Parametric Linguistic Systems: The Limits of Probabilistic Learning for Realistic Data

Lisa Pearl
University of California, Irvine: lpearl@uci.edu
March 5, 2009
Learning Meets Acquisition, DGfS 2009

Knowledge of language

Knowledge of multiple complex generative systems: phonology, morphology, syntax, …

Speakers use these systems to produce the observable data.

Children must discover the system that native speakers use to generate the observable data

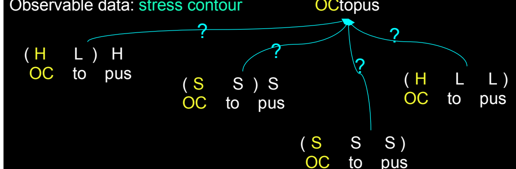
Knowledge of language

Knowledge of multiple complex generative systems: phonology, morphology, syntax, …

Speakers use these systems to produce the observable data.

Children must discover the system that native speakers use to generate the observable data

Observable data: stress contour OCtopus



Learning generative systems

The tricky part: Even if children know a generative system produces the observable data (useful knowledge), how do they know what variables are important to consider?

Observable data: stress contour OCtopus

Do individual segments matter? Do syllables matter? Are all segments/syllables involved? Does syllable weight matter? Does rhyming matter? …

Modeling acquisition

Important distinction: learnability vs. "acquirability" (Johnson 2004)

Acquirability (a more constrained form of learnability) = the ability of children to acquire the knowledge they do from the data they have, given the limitations they have

Practical matters: constraints on…
…what data children encounter
…when & how long they have to learn
…how they integrate information

Modeling acquisition: today

The data children encounter: estimated from child-directed speech (CHILDES database)

When & how long children have to learn: estimated from experimental studies of children's knowledge at certain ages

How children integrate information: assumes children have memory limitations and process data incrementally

Main idea: If the model reasonably reflects process of acquisition in children, manipulations of the model inform us about what those same manipulations would do to the process of acquisition in children.

Goal: Understand this

Road Map

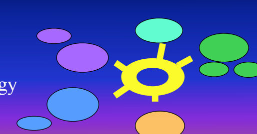
I. Background
  - Parametric systems
  - Parametric metrical phonology

II. Learning English metrical phonology
  - Analysis of data
  - Unbiased models & failure
  - Biased models & success

III. Implications for acquisition

Road Map

I. Background
  - Parametric systems
  - Parametric metrical phonology



II. Learning English metrical phonology
  - Analysis of data
  - Unbiased models & failure
  - Biased models & success

III. Implications for acquisition

Parametric systems & 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.

Parametric systems & the hypothesis space

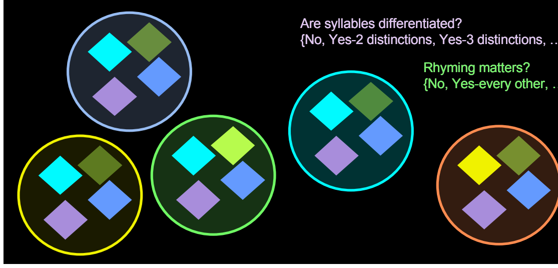
Hypothesis for a language consists of a combination of generalizations about that language (grammar). But this leads to a theoretically infinite hypothesis space.

Which syllable of a larger unit is stressed?
{Leftmost, Rightmost, Second from Left,…}

Are all syllables included?
{Yes, No-not leftmost, No-not rightmost, …}

Are syllables differentiated?
{No, Yes-2 distinctions, Yes-3 distinctions, …}

Rhyming matters?
{No, Yes-every other, …}



Parametric systems & 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.

Which syllable of a larger unit is stressed?
{Leftmost, Rightmost, ~~Second from Left,…}~~

Are all syllables included?
{Yes, No-not leftmost, No-not rightmost, ~~…~~}

Are syllables differentiated?
{No, Yes-2 distinctions, Yes-3 distinctions, ~~…~~}

~~Rhyming matters?~~
~~{No, Yes-every other, …}~~

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

Parametric systems & the hypothesis space

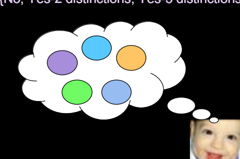
Hypothesis for a language consists of a combination of generalizations about that language (grammar). But this leads to a theoretically infinite hypothesis space.

Which syllable of a larger unit is stressed?
{Leftmost, Rightmost}

Are all syllables included?
{Yes, No-not leftmost, No-not rightmost}

Are syllables differentiated?
{No, Yes-2 distinctions, Yes-3 distinctions}

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



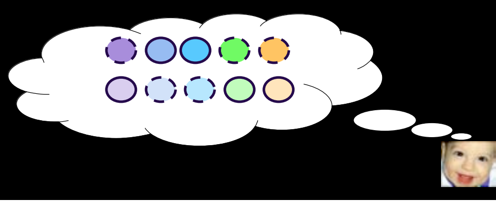
Idea: Constraint on hypothesis space - children's hypotheses are constrained so they only consider generalizations that are possible in the world's languages.

Chomsky (1981), Halle & Vergnaud (1987), Tesar & Smolensky (2000)

Parametric systems & the hypothesis space

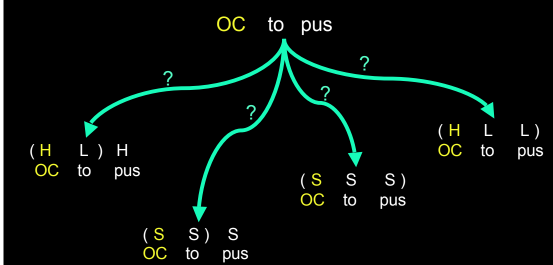
Today: Binary linguistic parameters chosen as implementation of constraints on hypothesis space.

What the learner must do: Set the appropriate value for the parameters of the system.



Learning parametric linguistic systems

Data are often ambiguous between competing hypotheses, since multiple grammars can account for the same data point. Knowing the parametric system doesn't solve the acquisition problem.



Learning parametric linguistic systems: today

Tractable case study of a parametric system of metrical phonology (adapted from Dresher (1999), Halle & Vergnaud (1987), and Hayes (1995))

Compared to prior computational models of parametric systems:

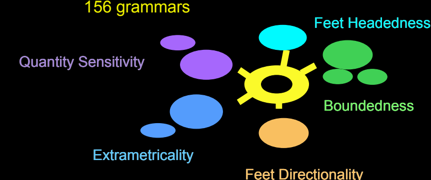
♦ involves more parameters than previous work (Gibson & Wexler 1994, Niyogi & Berwick 1996, Pearl & Weinberg 2007)

♦ input for the model is derived from child-directed speech distributions, while input for previous models often has not been (Dresher & Kaye 1990, Dresher 1999, Sakas & Nishimoto 2002, Sakas 2003, Fodor & Sakas 2004)

Parametric metrical phonology

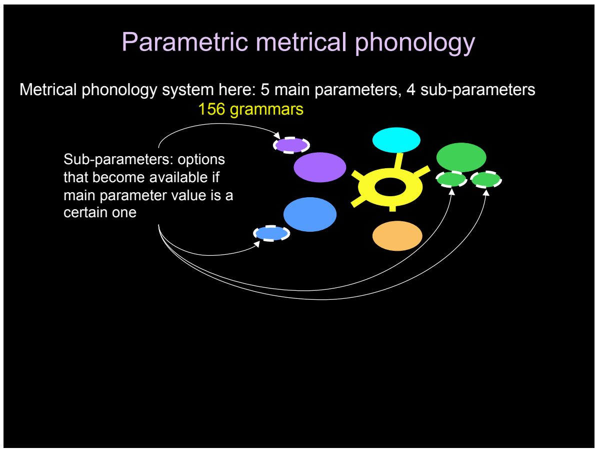
Metrical phonology system here: 5 main parameters, 4 sub-parameters

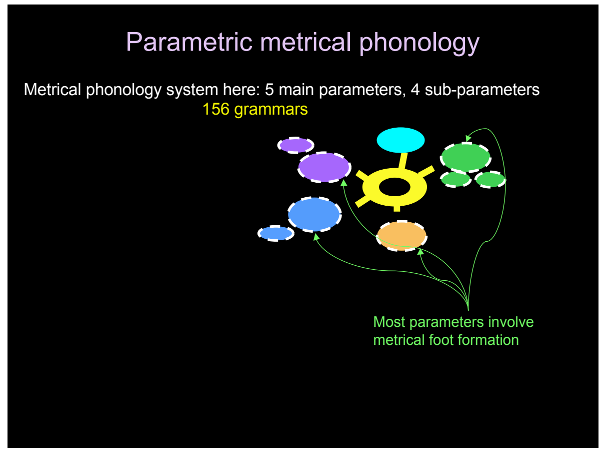
156 grammars

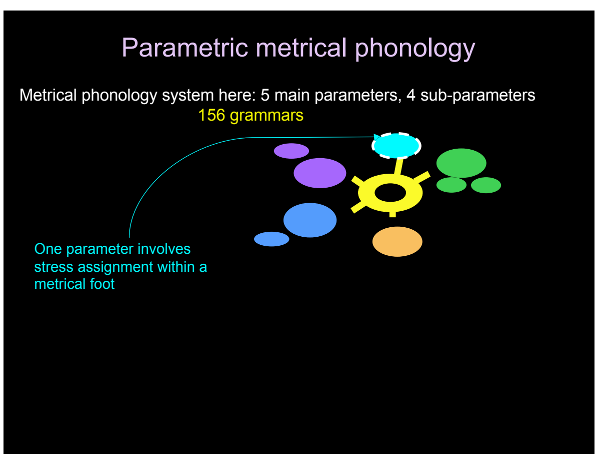


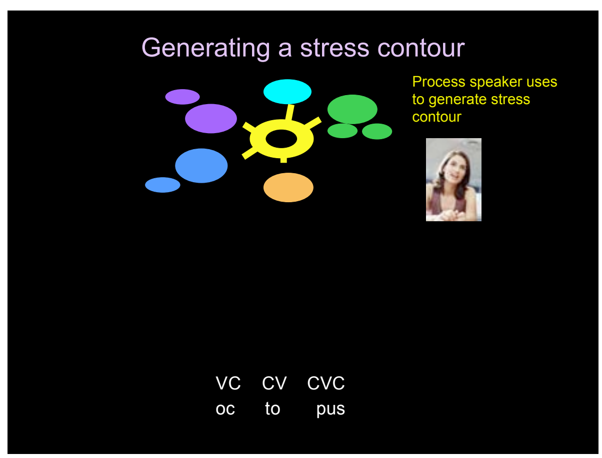
All combine to generate stress contour output

Note: Does not include interactions with the morphology system, due to learner's likely initial knowledge state when first acquiring the metrical phonology system (learner is under a year old and knows little morphology)


Parametric metrical phonology
Metrical phonology system here: 5 main parameters, 4 sub-parameters
156 grammars
Sub-parameters: options that become available if main parameter value is a certain one

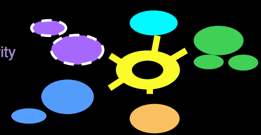

Parametric metrical phonology
Metrical phonology system here: 5 main parameters, 4 sub-parameters
156 grammars
Most parameters involve metrical foot formation


Parametric metrical phonology
Metrical phonology system here: 5 main parameters, 4 sub-parameters
156 grammars
One parameter involves stress assignment within a metrical foot


Generating a stress contour
Process speaker uses to generate stress contour
VC CV CVC
oc to pus

Generating a stress contour

Quantity Sensitivity



Process speaker uses to generate stress contour

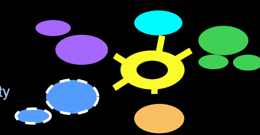
Are syllables differentiated?

Yes - by rime.

2 types:
VC & VV syllables are <u>H</u>eavy, V syllables are <u>L</u>ight.

H	L	H
VC	CV	CVC
oc	to	pus

Generating a stress contour



Process speaker uses to generate stress contour

Extrametricality

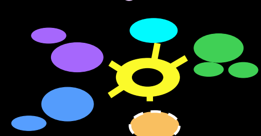
Are any syllables extrametrical?

Yes.

Rightmost syllable is not included in metrical foot.

(	...	)
H	L	H
VC	CV	CVC
oc	to	pus

Generating a stress contour



Process speaker uses to generate stress contour

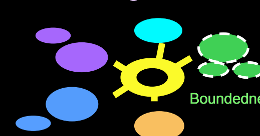
Feet Directionality

Which direction are feet constructed from?

From the right.

H	L)	H
VC	CV	CVC
oc	to	pus

Generating a stress contour



Process speaker uses to generate stress contour

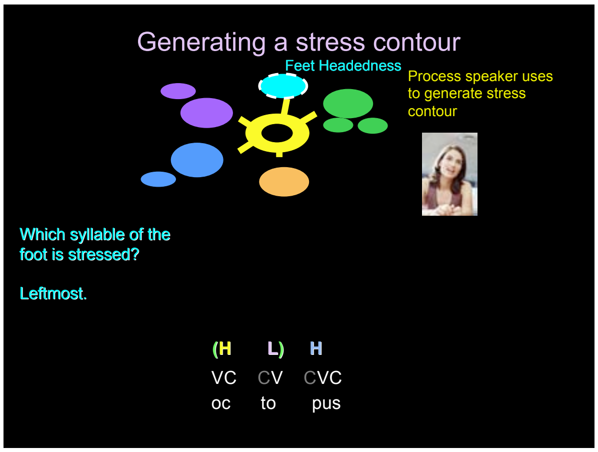
Boundedness

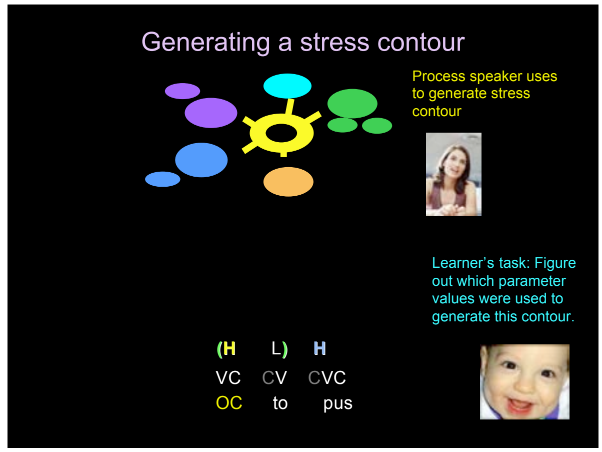
Are feet unrestricted in size?

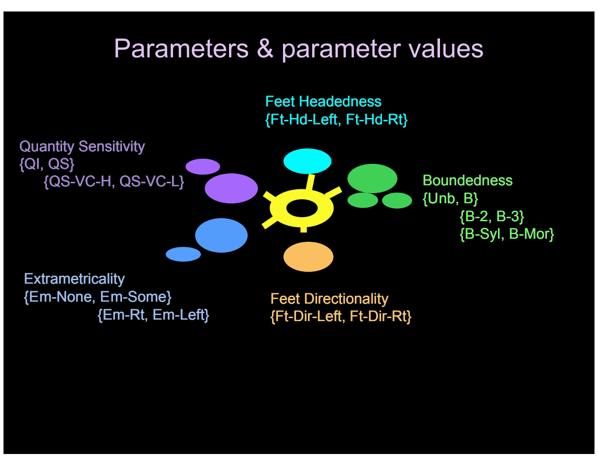
No.

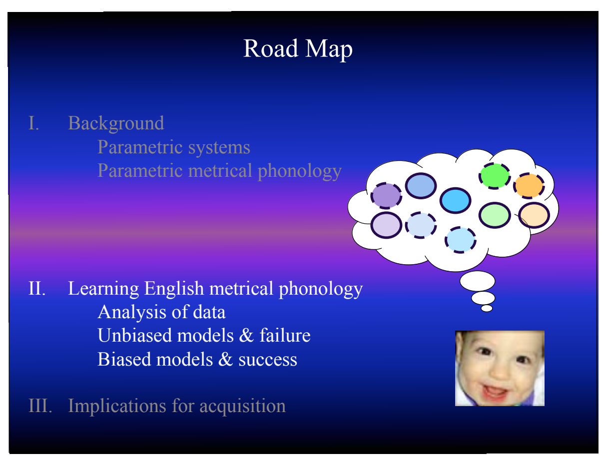
2 syllables per foot.

(H	L)	H
VC	CV	CVC
oc	to	pus


Generating a stress contour
Feet Headedness
Process speaker uses to generate stress contour
Which syllable of the foot is stressed?
Leftmost.
(H L) H
VC CV CVC
oc to pus


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.
(H L) H
VC CV CVC
OC to pus


Parameters & parameter values
Feet Headedness
{Ft-Hd-Left, Ft-Hd-Rt}
Quantity Sensitivity
{QI, QS}
{QS-VC-H, QS-VC-L}
Boundedness
{Unb, B}
{B-2, B-3}
{B-Syl, B-Mor}
Extrametricality
{Em-None, Em-Some}
{Em-Rt, Em-Left}
Feet Directionality
{Ft-Dir-Left, Ft-Dir-Rt}


Road Map
I. Background
Parametric systems
Parametric metrical phonology
II. Learning English metrical phonology
Analysis of data
Unbiased models & failure
Biased models & success
III. Implications for acquisition

Looking at English

Why English?
Modeling with realistic data is easier:
(1) English child-directed speech available (CHILDES)

Acquisition is non-trivial:
(1) English data are very ambiguous
(2) English data contain many exceptions (27% tokens)
exception = data point incompatible with English grammar on at least one parameter value
(partially due to interaction with morphology system)

English parameter values
= {QS, QS-VC-H, Em-Some, Em-Rt, Ft-Dir-Rt, B,B-2, B-Syl, Ft-Hd-Left }
(drawing from Dresher (1999) & Hayes (1995))

Model input: English child-directed speech data

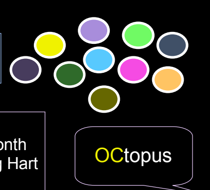
Estimate of child input: caretaker speech to children between the ages of 6 months and 2 years (CHILDES [Brent & Bernstein corpora]: MacWhinney 2000)

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)

Modeling framework

Model's hypothesis space:
Set of 156 grammars in parametric system



Model's data intake based on the number of words likely to be heard on average in a 6 month period: 1,666,667. (Akhtar et al. (2004), citing Hart & Risley (1995))

Model's update procedure:
Incremental update, since any procedure that children are likely to use should be incremental/online (Vallabha et al. 2007). Why? Humans (especially human children) don't have infinite memory, so they are more likely to integrate information into their generative system as it comes in.

Unbiased models

Probabilistic generation and testing of grammars (Yang 2002)

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

QI = 0.5	QS = 0.5
QS-VC-L = 0.5	QS-VC-H = 0.5
Em-Some = 0.5	Em-None = 0.5
Em-Left = 0.5	Em-Rt = 0.5
Ft-Dir-Left = 0.5	Ft-Dir-Rt = 0.5
B = 0.5	Unb = 0.5
B-2 = 0.5	B-3 = 0.5
B-Syl = 0.5	B-Mor = 0.5
Ft-Hd-Left = 0.5	Ft-Hd-Rt = 0.5

Initially all are equiprobable

Unbiased models

Probabilistic generation and testing of grammars (Yang 2002)

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

AFterNOON

QI = 0.5	QS = 0.5
QS-VC-L = 0.5	QS-VC-H = 0.5
Em-Some = 0.5	Em-None = 0.5
Em-Left = 0.5	Em-Rt = 0.5
Ft-Dir-Left = 0.5	Ft-Dir-Rt = 0.5
B = 0.5	Unb = 0.5
B-2 = 0.5	B-3 = 0.5
B-Syl = 0.5	B-Mor = 0.5
Ft-Hd-Left = 0.5	Ft-Hd-Rt = 0.5

QI/QS?…if QS, QS-VC-L or QS-VC-H?
Em-None/Em-Some?…
…

→ QS, QS-VC-L, Em-None, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Rt

Unbiased models

Probabilistic generation and testing of grammars (Yang 2002)

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

AFterNOON

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

QS, QS-VC-L, Em-None, Ft-Dir-Rt, B, B-2, B-Syl, Ft-Hd-Rt

→ (L) (L H)
VC CVC CVVC
AF ter NOON

Reward all parameter values
(don't attempt credit assignment)

Unbiased models

Probabilistic generation and testing of grammars (Yang 2002)

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

AFterNOON

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

QS, QS-VC-L, Em-None, Ft-Dir-Left, B, B-2, B-Syl, Ft-Hd-Rt

→ (L L) (H)
VC CVC CVVC
af TER NOON

Punish all parameter values
(don't attempt blame assignment)

Unbiased models

After seeing many data: expect probabilities of parameter values to increase or decrease until past some threshold, and then that parameter value is set (probability = 1.0).

Threshold chosen: Above 0.8 or below 0.2, based on estimates of when children generalize (Gómez & Lakusta (2004), Hudson Kam & Newport (2005))

QI = 0.3	QS = 0.7
QS-VC-L = 0.6	QS-VC-H = 0.4
Em-Some = 0.2	Em-None = 0.8

…

Em-None = 1.0 (Em-Some = 0.0)

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

Unbiased models: Update types

Naïve Parameter Learner (Yang 2002) [NParLearner]: Linear reward-penalty (Bush & Mosteller 1951)

Learning rate γ:
small = small changes
large = large changes

Parameter values v1 vs. v2

reward v1	punish v1
$p_{v1} = p_{v1} + \gamma(1 - p_{v1})$	$p_{v1} = (1 - \gamma)p_{v1}$
$p_{v2} = 1 - p_{v1}$	$p_{v2} = 1 - p_{v1}$

Bayesian Learner [BayesLearner]: Bayesian update of binomial distribution (Chew 1971)

Parameters α, β:

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

here: $\alpha = \beta = 0.5$

Parameter value v1

$$p_v = \frac{\alpha + 1 + successes}{\alpha + \beta + 2 + total\ data\ seen}$$

reward: success + 1 punish: success + 0

Unbiased models: Update types

Counting variants (Counting NParLearner & Counting BayesLearner)

These models keep count of how many successes or failures have occurred in a row for a given parameter value. Updating only occurs when the number of successes/failures goes over a threshold *c* (Yang 2002).

Useful for noisy data: a string of successes/failures is more indicative of actual success/failure on majority of data

Example Usage: c= 5
QI count: 0 QS count: 0

Unbiased models: Update types

Counting variants (Counting NParLearner & Counting BayesLearner)

These models keep count of how many successes or failures have occurred in a row for a given parameter value. Updating only occurs when the number of successes/failures goes over a threshold *c* (Yang 2002).

Useful for noisy data: a string of successes/failures is more indicative of actual success/failure on majority of data

Example Usage: c= 5
QI count: 0 QS count: 0
(1) QI grammar mismatches
QI count: -1 QS count: 0

Unbiased models: Update types

Counting variants (Counting NParLearner & Counting BayesLearner)

These models keep count of how many successes or failures have occurred in a row for a given parameter value. Updating only occurs when the number of successes/failures goes over a threshold *c* (Yang 2002).

Useful for noisy data: a string of successes/failures is more indicative of actual success/failure on majority of data

Example Usage: c= 5
QI count: 0 QS count: 0
(1) QI grammar mismatches
QI count: -1 QS count: 0
(2) QS grammar matches 3 data points in a row
QI count: -1 QS count: 3

Unbiased models: Update types

Counting variants (Counting NParLearner & Counting BayesLearner)

These models keep count of how many successes or failures have occurred in a row for a given parameter value. Updating only occurs when the number of successes/failures goes over a threshold *c* (Yang 2002).

Useful for noisy data: a string of successes/failures is more indicative of actual success/failure on majority of data

Example Usage: c= 5
QI count: 0 QS count: 0
(1) QI grammar mismatches
QI count: -1 QS count: 0
(2) QS grammar matches 3 data points in a row
QI count: -1 QS count: 3
(3) QI grammar matches 6 data points in a row
QI count: 5 QS count: 3

Unbiased models: Update types

Counting variants (Counting NParLearner & Counting BayesLearner)

These models keep count of how many successes or failures have occurred in a row for a given parameter value. Updating only occurs when the number of successes/failures goes over a threshold *c* (Yang 2002).

Useful for noisy data: a string of successes/failures is more indicative of actual success/failure on majority of data

Example Usage: c= 5
QI count: 0 QS count: 0
(1) QI grammar mismatches
QI count: -1 QS count: 0
(2) QS grammar matches 3 data points in a row
QI count: -1 QS count: 3
(3) QI grammar matches 6 data points in a row
QI count: 5 QS count: 3

QI value rewarded (only 1 update vs. 10 updates in non-counting variants)

Processing the input

Words are processed by the model one at a time, which assumes word segmentation is operational. Evidence from Jusczyk, Houston, & Newsome (1999) that 7-month-olds can segment words successfully.

Words are divided into syllables, with syllable rime identified as VC, VV, or V. Evidence from Jusczyk, Goodman, & Baumann (1999) and Turk, Jusczyk, & Gerken (1995) suggests young infants are sensitive to syllables and properties of syllable structure.

Sub-parameters (ex: QS-VC-H vs. QS-VC-L) are not set until the main parameter is set (ex: QS). This is based on the idea that children only consider information about a sub-parameter if they have to.

Unbiased model results

Goal: Converge on English values after learning period is over

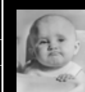
English parameter values
= {**QS**, **QS-VC-H**, **Em-Some**, **Em-Rt**, **Ft-Dir-Rt**, **B,B-2**, **B-Syl**, **Ft-Hd-Left** }

Unbiased model results: Not so good

Goal: Converge on English values after learning period is over

English parameter values
= {QS, QS-VC-H, Em-Some, Em-Rt, Ft-Dir-Rt, B,B-2, B-Syl, Ft-Hd-Left }

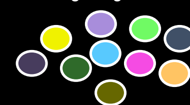
Model	Average success rate (1000 runs each condition)
NParLearner, γ = 0.001, 0.0025, 0.01, or 0.025	0.000
BayesLearner	0.000
Counting NParLearner, γ = 0.001, 0.0025, 0.01, or 0.025; c = 2, 5, 7, 10, 15, or 20	0.000333
Counting BayesLearner, c = 2, 5, 7, 10, 15, or 20	0.000



Examining why

Is it just these models or is there some underlying issue that will cause all unbiased models to fail?

Let's consider the hypothesis space, where the English grammar is 1 of 156 grammars under consideration.



How compatible are each of these competing grammars with the English child-directed speech data?

Examining why

Is it just these models or is there some underlying issue that will cause all unbiased models to fail?

Let's consider the hypothesis space, where the English grammar is 1 of 156 grammars under consideration.

It turns out that there are 51 other grammars more compatible than the English grammar with the data tokens (56 are more compatible by data types).

Implication: The English grammar is not the optimal grammar for this data set!

Examining why

Is it just these models or is there some underlying issue that will cause all unbiased models to fail?

Are the unbiased models finding the more optimal grammars?

English grammar compatibility:
72.97% by tokens, 62.14% by types

Unbiased models choose grammars with average compatibility of
73.56% by tokens, 63.30% by types

Implication: Unbiased models *are* finding the more optimal grammar for the data.

Examining why

Is it just these models or is there some underlying issue that will cause all unbiased models to fail?

The problem seems not to be that the unbiased models cannot find the more optimal grammars for the data given, but rather the problem is *because* the unbiased models find the more optimal grammars for the data given…and those grammars are not the English grammar.

Implication: This means any unbiased learning model should fail.

Larger implication: English children are not unbiased learners. They have some biases that constrain their learning.

Biased models: Bias on hypothesis space

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**.

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

Ft-Hd-Left
S S

Ft-Hd-Rt
S S

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

QS
VV V

QI
S S

Biased models: Bias on hypothesis space

Learner hypothesis bias: Ft-Hd-Left = 1.0, QS = 1.0
Hypothesis space is smaller (60 grammars)

Biased models: Bias on hypothesis space

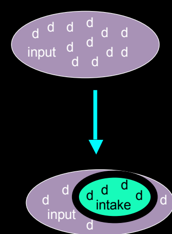
Learner hypothesis bias: Ft-Hd-Left = 1.0, QS = 1.0
Hypothesis space is smaller (60 grammars)

Model	Average success rate (1000 runs each condition)
NParLearner, γ = 0.001, 0.0025, 0.01, or 0.025	0.000
BayesLearner	0.001
Counting NParLearner, γ = 0.001, 0.0025, 0.01, or 0.025, c = 2, 5, 7, 10, 15, or 20	0.0165
Counting BayesLearner, c = 2, 5, 7, 10, 15, or 20	0.0178

Biased models: Bias on data intake

Pearl (2008): Selective learning bias

Modify the data the learner uses
(children learn only from certain data)



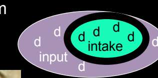
"Selective" Intuition: Use the really good data only.

One instantiation of "really good" = highly informative.

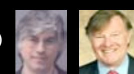
One instantiation of "highly informative" = data viewed by the learner as unambiguous (Fodor, 1998; Dresher, 1999; Lightfoot, 1999; Pearl & Weinberg, 2007)

Biased models: Bias on data intake

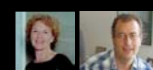
Identifying unambiguous data for a parametric system



Cues (Dresher, 1999; Lightfoot, 1999)

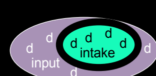


Parsing (Fodor, 1998; Sakas & Fodor, 2001)

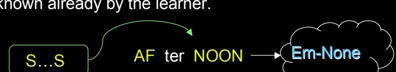


Biased models: Bias on data intake

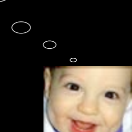
Identifying unambiguous data for a parametric system



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.

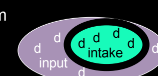


Cue for Em-None

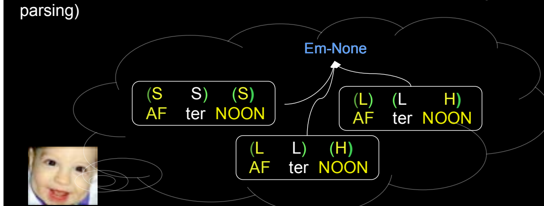


Biased models: Bias on data intake

Identifying unambiguous data for a parametric system



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



Biased models: Bias on data intake

Pearl (2008): A general class of probabilistic models learning from unambiguous data is *guaranteed* to succeed at acquiring the English grammar from English child-directed speech, provided the parameters are learned in certain orders.

Why learning from unambiguous data works: The unambiguous data favor the English grammar, so English becomes the optimal grammar.

However, they make up a small percentage of the available data (never more than 5%) so their effect can be washed away in the wake of ambiguous data if the ambiguous data are learned from as well and the parameters are not learned in an appropriate order.

Road Map

I. Background
  Parametric systems
  Parametric metrical phonology

II. Learning English metrical phonology
  Analysis of data
  Unbiased models & failure
  Biased models & success

III. Implications for acquisition

Today

Case study of acquiring a parametric system of metrical phonology, constraining the learning model to be a model of acquisition
  Input = realistic distributions of child-directed speech
  Learning period = limited to a plausible amount of time for children to acquire the system (6 months)
  Updating = incremental to reflect limited memory

What we found: Unbiased learning is not viable due to the data themselves.

Some kind of bias is required.

One that works: a plausible bias on the data intake of the learner
(learn from unambiguous data)
One that doesn't: a plausible bias on the hypothesis space
(use prior knowledge of the language's rhythmical properties)

Tomorrow?

When are biases necessary for acquisition, what biases are necessary, and what is the nature of those necessary biases?

Domain-specific biases: English metrical phonology (Pearl (2008)), English anaphoric *one* (Pearl & Lidz (submitted)), Object-Verb word order (Pearl & Weinberg (2007))

Domain-general biases: English anaphoric *one* (Pearl & Lidz (submitted), Regier & Gahl (2004)), Object-Verb word order (Pearl & Weinberg (2007)), structure-dependency (Perfors, Tenenbaum, & Regier (2006))

Tomorrow?

When we find successful biases, are they generally useful biases?

Metrical phonology, Object-Verb word order: Learning from unambiguous data is useful (Pearl (2008), Pearl & Weinberg (2007)).

English anaphoric *one*: Learning from unambiguous data is not so useful because of data sparseness. Ambiguous data must be leveraged. (Pearl & Lidz (submitted), Regier & Gahl (2004))

Tomorrow?

Can we test theories of knowledge instantiation (parametric, constraint-based, etc.) by how acquirable they are? Only acquirable knowledge instantiations are viable as representations of what children have in their minds.

One parametric system of metrical phonology is acquirable (Pearl (2008)), but only with certain biases.

Are other parametric systems also acquirable? What about constraint-based systems? What biases (if any) do they need?

Thank You

Jeff Lidz
Bill Idsardi

Charles Yang
Amy Weinberg

The audiences at

UC San Diego Linguistics Department
UC Irvine Machine Learning Group
UC Los Angeles Linguistics Department
University of Southern California Linguistics Department
GALANA 2008

Why not just do manipulations with real children?

Some manipulations are very difficult to do with children in a realistic language acquisition environment.

How do we control…
- …what hypotheses children consider?
- …what data children learn from?
- …how children change their beliefs in different hypotheses?

Ambiguity makes life hard: example

Credit problem (Dresher 1999): it's hard to know which parameter value is responsible for a particular stress contour because the data can be compatible with multiple grammars

(S)	(S	S)
CCVV	CVC	CVC
CU	CUM	ber

Grammar1 = {QI, Em-None, Ft-Dir-Rt, B,B-2, B-Syl, Ft-Hd-Left }
Grammar2 = {QS, QS-VC-H, Em-Some, Em-Rt, Ft-Dir-Left, Unb, Ft-Hd-Rt }

(H)	(H)	H
CCVV	CVC	CVC
CU	CUM	ber

No values are the same - which get(s) the credit?

Exceptional English

How many exceptions are there in the child-directed speech?
27.03% tokens (38.86% types)

Reasonable question: Is this the right parametric system to be using if the English grammar has this many exceptions?

Yes, if we believe being able to account for ~73% of the tokens (~62% of the types) with one system is better than not having a system at all to generate the observable data.

Learning trajectory:
(1) Start by learning the system that doesn't interact with morphology
(2) Realize there is interaction with the morphology system
(3) Enrich/expand the existing system to include these interactions and therefore account for more of the data

Unbiased models: Update types

Naïve Parameter Learner (Yang 2002) [NParLearner]: Linear reward-penalty (Bush & Mosteller 1951)

Learning rate γ:
small = small changes
large = large changes

Parameter values v1 vs. v2

reward v1	punish v1
$p_{v1} = p_{v1} + \gamma(1 - p_{v1})$	$p_{v1} = (1 - \gamma)p_{v1}$
$p_{v2} = 1 - p_{v1}$	$p_{v2} = 1 - p_{v1}$

Example Usage: The first data point is seen, and the grammar generated uses the QI value. That grammar fails to generate a contour that matches the observed contour. Let γ = 0.01.

Old QI value = 0.5
(Old QS value = 0.5)

New QI value = 0.495
(New QS value = 0.505)

Unbiased models: Update types

Bayesian Learner [BayesLearner]: Bayesian update of binomial distribution (Chew 1971)

Parameters α, β:

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

here: $\alpha = \beta = 0.5$

Parameter value v1

$$p_v = \frac{\alpha + 1 + successes}{\alpha + \beta + 2 + total\ data\ seen}$$

reward: success + 1 punish: success + 0

Example Usage: The first data point is seen, and the grammar generated uses the QI value. That grammar fails to generate a contour that matches the observed contour.

Old QI value = 0.5 (Old QS value = 0.5)

New QI value = 0.429 (New QS value = 0.571)

Examples of erroneous grammars chosen by unbiased models

QI, Em-Some, Em-Rt, Ft-Dir-Left, Unb, Ft-Hd-Left

QS, QS-VC-H, Em-Some, Em-Rt, Ft-Dir-Rt, Unb, Ft-Hd-Rt

QS, QS-VC-H, Em-Some, Em-Rt, Ft-Dir-Rt, B, B-2, B-Mor, Ft-Hd-Left

QS, QS-VC-L, Em-Some, Em-Rt, Ft-Dir-Rt, Unb, Ft-Hd-Rt

Required parameter-setting orders for unambiguous data

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

Cues

(a) QS-VC-Heavy before Em-Rt
(b) Em-Rt before B-Syl
(c) B-2 before B-Syl

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

Completely derivable from data saliency, data quantity, & default values

Parsing

Group 1:
QS, Ft-Hd-Left, B
Group 2:
Ft-Dir-Rt, QS-VC-Heavy
Group 3:
Em-Some, Em-Rt, B-2, B-Syl

The parameters are freely ordered w.r.t. each other within each group.

Only partially derivable from data saliency, data quantity, & default values