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: stress contour, Emphasis
Complex Linguistic 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
EM pha sis
(S S ) S
EM pha sis
( S S )
EM pha sis

General Problems with Learning Complex Linguistic Systems

What children encounter: the output of the generative linguistic system

EMphasis

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.

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 = finite (if large) hypothesis space of possible grammars
Learning Parametric Linguistic Systems

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.

\[ (\text{Clark 1994}) \]

Exponentially growing hypothesis space

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
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

Emphasis system parameters of variation - to be determined by learner from available data

Observable Data: stress contour of word

Emphasis
Parametric Metrical Phonology

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

- Quantity Sensitivity
- Extrametricality
- Boundedness
- Feet Directionality

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

All combine to generate stress contour output

---

A Brief Tour of Parametric Metrical Phonology

Are syllables differentiated?

No: system is quantity-insensitive (QI)

lu di

CVV CV CCVC

Most parameters involve metrical foot formation

All combine to generate stress contour output
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

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
A Brief Tour of Parametric Metrical Phonology

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

Leftmost or Rightmost syllable

Leftmost syllable

Rightmost syllable

Leftmost syllable excluded: Em-Left

Rightmost syllable excluded: Em-Right

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).

No: Metrical feet are unrestricted, delimited only by Heavy syllables if there are any (Unbounded).

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).

No: Metrical feet are restricted (Bounded). The size is restricted to 2 options: 2 or 3. narrowing of hypothesis space

Fl Dir Left 2 units per foot (Bounded-2) 3 units per foot (Bounded-3)

\[
\begin{array}{ccc}
\text{x} & \text{x} & \text{x} \\
\text{x} & \text{x} & \text{x} \\
\text{x} & \text{x} & \text{x} \\
\text{x} & \text{x} & \text{x} \\
\text{x} & \text{x} & \text{x} \\
\end{array}
\]
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.

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.

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.

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.

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.

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.

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?
A Brief Tour of Parametric Metrical Phonology

Within a metrical foot, which syllable is stressed?

Two options: hypothesis space restriction

Leftmost:
Stress the leftmost syllable (Ft H L H)

\( (H) (L) (L) (H) \)

Rightmost:
Stress the rightmost syllable (Ft H H L)

\( (H) (L) (L) (H) \)

Are syllables differentiated?
Yes - by rime.

VC & VV syllables are Heavy, V syllables are Light.

Quantity Sensitivity

Generating a Stress Contour

Process speaker uses to generate stress contour

Extrametricality

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.

Are feet unrestricted in size?
No.
2 syllables per foot.

Which syllable of the foot is stressed?
Leftmost.

Learner’s task: Figure out which parameter values were used to generate this contour.
Case study: English metrical phonology

Estimate of child input: caretaker speech to children between the ages of 6 months and 2 years (CHILDES (Brent & Bernstein 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)

...have many exceptions

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

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 (Vallabha et al. 2007).

Why? Humans (especially human children) don’t have infinite memory.
Unlikely: 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
(3) Update procedure

Modify the data the learner uses
Data Intake Filtering
“Selective Learning”, “Interpretive Bias”

“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)

Previous modeling work (Pearl 2008)

(1) Hypothesis space

(2) Data

(3) Update procedure

Modify the data the learner uses

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
(Bush & Mosteller 1951)

Yang (2002)

MAP Bayesian Learner (BayesLearner)
Probabilistic generation & testing of grammars. (incremental)
Hypothesis update: Bayesian updating
(Chew 1971: binomial distribution)

Yang (2002)

Probabilistic learning for English
Probabilistic generation and testing of grammars (Yang 2002)
For each parameter, the learner associates a probability with each of the competing parameter values.

\[
\begin{align*}
QI &= 0.5 \\
QS &= 0.5 \\
QSVCL &= 0.5 \\
QSVCH &= 0.5 \\
QI &= 0.5 \\
QS &= 0.5 \\
QSVCL &= 0.5 \\
QSVCH &= 0.5 \\
\end{align*}
\]

Initially all are equiprobable

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

Initially all are equiprobable
Probabilistic learning for English

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

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

\[ p_v = p_{v1} \cdot (1 - p_{v2}) \]

\[ p_v = 1 - p_v \]

Probabilistic generation and testing of grammars (Yang 2002)

\[ \text{Learning rate } \gamma \]

Small = small changes

Large = large changes

\[ \text{NParLearner (Yang 2002): Linear Reward-Penalty} \]

Update parameter value probabilities
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

$p_{v1} = p_{v1} \cdot \gamma(1-p_{v1})$
$p_{v2} = (1-\gamma)p_{v1}$

reward $v1$
punish $v1$

BayesLearner: Bayesian update of binomial distribution (Chew 1971)

Parameters $\alpha, \beta$:
$\alpha = \beta$: initial bias at $p = 0.5$
$\alpha, \beta < 1$: initial bias toward endpoints ($p = 0.0, 1.0$)

here: $\alpha = \beta = 0.5$

Probabilistic generation and testing of grammars

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)).

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

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

QS, QSVCL, Em-None, Ft Dir Right

...
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)).

<table>
<thead>
<tr>
<th>Model</th>
<th>Success rate (1000 runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NParLearner, γ = .001, .0025, .01, .025</td>
<td>0.0%</td>
</tr>
<tr>
<td>BayesLearner</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Examples of incorrect target grammars:

NParLearner:
- Em-None
- Ft Hd Left
- Unb
- Fl Dr Left
- Fl

BayesLearner:
- QI
- QSVCH
- Ft Dir Left
- Bounded
- Bounded-2
- Bounded-Syllabic
- Ft
- Hd Left
- Em-None
- Ft Dir Left
- Ft Dir Right
- QI
- QSVCH
- Ft
- Hd Left
- Em-None
- Ft Dir Left
- Bounded
- Bounded-2

The learning framework: where can we modify?

(1) Hypothesis space

(2) Data

(3) Update procedure

Probabilistic learning for English: Modifications

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities

Count-learning: 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 count limit c is reached.
Probabilistic learning for English: Modifications

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities + Count Learning

NParLearner (Yang 2002): Linear Reward-Penalty

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

Parameter values \( v1 \) vs. \( v2 \)

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

\[
p_{v1} = 1 \cdot p_{v1} \quad p_{v2} = 1 \cdot p_{v2}
\]

BayesLearner: Bayesian update of binomial distribution (Chew 1971)

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

Note: total data seen + 1

Parameter value \( v1 \)

\[
p = \frac{a + 1}{a + b + 1 + \text{total data seen}}
\]

reward: success + 1, punish: success + 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)).

Model | Success rate (1000 runs)
--- | ---
NParLearner \( \gamma \) = .001, .0025, .01, .025 | 0.0%
BayesLearner | 0.0%

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:
Q5, 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 and 56th 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 73.56% 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!

Since we believe children converge on the English grammar, maybe children aren't unbiased learners. So what kind of bias might work?
Probabilistic learning for English: Bias

Probabilistic generation and testing of grammars (Yang 2002)

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

- English infants may already have knowledge of Ft-Hd Left and QS.
- Turk, Jusczyk, & Gerken (1995): English infants are sensitive to the difference between long vowels and short vowels in syllables.

Probabilistic generation and testing of grammars (Yang 2002)

Where else can we modify?

(1) Hypothesis space
Prior knowledge, biases: QS, Ft-Hd Left known...

(2) Data
Hypothesis space

(3) Update procedure
Data

Why not?
Because English is still 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 60 available grammars, English is ranked
18th by token compatibility
18th by type compatibility
Other Biases for Children

One idea: selective learning bias to heed only unambiguous data (Pearl 2007, 2008)

A general class of probabilistic learning models is guaranteed to succeed as long as parameters are acquired in particular orders. (Many of these special orders can be derived from properties of the data.)

Why does learning from unambiguous data work?
If parameters are set in these special orders, the unambiguous data distribution favors the correct parameter value for English and so English parameter values are in fact optimal.

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 can go

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

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

(1) Selective bias:
   - How successful on other difficult learning cases (noisy data sets, other complex systems)?
   - Are there other methods of implementing selective biases that lead to successful learning (productive data: Yang 2005)?
   - How necessary is a selective bias? Are there other biases that might cause the correct grammar to be the optimal grammar for the English data?
Where we can go

(1) Selective bias:
   - How successful on other difficult learning cases (noisy data sets, other complex systems)?
   - How necessary is a selective 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)