \text{pha} \ \text{sis}

Case study: English metrical phonology

Estimate of child input: caretaker speech to children between the ages of 6 months and 2 years (CHILDES Brent & Beranek corpora; MacWhinney 2000)

Total Words: 540605 \ 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 (Canavati et al., 1997)
Case study: English metrical phonology

Non-trivial language: English (full of exceptions)

Noisy data: 27.03% tokens (38.86% types) incompatible with English grammar on at least one parameter value

Adult English system values:
- QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hd Left

Exceptions:
- QI: QSVCL, Em-None, Ft Dir Left, Unbounded, Bounded-3, Bounded-Moraic, Ft Hd Right

Case study: English metrical phonology

Non-trivial language: English (full of exceptions)

Noisy data: 27.03% tokens (38.86% types) incompatible with English grammar on at least one parameter value

English child-directed speech has a significant quantity of misleading data and is comprised mostly of ambiguous data - it's hard to learn, and therefore interesting!
Key point for cognitive modeling: psychological plausibility

Any probabilistic update procedure that children are likely to use must, at the very least, be incremental/online.

Why? Humans (especially human children) don’t have infinite memory.

Unlike: human children can hold a whole corpus’s worth of data in their minds for analysis later on.

Learning algorithms that operate over an entire data set do not have this property.
(ex: Foraker et al. 2007, Goldwater et al. 2007)

Desired: Learn from a single data point, or perhaps a small number of data points at most.
Previous modeling work (Pearl 2008)

(1) Hypothesis space

(2) Data

Modify the data the learner uses

(3) Update procedure

Data Intake Filtering

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

“Selective” Intuition: Use the really good data only.

One instantiation of “really good” = highly informative.

One instantiation of “highly informative” = data viewed by the learner as unambiguous (Fodor, 1998; Dresher, 1999; Lightfoot, 1999; Pearl & Weinberg, 2007)
Biased learner, using only unambiguous data

Pearl (2008): Success is guaranteed as long as the parameters are learned in a particular order.

However...this requires the learner to identify unambiguous data and know/derive the appropriate parameter-setting order, which may not be trivial.

So...is this selective learning bias really necessary? How well do unbiased learners do?

Two psychologically plausible probabilistic update procedures

Naïve Parameter Learner (NParLearner)

Probabilistic generation & testing of grammars. (incremental)
Hypothesis update: Linear reward-penalty

Yang (2002)

Probabilistic learning for English

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

Initially all are equiprobable
Probabilistic learning for English

For each data point encountered, the learner probabilistically generates a grammar.

After NOON

QI = 0.5
QSVCL = 0.5
Em-Some = 0.5
Em-Left = 0.5
Ft Dir Left = 0.5
Bounded = 0.5
Bounded-Syl = 0.5
Ft Hd Left = 0.5

QS = 0.5
QSVCH = 0.5
Em-None = 0.5
Em-Right = 0.5
Ft Dir RI = 0.5
Bounded-L = 0.5
Bounded-Mor = 0.5
Ft Hd RI = 0.5

If the generated stress contour matches the observed stress contour, all participating parameter values are rewarded.

Match (success): reward all

QI/QS?

QSVCL or QSVCH?

Em-None/Em-Some?

... if QS, QSVCL or QSVCH?

... if QS, QSVCL or QSVCH?

If the generated stress contour does not match the observed stress contour, all participating parameter values are punished.

Mismatch (failure): punish all

The learner then uses this grammar to generate a stress contour for the observed data point.

Matching (success): reward all

Mismatch (failure): punish all

Probabilistic generation and testing of grammars (Yang 2002)
Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities

NParLearner (Yang 2002): Linear Reward-Penalty

Learning rate $\gamma$:
- small = small changes
- large = large changes

<table>
<thead>
<tr>
<th>Parameter value $v_1$ vs. $v_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1 = p_1 + \gamma (1-p_1)$</td>
</tr>
<tr>
<td>$p_1 = 1-p_1$</td>
</tr>
</tbody>
</table>

After learning: expect probabilities of parameter values to converge near endpoints (above/below some threshold).

- QI = 0.3
- QS = 0.7
- QSVCL = 0.6
- Em-Some = 0.1
- Em-None = 0.9

Once set, a parameter value is always used during generation, since its probability is 1.0.

Probabilistic learning for English

Goal: Converge on English values after learning period is over

Learning Period Length: 1,666,667 words (based on estimates of words heard in a 6 month period, using Akhtar et al. (2004)).

Examples of incorrect target grammars

NParLearner:
Em-None, Ft-Hd Left, Unb, Ft Dir Left, QI
QS, Em-None, QSVCH, Ft Dir Rl, Ft-Hd Left, B-Mor, Bounded, Bounded-2
BayesLearner:
QS, Em-Some, Em-Right, QSVCH, Ft Dir Rl, Ft-Hd Left, Em-None, Ft Dir Left, B-Syl

The learning framework: where can we modify?

(1) Hypothesis space
(2) Data
(3) Update procedure

Linear Reward-Penalty, Bayesian...?
**Probabilistic learning for English: Modifications**

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities

**Batch learning** (for very small batch sizes): smooth out some of the irregularities in the data, better deal with complex systems (Yang 2002)

Implementation (Yang 2002):

- Matching contour = increase parameter value’s batch counter by 1
- Mismatching contour = decrease parameter value’s batch counter by 1

Invoke update procedure (Linear Reward-Penalty or Bayesian Updating) when batch limit $b$ is reached.

---

**Goal:** Converge on English values after learning period is over

<table>
<thead>
<tr>
<th>QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hd Left</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>NParLearner, $0.01 \times \gamma \times 0.05$</td>
</tr>
<tr>
<td>BayesLearner</td>
</tr>
</tbody>
</table>

---

**Learning Period Length:** 1,666,667 words (based on estimates of words heard in a 6 month period, using Akhtar et al. (2004)).

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

**Model** | **Success rate (1000 runs)**
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NParLearner, $0.01 \times \gamma \times 0.05$</td>
<td>1.2%</td>
</tr>
<tr>
<td>BayesLearner</td>
<td>0.0%</td>
</tr>
<tr>
<td>NParLearner + Batch, $0.01 \times \gamma \times 0.05, 2 \leq b \leq 10$</td>
<td>0.8%</td>
</tr>
<tr>
<td>BayesLearner + Batch, $2 \leq b \leq 10$</td>
<td>1.0%</td>
</tr>
</tbody>
</table>
**Probabilistic learning for English: Modifications**

Probabilistic generation and testing of grammars (Yang 2002)

**Learner hypothesis bias:** metrical phonology relies in part on knowledge of rhythmic properties of the language

Human infants may already have knowledge of Ft,Hd, and QS.


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

**Where else can we modify?**

1. **Hypothesis space**
   - Prior knowledge, biases: QS, Ft,Hd known...

2. **Data**
   - Linear Reward-Penalty, Bayesian, Batch...

3. **Update procedure**
   - BayesLearner + Batch, 2 \leq b \leq 10
   - NParLearner + Batch, 0.01 \leq \gamma \leq 0.05, 2 \leq b \leq 10

---

**Probabilistic learning for English**

Goal: Converge on English values after learning period is over

<p>| QS, QSVC, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hд Left |</p>
<table>
<thead>
<tr>
<th>Model</th>
<th>Success rate (1000 runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NParLearner, 0.01 x y = 0.05</td>
<td>1.2%</td>
</tr>
<tr>
<td>BayesLearner</td>
<td>0.0%</td>
</tr>
<tr>
<td>NParLearner + Batch, 0.01 x y = 0.05, 2 \leq b \leq 10</td>
<td>0.8%</td>
</tr>
<tr>
<td>BayesLearner + Batch, 2 \leq b \leq 10</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

**Goal:** Converge on English values after learning period is over

<p>| QS, QSVC, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hд Left |</p>
<table>
<thead>
<tr>
<th>Model</th>
<th>Success rate (1000 runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NParLearner, 0.01 x y = 0.05</td>
<td>1.2%</td>
</tr>
<tr>
<td>BayesLearner</td>
<td>0.0%</td>
</tr>
<tr>
<td>NParLearner + Batch, 0.01 x y = 0.05, 2 \leq b \leq 10</td>
<td>0.8%</td>
</tr>
<tr>
<td>BayesLearner + Batch, 2 \leq b \leq 10</td>
<td>1.0%</td>
</tr>
<tr>
<td>NParLearner + Batch + Bias, 0.01 x y = 0.05, 2 \leq b \leq 10</td>
<td>5.0%</td>
</tr>
<tr>
<td>BayesLearner + Batch + Bias, 2 \leq b \leq 10</td>
<td>1.0%</td>
</tr>
</tbody>
</table>
Goal: Converge on English values after learning period is over

**Probabilistic learning for English**

<table>
<thead>
<tr>
<th>Model</th>
<th>Success rate (1000 runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesLearner</td>
<td>1.2%</td>
</tr>
<tr>
<td>BayesLearner + Batch, 0.01 ≥ γ ≥ 0.05, 2 ≥ b ≥ 10</td>
<td>0.8%</td>
</tr>
<tr>
<td>BayesLearner + Bias, 2 ≤ b ≤ 10</td>
<td>1.0%</td>
</tr>
<tr>
<td>BayesLearner + Batch + Bias, 0.01 ≥ γ ≥ 0.05, 2 ≥ b ≥ 10</td>
<td>5.0%</td>
</tr>
<tr>
<td>BayesLearner + Batch + Bias, 2 ≤ b ≤ 10</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

Learning Period Length: 1,666,667 words (based on estimates of words heard in a 6 month period, using Akhtar et al. (2004)).

**What gives?**

Metrical phonology system here: 5 main parameters, 4 sub-parameters (adapted from Dresher 1999 and Hayes 1995)

156 viable grammars

**English is not the optimal grammar**

Adult English system values:
- QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hd Left

Of the 156 available grammars, English is ranked
- 52nd by token compatibility
- 56th by type compatibility

If prior knowledge of the hypothesis space is assumed (Ft Hd Left and QS), there are 60 available grammars.

English is ranked
- 18th by token compatibility
- 18th by type compatibility

**Unbiased probabilistic learning is more likely to find the optimal grammar**

English is compatible with 72.97% of the data by tokens, and 62.14% of the data by types.

The average compatibility of the grammars selected by unbiased probabilistic learning (using batch learning) was 73.50% of the data by tokens and 63.93% of the data by types.
Unbiased probabilistic learning is more likely to find the optimal grammar.

English is compatible with 72.97% of the data by tokens, and 62.14% of the data by types.

The average compatibility of the grammars selected by unbiased probabilistic learning (using batch learning) was 73.56% of the data by tokens and 63.3% of the data by types.

Unbiased probabilistic learning works just fine - it’s the English child-directed speech that’s the problem!

Biased Children

The data actually lead an unbiased probabilistic learner to more optimal grammars than the English grammar.

Yet English children seem to learn the English grammar.

Conclusion: Children must have some additional bias that causes the sub-optimal English grammar to become the optimal grammar for this data set.

One idea: selective learning bias to heed only unambiguous data (Pearl 2008)

Road Map

Complex linguistic systems
- General problems
- Parametric systems
- Parametric metrical phonology
- Case study: English metrical phonology

Learnability of complex linguistic systems
- General learnability framework
- Previous learning successes: biased learners
- Unbiased probabilistic learning
- Where the problem lies: tricky data

Where next? Implications & Extensions

Where we are now

Modeling: aimed at understanding how children learn language, generating child behavior by using psychologically plausible methods

Learning complex systems: difficult. Correct grammar is not the optimal grammar for child’s input data without some kind of additional bias.
Where we are now

Modeling: aimed at understanding how children learn language, generating child behavior by using psychologically plausible methods.

Learning complex systems: difficult. Correct grammar is not the optimal grammar for child’s input data without some kind of additional bias.

Bias on data (Pearl 2008): interpretive bias to use highly informative data.

Bias on hypothesis space: linguistic parameters already known, some values already known.

Where we can go

(1) Interpretive bias:
How successful on other difficult learning cases (noisy data sets, other complex systems)?
Are there other methods of implementing interpretative biases that lead to successful learning (productive data: Yang 2005)?
How necessary is an interpretive bias? Are there other biases that might cause the correct grammar to be the optimal grammar for the English data?

(2) Hypothesis space bias:
Will other hypothesis space instantiations allow the correct grammar to be the optimal grammar (constraints (Tesar & Smolensky 2000))? What learning mechanisms make the correct grammar learnable in these hypothesis spaces?
Is it possible to converge on the correct grammar given a less well-defined hypothesis space a priori (e.g. only knowing that units larger than syllables are required)?

The big idea

Complex linguistic systems may well require something beyond probabilistic methods in order to be learned as well as children learn them given the data children are given.

What this likely is: learner biases in hypothesis space and data intake (how to deploy probabilistic learning).

What we can do with computational modeling:
(a) empirically test learning strategies that would be difficult to investigate with standard techniques
(b) generate experimentally testable predictions about learning (Pearl 2008: learning trajectory)
A Brief Tour of Parametric Metrical Phonology

Are syllables differentiated?

Yes: system is quantity-sensitive (QS)

Only allowed method: differ by rime weight

CVV CV CCVC
lu di crous
A Brief Tour of Parametric Metrical Phonology

Are syllables differentiated? 

No: system is quantity-insensitive (QI)

<table>
<thead>
<tr>
<th>CVV</th>
<th>CV</th>
<th>CCVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>lu</td>
<td>di</td>
<td>crous</td>
</tr>
</tbody>
</table>

Yes: system is quantity-sensitive (QS)

Only allowed method: differ by rime weight

Only allowed number of divisions: 2

Heavy vs. Light: VV always Heavy
V always Light

Option 1: VC Heavy (QS-VC-H)

Option 2: VC Light (QS-VC-L)

Are all syllables included in metrical feet?

Yes: system has no extrametricality (Em-None)

No: system has extrametricality (Em-Some)

Only allowed # of exclusions: 1

Only allowed exclusions:
Leftmost or Rightmost syllable

Narrowing of hypothesis space
A Brief Tour of Parametric Metrical Phonology

What direction are metrical feet constructed?

Two logical options:

- From the left:
  Metrical feet are constructed from the left edge of the word (Ft Dir Left)
  
- From the right:
  Metrical feet are constructed from the right edge of the word (Ft Dir Right)

Are metrical feet unrestricted in size?

Yes: Metrical feet are unrestricted, delimited only by Heavy syllables if there are any (Unbounded).

narrowing of hypothesis space
A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

Yes: Metrical feet are unrestricted, delimited only by Heavy syllables if there are any (Unbounded).

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3.

The counting units are restricted to 2 options: syllables or moras.

narrowing of hypothesis space
A Brief Tour of Parametric Metrical Phonology

**Are metrical feet unrestricted in size?**

Yes: Metrical feet are unrestricted, delimited only by Heavy syllables if there are any (Unbounded).

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3.

The counting units are restricted to 2 options: syllables or moras.

**Count by syllables (Bounded-Syllabic)**

- Ft Dir Left Bounded-2
  - (H L) (L H)
  - (L L) (L H)
  - (S S) (S S)

**Count by moras (Bounded-Moraic)**

- Ft Dir Left Bounded-2
  - (H) (L) (L)
  - (S) (S)

---

**Within a metrical foot, which syllable is stressed?**

**Leftmost:** Stress the leftmost syllable (Ft H'd Left)

**Rightmost:** Stress the rightmost syllable (Ft H'd Right)

**Hypothesis space narrowing:**

- Mora (unit of weight): $H = 2$ moras, $L = 1$ mora

---

**Brief Tour of Parametric Metrical Phonology**

**Are metrical feet unrestricted in size?**

Yes: Metrical feet are unrestricted, delimited only by Heavy syllables if there are any (Unbounded).

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3.

The counting units are restricted to 2 options: syllables or moras.

**Count by syllables (Bounded-Syllabic)**

- Ft Dir Left Bounded-2
  - (H L) (L H)
  - (L L) (L H)
  - (S S) (S S)

**Count by moras (Bounded-Moraic)**

- Ft Dir Left Bounded-2
  - (H) (L) (L)
  - (S) (S)
Choosing among grammars

Human learning seems to be gradual and somewhat robust to noise - need some probabilistic learning component.

Since grammars are parameterized, child can make use of this information to constrain hypothesis space. Learn over parameters, not entire parameter value sets.

A caveat about learning parameters separately

Parameters are system components that combine together to generate output. Choice of one parameter may influence choice of subsequent parameters.
A caveat about learning parameters separately

Parameters are system components that combine together to generate output.
Choice of one parameter may influence choice of subsequent parameters.

Point: The order in which parameters are set may determine if they are set correctly from the data.

Dresher 1999

Practical matters:
Feasibility of unambiguous data

Existence?
Depends on data set (empirically determined).

Identification?

Identifying unambiguous data:

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

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

Both operate over a single data point at a time: compatible with incremental learning.
Each parameter has 2 values.

**Advantage** in data: How much more unambiguous data there is for one value over the other in the data distribution.

Assumption (Yang 2002): The value with the greater advantage will be the one a probabilistic learner will converge on over time.

Allows us to be fairly agnostic about the exact nature of the probabilistic learning, provided it has this behavior.

The order in which parameters are set may determine if they are set correctly from the data.

Success guaranteed as long as parameter-setting order constraints are followed.

### Parsing

**Group 1:**
- QS, Ft-Hl 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.

The rest of the parameters are freely ordered w.r.t. each other.

### Cues

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

(b) Em-Right before Bounded-Syl

(c) Bounded-2 before Bounded-Syl
Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

Identification?

Identifying unambiguous data:

- **Cues** (Dresher, 1999; Lightfoot, 1999)
  - heuristics of matching to observable form of the data.
  - Cues are available for each parameter value, known already by the learner.

- **Parsing** (Fodor, 1998; Sakas & Fodor, 2001)
  - Both operate over a single data point at a time: compatible with incremental learning.
Practical matters:
Feasibility of unambiguous data
Existence? Depends on data set (empirically determined).
Identification?

Identifying unambiguous data:

**Cues** (Dresher 1999; Lightfoot 1999): heuristic pattern-matching to observable form of the data. Cues are available for each parameter value, known already by the learner.

- **QS**: 2 syllable word with 2 stresses
  - \( \text{V}_1 \text{V}_2 \)
- **Em-Right**: Rightmost syllable is Heavy and unstressed
  - \( \text{H} \)
- **Unb**: 3+ unstressed S/L syllables in a row
  - \( \text{S} \text{S} \text{S} \ldots \text{L} \text{L} \text{L} \)
- **Ft Hd Left**: Leftmost foot has stress on leftmost syllable
  - \( \text{S} \text{S} \text{S} \ldots \text{H} \text{L} \text{L} \text{L} \)

**Parsing** (Fodor 1998; Sakas & Fodor 2001): extract necessary parameter values from all successful parses of data point (strongest form of parsing).

**Combinations leading to successful parses of afternoon**
- Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl
- Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl
- Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl
- Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl
- Em-None, Ft Dir Right, Ft Hd Right, Unb
- Em-None, Ft Dir Left, Ft Hd Left, Unb
Probabilistic learning for English: Modifications

Probabilistic generation and testing of parameter values (Yang 2002)

Learner bias: metrical phonology relies in part on knowledge of rhythmical properties of the language

Human infants may already have knowledge of Ft Hd Left and QS.

Build this bias into a model: set probability of QS = Ft Hd Left = 1.0.
These will always be chosen during generation.
QS...QSVC or QSVCH?
Ft Hd Left

Update parameter value probabilities + Batch Learning

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>Qt: 0.00398</td>
<td>QS: 0.0205</td>
</tr>
<tr>
<td></td>
<td>None: 0.0294</td>
</tr>
<tr>
<td></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: 0.000000925</td>
</tr>
<tr>
<td>Unbounded: 0.000000370</td>
<td>Bounded: 0.00435</td>
</tr>
</tbody>
</table>

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

Moving Targets & Unambiguous Data: What Happens After Parameter-Setting

<table>
<thead>
<tr>
<th>Quantity Sensitivity</th>
<th>Extrametricality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qt: 0.00398</td>
<td>QS: 0.0205</td>
</tr>
<tr>
<td></td>
<td>None: 0.0294</td>
</tr>
<tr>
<td></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: 0.000000925</td>
</tr>
<tr>
<td>Unbounded: 0.000000370</td>
<td>Bounded: 0.00435</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feet Headedness</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Left: 0.00148</td>
<td>Right: 0.000</td>
</tr>
</tbody>
</table>
Moving Targets & Unambiguous Data: What Happens After Parameter-Setting

<table>
<thead>
<tr>
<th>QS</th>
<th>Extrametricality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feats</td>
<td>Boundedness</td>
</tr>
<tr>
<td>Left: 0.000</td>
<td>Right: 0.0000037</td>
</tr>
<tr>
<td></td>
<td>Unbounded: 0.00125</td>
</tr>
</tbody>
</table>

Getting to English

The child must set all the parameter values in order to converge on a language system. Current knowledge of the system (parameters set) influences the perception of unambiguous data (subsequent parameters set).

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.

Existence

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.

Identification


text

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.

Existence

Identification

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

2) The selective learning 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.
Where we can go: 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, Fr/Hd/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, QSVCH) 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)