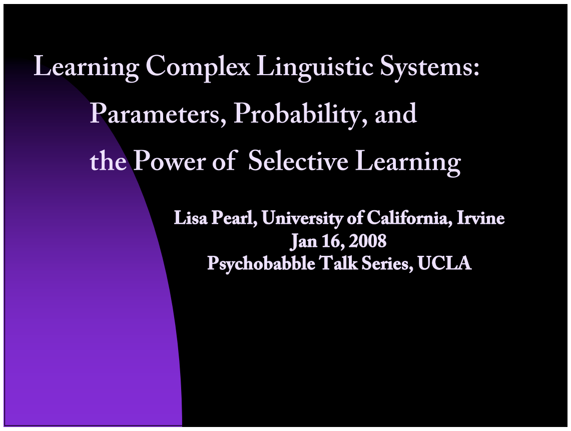
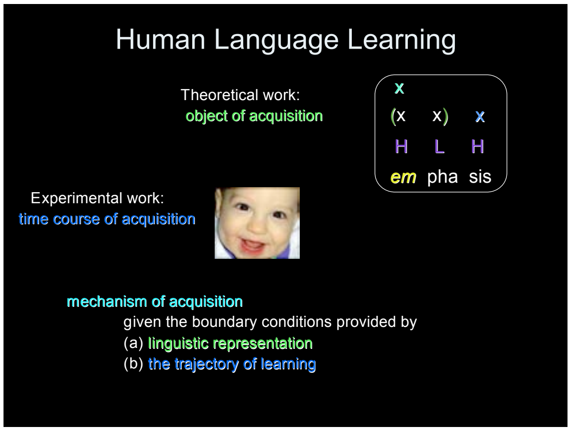


Learning Complex Linguistic Systems: Parameters, Probability, and the Power of Selective Learning

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



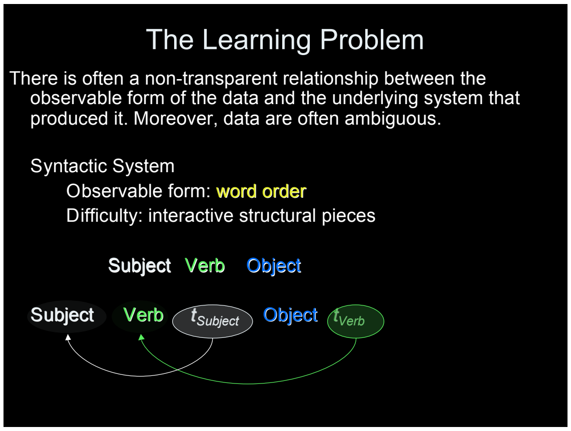
Human Language Learning

Theoretical work:
object of acquisition

x
(x x) x
H L H
em pha sis

Experimental work:
time course of acquisition

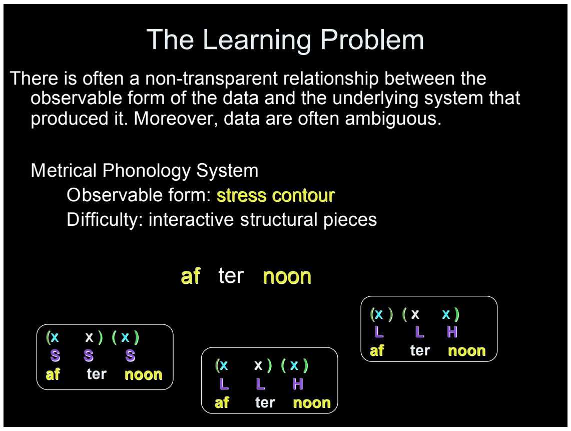
mechanism of acquisition
given the boundary conditions provided by
(a) linguistic representation
(b) the trajectory of learning



The Learning Problem

There is often a non-transparent relationship between the observable form of the data and the underlying system that produced it. Moreover, data are often ambiguous.

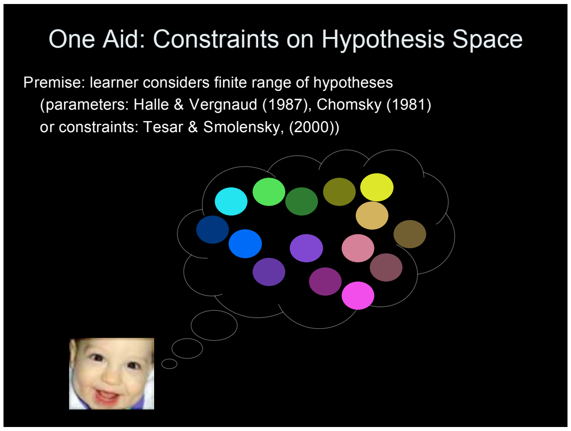
Syntactic System
Observable form: word order
Difficulty: interactive structural pieces



The Learning Problem

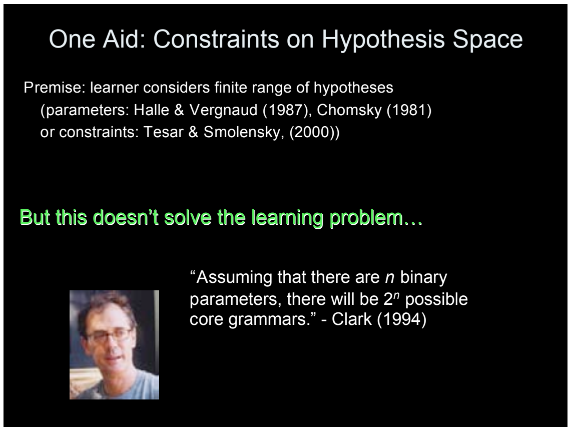
There is often a non-transparent relationship between the observable form of the data and the underlying system that produced it. Moreover, data are often ambiguous.

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



One Aid: Constraints on Hypothesis Space

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

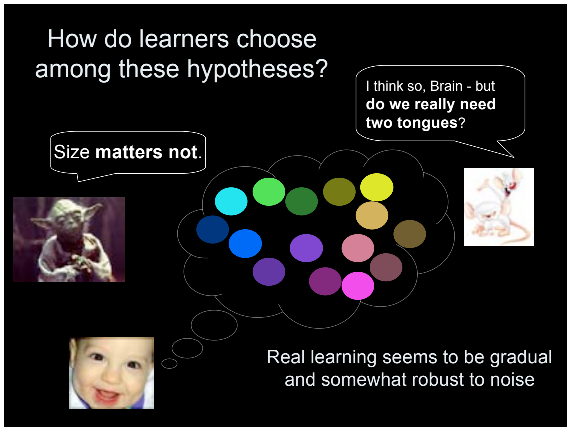


One Aid: Constraints on Hypothesis Space

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

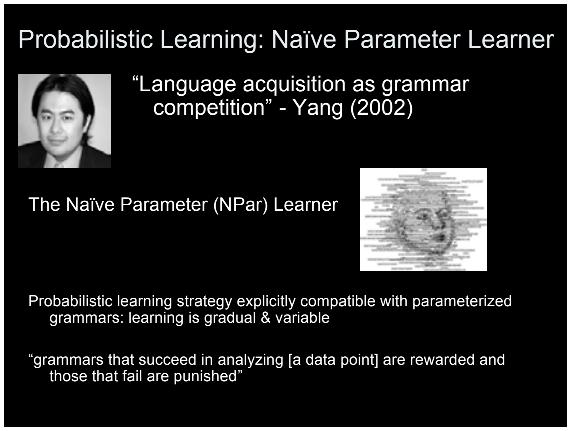
But this doesn't solve the learning problem…

"Assuming that there are *n* binary parameters, there will be 2^n possible core grammars." - Clark (1994)



How do learners choose among these hypotheses?

Real learning seems to be gradual and somewhat robust to noise



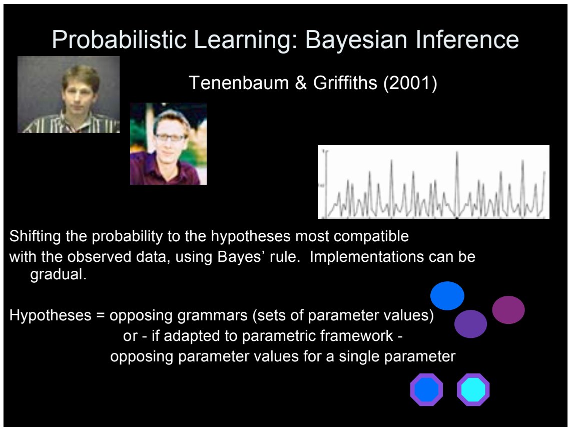
Probabilistic Learning: Naïve Parameter Learner

"Language acquisition as grammar competition" - Yang (2002)

The Naïve Parameter (NPar) Learner

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

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

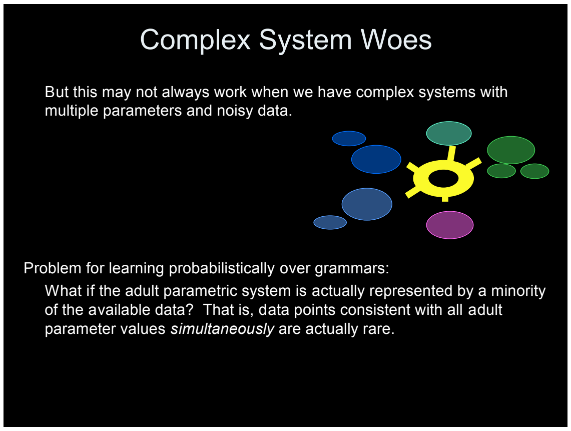


Probabilistic Learning: Bayesian Inference

Tenenbaum & Griffiths (2001)

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

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

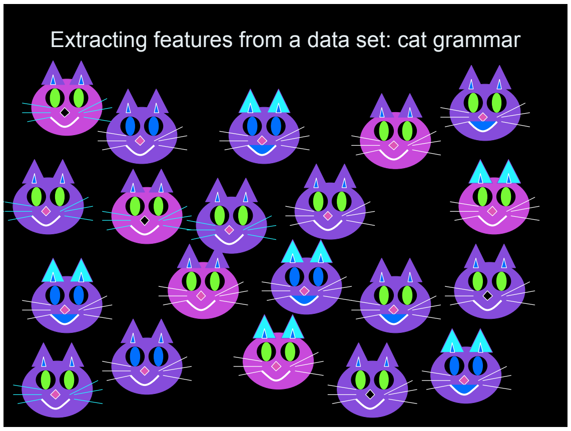


Complex System Woes

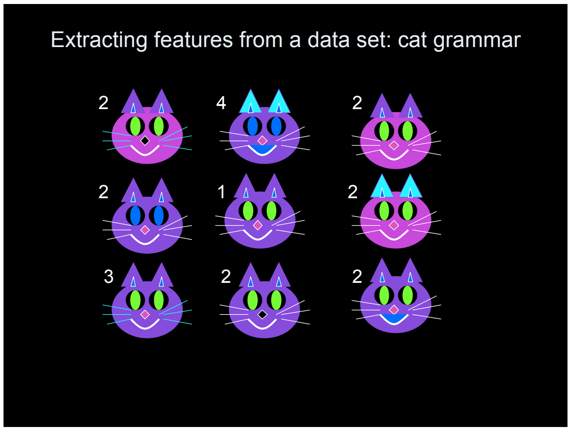
But this may not always work when we have complex systems with multiple parameters and noisy data.

Problem for learning probabilistically over grammars:

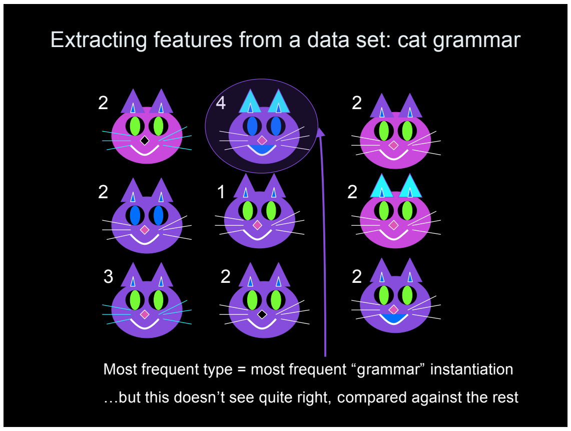
What if the adult parametric system is actually represented by a minority of the available data? That is, data points consistent with all adult parameter values *simultaneously* are actually rare.



Extracting features from a data set: cat grammar



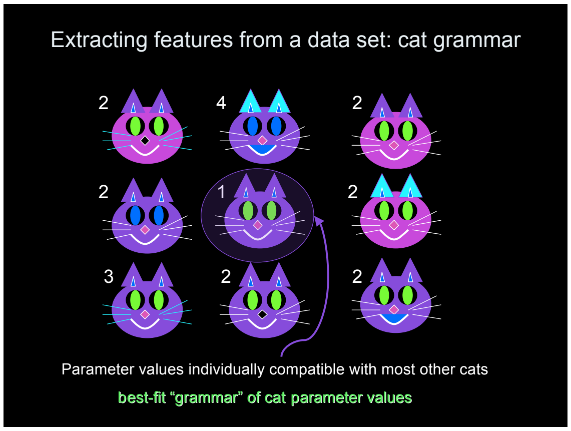
Extracting features from a data set: cat grammar



Extracting features from a data set: cat grammar

Most frequent type = most frequent “grammar” instantiation

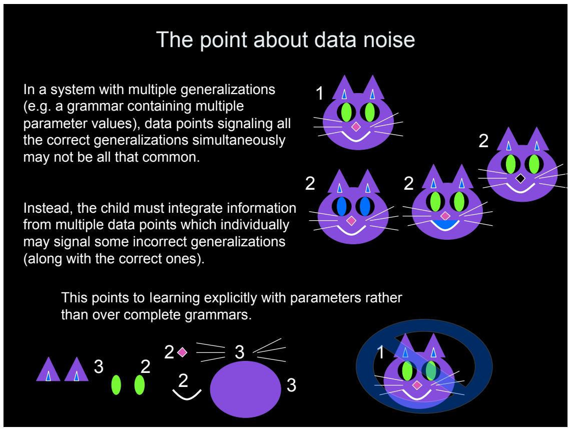
…but this doesn’t see quite right, compared against the rest



Extracting features from a data set: cat grammar

Parameter values individually compatible with most other cats

best-fit “grammar” of cat parameter values

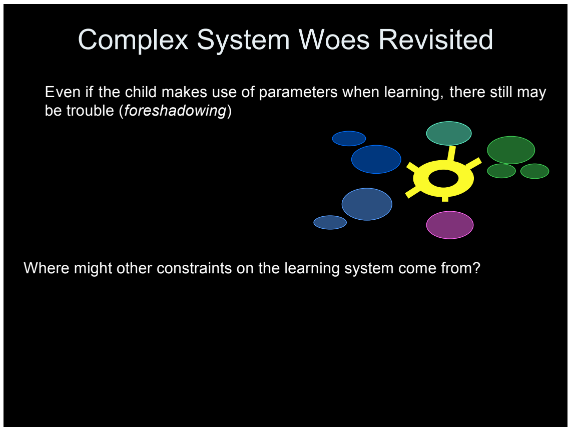


The point about data noise

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

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

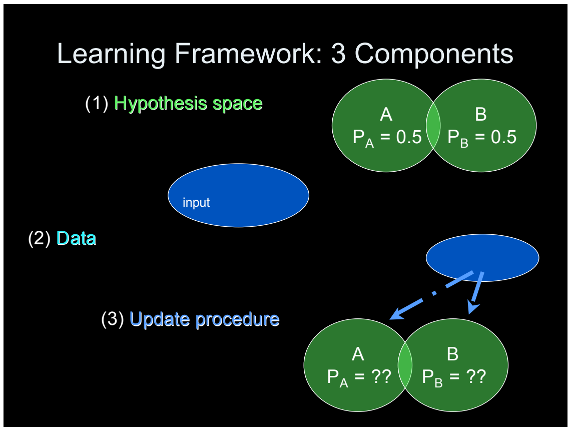
This points to learning explicitly with parameters rather than over complete grammars.



Complex System Woes Revisited

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

Where might other constraints on the learning system come from?

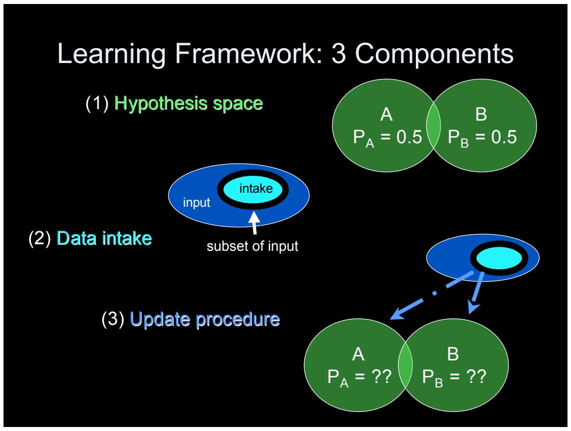


Learning Framework: 3 Components

(1) Hypothesis space

(2) Data

(3) Update procedure

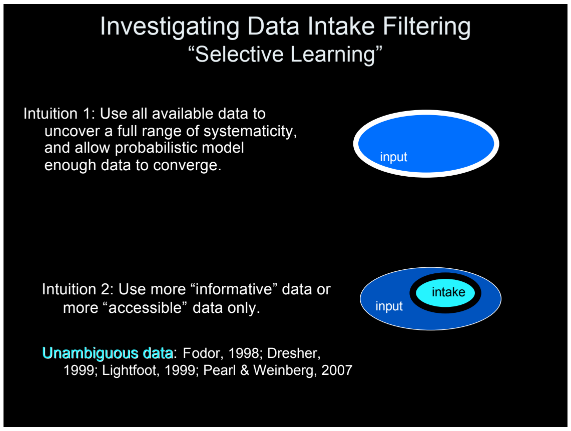


Learning Framework: 3 Components

(1) Hypothesis space

(2) Data intake

(3) Update procedure

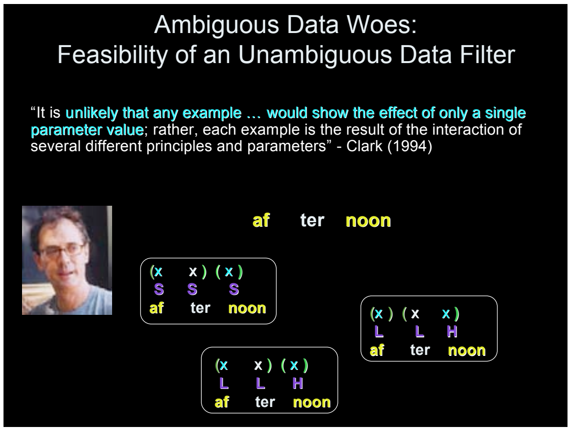


Investigating Data Intake Filtering
“Selective Learning”

Intuition 1: Use all available data to uncover a full range of systematicity, and allow probabilistic model enough data to converge.

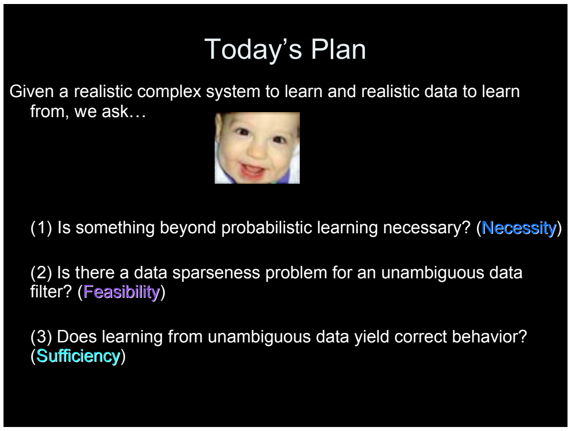
Intuition 2: Use more “informative” data or more “accessible” data only.

Unambiguous data: Fodor, 1998; Dresher, 1999; Lightfoot, 1999; Pearl & Weinberg, 2007



Ambiguous Data Woes: Feasibility of an Unambiguous Data Filter

“It is unlikely that any example … would show the effect of only a single parameter value; rather, each example is the result of the interaction of several different principles and parameters” - Clark (1994)



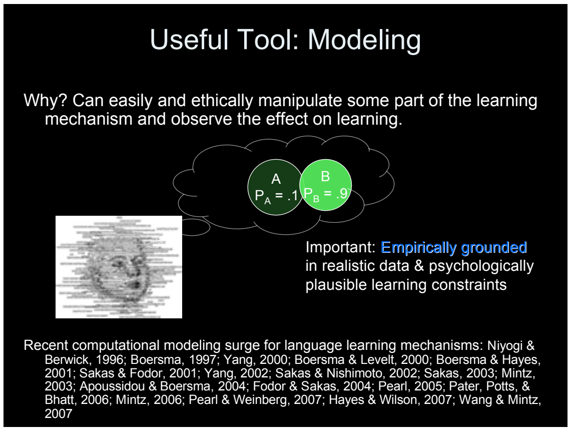
Today’s Plan

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

(1) Is something beyond probabilistic learning necessary? (Necessity)

(2) Is there a data sparseness problem for an unambiguous data filter? (Feasibility)

(3) Does learning from unambiguous data yield correct behavior? (Sufficiency)

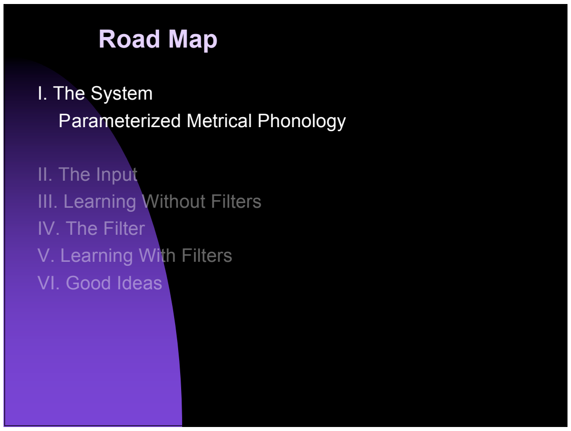


Useful Tool: Modeling

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

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

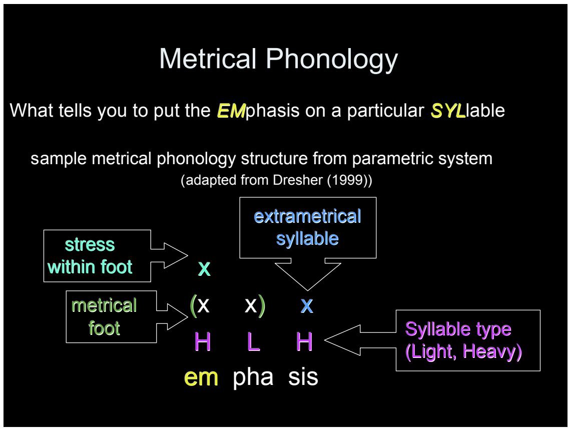
Recent computational modeling surge for language learning mechanisms: Niyogi & Berwick, 1996; Boersma, 1997; Yang, 2000; Boersma & Levelt, 2000; Boersma & Hayes, 2001; Sakas & Fodor, 2001; Yang, 2002; Sakas & Nishimoto, 2002; Sakas, 2003; Mintz, 2003; Apoussidou & Boersma, 2004; Fodor & Sakas, 2004; Pearl, 2005; Pater, Potts, & Bhatt, 2006; Mintz, 2006; Pearl & Weinberg, 2007; Hayes & Wilson, 2007; Wang & Mintz, 2007



Road Map

I. The System
Parameterized Metrical Phonology

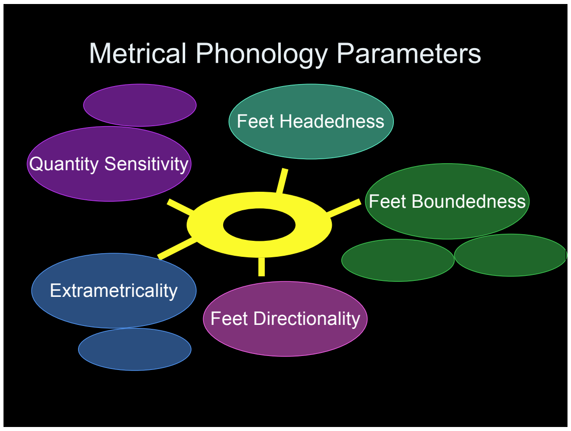
II. The Input
III. Learning Without Filters
IV. The Filter
V. Learning With Filters
VI. Good Ideas



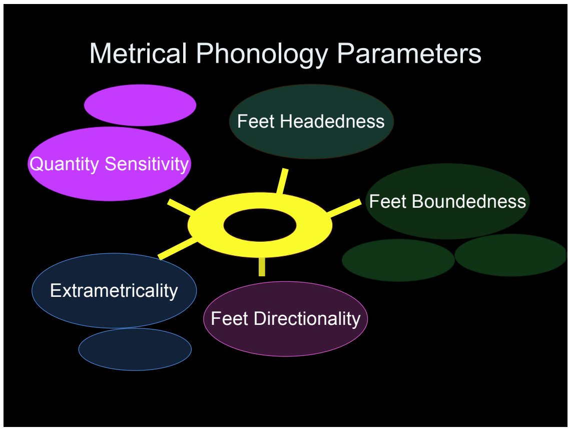
Metrical Phonology

What tells you to put the ***EM***phasis on a particular ***SYL***lable

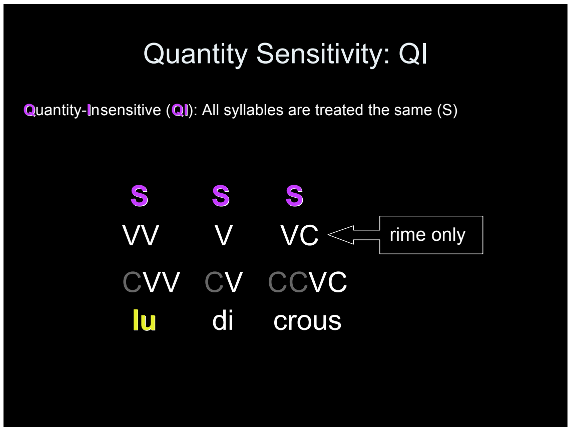
sample metrical phonology structure from parametric system
(adapted from Dresher (1999))



Metrical Phonology Parameters



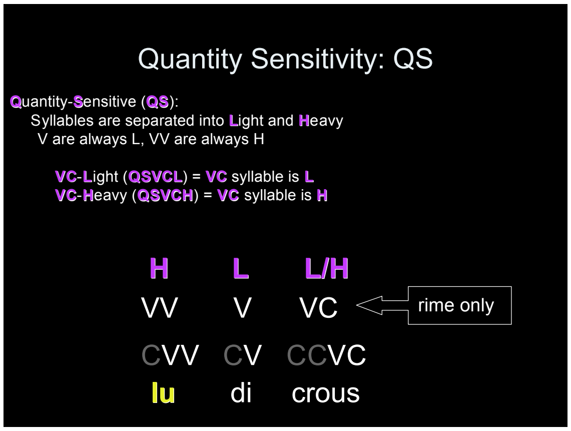
Metrical Phonology Parameters



Quantity Sensitivity: QI

Quantity-**I**nsensitive (**QI**): All syllables are treated the same (S)

S	**S**	**S**	
VV	V	VC	← rime only
CVV	CV	CCVC	
lu	di	crous	

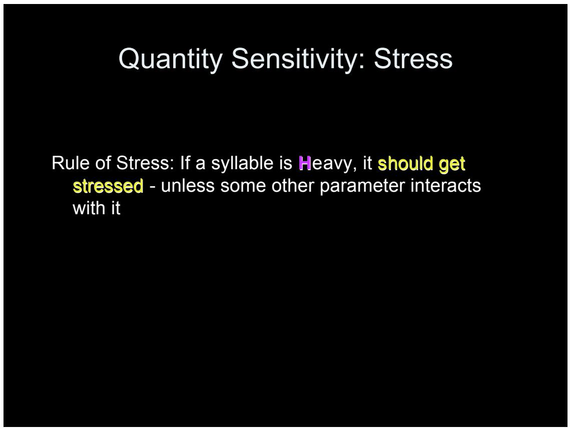


Quantity Sensitivity: QS

Quantity-**S**ensitive (**QS**):
Syllables are separated into **L**ight and **H**eavy
V are always L, VV are always H

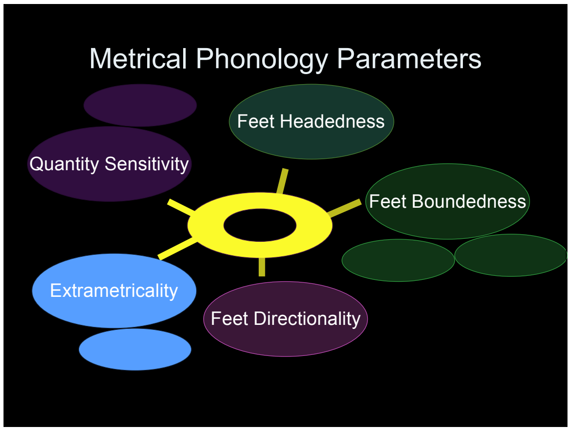
VC-**L**ight (**QSVCL**) = **VC** syllable is **L**
VC-**H**eavy (**QSVCH**) = **VC** syllable is **H**

H	**L**	**L/H**	
VV	V	VC	← rime only
CVV	CV	CCVC	
lu	di	crous	

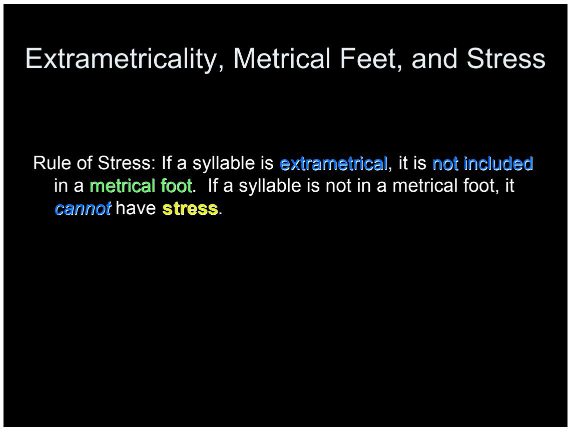


Quantity Sensitivity: Stress

Rule of Stress: If a syllable is **H**eavy, it **should get stressed** - unless some other parameter interacts with it

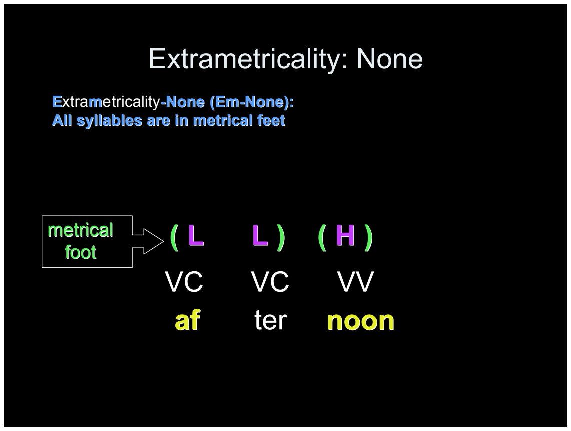


Metrical Phonology Parameters



Extrametricality, Metrical Feet, and Stress

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



Extrametricality: None

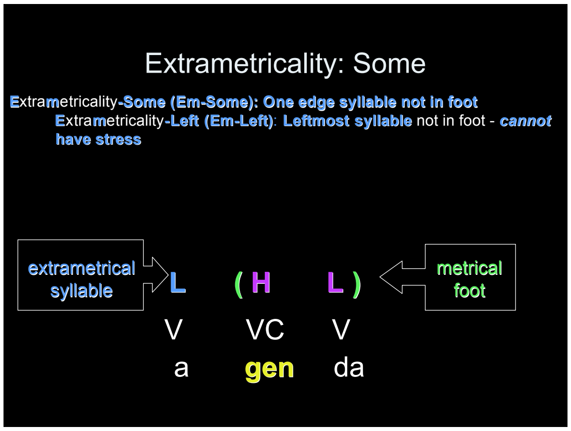
Extrametricality-None (Em-None):
All syllables are in metrical feet

metrical foot

(L   L)   (H)

VC   VC   VV

af   ter   **noon**



Extrametricality: Some

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

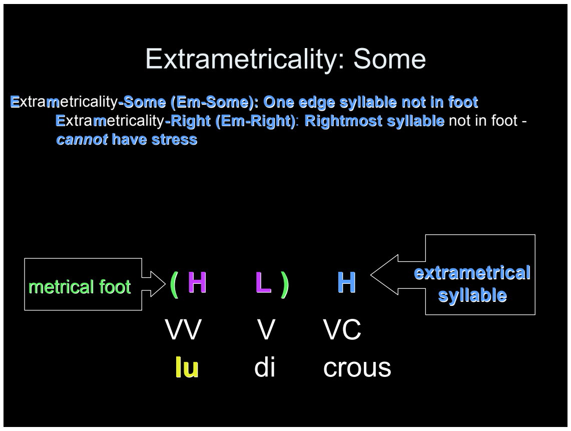
extrametrical syllable

L   (H   L)

V   VC   V

a   **gen**   da

metrical foot



Extrametricality: Some

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

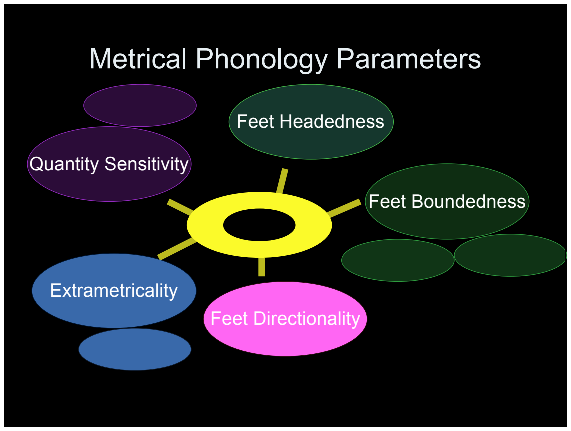
metrical foot

(H   L)   H

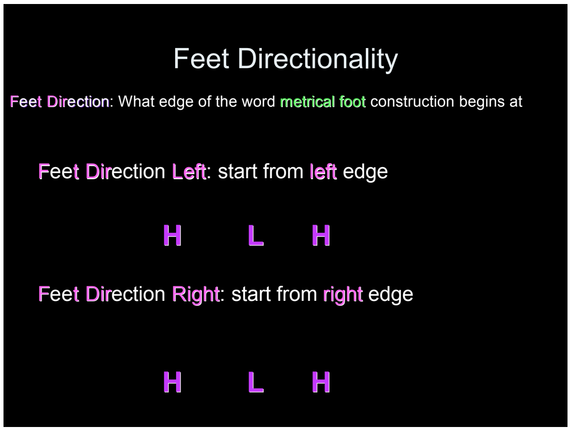
VV   V   VC

lu   di   crous

extrametrical syllable



Metrical Phonology Parameters



Feet Directionality

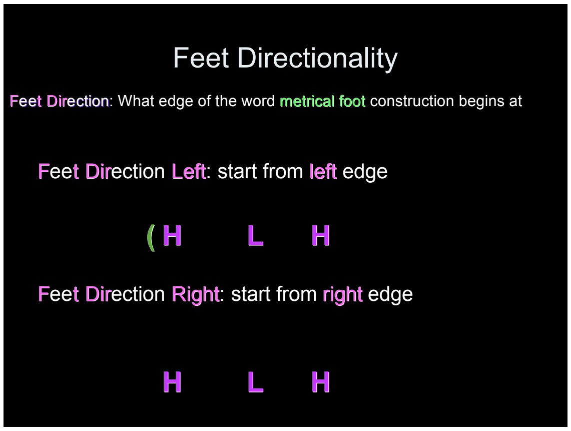
Feet Direction: What edge of the word metrical foot construction begins at

Feet Direction Left: start from left edge

H   L   H

Feet Direction Right: start from right edge

H   L   H



Feet Directionality

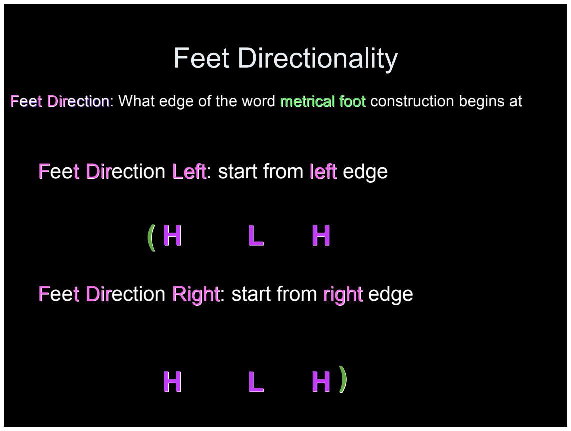
Feet Direction: What edge of the word metrical foot construction begins at

Feet Direction Left: start from left edge

(H L H

Feet Direction Right: start from right edge

H L H



Feet Directionality

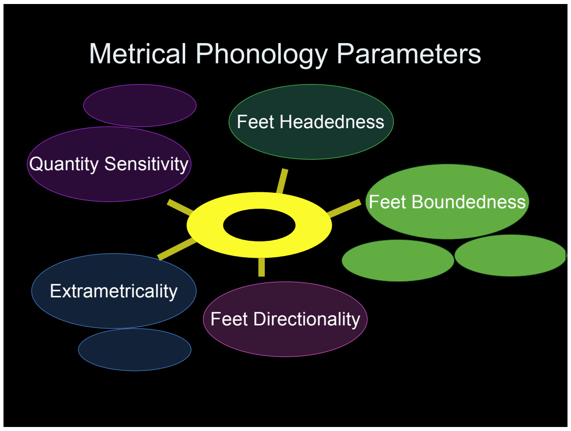
Feet Direction: What edge of the word metrical foot construction begins at

Feet Direction Left: start from left edge

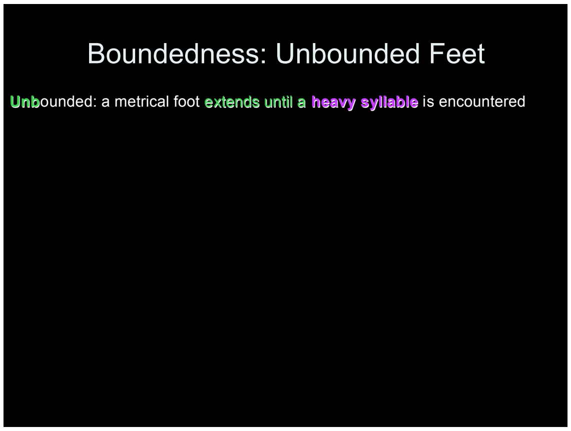
(H L H

Feet Direction Right: start from right edge

H L H)

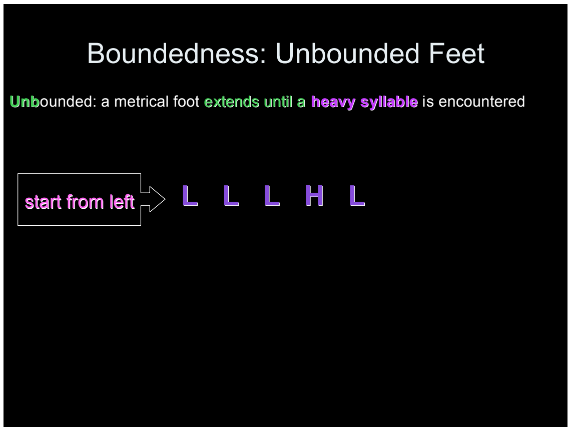


Metrical Phonology Parameters



Boundedness: Unbounded Feet

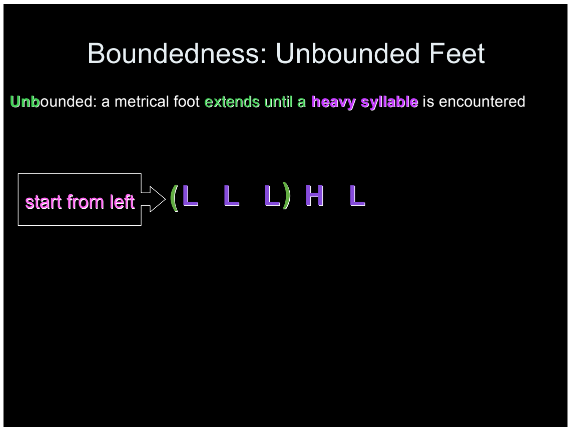
Unbounded: a metrical foot extends until a heavy syllable is encountered



Boundedness: Unbounded Feet

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

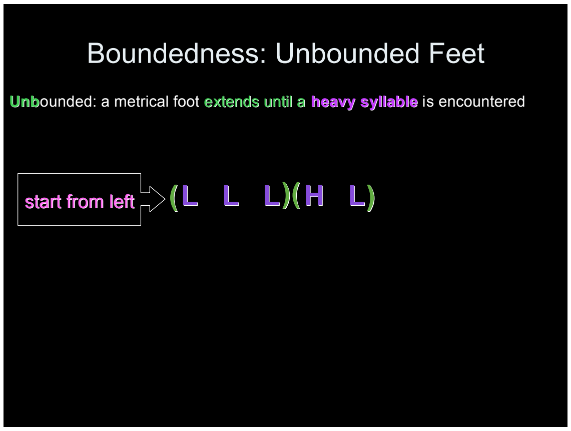
start from left ⇨ L L L H L



Boundedness: Unbounded Feet

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

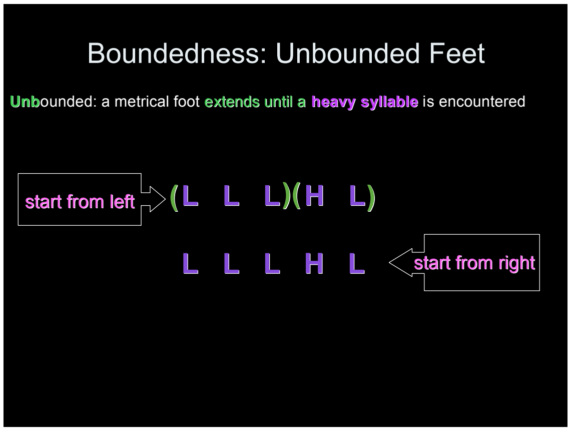
start from left ⇨ (L L L) H L



Boundedness: Unbounded Feet

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

start from left ⇨ (L L L)(H L)

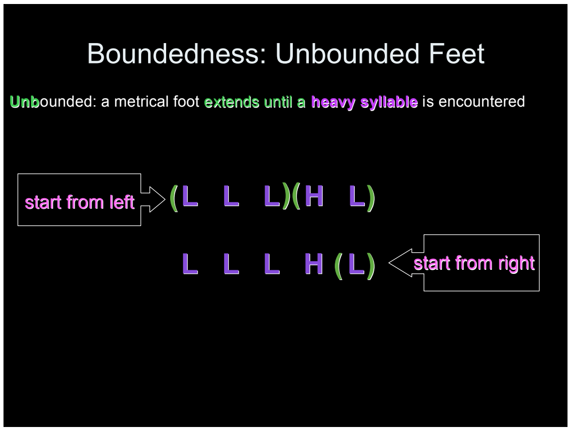


Boundedness: Unbounded Feet

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

start from left ⇨ (L L L)(H L)

L L L H L ⇦ start from right

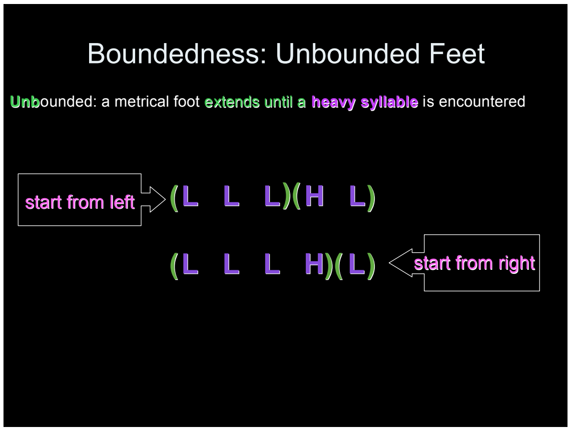


Boundedness: Unbounded Feet

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

start from left ⇨ (L L L)(H L)

L L L H (L) ⇦ start from right

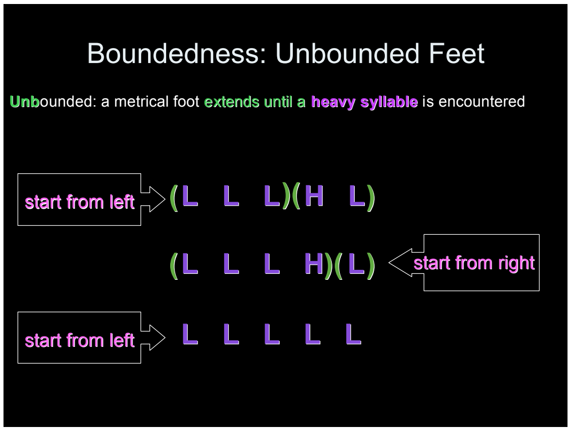


Boundedness: Unbounded Feet

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

start from left ⇨ (L L L)(H L)

(L L L H)(L) ⇦ start from right



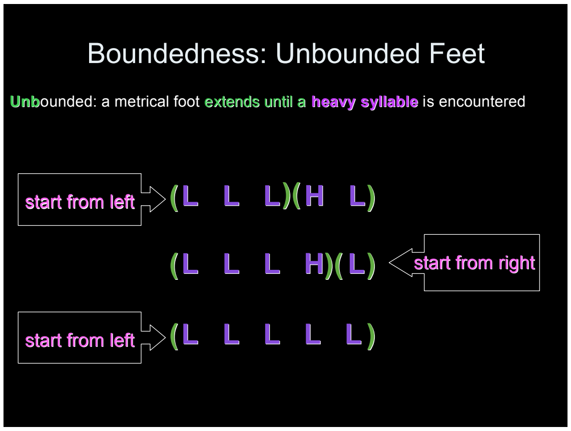
Boundedness: Unbounded Feet

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

start from left ⇨ (L L L)(H L)

(L L L H)(L) ⇦ start from right

start from left ⇨ L L L L L



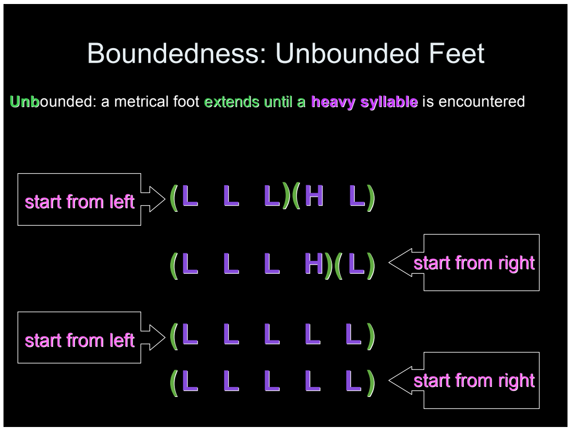
Boundedness: Unbounded Feet

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

start from left ⇨ (L L L)(H L)

(L L L H)(L) ⇦ start from right

start from left ⇨ (L L L L L)



Boundedness: Unbounded Feet

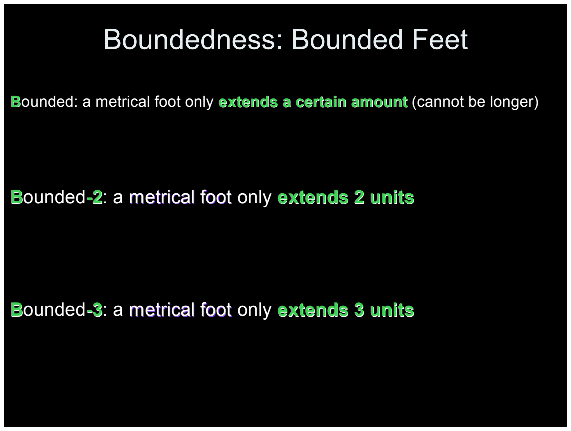
Unbounded: a metrical foot **extends until a heavy syllable** is encountered

start from left ⇨ (L L L)(H L)

(L L L H)(L) ⇦ start from right

start from left ⇨ (L L L L L)

(L L L L L) ⇦ start from right

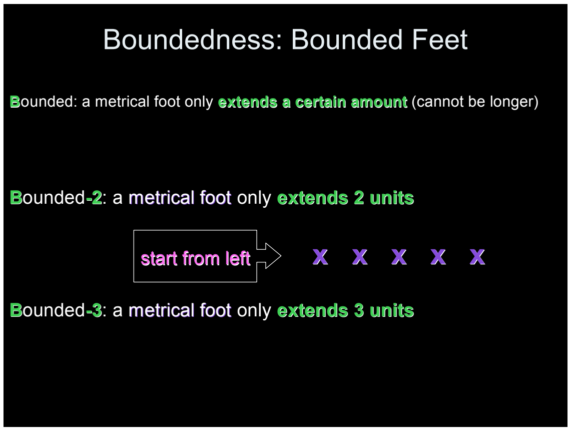


Boundedness: Bounded Feet

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

Bounded**-2**: a metrical foot only **extends 2 units**

Bounded**-3**: a metrical foot only **extends 3 units**



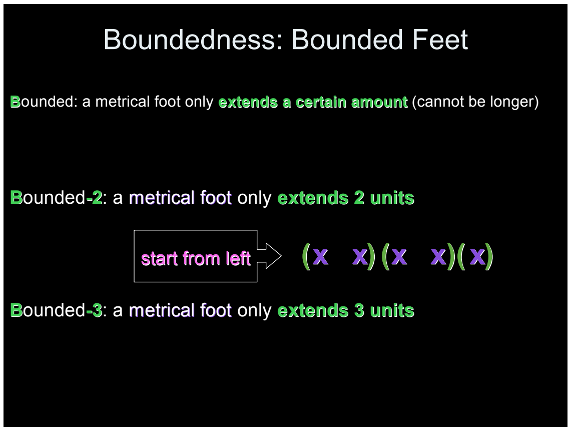
Boundedness: Bounded Feet

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

Bounded**-2**: a metrical foot only **extends 2 units**

start from left ⇨ x x x x x

Bounded**-3**: a metrical foot only **extends 3 units**



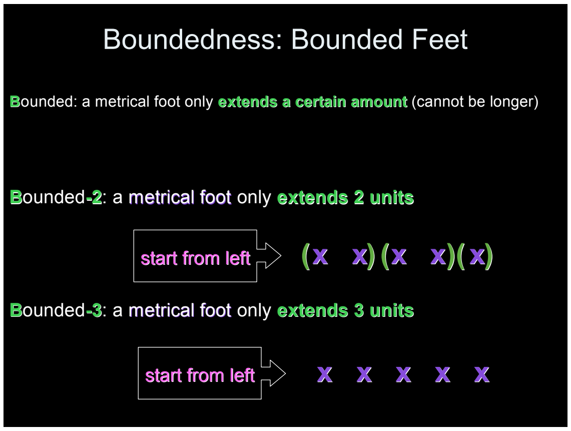
Boundedness: Bounded Feet

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

Bounded**-2**: a metrical foot only **extends 2 units**

start from left ⇨ (x x)(x x)(x)

Bounded**-3**: a metrical foot only **extends 3 units**



Boundedness: Bounded Feet

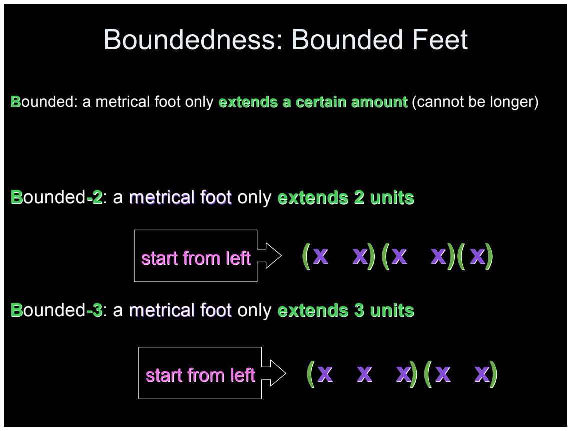
Bounded: a metrical foot only **extends a certain amount** (cannot be longer)

Bounded**-2**: a metrical foot only **extends 2 units**

start from left ⇨ (x x)(x x)(x)

Bounded**-3**: a metrical foot only **extends 3 units**

start from left ⇨ x x x x x



Boundedness: Bounded Feet

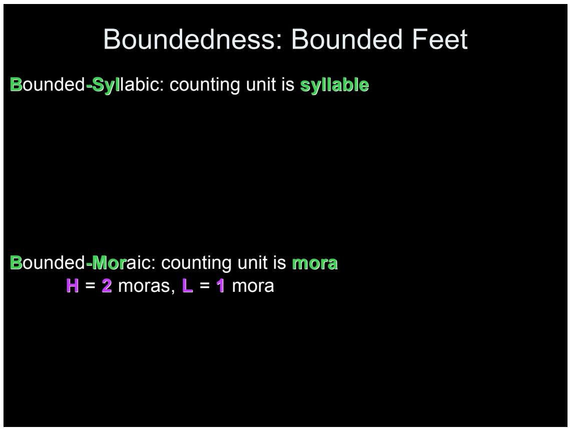
Bounded: a metrical foot only **extends a certain amount** (cannot be longer)

Bounded**-2**: a metrical foot only **extends 2 units**

start from left ⇨ (x x)(x x)(x)

Bounded**-3**: a metrical foot only **extends 3 units**

start from left ⇨ (x x x)(x x)

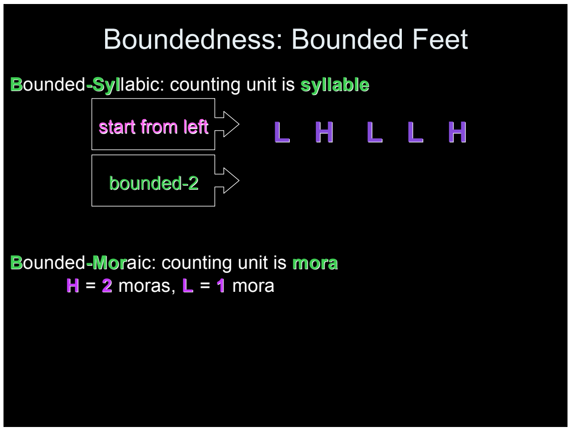


Boundedness: Bounded Feet

Bounded-**Syl**labic: counting unit is **syllable**

Bounded-**Mor**aic: counting unit is **mora**

H = **2** moras, **L** = **1** mora



Boundedness: Bounded Feet

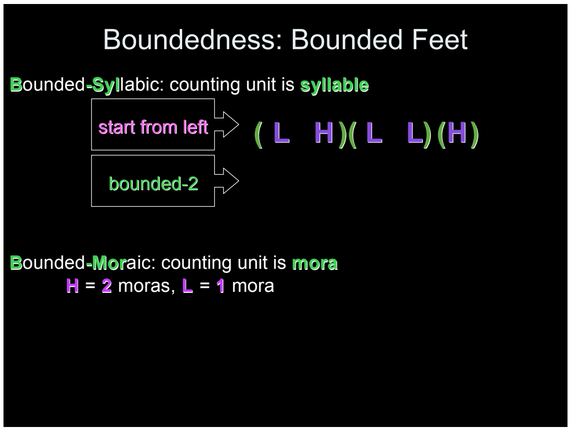
Bounded-**Syl**labic: counting unit is **syllable**

start from left → L H L L H

bounded-2 →

Bounded-**Mor**aic: counting unit is **mora**

H = **2** moras, **L** = **1** mora



Boundedness: Bounded Feet

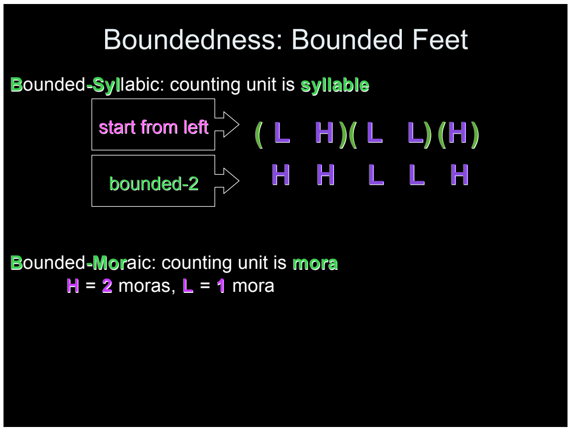
Bounded-**Syl**labic: counting unit is **syllable**

start from left → (L H)(L L)(H)

bounded-2 →

Bounded-**Mor**aic: counting unit is **mora**

H = **2** moras, **L** = **1** mora



Boundedness: Bounded Feet

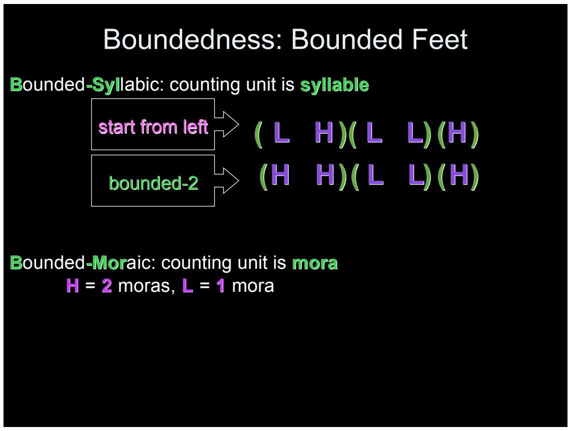
Bounded-**Syl**labic: counting unit is **syllable**

start from left → (L H)(L L)(H)

bounded-2 → H H L L H

Bounded-**Mor**aic: counting unit is **mora**

H = **2** moras, **L** = **1** mora



Boundedness: Bounded Feet

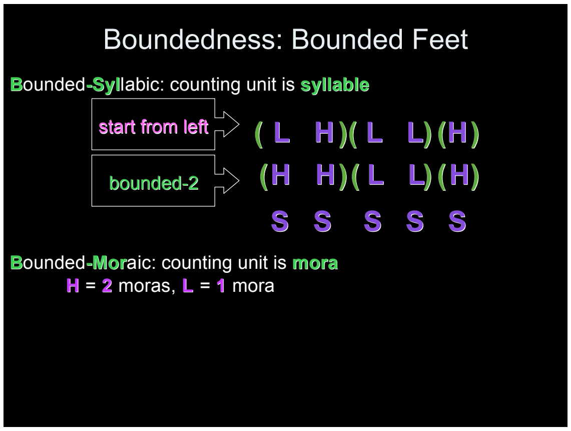
Bounded-**Syl**labic: counting unit is **syllable**

start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

Bounded-**Mor**aic: counting unit is **mora**

H = **2** moras, **L** = **1** mora



Boundedness: Bounded Feet

Bounded-**Syl**labic: counting unit is **syllable**

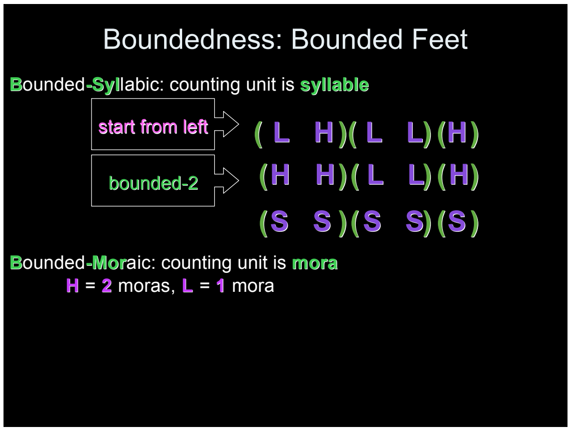
start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

S S S S S

Bounded-**Mor**aic: counting unit is **mora**

H = **2** moras, **L** = **1** mora



Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is syllable

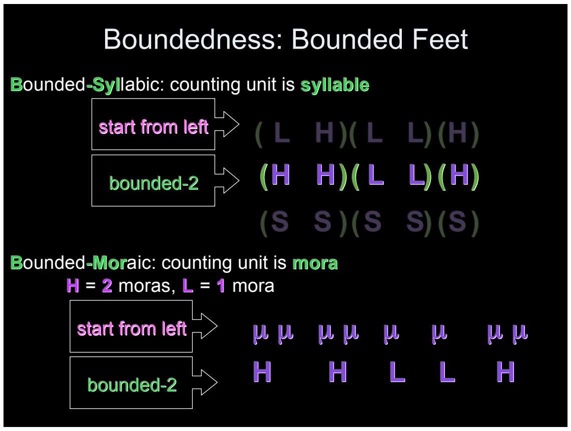
start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

(S S)(S S)(S)

Bounded-Moraic: counting unit is mora

H = 2 moras, L = 1 mora



Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is syllable

start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

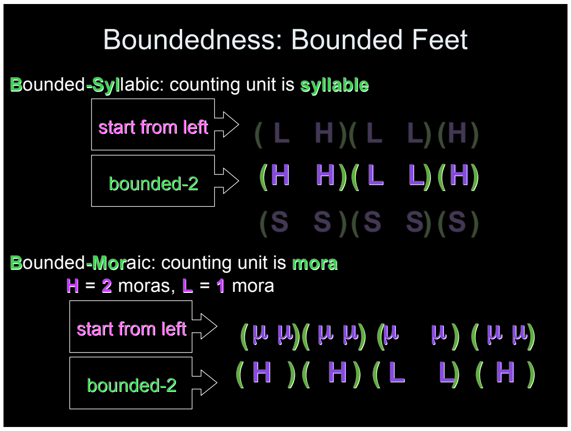
(S S)(S S)(S)

Bounded-Moraic: counting unit is mora

H = 2 moras, L = 1 mora

start from left → μ μ μ μ μ μ μ μ

bounded-2 → H H L L H



Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is syllable

start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

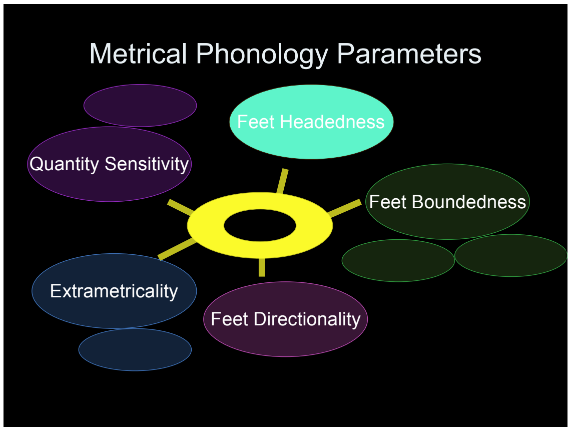
(S S)(S S)(S)

Bounded-Moraic: counting unit is mora

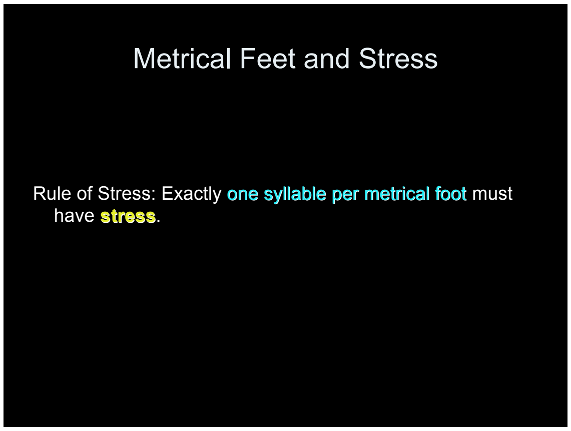
H = 2 moras, L = 1 mora

start from left → (μ μ)(μ μ) (μ μ) (μ μ)

bounded-2 → (H)(H) (L L) (H)

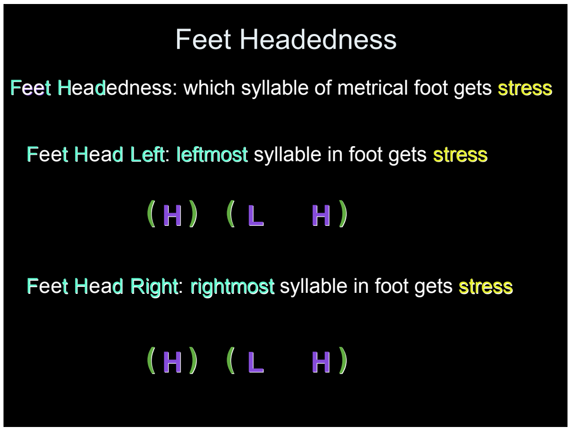


Metrical Phonology Parameters



Metrical Feet and Stress

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



Feet Headedness

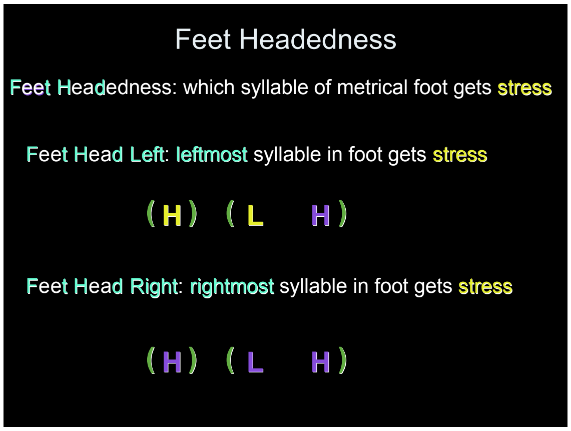
Feet Headedness: which syllable of metrical foot gets stress

Feet Head Left: leftmost syllable in foot gets stress

(H) (L H)

Feet Head Right: rightmost syllable in foot gets stress

(H) (L H)



Feet Headedness

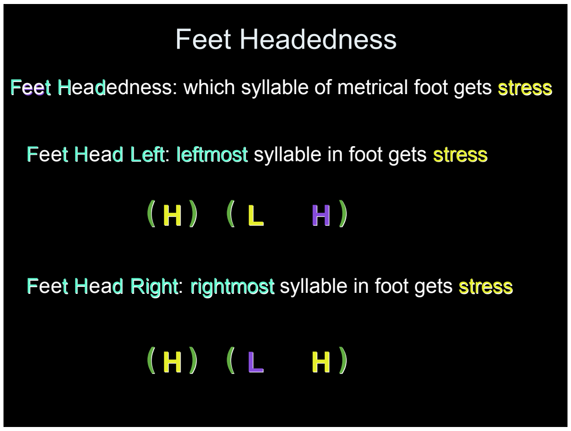
Feet Headedness: which syllable of metrical foot gets stress

Feet Head Left: leftmost syllable in foot gets stress

(H) (L H)

Feet Head Right: rightmost syllable in foot gets stress

(H) (L H)



Feet Headedness

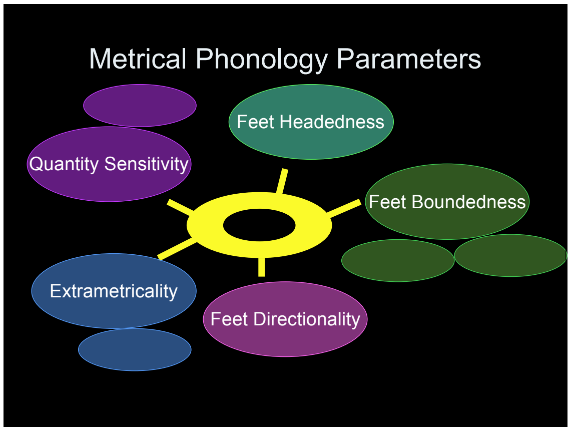
Feet Headedness: which syllable of metrical foot gets stress

Feet Head Left: leftmost syllable in foot gets stress

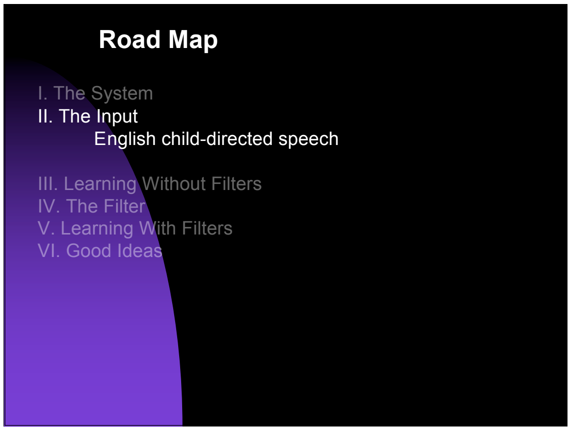
(H) (L H)

Feet Head Right: rightmost syllable in foot gets stress

(H) (L H)

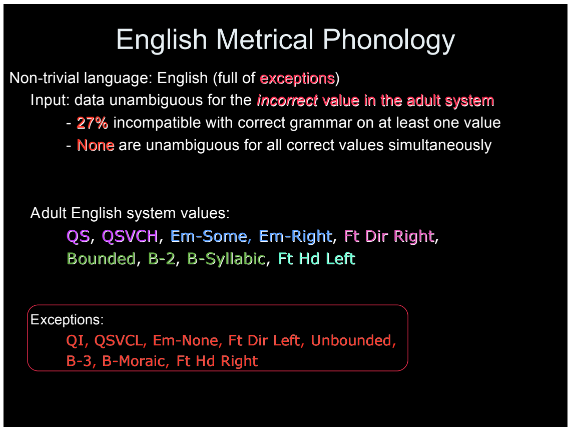


Metrical Phonology Parameters



Road Map

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



English Metrical Phonology

Non-trivial language: English (full of exceptions)

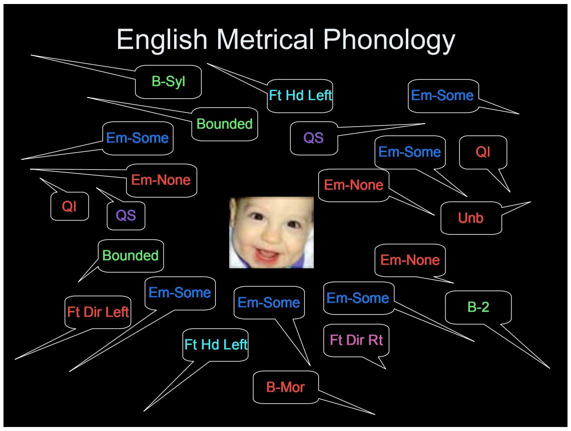
Input: data unambiguous for the *incorrect* value in the adult system

- 27% incompatible with correct grammar on at least one value
- None are unambiguous for all correct values simultaneously

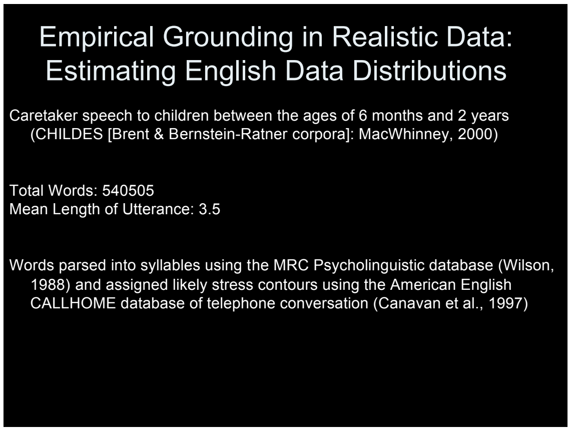
Adult English system values:

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

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



English Metrical Phonology

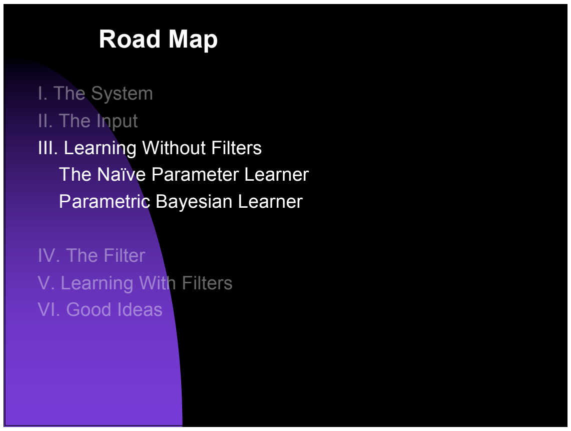


Empirical Grounding in Realistic Data: Estimating English Data Distributions

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

Total Words: 540505
Mean Length of Utterance: 3.5

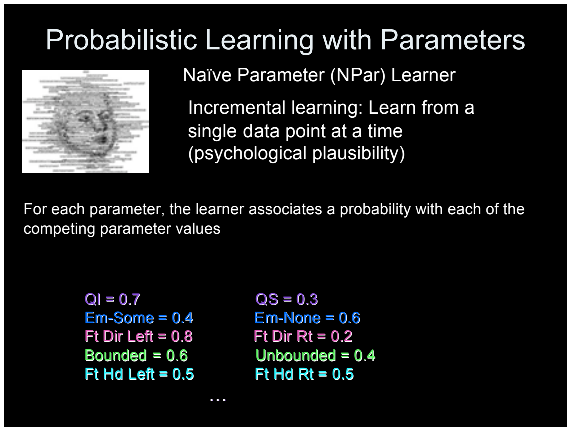
Words parsed into syllables using the MRC Psycholinguistic database (Wilson, 1988) and assigned likely stress contours using the American English CALLHOME database of telephone conversation (Canavan et al., 1997)



Road Map

I. The System
II. The Input
III. Learning Without Filters
    The Naïve Parameter Learner
    Parametric Bayesian Learner

IV. The Filter
V. Learning With Filters
VI. Good Ideas



Probabilistic Learning with Parameters

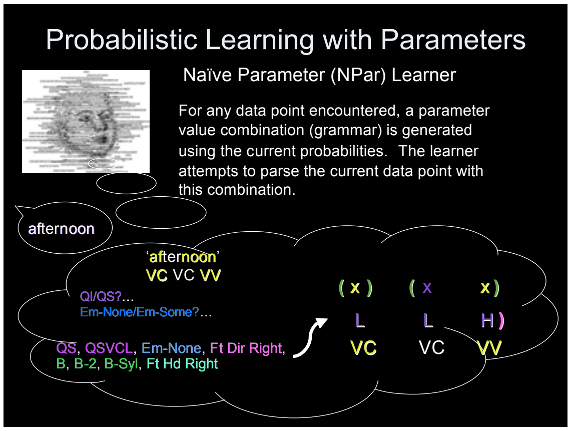
Naïve Parameter (NPar) Learner

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

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

QI = 0.7	QS = 0.3
Em-Some = 0.4	Em-None = 0.6
Ft Dir Left = 0.8	Ft Dir Rt = 0.2
Bounded = 0.6	Unbounded = 0.4
Ft Hd Left = 0.5	Ft Hd Rt = 0.5

…



Probabilistic Learning with Parameters

Naïve Parameter (NPar) Learner

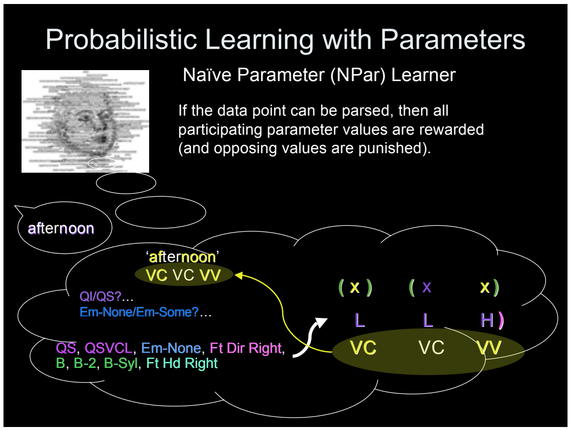
For any data point encountered, a parameter value combination (grammar) is generated using the current probabilities. The learner attempts to parse the current data point with this combination.

afternoon

'afternoon'
VC VC VV

QI/QS?…
Em-None/Em-Some?…

QS, QSVCL, Em-None, Ft Dir Right, B, B-2, B-Syl, Ft Hd Right



Probabilistic Learning with Parameters

Naïve Parameter (NPar) Learner

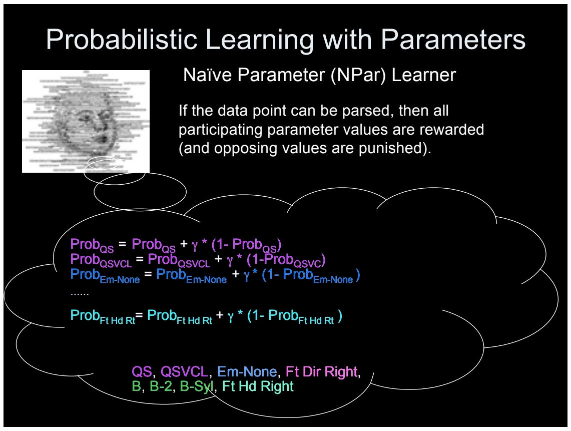
If the data point can be parsed, then all participating parameter values are rewarded (and opposing values are punished).

afternoon

'afternoon'
VC VC VV

QI/QS?…
Em-None/Em-Some?…

QS, QSVCL, Em-None, Ft Dir Right, B, B-2, B-Syl, Ft Hd Right



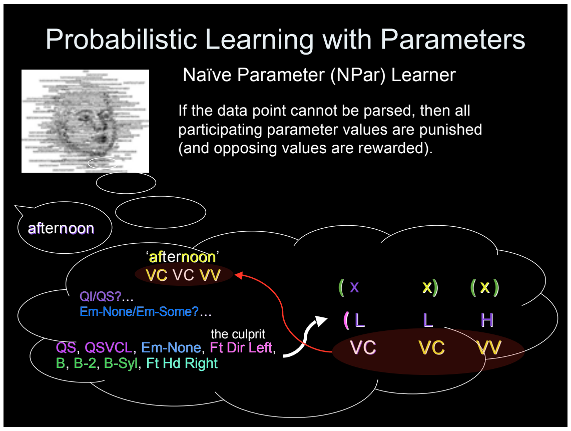
Probabilistic Learning with Parameters

Naïve Parameter (NPar) Learner

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

$\text{Prob}_{QS} = \text{Prob}_{QS} + \gamma * (1- \text{Prob}_{QS})$
$\text{Prob}_{QSVCL} = \text{Prob}_{QSVCL} + \gamma * (1-\text{Prob}_{QSVC})$
$\text{Prob}_{Em\text{-}None} = \text{Prob}_{Em\text{-}None} + \gamma * (1- \text{Prob}_{Em\text{-}None})$
……
$\text{Prob}_{Ft\ Hd\ Rt} = \text{Prob}_{Ft\ Hd\ Rt} + \gamma * (1- \text{Prob}_{Ft\ Hd\ Rt})$

QS, QSVCL, Em-None, Ft Dir Right, B, B-2, B-Syl, Ft Hd Right



Probabilistic Learning with Parameters

Naïve Parameter (NPar) Learner

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

afternoon

'afternoon'

VC VC VV

QI/QS?…

Em-None/Em-Some?…

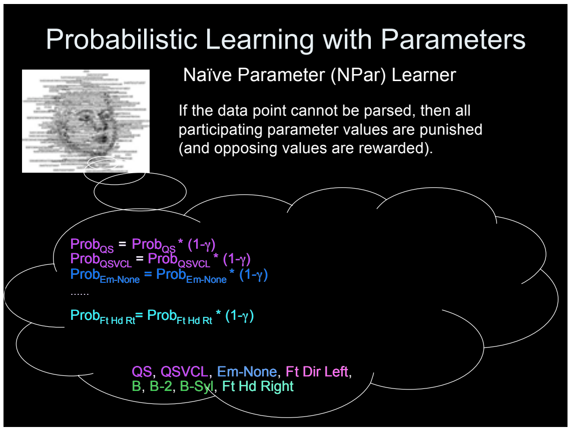
the culprit

QS, QSVCL, Em-None, Ft Dir Left,
B, B-2, B-Syl, Ft Hd Right

(x x) (x)

(L L H

VC VC VV



Probabilistic Learning with Parameters

Naïve Parameter (NPar) Learner

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

$Prob_{QS} = Prob_{QS} * (1-\gamma)$

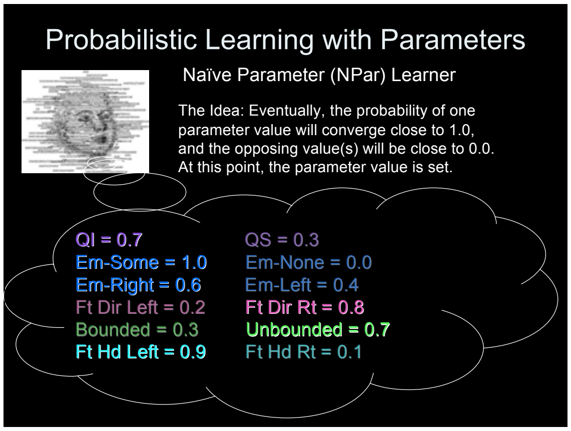
$Prob_{QSVCL} = Prob_{QSVCL} * (1-\gamma)$

$Prob_{Em\text{-}None} = Prob_{Em\text{-}None} * (1-\gamma)$

……

$Prob_{Ft\ Hd\ Rt} = Prob_{Ft\ Hd\ Rt} * (1-\gamma)$

QS, QSVCL, Em-None, Ft Dir Left,
B, B-2, B-Syl, Ft Hd Right

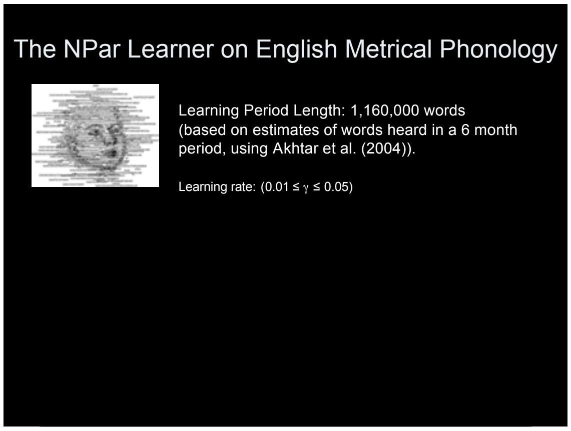


Probabilistic Learning with Parameters

Naïve Parameter (NPar) Learner

The Idea: Eventually, the probability of one parameter value will converge close to 1.0, and the opposing value(s) will be close to 0.0. At this point, the parameter value is set.

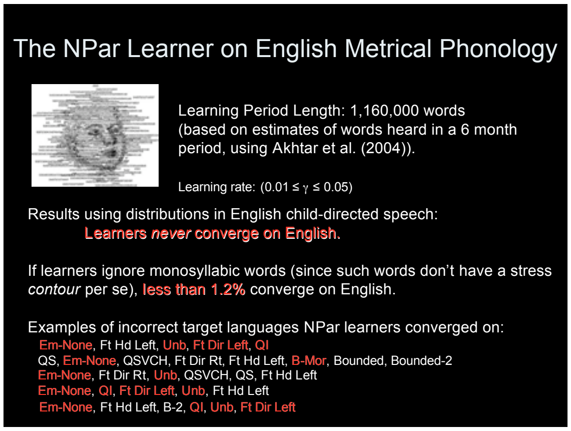
QI = 0.7	QS = 0.3
Em-Some = 1.0	Em-None = 0.0
Em-Right = 0.6	Em-Left = 0.4
Ft Dir Left = 0.2	Ft Dir Rt = 0.8
Bounded = 0.3	Unbounded = 0.7
Ft Hd Left = 0.9	Ft Hd Rt = 0.1



The NPar Learner on English Metrical Phonology

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

Learning rate: ($0.01 \leq \gamma \leq 0.05$)



The NPar Learner on English Metrical Phonology

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

Learning rate: ($0.01 \leq \gamma \leq 0.05$)

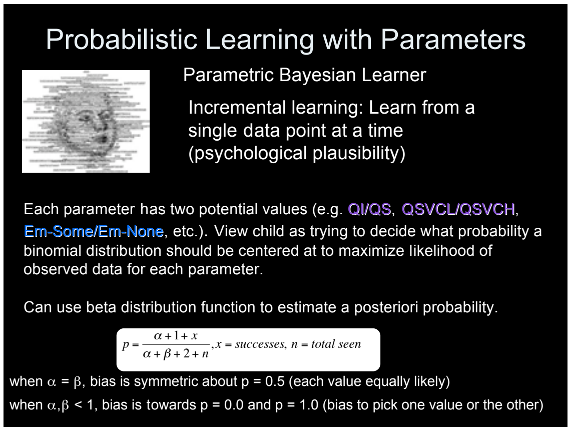
Results using distributions in English child-directed speech:

Learners *never* converge on English.

If learners ignore monosyllabic words (since such words don't have a stress *contour* per se), less than 1.2% converge on English.

Examples of incorrect target languages NPar learners converged on:

Em-None, Ft Hd Left, Unb, Ft Dir Left, QI
QS, Em-None, QSVCH, Ft Dir Rt, Ft Hd Left, B-Mor, Bounded, Bounded-2
Em-None, Ft Dir Rt, Unb, QSVCH, QS, Ft Hd Left
Em-None, QI, Ft Dir Left, Unb, Ft Hd Left
Em-None, Ft Hd Left, B-2, QI, Unb, Ft Dir Left



Probabilistic Learning with Parameters

Parametric Bayesian Learner

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

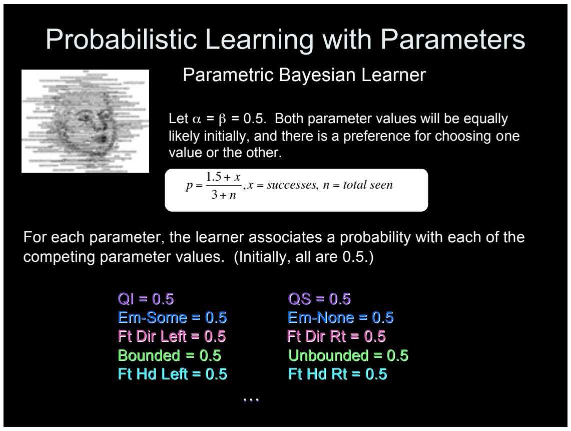
Each parameter has two potential values (e.g. QI/QS, QSVCL/QSVCH, Em-Some/Em-None, etc.). View child as trying to decide what probability a binomial distribution should be centered at to maximize likelihood of observed data for each parameter.

Can use beta distribution function to estimate a posteriori probability.

$$p = \frac{\alpha + 1 + x}{\alpha + \beta + 2 + n}, x = successes,\ n = total\ seen$$

when $\alpha = \beta$, bias is symmetric about p = 0.5 (each value equally likely)

when $\alpha,\beta < 1$, bias is towards p = 0.0 and p = 1.0 (bias to pick one value or the other)



Probabilistic Learning with Parameters

Parametric Bayesian Learner

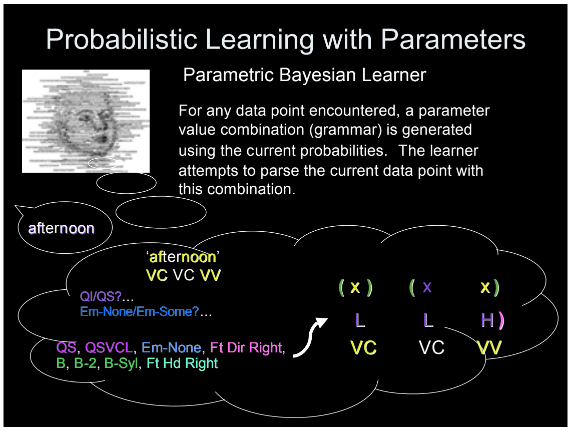
Let α = β = 0.5. Both parameter values will be equally likely initially, and there is a preference for choosing one value or the other.

$$p = \frac{1.5 + x}{3 + n}, x = successes,\ n = total\ seen$$

For each parameter, the learner associates a probability with each of the competing parameter values. (Initially, all are 0.5.)

QI = 0.5	QS = 0.5
Em-Some = 0.5	Em-None = 0.5
Ft Dir Left = 0.5	Ft Dir Rt = 0.5
Bounded = 0.5	Unbounded = 0.5
Ft Hd Left = 0.5	Ft Hd Rt = 0.5

…



Probabilistic Learning with Parameters

Parametric Bayesian Learner

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

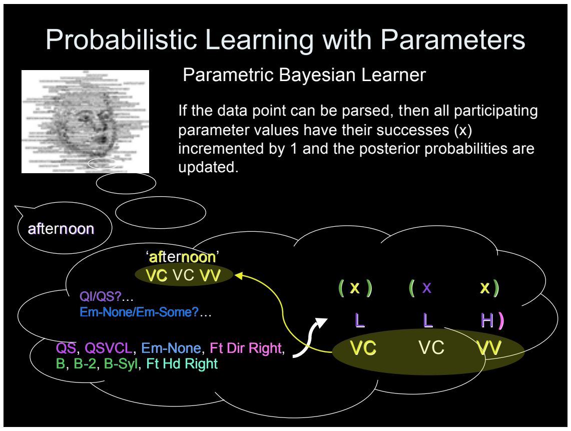
afternoon

'afternoon'
VC VC VV

QI/QS?...
Em-None/Em-Some?...

QS, QSVCL, Em-None, Ft Dir Right,
B, B-2, B-Syl, Ft Hd Right

(x) (x x)
L L H)
VC VC VV



Probabilistic Learning with Parameters

Parametric Bayesian Learner

If the data point can be parsed, then all participating parameter values have their successes (x) incremented by 1 and the posterior probabilities are updated.

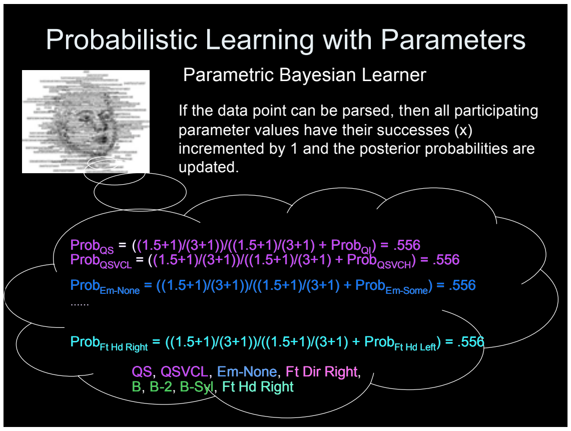
afternoon

'afternoon'
VC VC VV

QI/QS?...
Em-None/Em-Some?...

QS, QSVCL, Em-None, Ft Dir Right,
B, B-2, B-Syl, Ft Hd Right

(x) (x x)
L L H)
VC VC VV



Probabilistic Learning with Parameters

Parametric Bayesian Learner

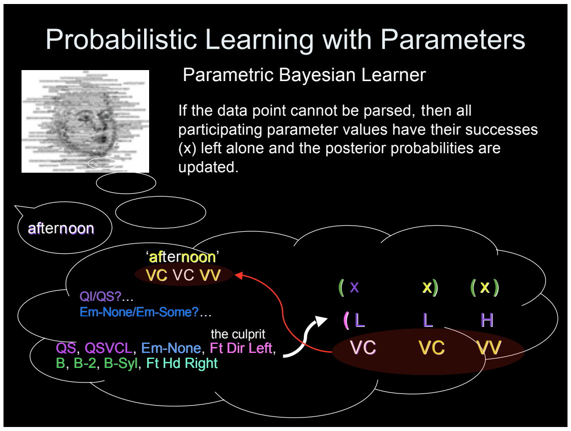
If the data point can be parsed, then all participating parameter values have their successes (x) incremented by 1 and the posterior probabilities are updated.

Prob_{QS} = ((1.5+1)/(3+1))/((1.5+1)/(3+1) + Prob_{QI}) = .556
Prob_{QSVCL} = ((1.5+1)/(3+1))/((1.5+1)/(3+1) + Prob_{QSVCH}) = .556

$\text{Prob}_{Em\text{-}None}$ = ((1.5+1)/(3+1))/((1.5+1)/(3+1) + $\text{Prob}_{Em\text{-}Some}$) = .556
……

$\text{Prob}_{Ft\ Hd\ Right}$ = ((1.5+1)/(3+1))/((1.5+1)/(3+1) + $\text{Prob}_{Ft\ Hd\ Left}$) = .556

QS, QSVCL, Em-None, Ft Dir Right,
B, B-2, B-Syl, Ft Hd Right



Probabilistic Learning with Parameters

Parametric Bayesian Learner

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

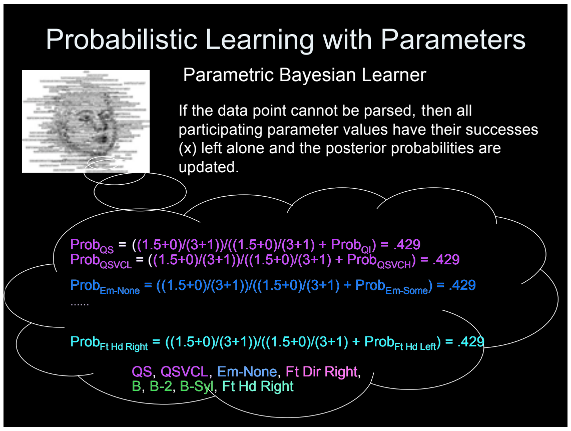
afternoon

'afternoon'
VC VC VV

QI/QS?...
Em-None/Em-Some?...

the culprit
QS, QSVCL, Em-None, Ft Dir Left,
B, B-2, B-Syl, Ft Hd Right

(x x) (x)
(L L H
VC VC VV



Probabilistic Learning with Parameters

Parametric Bayesian Learner

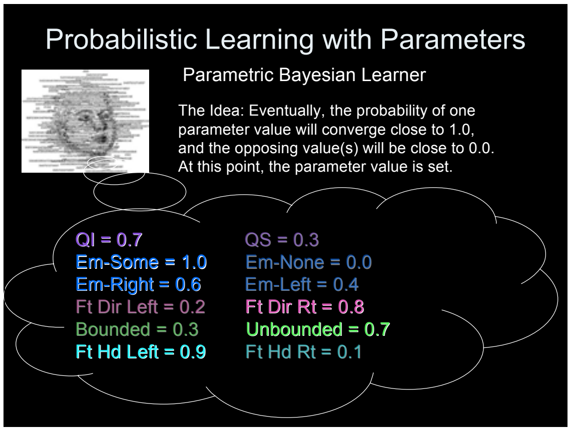
If the data point cannot be parsed, then all participating parameter values have their successes (x) left alone and the posterior probabilities are updated.

Prob_{QS} = ((1.5+0)/(3+1))/((1.5+0)/(3+1) + Prob_{QI}) = .429
Prob_{QSVCL} = ((1.5+0)/(3+1))/((1.5+0)/(3+1) + Prob_{QSVCH}) = .429

$\text{Prob}_{Em\text{-}None}$ = ((1.5+0)/(3+1))/((1.5+0)/(3+1) + $\text{Prob}_{Em\text{-}Some}$) = .429
……

$\text{Prob}_{Ft\ Hd\ Right}$ = ((1.5+0)/(3+1))/((1.5+0)/(3+1) + $\text{Prob}_{Ft\ Hd\ Left}$) = .429

QS, QSVCL, Em-None, Ft Dir Right,
B, B-2, B-Syl, Ft Hd Right

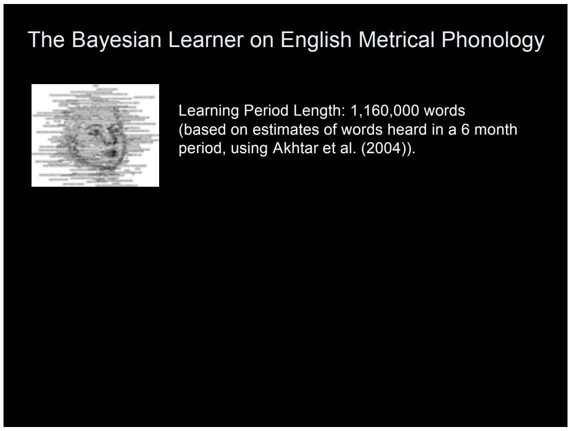


Probabilistic Learning with Parameters

Parametric Bayesian Learner

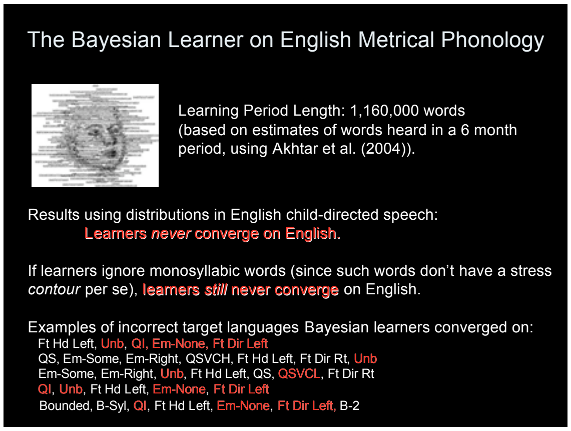
The Idea: Eventually, the probability of one parameter value will converge close to 1.0, and the opposing value(s) will be close to 0.0. At this point, the parameter value is set.

QI = 0.7	QS = 0.3
Em-Some = 1.0	Em-None = 0.0
Em-Right = 0.6	Em-Left = 0.4
Ft Dir Left = 0.2	Ft Dir Rt = 0.8
Bounded = 0.3	Unbounded = 0.7
Ft Hd Left = 0.9	Ft Hd Rt = 0.1



The Bayesian Learner on English Metrical Phonology

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



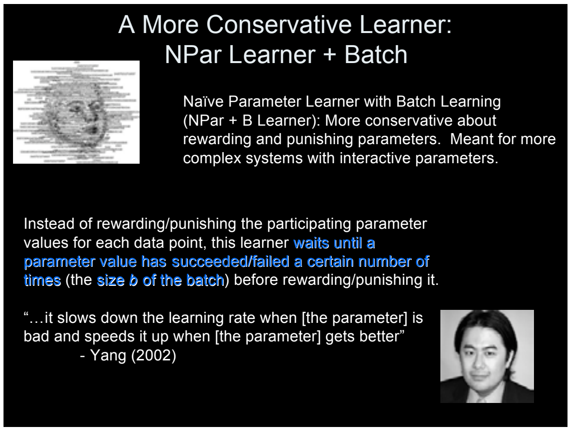
The Bayesian Learner on English Metrical Phonology

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

Results using distributions in English child-directed speech:
Learners *never* converge on English.

If learners ignore monosyllabic words (since such words don't have a stress *contour* per se), learners *still* never converge on English.

Examples of incorrect target languages Bayesian learners converged on:
Ft Hd Left, Unb, QI, Em-None, Ft Dir Left
QS, Em-Some, Em-Right, QSVCH, Ft Hd Left, Ft Dir Rt, Unb
Em-Some, Em-Right, Unb, Ft Hd Left, QS, QSVCL, Ft Dir Rt
QI, Unb, Ft Hd Left, Em-None, Ft Dir Left
Bounded, B-Syl, QI, Ft Hd Left, Em-None, Ft Dir Left, B-2

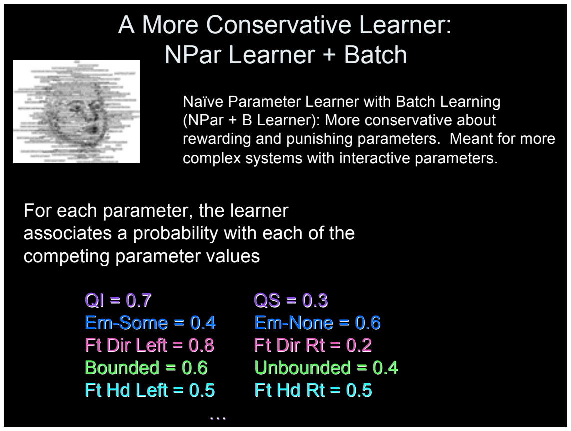


A More Conservative Learner: NPar Learner + Batch

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

Instead of rewarding/punishing the participating parameter values for each data point, this learner waits until a parameter value has succeeded/failed a certain number of times (the size *b* of the batch) before rewarding/punishing it.

"…it slows down the learning rate when [the parameter] is bad and speeds it up when [the parameter] gets better"
- Yang (2002)



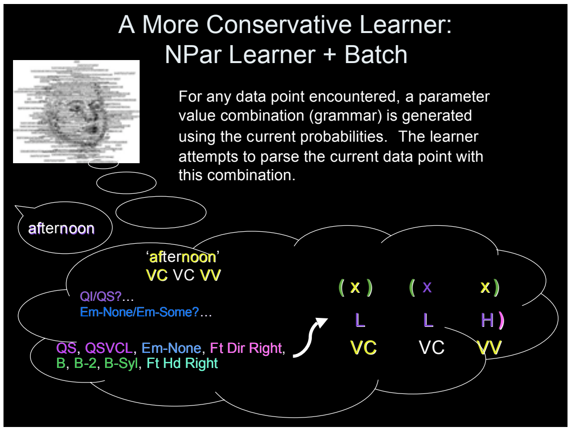
A More Conservative Learner: NPar Learner + Batch

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

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

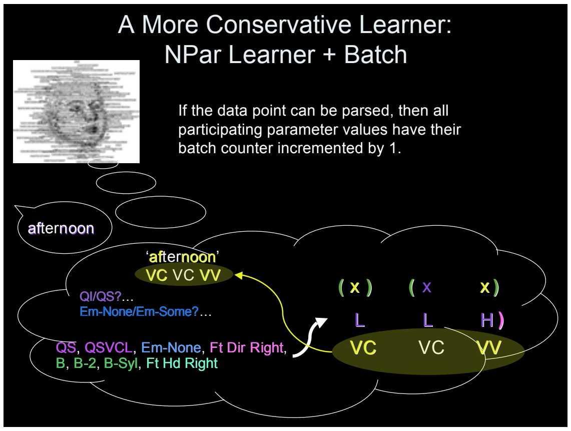
QI = 0.7	QS = 0.3
Em-Some = 0.4	Em-None = 0.6
Ft Dir Left = 0.8	Ft Dir Rt = 0.2
Bounded = 0.6	Unbounded = 0.4
Ft Hd Left = 0.5	Ft Hd Rt = 0.5

…



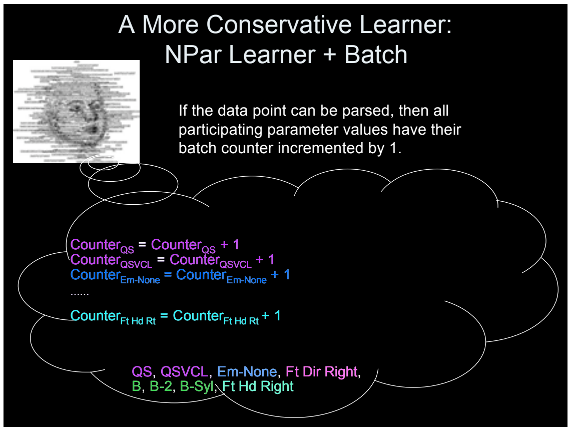
A More Conservative Learner: NPar Learner + Batch

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



A More Conservative Learner: NPar Learner + Batch

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

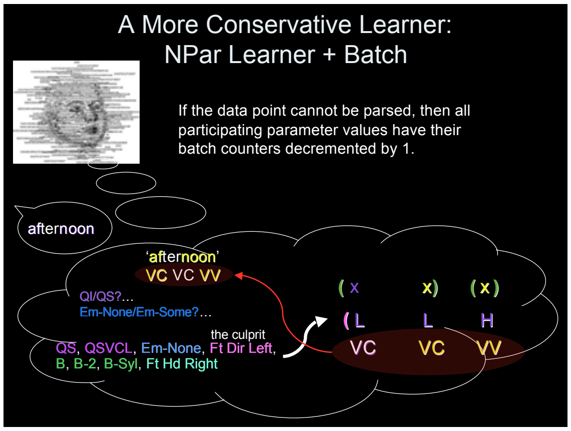


A More Conservative Learner: NPar Learner + Batch

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

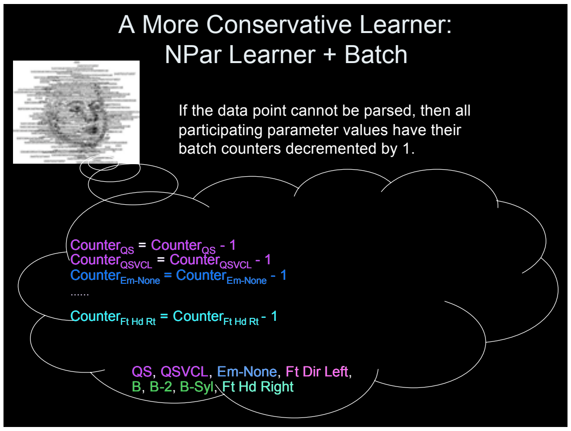
$\text{Counter}_{QS} = \text{Counter}_{QS} + 1$
$\text{Counter}_{QSVCL} = \text{Counter}_{QSVCL} + 1$
$\text{Counter}_{\text{Em-None}} = \text{Counter}_{\text{Em-None}} + 1$
......
$\text{Counter}_{\text{Ft Hd Rt}} = \text{Counter}_{\text{Ft Hd Rt}} + 1$

QS, QSVCL, Em-None, Ft Dir Right,
B, B-2, B-Syl, Ft Hd Right



A More Conservative Learner: NPar Learner + Batch

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

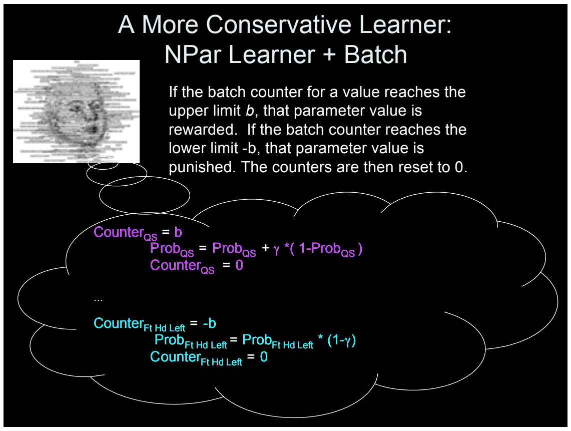


A More Conservative Learner: NPar Learner + Batch

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

$\text{Counter}_{QS} = \text{Counter}_{QS} - 1$
$\text{Counter}_{QSVCL} = \text{Counter}_{QSVCL} - 1$
$\text{Counter}_{\text{Em-None}} = \text{Counter}_{\text{Em-None}} - 1$
......
$\text{Counter}_{\text{Ft Hd Rt}} = \text{Counter}_{\text{Ft Hd Rt}} - 1$

QS, QSVCL, Em-None, Ft Dir Left,
B, B-2, B-Syl, Ft Hd Right



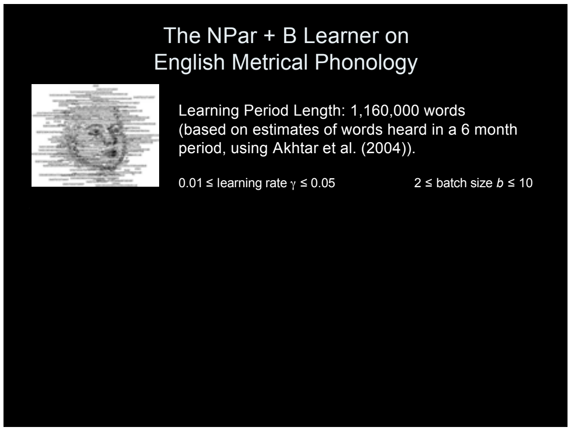
A More Conservative Learner: NPar Learner + Batch

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

$\text{Counter}_{QS} = b$
$\text{Prob}_{QS} = \text{Prob}_{QS} + \gamma *(1-\text{Prob}_{QS})$
$\text{Counter}_{QS} = 0$

...

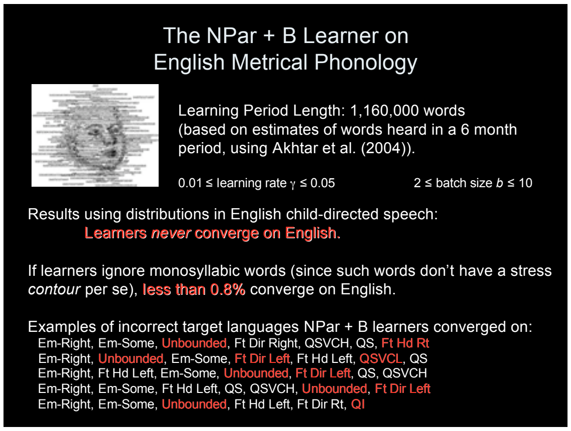
$\text{Counter}_{\text{Ft Hd Left}} = -b$
$\text{Prob}_{\text{Ft Hd Left}} = \text{Prob}_{\text{Ft Hd Left}} * (1-\gamma)$
$\text{Counter}_{\text{Ft Hd Left}} = 0$



The NPar + B Learner on English Metrical Phonology

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

$0.01 \le$ learning rate $\gamma \le 0.05$ $2 \le$ batch size $b \le 10$



The NPar + B Learner on English Metrical Phonology

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

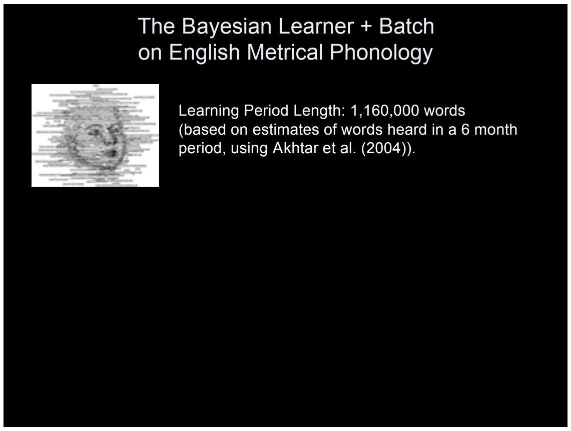
0.01 ≤ learning rate γ ≤ 0.05   2 ≤ batch size b ≤ 10

Results using distributions in English child-directed speech:

Learners *never* converge on English.

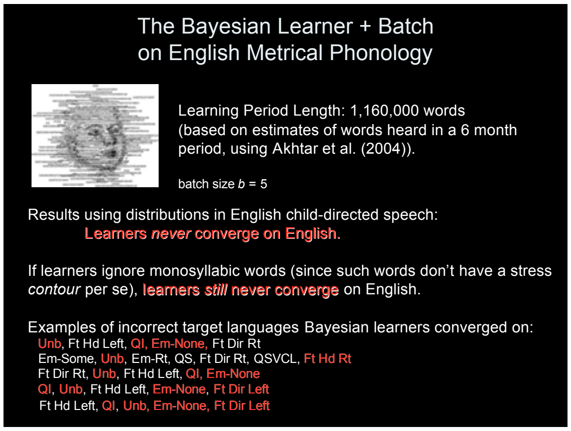
If learners ignore monosyllabic words (since such words don't have a stress *contour* per se), less than 0.8% converge on English.

Examples of incorrect target languages NPar + B learners converged on:
Em-Right, Em-Some, Unbounded, Ft Dir Right, QSVCH, QS, Ft Hd Rt
Em-Right, Unbounded, Em-Some, Ft Dir Left, Ft Hd Left, QSVCL, QS
Em-Right, Ft Hd Left, Em-Some, Unbounded, Ft Dir Left, QS, QSVCH
Em-Right, Em-Some, Ft Hd Left, QS, QSVCH, Unbounded, Ft Dir Left
Em-Right, Em-Some, Unbounded, Ft Hd Left, Ft Dir Rt, QI



The Bayesian Learner + Batch on English Metrical Phonology

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



The Bayesian Learner + Batch on English Metrical Phonology

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

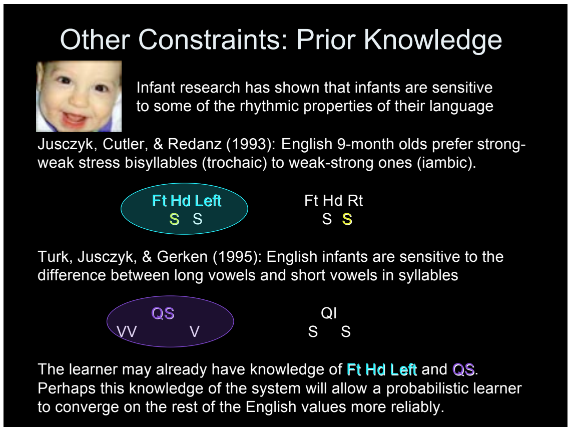
batch size b = 5

Results using distributions in English child-directed speech:

Learners *never* converge on English.

If learners ignore monosyllabic words (since such words don't have a stress *contour* per se), learners *still* never converge on English.

Examples of incorrect target languages Bayesian learners converged on:
Unb, Ft Hd Left, QI, Em-None, Ft Dir Rt
Em-Some, Unb, Em-Rt, QS, Ft Dir Rt, QSVCL, Ft Hd Rt
Ft Dir Rt, Unb, Ft Hd Left, QI, Em-None
QI, Unb, Ft Hd Left, Em-None, Ft Dir Left
Ft Hd Left, QI, Unb, Em-None, Ft Dir Left



Other Constraints: Prior Knowledge

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

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

Ft Hd Left
S S

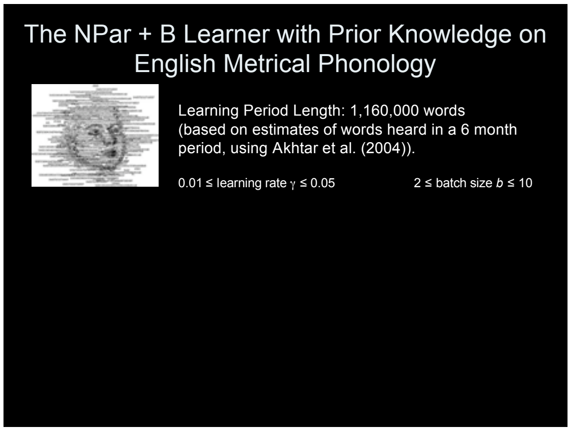
Ft Hd Rt
S S

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

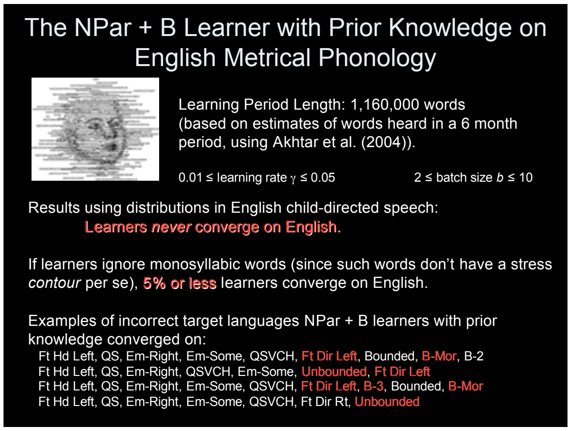
The learner may already have knowledge of Ft Hd Left and QS. Perhaps this knowledge of the system will allow a probabilistic learner to converge on the rest of the English values more reliably.



The NPar + B Learner with Prior Knowledge on English Metrical Phonology

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

0.01 ≤ learning rate γ ≤ 0.05   2 ≤ batch size b ≤ 10



The NPar + B Learner with Prior Knowledge on English Metrical Phonology

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

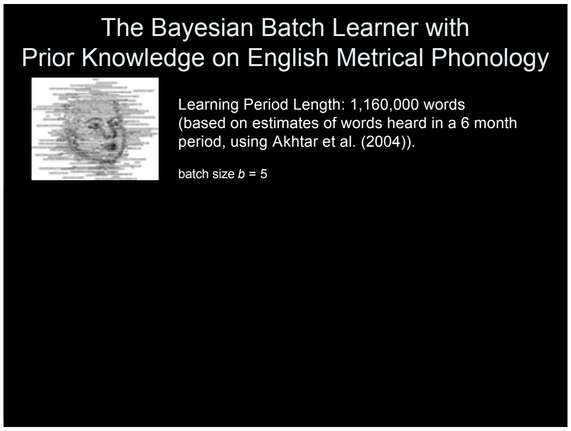
0.01 ≤ learning rate γ ≤ 0.05   2 ≤ batch size b ≤ 10

Results using distributions in English child-directed speech:

Learners *never* converge on English.

If learners ignore monosyllabic words (since such words don't have a stress *contour* per se), 5% or less learners converge on English.

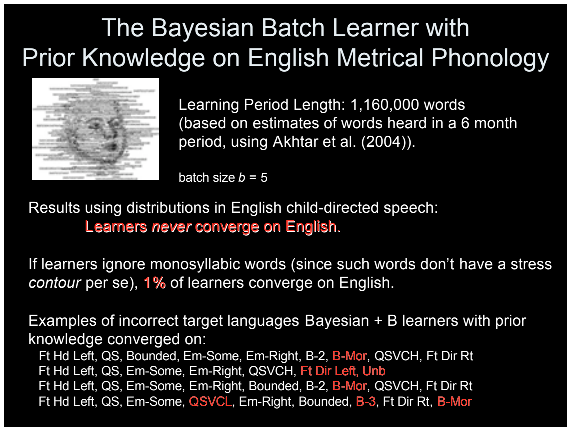
Examples of incorrect target languages NPar + B learners with prior knowledge converged on:
Ft Hd Left, QS, Em-Right, Em-Some, QSVCH, Ft Dir Left, Bounded, B-Mor, B-2
Ft Hd Left, QS, Em-Right, QSVCH, Em-Some, Unbounded, Ft Dir Left
Ft Hd Left, QS, Em-Right, Em-Some, QSVCH, Ft Dir Left, B-3, Bounded, B-Mor
Ft Hd Left, QS, Em-Right, Em-Some, QSVCH, Ft Dir Rt, Unbounded



The Bayesian Batch Learner with Prior Knowledge on English Metrical Phonology

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

batch size *b* = 5



The Bayesian Batch Learner with Prior Knowledge on English Metrical Phonology

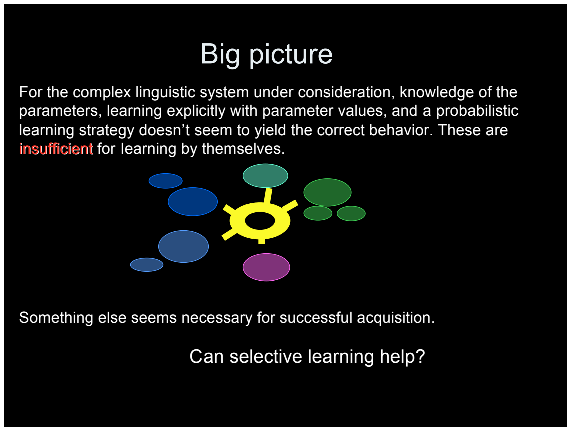
Learning Period Length: 1,160,000 words (based on estimates of words heard in a 6 month period, using Akhtar et al. (2004)).

batch size *b* = 5

Results using distributions in English child-directed speech:
Learners *never* converge on English.

If learners ignore monosyllabic words (since such words don't have a stress *contour* per se), 1% of learners converge on English.

Examples of incorrect target languages Bayesian + B learners with prior knowledge converged on:
Ft Hd Left, QS, Bounded, Em-Some, Em-Right, B-2, B-Mor, QSVCH, Ft Dir Rt
Ft Hd Left, QS, Em-Some, Em-Right, QSVCH, Ft Dir Left, Unb
Ft Hd Left, QS, Em-Some, Em-Right, Bounded, B-2, B-Mor, QSVCH, Ft Dir Rt
Ft Hd Left, QS, Em-Some, QSVCL, Em-Right, Bounded, B-3, Ft Dir Rt, B-Mor

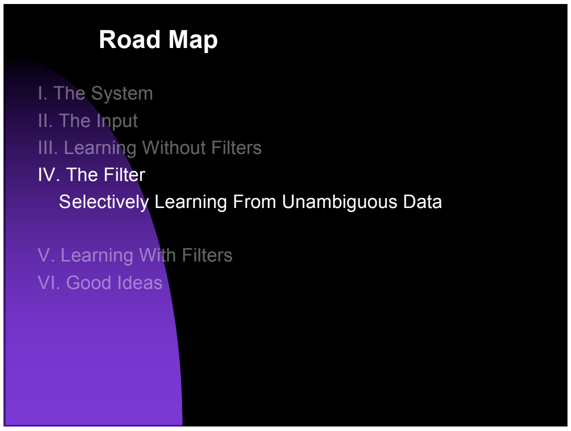


Big picture

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

Something else seems necessary for successful acquisition.

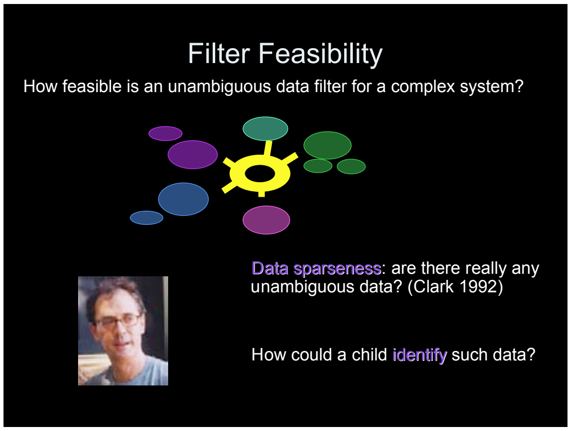
Can selective learning help?



Road Map

I. The System
II. The Input
III. Learning Without Filters
IV. The Filter
Selectively Learning From Unambiguous Data

V. Learning With Filters
VI. Good Ideas

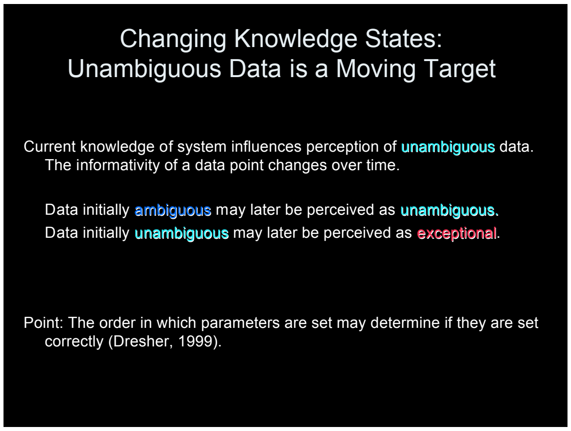


Filter Feasibility

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

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

How could a child identify such data?

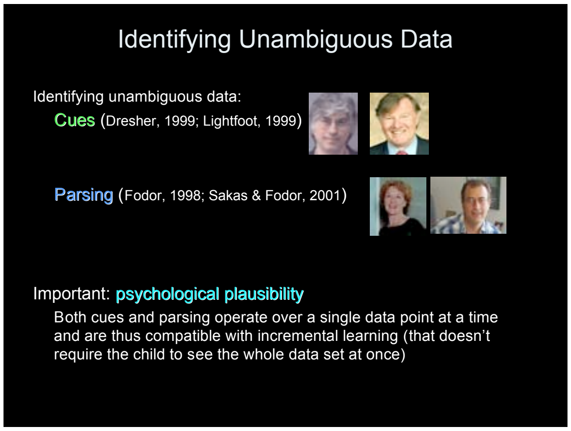


Changing Knowledge States: Unambiguous Data is a Moving Target

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

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

Point: The order in which parameters are set may determine if they are set correctly (Dresher, 1999).



Identifying Unambiguous Data

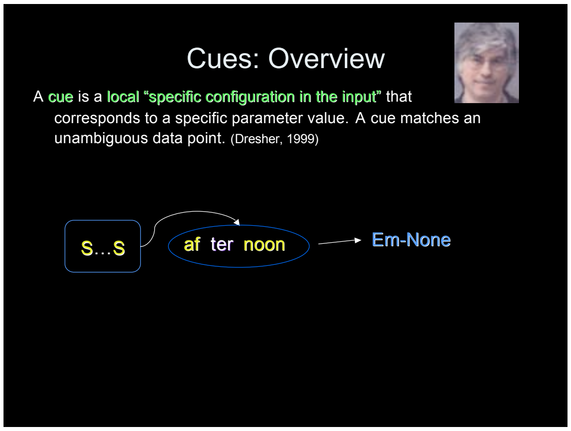
Identifying unambiguous data:

Cues (Dresher, 1999; Lightfoot, 1999)

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

Important: psychological plausibility

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



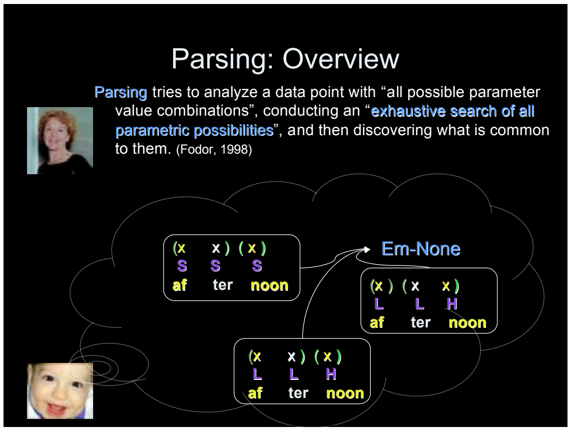
Cues: Overview

A cue is a local "specific configuration in the input" that corresponds to a specific parameter value. A cue matches an unambiguous data point. (Dresher, 1999)

Cues for Metrical Phonology Parameters

Recall: Cues match local surface structure (sample cues below)

QS: 2 syllable word with 2 stresses	VV VV
Em-Right: Rightmost syllable is Heavy and unstressed	… H
Unb: 3+ unstressed S/L syllables in a row	…S S S… … L L L L
Ft Hd Left: Leftmost foot has stress on leftmost syllable	S S S… H L L …



Parsing: Overview

Parsing tries to analyze a data point with "all possible parameter value combinations", conducting an "exhaustive search of all parametric possibilities", and then discovering what is common to them. (Fodor, 1998)

Parsing with Metrical Phonology Parameters

Sample Datum: VC VC VV ('afternoon')

(QS, QSVCL, Em-None, Ft Dir Right, B, B-2, B-Syl, Ft Hd Right)

(x) (x x)
L L H)
VC VC VV

(QS, QSVCL, Em-None, Ft Dir Left, B, B-2, B-Syl, Ft Hd Left)

(x x) (x)
(L L H
VC VC VV

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

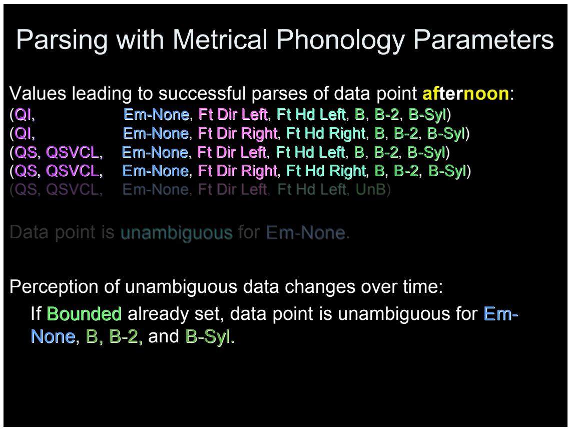
(x x) (x)
S S S)
VC VC VV

Parsing with Metrical Phonology Parameters

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

(QI, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QI, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, UnB)

Data point is unambiguous for Em-None.



Parsing with Metrical Phonology Parameters

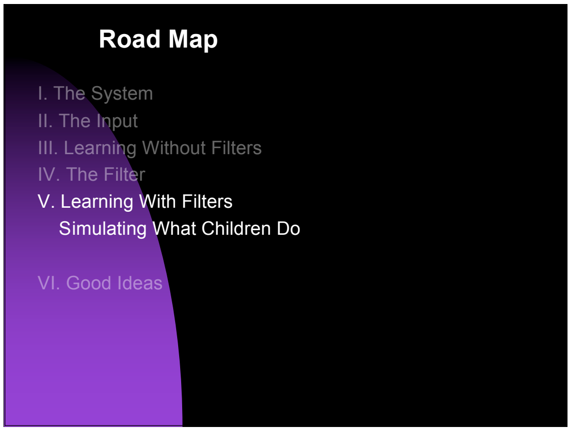
Values leading to successful parses of data point afternoon:

(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)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, UnB)

Data point is unambiguous for Em-None.

Perception of unambiguous data changes over time:

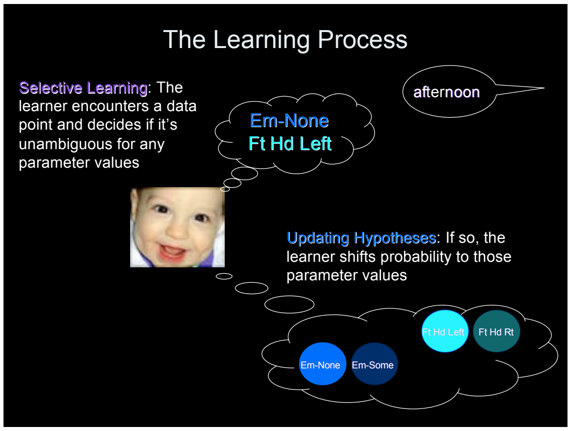
If Bounded already set, data point is unambiguous for Em-None, B, B-2, and B-Syl.



Road Map

I. The System
II. The Input
III. Learning Without Filters
IV. The Filter
V. Learning With Filters
Simulating What Children Do

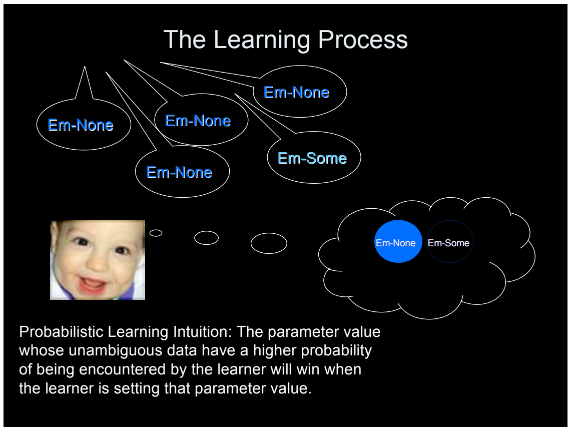
VI. Good Ideas



The Learning Process

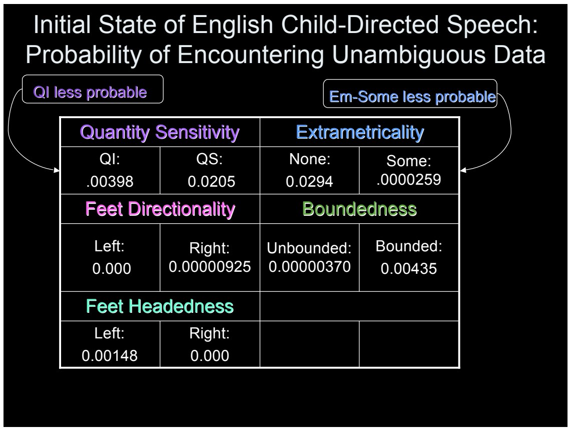
Selective Learning: The learner encounters a data point and decides if it's unambiguous for any parameter values

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



The Learning Process

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

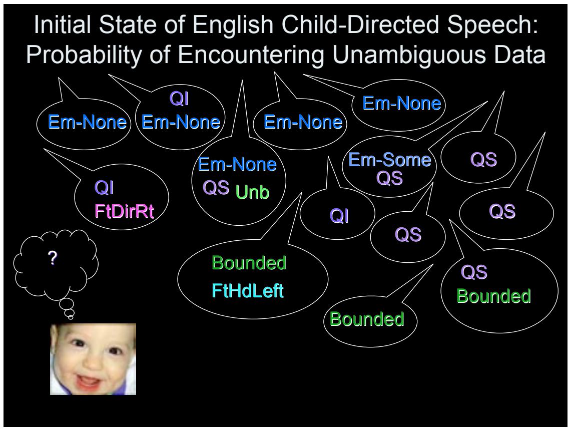


Initial State of English Child-Directed Speech: Probability of Encountering Unambiguous Data

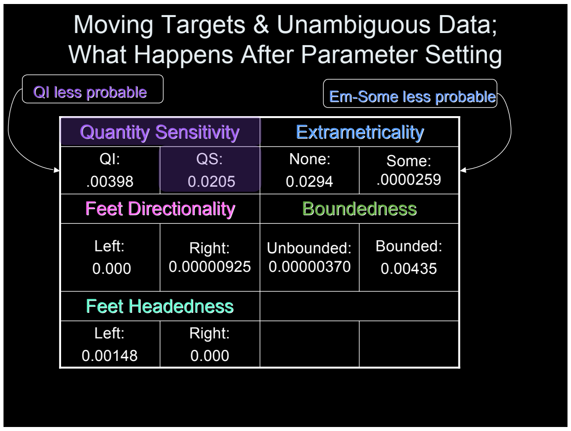
QI less probable

Em-Some less probable

Quantity Sensitivity		Extrametricality	
QI: .00398	QS: 0.0205	None: 0.0294	Some: .0000259
Feet Directionality		Boundedness	
Left: 0.000	Right: 0.00000925	Unbounded: 0.00000370	Bounded: 0.00435
Feet Headedness			
Left: 0.00148	Right: 0.000		



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

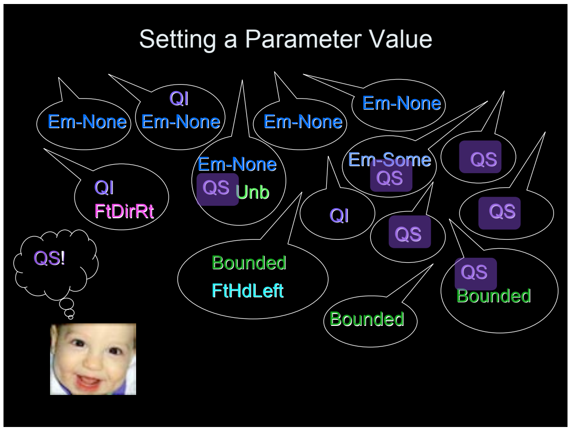


Moving Targets & Unambiguous Data; What Happens After Parameter Setting

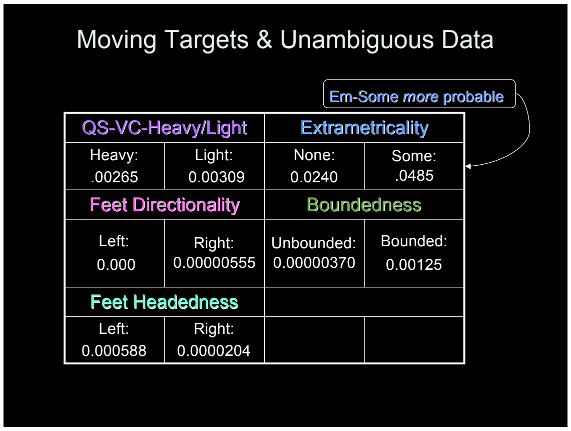
QI less probable

Em-Some less probable

Quantity Sensitivity		Extrametricality	
QI: .00398	QS: 0.0205	None: 0.0294	Some: .0000259
Feet Directionality		Boundedness	
Left: 0.000	Right: 0.00000925	Unbounded: 0.00000370	Bounded: 0.00435
Feet Headedness			
Left: 0.00148	Right: 0.000		



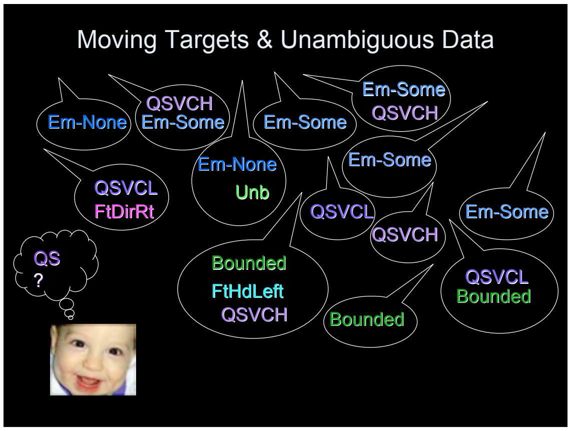
Setting a Parameter Value



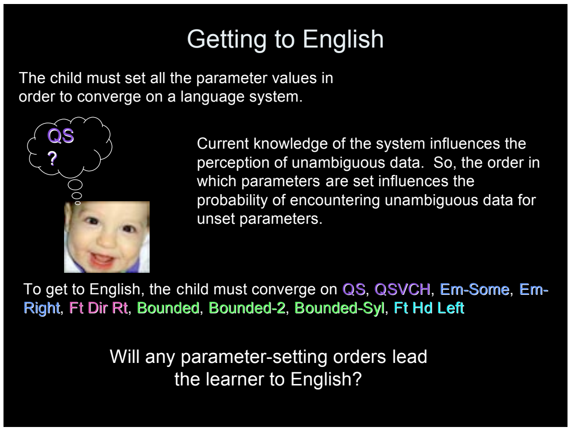
Moving Targets & Unambiguous Data

Em-Some *more* probable

QS-VC-Heavy/Light		Extrametricality	
Heavy: .00265	Light: 0.00309	None: 0.0240	Some: .0485
Feet Directionality		Boundedness	
Left: 0.000	Right: 0.00000555	Unbounded: 0.00000370	Bounded: 0.00125
Feet Headedness			
Left: 0.000588	Right: 0.0000204		



Moving Targets & Unambiguous Data



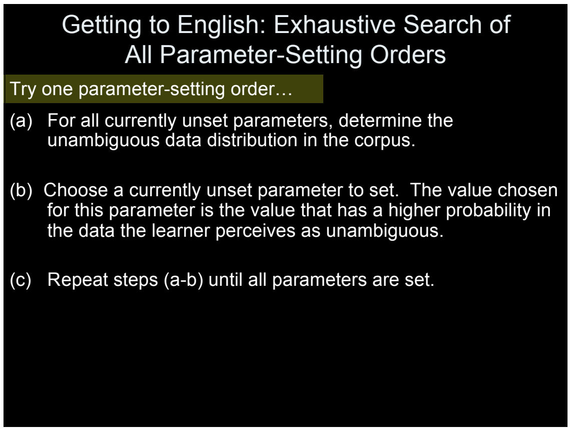
Getting to English

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

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

To get to English, the child must converge on QS, QSVCH, Em-Some, Em-Right, Ft Dir Rt, Bounded, Bounded-2, Bounded-Syl, Ft Hd Left

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



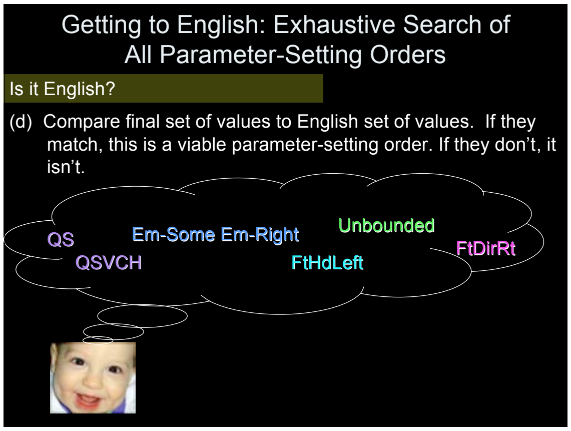
Getting to English: Exhaustive Search of All Parameter-Setting Orders

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.

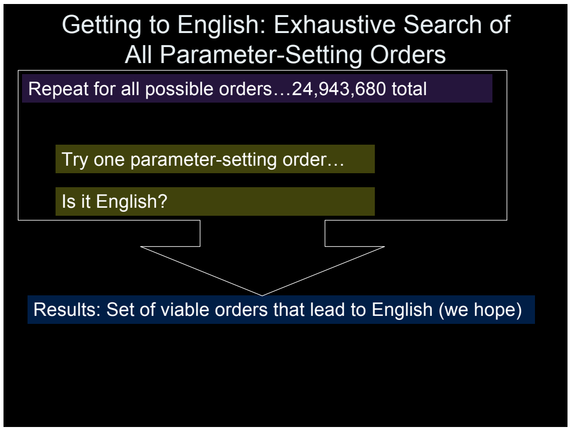


Getting to English: Exhaustive Search of All Parameter-Setting Orders

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.

QS
Em-Some Em-Right
Unbounded
FtDirRt
QSVCH
FtHdLeft



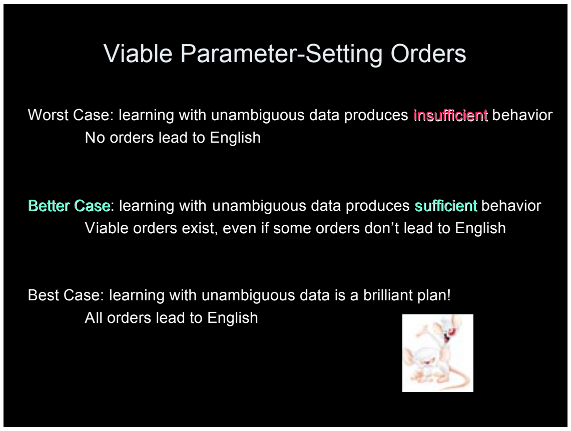
Getting to English: Exhaustive Search of All Parameter-Setting Orders

Repeat for all possible orders…24,943,680 total

Try one parameter-setting order…

Is it English?

Results: Set of viable orders that lead to English (we hope)

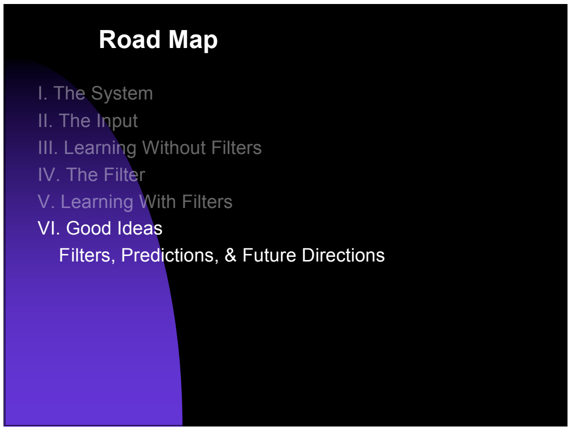


Viable Parameter-Setting Orders

Worst Case: learning with unambiguous data produces insufficient behavior
No orders lead to English

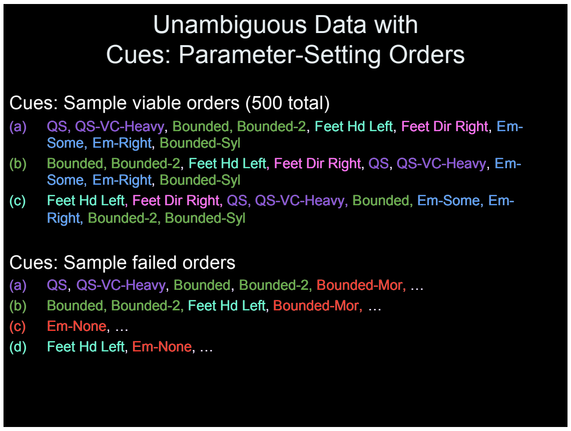
Better Case: learning with unambiguous data produces sufficient behavior
Viable orders exist, even if some orders don't lead to English

Best Case: learning with unambiguous data is a brilliant plan!
All orders lead to English



Road Map

I. The System
II. The Input
III. Learning Without Filters
IV. The Filter
V. Learning With Filters
VI. Good Ideas
Filters, Predictions, & Future Directions



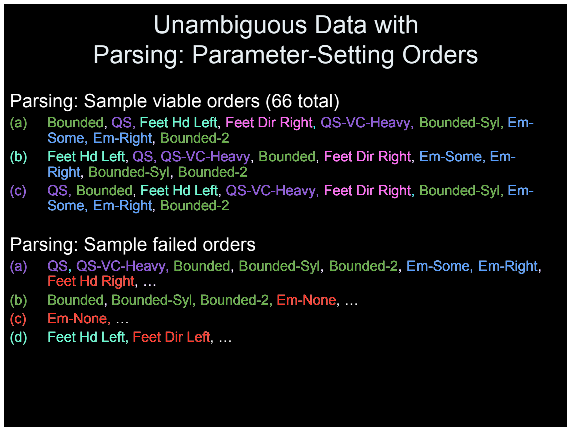
Unambiguous Data with Cues: Parameter-Setting Orders

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-Syl
(b) Bounded, Bounded-2, Feet Hd Left, Feet Dir Right, QS, QS-VC-Heavy, Em-Some, Em-Right, Bounded-Syl
(c) Feet Hd Left, Feet Dir Right, QS, QS-VC-Heavy, Bounded, Em-Some, Em-Right, Bounded-2, Bounded-Syl

Cues: Sample failed orders

(a) QS, QS-VC-Heavy, Bounded, Bounded-2, Bounded-Mor, …
(b) Bounded, Bounded-2, Feet Hd Left, Bounded-Mor, …
(c) Em-None, …
(d) Feet Hd Left, Em-None, …



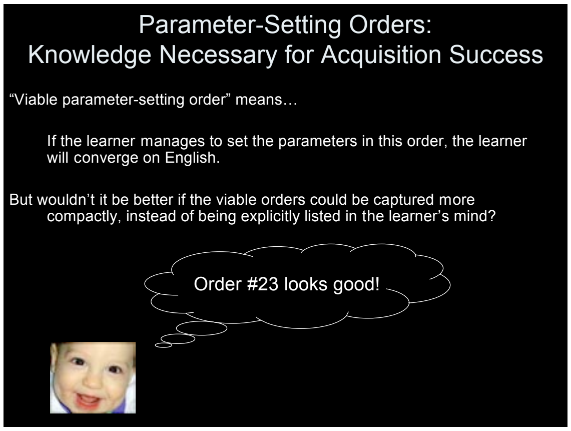
Unambiguous Data with Parsing: Parameter-Setting Orders

Parsing: Sample viable orders (66 total)

(a) Bounded, QS, Feet Hd Left, Feet Dir Right, QS-VC-Heavy, Bounded-Syl, 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, …

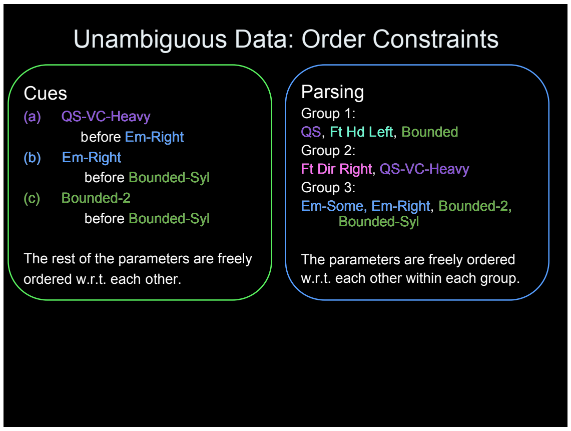


Parameter-Setting Orders: Knowledge Necessary for Acquisition Success

"Viable parameter-setting order" means…

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

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



Unambiguous Data: Order Constraints

Cues

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

(b) Em-Right
before Bounded-Syl

(c) Bounded-2
before Bounded-Syl

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

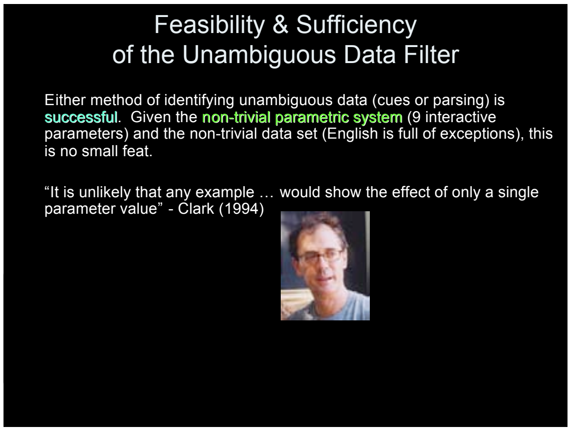
Parsing

Group 1:
QS, Ft Hd Left, Bounded

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

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

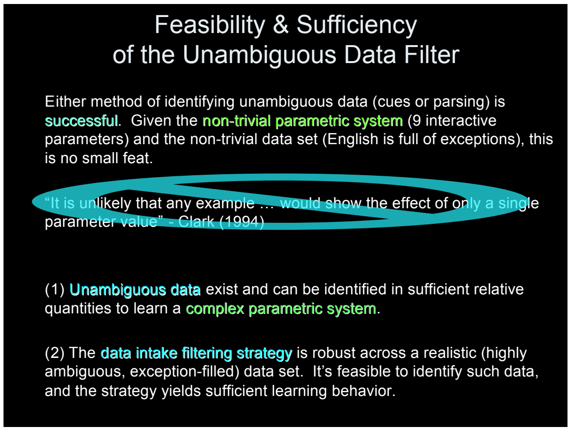
The parameters are freely ordered w.r.t. each other within each group.



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.

"It is unlikely that any example … would show the effect of only a single parameter value" - Clark (1994)



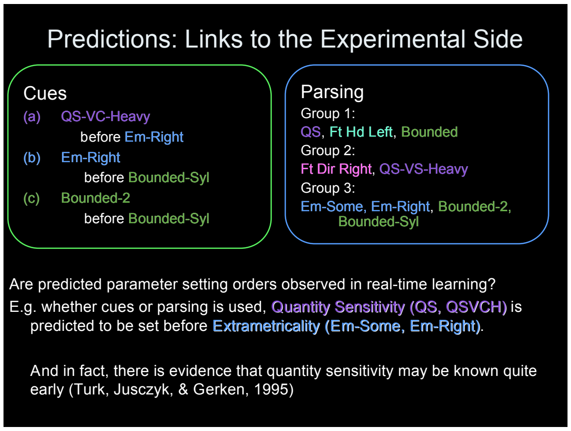
Feasibility & Sufficiency of the Unambiguous Data Filter

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

"It is unlikely that any example … would show the effect of only a single parameter value" - Clark (1994)

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

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



Predictions: Links to the Experimental Side

Cues

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

(b) Em-Right
before Bounded-Syl

(c) Bounded-2
before Bounded-Syl

Parsing

Group 1:
QS, Ft Hd Left, Bounded

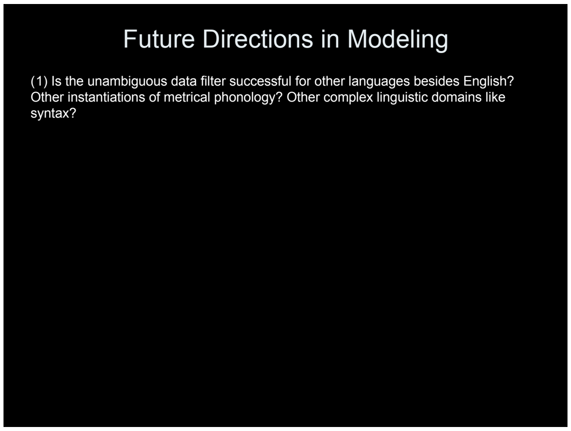
Group 2:
Ft Dir Right, QS-VS-Heavy

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

Are predicted parameter setting orders observed in real-time learning?

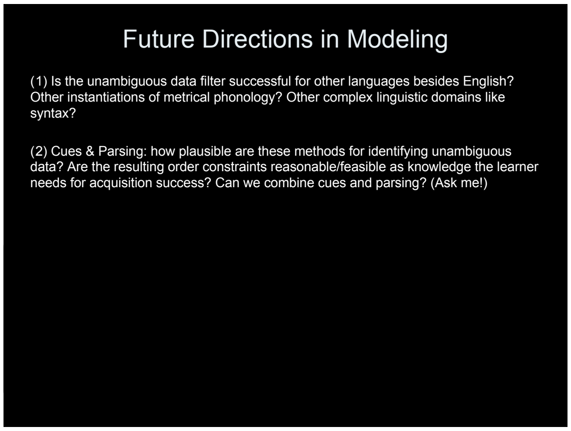
E.g. whether cues or parsing is used, Quantity Sensitivity (QS, QSVCH) is predicted to be set before Extrametricality (Em-Some, Em-Right).

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



Future Directions in Modeling

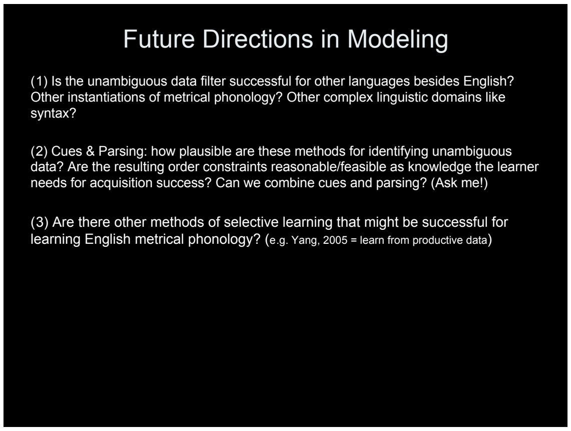
(1) Is the unambiguous data filter successful for other languages besides English? Other instantiations of metrical phonology? Other complex linguistic domains like syntax?



Future Directions in Modeling

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

(2) Cues & Parsing: how plausible are these methods for identifying unambiguous data? Are the resulting order constraints reasonable/feasible as knowledge the learner needs for acquisition success? Can we combine cues and parsing? (Ask me!)

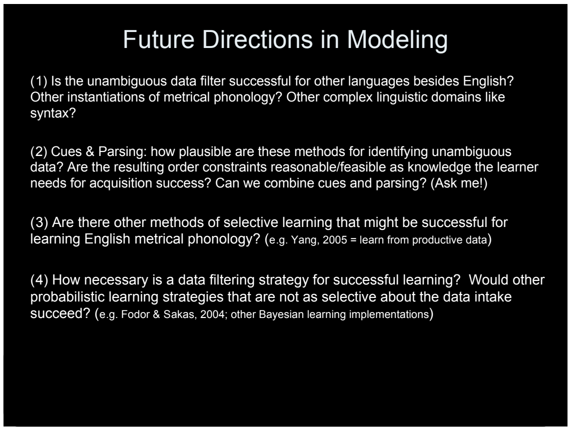


Future Directions in Modeling

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

(2) Cues & Parsing: how plausible are these methods for identifying unambiguous data? Are the resulting order constraints reasonable/feasible as knowledge the learner needs for acquisition success? Can we combine cues and parsing? (Ask me!)

(3) Are there other methods of selective learning that might be successful for learning English metrical phonology? (e.g. Yang, 2005 = learn from productive data)



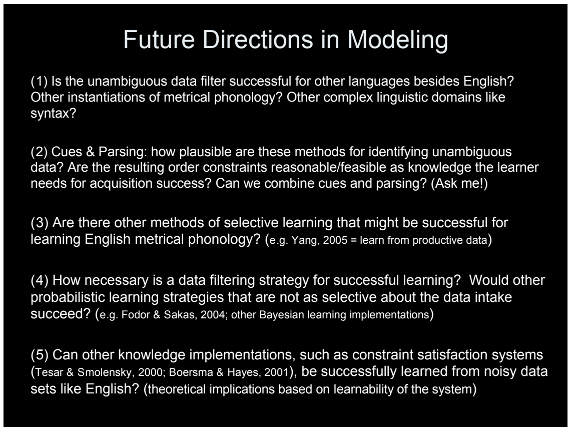
Future Directions in Modeling

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

(2) Cues & Parsing: how plausible are these methods for identifying unambiguous data? Are the resulting order constraints reasonable/feasible as knowledge the learner needs for acquisition success? Can we combine cues and parsing? (Ask me!)

(3) Are there other methods of selective learning that might be successful for learning English metrical phonology? (e.g. Yang, 2005 = learn from productive data)

(4) How necessary is a data filtering strategy for successful learning? Would other probabilistic learning strategies that are not as selective about the data intake succeed? (e.g. Fodor & Sakas, 2004; other Bayesian learning implementations)



Future Directions in Modeling

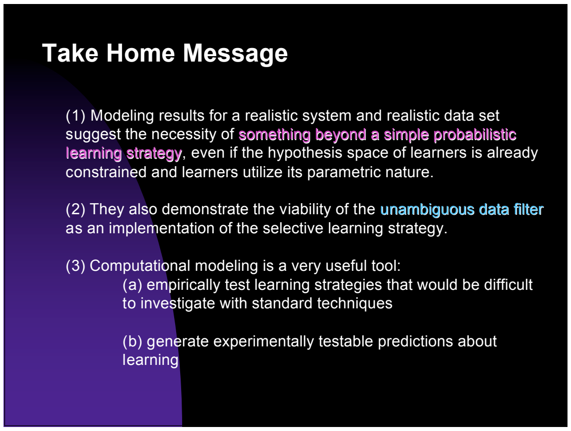
(1) Is the unambiguous data filter successful for other languages besides English? Other instantiations of metrical phonology? Other complex linguistic domains like syntax?

(2) Cues & Parsing: how plausible are these methods for identifying unambiguous data? Are the resulting order constraints reasonable/feasible as knowledge the learner needs for acquisition success? Can we combine cues and parsing? (Ask me!)

(3) Are there other methods of selective learning that might be successful for learning English metrical phonology? (e.g. Yang, 2005 = learn from productive data)

(4) How necessary is a data filtering strategy for successful learning? Would other probabilistic learning strategies that are not as selective about the data intake succeed? (e.g. Fodor & Sakas, 2004; other Bayesian learning implementations)

(5) Can other knowledge implementations, such as constraint satisfaction systems (Tesar & Smolensky, 2000; Boersma & Hayes, 2001), be successfully learned from noisy data sets like English? (theoretical implications based on learnability of the system)



Take Home Message

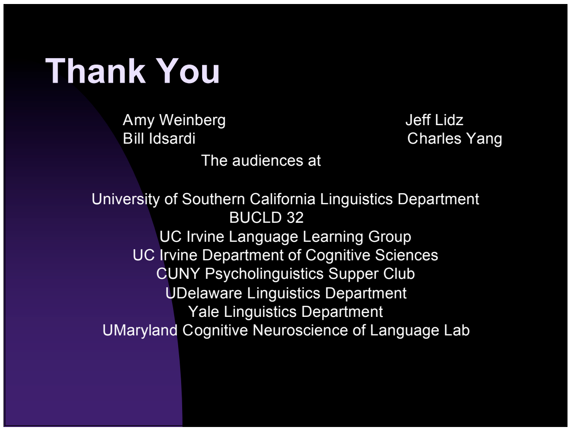
(1) Modeling results for a realistic system and realistic data set suggest the necessity of something beyond a simple probabilistic learning strategy, even if the hypothesis space of learners is already constrained and learners utilize its parametric nature.

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

(3) Computational modeling is a very useful tool:

(a) empirically test learning strategies that would be difficult to investigate with standard techniques

(b) generate experimentally testable predictions about learning



Thank You

Amy Weinberg	Jeff Lidz
Bill Idsardi	Charles Yang

The audiences at

University of Southern California Linguistics Department
BUCLD 32
UC Irvine Language Learning Group
UC Irvine Department of Cognitive Sciences
CUNY Psycholinguistics Supper 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):

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.

Why Parameters?

Why posit parameters instead of just associating stress contours with words?

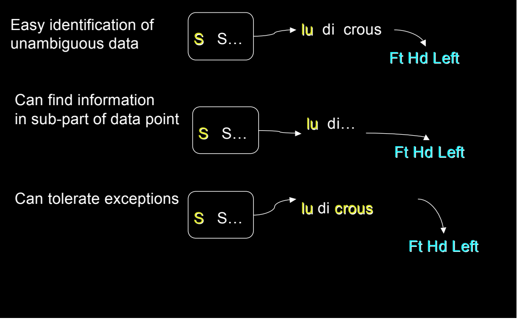
Arguments from stress change over time (Dresher & Lahiri, 2003):

(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

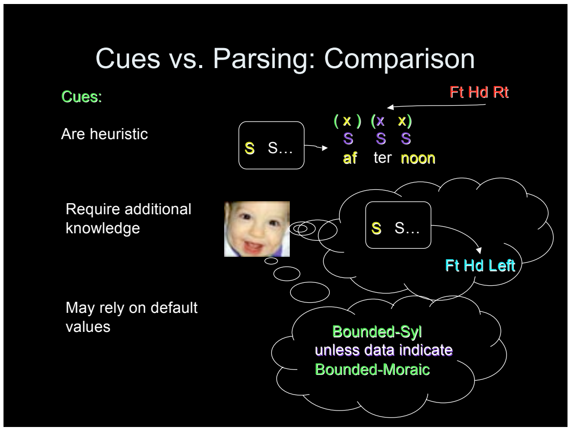
Cues:



Easy identification of unambiguous data

Can find information in sub-part of data point

Can tolerate exceptions



Cues vs. Parsing: Comparison

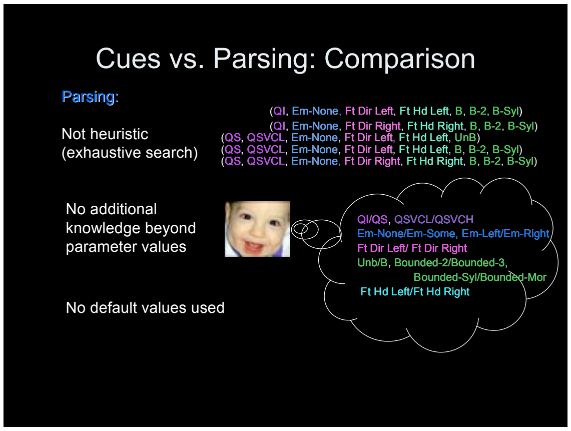
Cues:

Are heuristic

Require additional knowledge

May rely on default values

Ft Hd Rt
S S…
(x) (x x)
S S S
af ter noon
S S…
Ft Hd Left
Bounded-Syl
unless data indicate
Bounded-Moraic



Cues vs. Parsing: Comparison

Parsing:

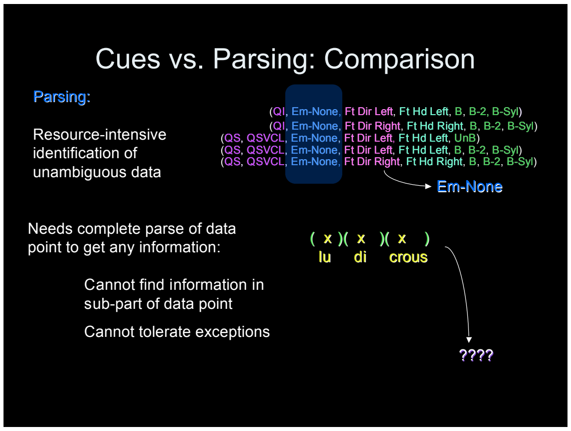
Not heuristic (exhaustive search)

(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)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, UnB)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)

No additional knowledge beyond parameter values

No default values used

QI/QS, QSVCL/QSVCH
Em-None/Em-Some, Em-Left/Em-Right
Ft Dir Left/ Ft Dir Right
Unb/B, Bounded-2/Bounded-3,
Bounded-Syl/Bounded-Mor
Ft Hd Left/Ft Hd Right



Cues vs. Parsing: Comparison

Parsing:

Resource-intensive identification of unambiguous data

(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)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, UnB)
(QS, QSVCL, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl)
(QS, QSVCL, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl)

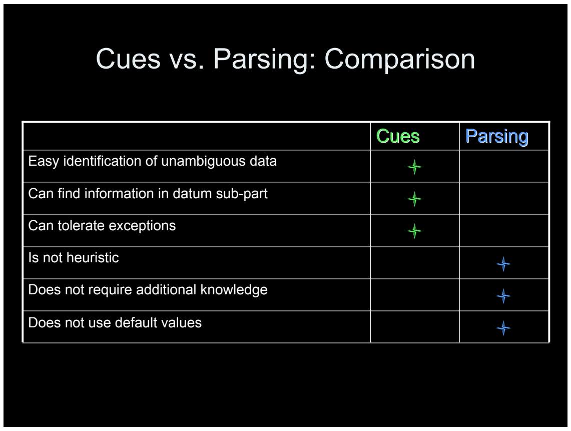
Em-None

Needs complete parse of data point to get any information:

Cannot find information in sub-part of data point

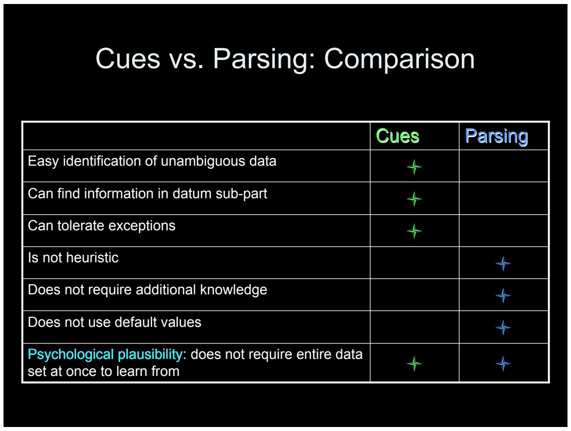
Cannot tolerate exceptions

(x)(x)(x)
lu di crous
????



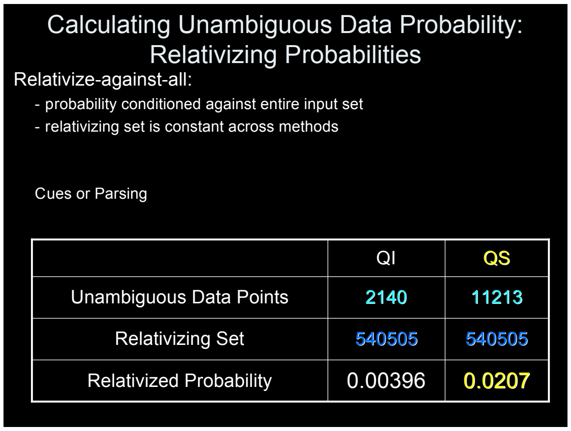
Cues vs. Parsing: Comparison

	Cues	Parsing
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		✦



Cues vs. Parsing: Comparison

	Cues	Parsing
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		✦
Psychological plausibility: does not require entire data set at once to learn from	✦	✦



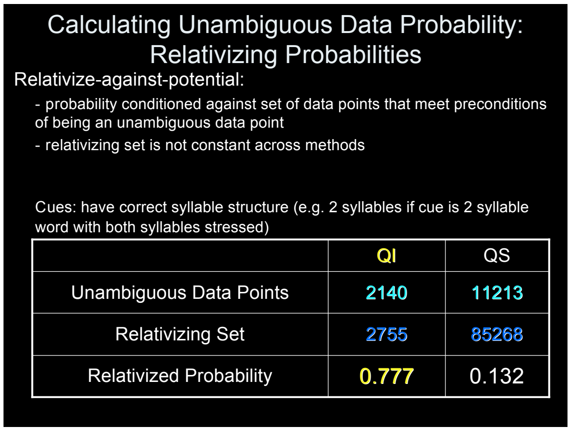
Calculating Unambiguous Data Probability: Relativizing Probabilities

Relativize-against-all:

- probability conditioned against entire input set
- relativizing set is constant across methods

Cues or Parsing

	QI	QS
Unambiguous Data Points	2140	11213
Relativizing Set	540505	540505
Relativized Probability	0.00396	0.0207



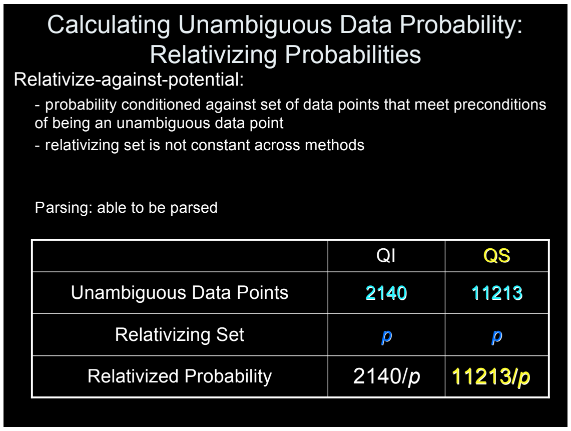
Calculating Unambiguous Data Probability: Relativizing Probabilities

Relativize-against-potential:

- probability conditioned against set of data points that meet preconditions of being an unambiguous data point
- relativizing set is not constant across methods

Cues: have correct syllable structure (e.g. 2 syllables if cue is 2 syllable word with both syllables stressed)

	QI	QS
Unambiguous Data Points	2140	11213
Relativizing Set	2755	85268
Relativized Probability	0.777	0.132



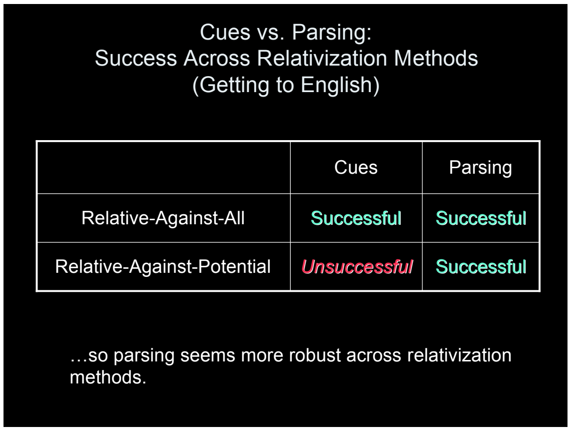
Calculating Unambiguous Data Probability: Relativizing Probabilities

Relativize-against-potential:

- probability conditioned against set of data points that meet preconditions of being an unambiguous data point
- relativizing set is not constant across methods

Parsing: able to be parsed

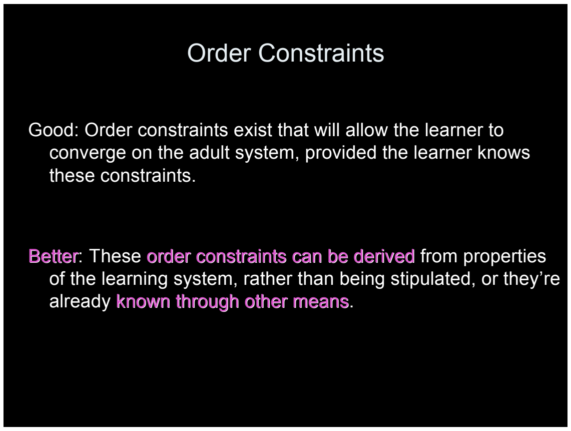
	QI	QS
Unambiguous Data Points	2140	11213
Relativizing Set	*p*	*p*
Relativized Probability	2140/*p*	11213/*p*



Cues vs. Parsing: Success Across Relativization Methods (Getting to English)

	Cues	Parsing
Relative-Against-All	Successful	Successful
Relative-Against-Potential	*Unsuccessful*	Successful

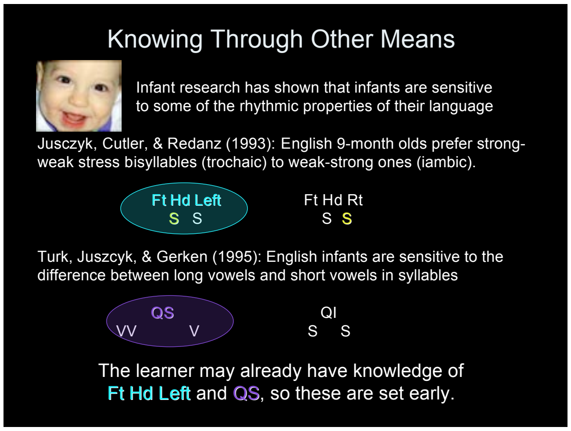
…so parsing seems more robust across relativization methods.



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): English 9-month olds prefer strong-weak stress bisyllables (trochaic) to weak-strong ones (iambic).

Ft Hd Left
S S

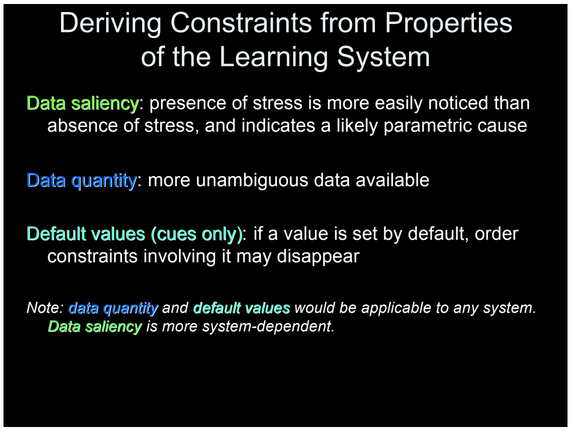
Ft Hd Rt
S S

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

QS
VV V

QI
S S

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



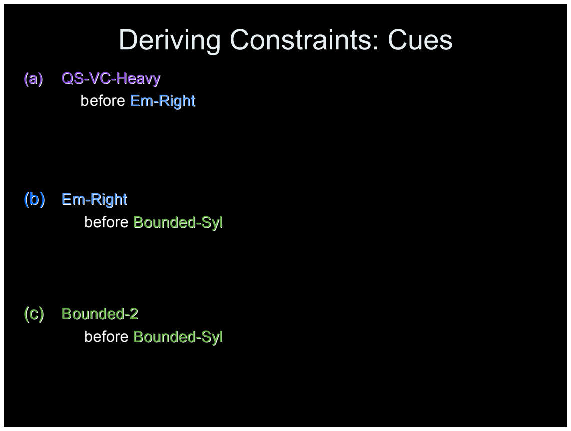
Deriving Constraints from Properties of the Learning System

Data saliency: presence of stress is more easily noticed than absence of stress, and indicates a likely parametric cause

Data quantity: more unambiguous data available

Default values (cues only): if a value is set by default, order constraints involving it may disappear

Note: data quantity and default values would be applicable to any system. Data saliency is more system-dependent.

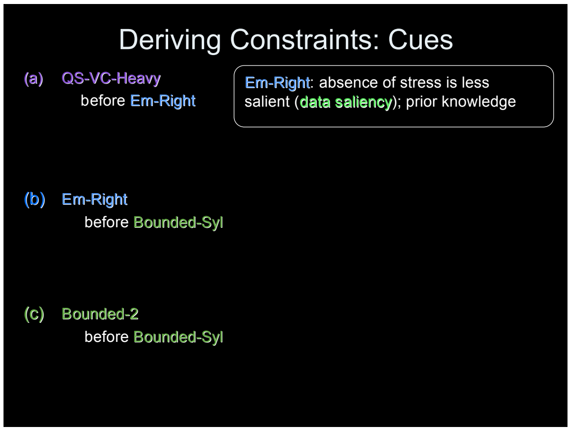


Deriving Constraints: Cues

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

(b) Em-Right
before Bounded-Syl

(c) Bounded-2
before Bounded-Syl



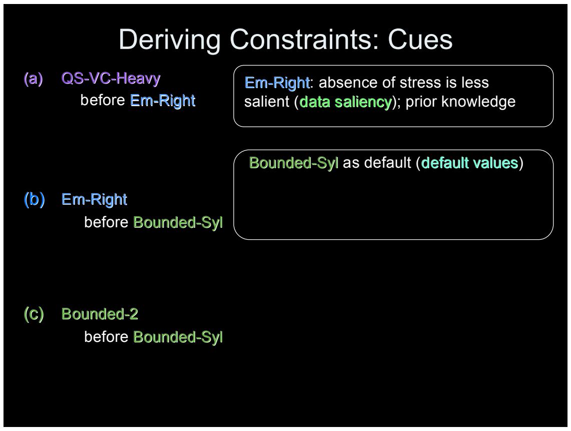
Deriving Constraints: Cues

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

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

(b) Em-Right
before Bounded-Syl

(c) Bounded-2
before Bounded-Syl



Deriving Constraints: Cues

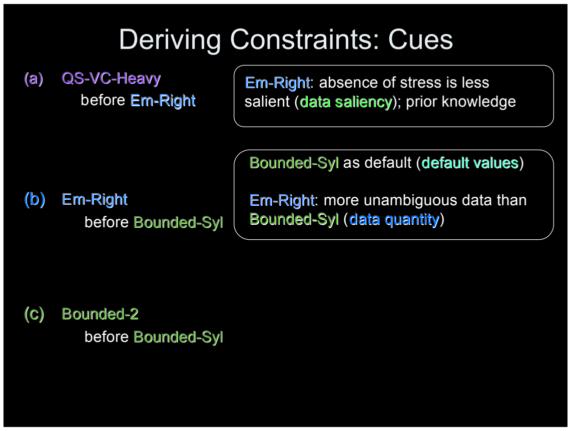
(a) QS-VC-Heavy
before Em-Right

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

(b) Em-Right
before Bounded-Syl

Bounded-Syl as default (default values)

(c) Bounded-2
before Bounded-Syl



Deriving Constraints: Cues

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

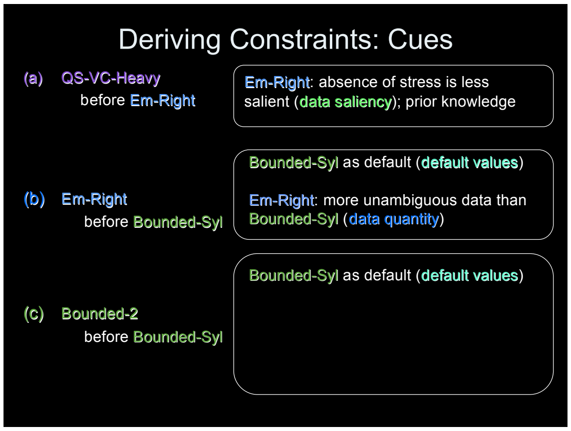
Em-Right: absence of stress is less salient (data saliency); prior knowledge

(b) Em-Right
before Bounded-Syl

Bounded-Syl as default (default values)

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

(c) Bounded-2
before Bounded-Syl



Deriving Constraints: Cues

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

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

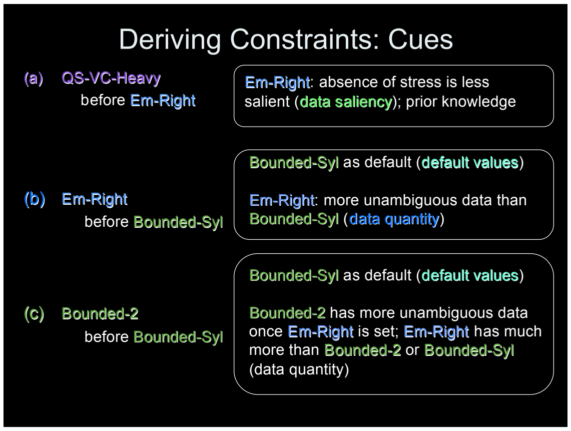
Bounded-Syl as default (default values)

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

(b) Em-Right
before Bounded-Syl

Bounded-Syl as default (default values)

(c) Bounded-2
before Bounded-Syl



Deriving Constraints: Cues

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

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

Bounded-Syl as default (default values)

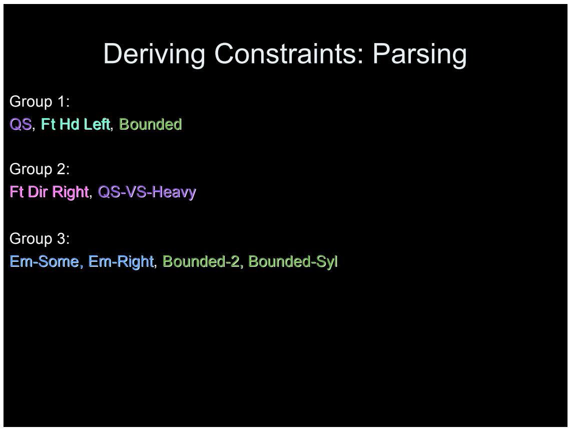
Em-Right: more unambiguous data than Bounded-Syl (data quantity)

(b) Em-Right
before Bounded-Syl

Bounded-Syl as default (default values)

Bounded-2 has more unambiguous data once Em-Right is set; Em-Right has much more than Bounded-2 or Bounded-Syl (data quantity)

(c) Bounded-2
before Bounded-Syl

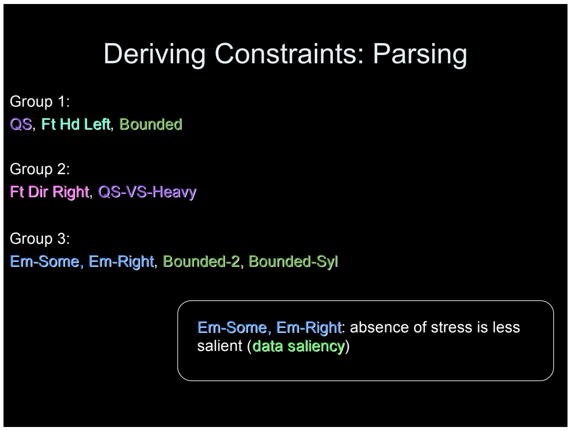


Deriving Constraints: Parsing

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



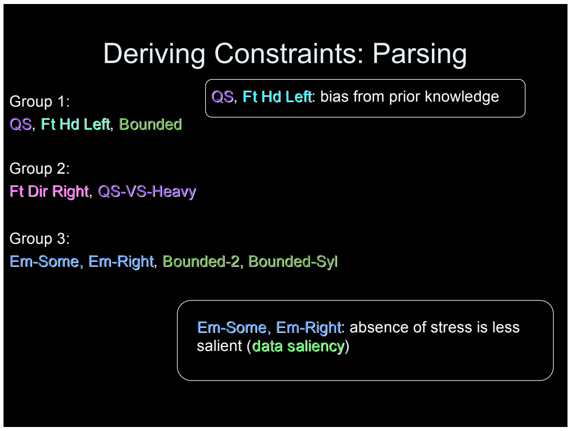
Deriving Constraints: Parsing

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)



Deriving Constraints: Parsing

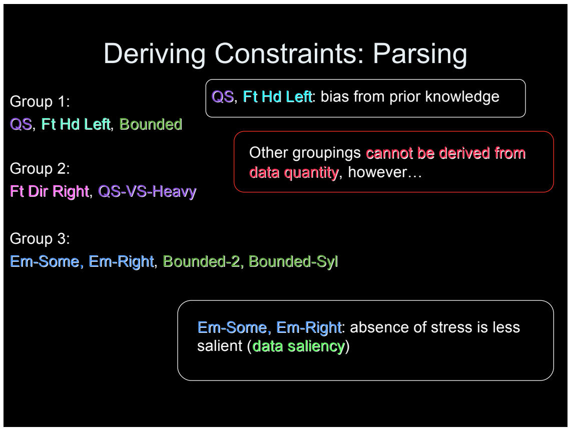
Group 1:
QS, Ft Hd Left, Bounded

QS, Ft Hd Left: bias from prior knowledge

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)



Deriving Constraints: Parsing

Group 1:
QS, Ft Hd Left, Bounded

QS, Ft Hd Left: bias from prior knowledge

Other groupings cannot be derived from data quantity, however…

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

Em-Right: Rightmost syllable is Heavy and unstressed …H H

If a syllable is Heavy, it should be stressed.
If an edge syllable is Heavy and unstressed, an immediate solution (given the available parametric system) is that the syllable is extrametrical.

Combining Cues and Parsing

Viable combination of cues & parsing:
parsing of data point subpart = derivation of cues?

Would partial parsing
(a) derive cues that lead to successful acquisition?
(b) retain the strengths that cues & parsing have separately?
(c) be a more psychologically plausible implementation of the unambiguous data filter?