Constrained Probabilistic Learning for Complex Linguistic Systems

Lisa Pearl, University of California, Irvine
April 7, 2008
Linguistics Colloquium, UCSD

Human Language Learning

Theoretical work:
object of acquisition

Experimental work:
time course of acquisition & data

mechanism of acquisition
given the boundary conditions provided by
(a) linguistic representation
(b) the trajectory of learning
The Nature of Linguistic Knowledge

Different aspects: **more** and **less** transparent from data

Categorization/Clustering

Ex: What are the contrastive sounds of a language?

Extraction

Ex: Where are words in fluent speech?

The Nature of Linguistic Knowledge

Different aspects: **more** and **less** transparent from data

Categorization/Clustering

Ex: What are the contrastive sounds of a language?

Extraction

Ex: Where are words in fluent speech?

"huwa afrjyd av la big ba’id wolf"

who’s afraid of the big bad wolf
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering
Ex: What are the contrastive sounds of a language?

Ex:
Confide
Confided

Mapping
What are the word affixes that signal meaning (e.g., past tense in English)?

Ex: Where are words in fluent speech?

hüwz střejd sv do břg ba’d wálf
Who’s afraid of the big bad wolf

Extraction
Where are words in fluent speech?

Mapping
What are the word affixes that signal meaning (e.g., past tense in English)?

Different aspects: more and less transparent from data

Complex systems: What is the generative system that creates the observed (structured) data of language (e.g., syntax, metrical phonology)?

Observable data: word order

Subject
Verb
Object

Kannada
Subject
Verb
Object

German
Subject
Verb
Object

English
Subject
Verb
Object
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Observable data: stress contour emphasis

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

Observable data: stress contour emphasis

(H L) H (S) S (H L) L
EM pha sis
EM pha sis
EM pha sis

Complex linguistic systems
General problems
Parametric systems
Parametric metrical phonology

Learnability of complex linguistic systems
General learnability framework
Case study: English metrical phonology
Available data & associated woes
Unconstrained probabilistic learning
Constrained probabilistic learning

Where next? Implications & Extensions

Road Map
General Problems with Learning Complex Linguistic Systems

What children encounter: the output of the generative linguistic system

**EMphasis**

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

Are syllables differentiated?

Are all syllables included?

Which syllable of a larger unit is stressed?

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.

Moreover, data are often ambiguous, even if parameters of variation are known.

Hypothesis for a language consists of a combination of generalizations about that language (grammar). But this leads to a theoretically infinite hypothesis space.
General Problems with Learning Complex Linguistic Systems

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:
Children’s hypotheses are constrained so they only consider generalizations that are possible in the world’s languages.


Linguistic parameters gives 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

(Clark 1994)
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. Only syllables in metrical feet can be stressed.  
Stress assigned within metrical feet  
The way this is done also varies from language to language.  

Observable Data: stress contour of word *emphasis*

A Brief Tour of Parametric Metrical Phonology

Are syllables differentiated?  
No: system is quantity-insensitive (QI)  

Yes: system is quantity-sensitive (QS)  
Only allowed method: differ by rime weight

Parametric Metrical Phonology

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

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

Most parameters involve metrical foot formation  
All combine to generate stress contour output
A Brief Tour of Parametric Metrical Phonology

**Are syllables differentiated?**

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

No: system is quantity-insensitive (QI)

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 |

**Are all syllables included inmetrical feet?**

Yes: system has no extrametricality (Em-None)

<table>
<thead>
<tr>
<th>VC</th>
<th>VC</th>
<th>VV</th>
</tr>
</thead>
<tbody>
<tr>
<td>af</td>
<td>ter</td>
<td>noon</td>
</tr>
</tbody>
</table>

No: system has extrametricality (Em-Some)

Only allowed # of exclusions: 1

Only allowed exclusions:

Leftmost or Rightmost syllable

Leftmost syllable excluded: Em-Left

Rightmost syllable excluded: Em-Right

- agen da

---
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).

Ft Dir Left

Ft Dir Right
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.
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)

Count by moras (Bounded-Moraic)

Within a metrical foot, which syllable is stressed?

Two options, hypothesis space restriction

Leftmost: Stress the leftmost syllable (Ft Dir Left)

Rightmost: Stress the rightmost syllable (Ft Dir Right)
Generating a Stress Contour

Are syllables differentiated?
Yes.
VC syllables are heavy.

Generating a Stress Contour

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

Generating a Stress Contour

Which direction are feet constructed from?
From the right.

Generating a Stress Contour

Are feet unrestricted?
No.
2 syllables per foot.
Generating a Stress Contour

Process speaker uses to generate stress contour

Which syllable of the foot is stressed?

Leftmost.

(H L) H
VC CV CVC
em pha sis

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

Road Map

Complex linguistic systems
- General problems
- Parametric systems
- Parametric metrical phonology

Learnability of complex linguistic systems
- General learnability framework
- Case study: English metrical phonology
- Available data & associated woes
- Unconstrained probabilistic learning
- Constrained probabilistic learning

Where next? Implications & Extensions

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.

2^n options
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
The learning framework: 3 components

(1) Hypothesis space

(2) Data

(3) Update procedure

Key point for cognitive modeling: psychological plausibility

Any probabilistic update procedure must, at the very least, be incremental/online.

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

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

Learning algorithms that operate over an entire data set do not have this property (see: 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.

Two psychologically plausible probabilistic update procedures

Naive Parameter Learner (NParLearner)

Yang (2002)

Probabilistic generation & testing of parameter value combinations.

Hypothesis update: Linear reward-penalty

(Bush & Mosteller 1951)

Two psychologically plausible probabilistic update procedures

Naive Parameter Learner (NParLearner)

Yang (2002)

Probabilistic generation & testing of parameter value combinations.

Hypothesis update: Linear reward-penalty

(Bush & Mosteller 1951)

Bayesian Learner (BayesLearner)

Probabilistic generation & testing of parameter value combinations.

Hypothesis update: Bayesian updating

(Chew 1971: binomial distribution)
Case study: English metrical phonology

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

Estimate of child input: caretaker speech to children between the ages of 6 months and 2 years (CHILDES [Brent & Bernstein-Ratner corpora]; MacWhitney 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)

Non-trivial language: English (full of exceptions)

Noisy data: 27% incompatible with correct English grammar on at least one parameter value

Hard - therefore interesting!

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

English Data
Probabilistic learning for English

Probabilistic generation and testing of parameter values (Yang 2002)

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

- QS = 0.5
- QSVCL = 0.5
- Em-None = 0.5
- QSVCH = 0.5
- Em-Left = 0.5
- Em-Right = 0.5
- Ft Dir Left = 0.5
- Ft Dir Right = 0.5
- Bounded = 0.5
- Bounded-2 = 0.5
- Bounded-Syl = 0.5
- Ft HD Left = 0.5
- Ft HD Right = 0.5

Initially all are equiprobable

Probabilistic learning for English

Probabilistic generation and testing of parameter values (Yang 2002)

For each data point encountered, the learner probabilistically generates a set of parameter values (grammar).

- QS = 0.5
- QSVCL = 0.5
- Em-None = 0.5
- QSVCH = 0.5
- Em-Left = 0.5
- Em-Right = 0.5
- Ft Dir Left = 0.5
- Ft Dir Right = 0.5
- Bounded = 0.5
- Bounded-2 = 0.5
- Bounded-Syl = 0.5
- Ft HD Left = 0.5
- Ft HD Right = 0.5

If the generated stress contour matches the observed stress contour, the grammar successfully "parses" the data point. All participating parameter values are rewarded.

- QS, QSVCL, Em-None, Ft Dir Right, Bounded, Bounded-2, Bounded-Syl, Ft HD Right
- VC, CVC, CVVC
- AF, ter, NOON

If the generated stress contour does not match the observed stress contour, the grammar does not successfully "parse" the data point. All participating parameter values are punished.

- QS, QSVCL, Em-None, Ft Dir Left, Bounded, Bounded-2, Bounded-Syl, Ft HD Right
- VC, CVC, CVVC
- AF, ter, NOON
Probabilistic learning for English

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

Probabilistic generation and testing of parameter values (Yang 2002)

Update parameter value probabilities

\[ p_{v1} = p_{v1} + \gamma (1 - p_{v1}) \]
\[ p_{v2} = 1 - p_{v1} \]

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

**BayesLearner:** Bayesian update of binomial distribution (Chew 1971)

Parameters \( \alpha, \beta \):

- Initial bias at \( p = 0.5 \)
- \( \alpha, \beta > 1 \), initial bias toward endpoints \( p = 0.0, 1.0 \)

Here: \( \alpha = \beta = 0.5 \)
Once set, a parameter value is always used during generation, since its probability is 1.0.

Probabilistic learning for English
Probabilistic generation and testing of parameter values (Yang 2002)
Update parameter value probabilities

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

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

Goal: Converge on English values after learning period is over

Proabilistic learning for English: Modifications
Probabilistic generation and testing of parameter values (Yang 2002)
Update parameter value probabilities

Batch-learning (for very small batch sizes): smooth out some of the irregularities in the data

Implementation (Yang 2002):
Success = increase parameter value’s batch counter by 1
Failure = decrease parameter value’s batch counter by 1

Invoke update procedure (Linear Reward-Penalty or Bayesian Updating) when batch limit $b$ is reached. Then, reset parameter’s batch counters.
Probabilistic learning for English: Modifications

Probabilistic generation and testing of parameter values (Yang 2002)

Update parameter value probabilities + Batch Learning

NParLearner (Yang 2002): Linear Reward-Penalty

Invoke when the batch counter for \( p_{v1} \) or \( p_{v2} \) equals \( b \).

\[
\begin{align*}
p_{v1} &= p_{v1} + \gamma (1-p_{v1}) \\
p_{v2} &= 1 - p_{v1}
\end{align*}
\]

BayesLearner: Bayesian update of binomial distribution (Chew 1971)

Invoke when the batch counter for \( p_{v1} \) or \( p_{v2} \) equals \( b \).

Note: total data seen + 1

Parameter value v1

\[
p = \frac{a + \gamma + 1}{a + \beta + 2 + \text{total data seen}}
\]

reward: success + 1
punish: success + 0

Probabilistic learning for English

Goal: Converge on English values after learning period is over

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Success rate (1000 runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NParLearner, 0.01 ≤ γ ≤ 0.05</td>
<td>1.2%</td>
</tr>
<tr>
<td>BayesLearner</td>
<td>0.0%</td>
</tr>
<tr>
<td>NParLearner + Batch, 0.01 ≤ γ ≤ 0.05, 2 ≤ b ≤ 10</td>
<td>0.8%</td>
</tr>
<tr>
<td>BayesLearner + Batch, 2 ≤ b ≤ 10</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

Probabilistic learning for English: Modifications

Goal: Converge on English values after learning period is over

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

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

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


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

Update parameter value probabilities + Batch Learning

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)).

<table>
<thead>
<tr>
<th>Model</th>
<th>Success rate (1000 runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NParLearner, 0.01 x γ ≤ 0.05</td>
<td>1.0%</td>
</tr>
<tr>
<td>BayesLearner</td>
<td>0.0%</td>
</tr>
<tr>
<td>NParLearner + Batch, 0.01 x γ ≤ 0.05, 2 x b ≤ 10</td>
<td>0.8%</td>
</tr>
<tr>
<td>BayesLearner + Batch, 2 x b ≤ 10</td>
<td>1.0%</td>
</tr>
<tr>
<td>NParLearner + Batch + Bias, 0.01 x γ ≤ 0.05, 2 x b ≤ 10</td>
<td>5.0%</td>
</tr>
<tr>
<td>BayesLearner + Batch + Bias, 2 x b ≤ 10</td>
<td>1.0%</td>
</tr>
</tbody>
</table>

The best isn’t so great
Where else can we modify?

(1) Hypothesis space

(2) Data

(3) Update procedure

What about the data the learner uses?

Prior knowledge, biases: QS, Ft H/L Left known...

Linear Reward-Penalty, Bayesian, Batch...

input

output
Data Intake Filtering
"Selective Learning"

"Equal Opportunity" Intuition: Use all available data to uncover a full range of systematically, 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)

Where else can we modify?

(1) Hypothesis space
Prior knowledge, biases: GS, FL H Left known...

(2) Data
What about the data the learner uses?

(3) Update procedure
Linear Reward-Penalty, Bayesian, Batch...

Practical matters:
Feasibility of unambiguous data

Existence?

Clark, 1994

AF ter NOON

What is the same here, other than the output?

Identification?

Even if unambiguous data points existed, how could a child identify them?
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.

**Parsing** (Fodor, 1998; Sakas & Fodor, 2001): Both operate over a single data point at a time: compatible with incremental learning.

- After noon: Em-None
- QS: 2 syllable word with 2 stresses
- Em-None: Rightmost syllable is Heavy and unstressed
- Unb: 3+ unstressed S/L syllables in a row
- Ft Hd Left: Leftmost foot has stress on leftmost syllable

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:

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
- QI, QSVCL
- Ft Dir Left, Ft Dir Right
- B

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.

Probabilistic learning from unambiguous data

Each parameter has 2 values.

Input

Advantage

Has advantage

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

(Pearl 2008)
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).

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

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

Dresher 1999
The order in which parameters are set may determine if they are set correctly from the data. (Pearl 2008)

Dresher 1999

The order in which parameters are set may determine if they are set correctly from the data. (Pearl 2008)

Dresher 1999

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

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, Ft, HD 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.

Existence? Identification

Clark 1994

"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."

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 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.
**Road Map**

Complex linguistic systems
- General problems
- Parametric systems
- Parametric metrical phonology

Learnability of complex linguistic systems
- General learnability framework
- Case study: English metrical phonology
- Available data & associated woes
- Unconstrained probabilistic learning
- Constrained probabilistic learning

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. Success comes from integrating biases into probabilistic learning models.

Bias on hypothesis space:
- Linguistic parameters already known, some values already known

Bias on data:
- Interpretive bias to use highly informative data

Input
- Informational bias
- Hypothesis space

**Where we can go**

(1) Interpretive bias:
- How successful on other difficult learning cases (noisy data sets, other complex systems)?
- How reasonable are cues/parsing for identifying unambiguous data? (Ask me!)
- 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 cleverer probabilistic learning methods than can succeed (Fodor & Sakas 2004, Bayesian strategies)?

**Where we can go: Links to the Experimental Side**

<table>
<thead>
<tr>
<th>Cues</th>
<th>Parsing</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) QS-VC-Heavy before Em-Right</td>
<td>Group 1: QS, Ft-HiLeft, Bounded</td>
</tr>
<tr>
<td>(b) Em-Right before Bounded-Syl</td>
<td>Group 2: Ft-DrRight, QS-VS-Heavy</td>
</tr>
<tr>
<td>(c) Bounded-2 before Bounded-Syl</td>
<td>Group 3: Em-Some, Em-Right, Bounded-2, Bounded-Syl</td>
</tr>
</tbody>
</table>

Are predicted parameter setting orders observed in real-time learning?
E.g., whether cues or parsing is used. Quantity Sensitivity (QS, QSVSCH) 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).
Where we can go

(1) Interpretive bias:
- How successful on other difficult learning cases (noisy data sets, other complex systems)?
- How reasonable are cues/parsing for identifying unambiguous data? (Ask me!)
- 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 cleverer probabilistic learning methods than can succeed (Fodor & Sakas 2004, Bayesian strategies)?

(2) Hypothesis space bias:
- Is it possible to infer the correct parameters of variation given less structured information a priori (e.g. larger units than syllables are required)? [Model Selection]
- Are other instantiations of hypothesis space restrictions learnable from realistic data (constraints (Tesar & Smolensky 2000))?+

The big idea

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

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

Thank You

Amy Weinberg                Jeff Lida
Bill Idiashki               Charles Yang
Bill Sakas                  Janet Fodor

The audiences at

UC Irvine Machine Learning Group
University of California, Los Angeles Linguistics Department
University of Southern California Linguistics Department
BUCLD 32
UC Irvine Language Learning Group
UC Irvine Department of Cognitive Sciences
CUNY Psycholinguistics Super Club
UDelaware Linguistics Department
Yale Linguistics Department
UMaryland Cognitive Neuroscience of Language Lab
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

Ft Ft Hd Left

Cues vs. Parsing: Comparison

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

Bounded-Syl unless data indicate Bounded-Moraic

Ft Ft Hd Left
### Cues vs. Parsing: Comparison

**Cues**
- Easy identification of unambiguous data
- Can find information in datum sub-part
- Can tolerate exceptions
- Is not heuristic
- Does not require additional knowledge
- Does not use 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

---

<table>
<thead>
<tr>
<th>Feature</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>
Practical matters:
Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

Identification?
Identifying unambiguous data:
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*:
- 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
- QI, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, UnB

If Bounded is known...

The effect of learning parameters

Getting to English: Exhaustive Search of All Parameter-Setting Orders

Try one parameter-setting order...
(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.
Getting to English: Exhaustive Search of All Parameter-Setting Orders

Is it English?
(d) Compare final set of values to English set of values. If they match, this is a viable parameter-setting order. If they don’t, it isn’t.

Unambiguous Data with Cues: Parameter-Setting Orders

Cues: Sample viable orders (500 total)
(a) QS, QS-VC-Heavy, Bounded, Bounded-2, Feet Hd Left, Feet Dir Right, Em-Some, Em-Right, Bounded-2, Bounded-2, Bounded-2, Bounded-Mor, ...
(d) Feet Hd Left, Em,None, ...

Cues: Sample failed orders
(a) QS, QS-VC-Heavy, Bounded, Bounded-2, Bounded-2, Bounded-Mor, ...
(b) Bounded, Bounded-2, Feet Hd Left, Bounded-Mor, ...
(c) Em-No ...
(d) Feet Hd Left, Em-None, ...
Unambiguous Data with Parsing: Parameter-Setting Orders

Parsing: Sample viable orders (66 total)
(a) Bounded, QS, Feet HD Left, Feet Dir Right, QS-VC-Heavy, 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
(c) QS, Bounded, Feet HD Left, QS-VC-Heavy, Feet Dir Right, Bounded-Syl, Em-Some, Em-Right, Bounded-2

Parsing: Sample failed orders
(a) QS, QS-VC-Heavy, Bounded, Bounded-Syl, Bounded-2, Em-Some, Em-Right, Feet HD Right...
(b) Bounded, Bounded-Syl, Bounded-2, Em-None, ...
(c) Em-None, ...
(d) Feet HD Left, Feet Dir Left, ...

Parameter-Setting Orders: Knowledge Necessary for Acquisition Success

“Viable parameter-setting order” means...

If the probabilistic learner manages to set the parameters in this order, the learner is guaranteed 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!

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. Jusczyk, Cutler, & Redanz (1993) found that English 9-month olds prefer strong-weak stress bisyllables (trochaic) to weak-strong ones (iambic).

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

The learner may already have knowledge of \( \text{Ft} \) and \( \text{QS} \), 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) \( \text{QS-VC-Heavy} \) before \( \text{Em-Right} \)

(b) \( \text{Em-Right} \) before \( \text{Bounded-Syl} \)

(c) \( \text{Bounded-2} \) before \( \text{Bounded-Syl} \)
Deriving Constraints: Cues

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

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

(b) Em-Right before Bounded-Syl

Bounded-Syl as default (default values)

(c) Bounded-2 before Bounded-Syl

Bounded-Syl as default (default values)

Deriving Constraints: Cues

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

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

(b) Em-Right before Bounded-Syl

Bounded-Syl as default (default values)

(c) Bounded-2 before Bounded-Syl

Bounded-Syl as default (default values)

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

Bounded-2 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

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

Parsing Constraints

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

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

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?