Psych215L: Language Acquisition

Lecture 19

Grammar & Complex Systems II

Complex Linguistic Systems

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

Observable data: word order Subject Verb Object

Complex Linguistic Systems

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

Observable data: word order Subject Verb Object

English

Subject Verb Object

Kannada

Subject t_{Object} Verb Object

German

Subject Verb $t_{Subject}$ Object t_{Verb}

Complex Linguistic Systems

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

Observable data: stress contour EMphasis

Complex Linguistic Systems

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

Observable data: stress contour EMphasis

(S S) S
EM pha sis

(H L) H
EM pha sis

(H L L)
EM pha sis

(S S S)
EM pha sis

General Problems with Learning Complex Linguistic Systems

What children encounter: the output of the generative linguistic system

EMphasis

General Problems with Learning Complex Linguistic Systems

What children encounter: the output of the generative linguistic system

EMphasis

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

Are syllables differentiated?

Are all syllables included in larger units?

Which syllable(s) of a larger unit is/are stressed?

EM pha sis

General Problems with Learning Complex Linguistic Systems

What children encounter: the output of the generative linguistic system

EMphasis

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

Are syllables differentiated?

Are all syllables included in larger units?

Which syllable(s) of a larger unit is/are stressed?

EM pha sis

Why this is tricky:

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

(H L) H
EM pha sis

(S S S)
EM pha sis

The Hypothesis Space

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

The Hypothesis Space

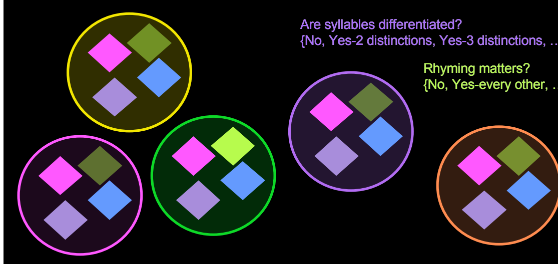
Hypothesis for a language consists of a combination of generalizations about that language (grammar). But this leads to a theoretically infinite hypothesis space.

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

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

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

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



The Hypothesis Space

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

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

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

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

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

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

The Hypothesis Space

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

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

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

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

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

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

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

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

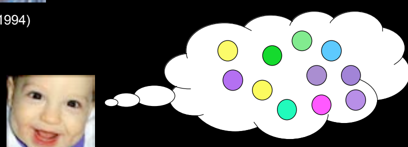
Learning Parametric Linguistic Systems

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

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

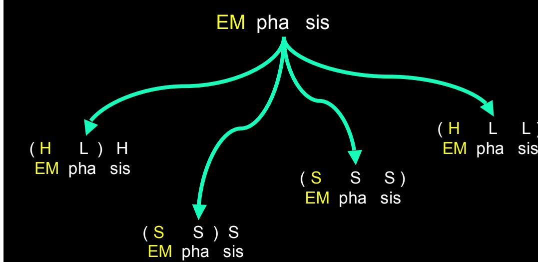
Exponentially growing hypothesis space

(Clark 1994)



Learning Parametric Linguistic Systems

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



Parametric Metrical Phonology

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

Process speakers use:
Basic input unit: syllables — em pha sis

Larger units formed: metrical feet — (em pha) sis
The way these are formed varies from language to language.

Stress assigned within metrical feet — (EM pha) sis
The way this is done also varies from language to language.

Observable Data: stress contour of word — EMphasis

Parametric Metrical Phonology

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

Process speakers use:
Basic input unit: syllables — em pha sis

Larger units formed: metrical feet — (em pha) sis
The way these are formed varies from language to language.

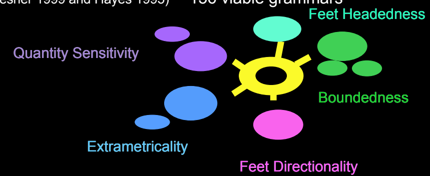
Stress assigned within metrical feet — (EM pha) sis
The way this is done also varies from language to language.

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

Observable Data: stress contour of word — EMphasis

Parametric Metrical Phonology

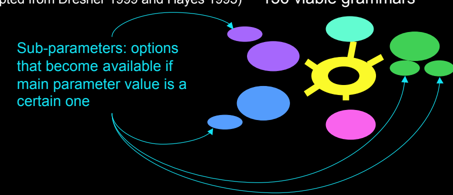
Metrical phonology system here: 5 main parameters, 4 sub-parameters
(adapted from Dresher 1999 and Hayes 1995) - 156 viable grammars



All combine to generate stress contour output

Parametric Metrical Phonology

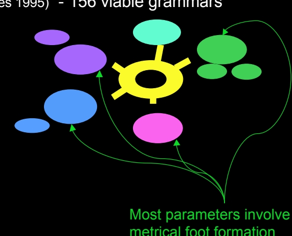
Metrical phonology system here: 5 main parameters, 4 sub-parameters
(adapted from Dresher 1999 and Hayes 1995) - 156 viable grammars



All combine to generate stress contour output

Parametric Metrical Phonology

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



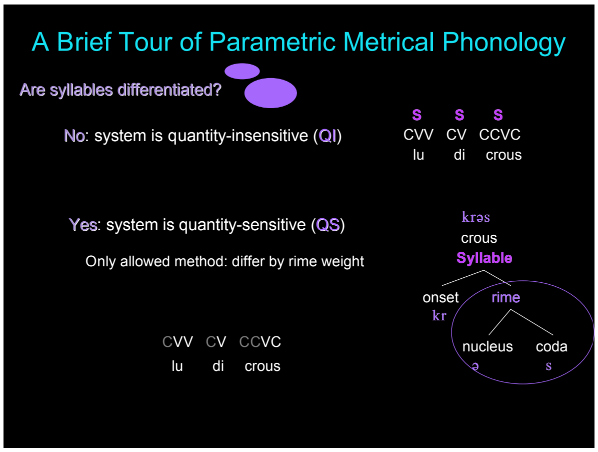
All combine to generate stress contour output

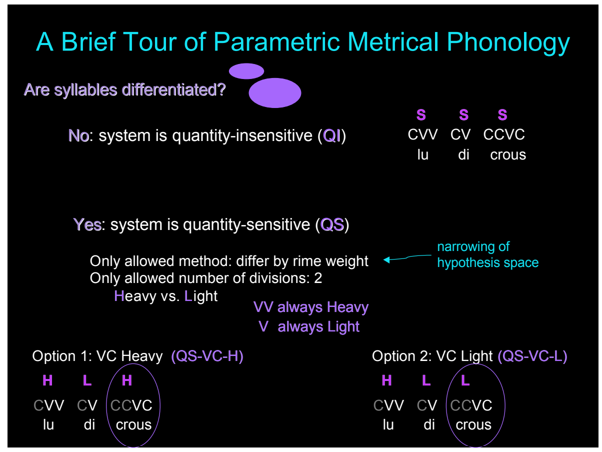
A Brief Tour of Parametric Metrical Phonology

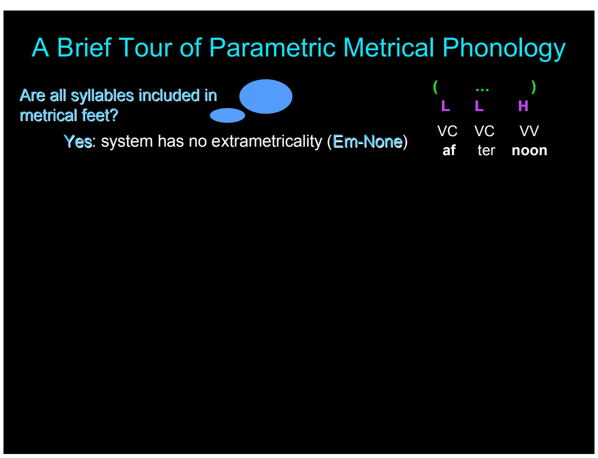
Are syllables differentiated?

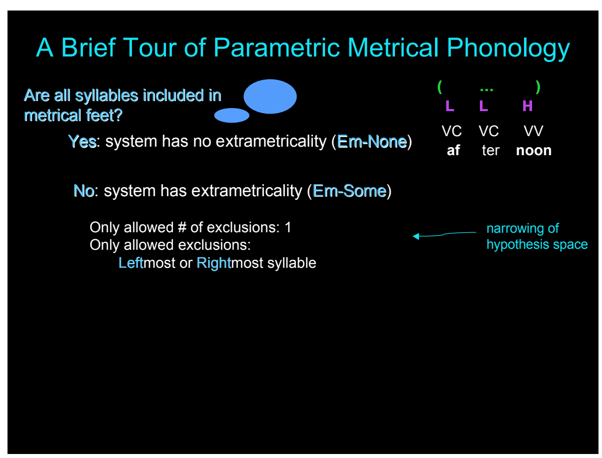
No: system is quantity-insensitive (QI)

S	S	S
CVV	CV	CCVC
lu	di	crous


A Brief Tour of Parametric Metrical Phonology
Are syllables differentiated?
No: system is quantity-insensitive (QI)
S S S
CVV CV CCVC
lu di crous
Yes: system is quantity-sensitive (QS)
Only allowed method: differ by rime weight
krəs
crous
Syllable
onset
kr
rime
nucleus
ə
coda
s
CVV CV CCVC
lu di crous


A Brief Tour of Parametric Metrical Phonology
Are syllables differentiated?
No: system is quantity-insensitive (QI)
S S S
CVV CV CCVC
lu di crous
Yes: system is quantity-sensitive (QS)
Only allowed method: differ by rime weight
Only allowed number of divisions: 2
Heavy vs. Light
narrowing of hypothesis space
VV always Heavy
V always Light
Option 1: VC Heavy (QS-VC-H)
H L H
CVV CV CCVC
lu di crous
Option 2: VC Light (QS-VC-L)
H L L
CVV CV CCVC
lu di crous


A Brief Tour of Parametric Metrical Phonology
Are all syllables included in metrical feet?
(...)
L L H
VC VC VV
af ter noon
Yes: system has no extrametricality (Em-None)


A Brief Tour of Parametric Metrical Phonology
Are all syllables included in metrical feet?
(...)
L L H
VC VC VV
af ter noon
Yes: system has no extrametricality (Em-None)
No: system has extrametricality (Em-Some)
Only allowed # of exclusions: 1
Only allowed exclusions:
Leftmost or Rightmost syllable
narrowing of hypothesis space

A Brief Tour of Parametric Metrical Phonology

Are all syllables included in metrical feet?

(…)
L L H
VC VC VV
af ter **noon**

Yes: system has no extrametricality (Em-None)

No: system has extrametricality (Em-Some)

Only allowed # of exclusions: 1
Only allowed exclusions:
Leftmost or Rightmost syllable

narrowing of hypothesis space

Leftmost syllable excluded: Em-Left

(…)
L H L
V VC V
a gen da

Rightmost syllable excluded: Em-Right

(…)
H L H
VV V VC
lu di crous

A Brief Tour of Parametric Metrical Phonology

What direction are metrical feet constructed?

Two logical options

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

(
H L H
VV V VC
lu di crous

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

)
H L H
VV V VC
lu di crous

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

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

narrowing of hypothesis space

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

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

Ft Dir Left

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

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

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

Ft Dir Left →

(L L L)(H L)

← Ft Dir Right

L L L H L
↓
L L L H L)
↓
L L L H) (L)
↓
(L L L H) (L)

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

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

Ft Dir Left →

(L L L)(H L)

← Ft Dir Right

(L L L H) (L)

Ft Dir Left/Right

(L L L L L
↓
(L L L L L)

S S S S S)
↓
(S S S S S)

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

(L L L)(H L)
(L L L H) (L)
(L L L L L)
(S S S S S)

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

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3. ← narrowing of hypothesis space

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

(L L L)(H L)
(L L L H) (L)
(L L L L L)
(S S S S S)

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

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3. ← narrowing of hypothesis space

Ft Dir Left →

2 units per foot (Bounded-2)

x x x x
↓
(x x) (x x
↓
(x x) (x x)

3 units per foot (Bounded-3)

x x x x
↓
(x x x) (x
↓
(x x x) (x)

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

(L L L)(H L)

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

(L L L H) (L)

(L L L L L)

(S S S S S)

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3. ← narrowing of hypothesis space

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

(x x) (x x) B-2

(x x x) (x) B-3

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

(L L L)(H L)

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

(L L L H) (L)

(L L L L L)

(S S S S S)

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3. ← narrowing of hypothesis space

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

(x x) (x x) B-2

(x x x) (x) B-3

Ft Dir Left
Bounded-2
→
x x

(H L)(L H)

(L L) (L H) ← Count by syllables (Bounded-Syllabic)

(S S) (S S)

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

(L L L)(H L)

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

(L L L H) (L)

(L L L L L)

(S S S S S)

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3. ← narrowing of hypothesis space

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

(x x) (x x) B-2

(x x x) (x) B-3

Count by syllables (Bounded-Syllabic)

(H L)(L H)

Ft Dir Left
Bounded-2
→
x x

Count by moras (Bounded-Moraic)

xx x x xx
H L L H

(H) (L L) (H)

Moras (unit of weight):
H = 2 moras xx
L = 1 mora x

A Brief Tour of Parametric Metrical Phonology

Are metrical feet unrestricted in size?

(L L L)(H L)

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

(L L L H) (L)

(L L L L L)

(S S S S S)

No: Metrical feet are restricted (Bounded).

The size is restricted to 2 options: 2 or 3. ← narrowing of hypothesis space

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

(x x) (x x) B-2

(x x x) (x) B-3

Count by syllables (Bounded-Syllabic)

Ft Dir Left
Bounded-2

Count by moras (Bounded-Moraic)

(H L)(L H) ←compare→ (H) (L L) (H)

A Brief Tour of Parametric Metrical Phonology

Within a metrical foot, which syllable is stressed?

Two options, hypothesis space restriction

Leftmost:
Stress the leftmost syllable (Ft Hd Left)

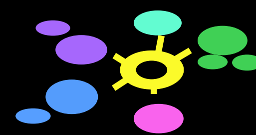
(H) (L L) (H)

(H) (L L) (H)

Rightmost:
Stress the rightmost syllable (Ft Hd Right)

(H) (L L) (H)

Generating a Stress Contour

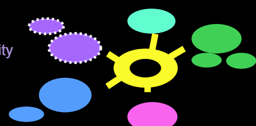


Process speaker uses to generate stress contour

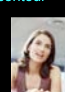
VC	CV	CVC
em	pha	sis

Generating a Stress Contour

Quantity Sensitivity



Process speaker uses to generate stress contour



Are syllables differentiated?

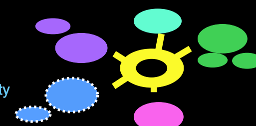
Yes - by rime.

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

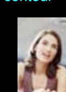
H	L	H
VC	CV	CVC
em	pha	sis

Generating a Stress Contour

Extrametricality



Process speaker uses to generate stress contour



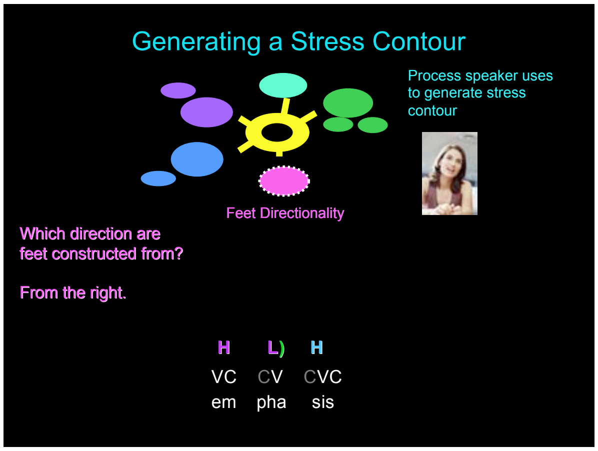
Are any syllables extrametrical?

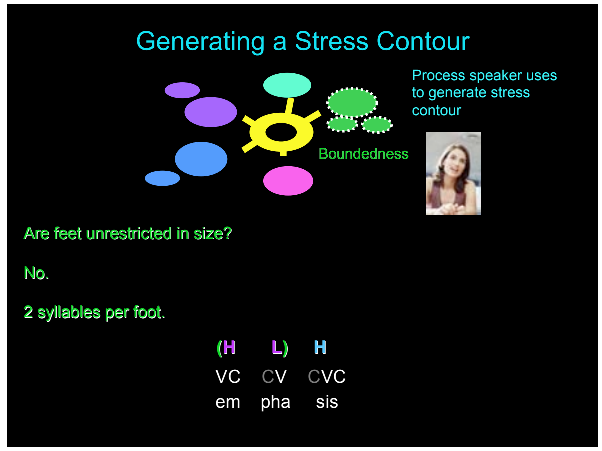
Yes.

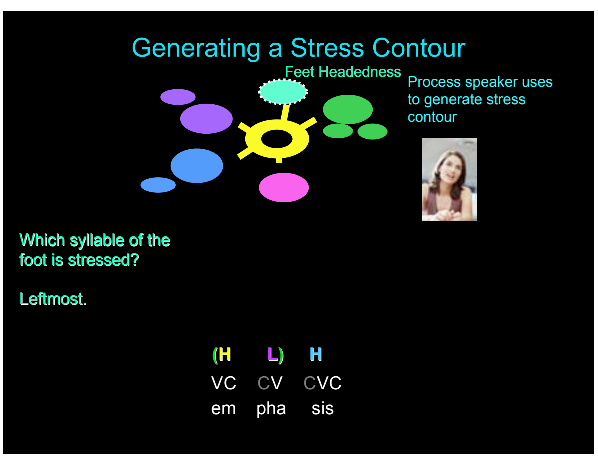
Rightmost syllable is not included in metrical foot.

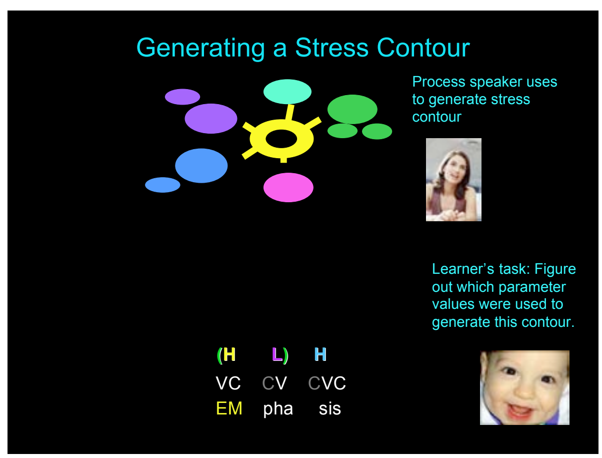
(...)

H	L	H
VC	CV	CVC
em	pha	sis


Generating a Stress Contour
Process speaker uses to generate stress contour
Feet Directionality
Which direction are feet constructed from?
From the right.
H L) H
VC CV CVC
em pha sis


Generating a Stress Contour
Process speaker uses to generate stress contour
Boundedness
Are feet unrestricted in size?
No.
2 syllables per foot.
(H L) H
VC CV CVC
em pha sis


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


Generating a Stress Contour
Process speaker uses to generate stress contour
Learner's task: Figure out which parameter values were used to generate this contour.
(H L) H
VC CV CVC
EM pha sis

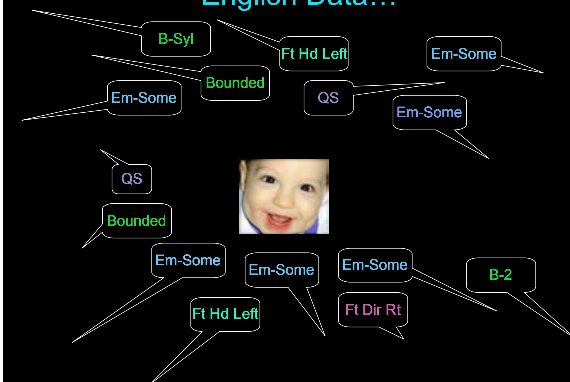
Case study: English metrical phonology

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

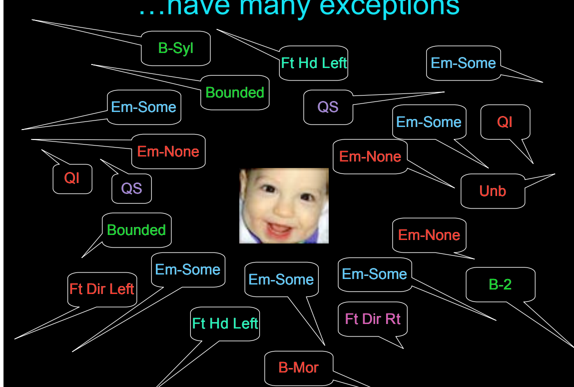
Total Words: 540505   Mean Length of Utterance: 3.5

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

English Data…



…have many exceptions



Case study: English metrical phonology

Non-trivial language: English (full of exceptions)

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

Adult English system values:

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

Exceptions:

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

Case study: English metrical phonology

Non-trivial language: English (full of exceptions)

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

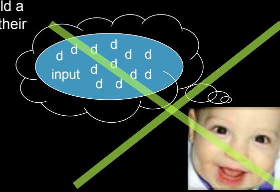
English child-directed speech has a significant quantity of misleading data and is comprised mostly of ambiguous data - it's hard to learn, and therefore interesting!

Key point for cognitive modeling: psychological plausibility

Any probabilistic update procedure that children are likely to use must, at the very least, be incremental/online (Vallabha et al. 2007).

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

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

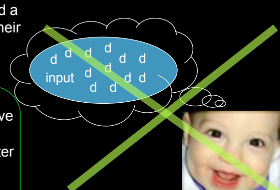


Key point for cognitive modeling: psychological plausibility

Any probabilistic update procedure that children are likely to use must, at the very least, be incremental/online (Vallabha et al. 2007).

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

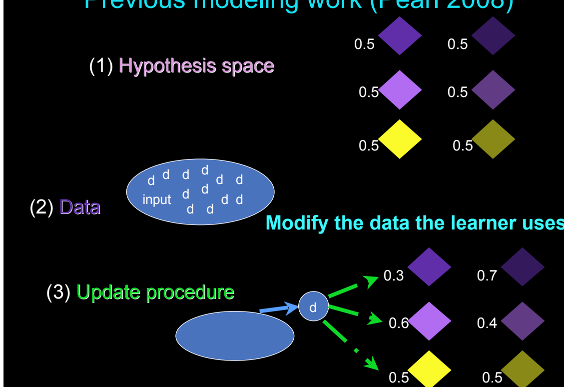
Unlikely: human children can hold a whole corpus's worth of data in their minds for analysis later on



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

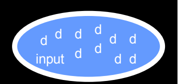
Desired: Learn from a single data point, or perhaps a small number of data points at most.

Previous modeling work (Pearl 2008)



Data Intake Filtering
"Selective Learning", "Interpretive Bias"

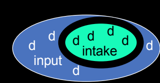
"Equal Opportunity" Intuition: Use all available data to uncover a full range of systematicity, and allow probabilistic model enough data to converge.



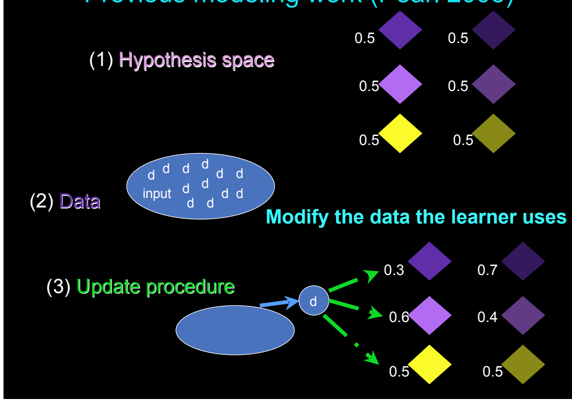
"Selective" Intuition: Use the really good data only.

One instantiation of "really good" = highly informative.

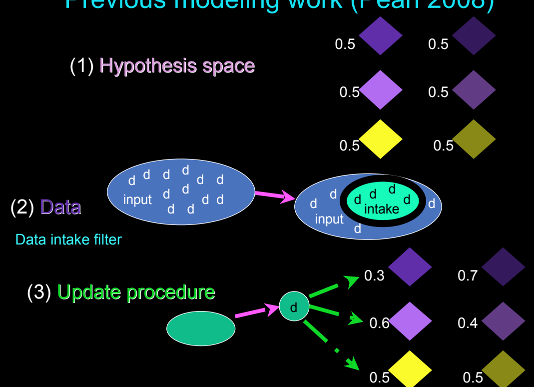
One instantiation of "highly informative" = data viewed by the learner as unambiguous (Fodor, 1998; Dresher, 1999; Lightfoot, 1999; Pearl & Weinberg, 2007)



Previous modeling work (Pearl 2008)



Previous modeling work (Pearl 2008)



Biased learner, using only unambiguous data

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

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

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

Two psychologically plausible probabilistic update procedures

Yang (2002)

Naïve Parameter Learner (NParLearner)

Probabilistic generation & testing of grammars. (incremental)
Hypothesis update: Linear reward-penalty
(Bush & Mosteller 1951)

Two psychologically plausible probabilistic update procedures

Yang (2002)

Naïve Parameter Learner (NParLearner)

Probabilistic generation & testing of grammars. (incremental)
Hypothesis update: Linear reward-penalty
(Bush & Mosteller 1951)

MAP Bayesian Learner (BayesLearner)

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

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

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

QI = 0.5	QS = 0.5
QSVCL = 0.5	QSVCH = 0.5
Em-Some = 0.5	Em-None = 0.5
Em-Left = 0.5	Em-Right = 0.5
Ft Dir Left = 0.5	Ft Dir Rt = 0.5
Bounded = 0.5	Unbounded = 0.5
Bounded-2 = 0.5	Bounded-3 = 0.5
Bounded-Syl = 0.5	Bounded-Mor = 0.5
Ft Hd Left = 0.5	Ft Hd Rt = 0.5

Initially all are equiprobable

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

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

AFterNOON

QI = 0.5	QS = 0.5
QSVCL = 0.5	QSVCH = 0.5
Em-Some = 0.5	Em-None = 0.5
Em-Left = 0.5	Em-Right = 0.5
Ft Dir Left = 0.5	Ft Dir Rt = 0.5
Bounded = 0.5	Unbounded = 0.5
Bounded-2 = 0.5	Bounded-3 = 0.5
Bounded-Syl = 0.5	Bounded-Mor = 0.5
Ft Hd Left = 0.5	Ft Hd Rt = 0.5

QI/QS?…if QS, QSVCL or QSVCH?
Em-None/Em-Some?…
…

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

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

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

AFterNOON

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

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

(L)	(L	H)
VC	CVC	CVVC
AF	ter	NOON

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

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

AFterNOON

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

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

(L	L)	(H)
VC	CVC	CVVC
af	TER	NOON

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

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

AFterNOON

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

(L)	(L	H)
VC	CVC	CVVC
AF	ter	NOON

Match (success): reward all

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

(L	L)	(H)
VC	CVC	CVVC
af	TER	NOON

Mismatch (failure): punish all

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities

NParLearner (Yang 2002): Linear Reward-Penalty

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

Parameter values v1 vs. v2

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

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities

NParLearner (Yang 2002): Linear Reward-Penalty

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

Parameter values v1 vs. v2

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

BayesLearner: Bayesian update of binomial distribution (Chew 1971)

Parameters α, β:

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

here: $\alpha = \beta = 0.5$

Parameter value v1

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

reward: success + 1 punish: success + 0

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities

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

QI = 0.3 QS = 0.7
QSVCL = 0.6 QSVCH = 0.4
Em-Some = 0.1 Em-None = 0.9
…

Probabilistic learning for English

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities

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

QI = 0.3 QS = 0.7
QSVCL = 0.6 QSVCH = 0.4
Em-Some = 0.1 Em-None = 0.9
…

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

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

QI/QS?…if QS, QSVCL or QSVCH?
Em-None
…

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

Probabilistic learning for English

Goal: Converge on English values after learning period is over

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

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

Probabilistic learning for English

Goal: Converge on English values after learning period is over

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

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

Model	Success rate (1000 runs)
NParLearner, γ = .001, .0025, .01, .025	0.0%
BayesLearner	0.0%

Examples of incorrect target grammars

NParLearner:

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

BayesLearner:

QS, Em-Some, Em-Right, QSVCH, Ft Hd Left, Ft Dir Rt, Unb

Bounded, B-Syl, QI, Ft Hd Left, Em-None, Ft Dir Left, B-2

The learning framework: where can we modify?

(1) Hypothesis space

0.5 0.5

0.5 0.5

0.5 0.5

(2) Data

d d d d d d input d d d d d d

(3) Update procedure

d

0.3 0.7

0.6 0.4

0.5 0.5

The learning framework: where can we modify?

(1) Hypothesis space

0.5 0.5

0.5 0.5

0.5 0.5

(2) Data

d d d d d d input d d d d d d

(3) Update procedure

Linear Reward-Penalty, Bayesian…?

d

0.3 0.7

0.6 0.4

0.5 0.5

Probabilistic learning for English: Modifications

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities

Count-learning: smooth out some of the irregularities in the data, better deal with complex systems (Yang 2002)

Implementation (Yang 2002):
Matching contour = increase parameter value's batch counter by 1
Mismatching contour = decrease parameter value's batch counter by 1

Invoke update procedure (Linear Reward-Penalty or Bayesian Updating) when count limit *c* is reached.

Probabilistic learning for English: Modifications

Probabilistic generation and testing of grammars (Yang 2002)

Update parameter value probabilities + Count Learning

NParLearner (Yang 2002): Linear Reward-Penalty

Invoke when the batch counter for p_{v1} or p_{v2} equals *c*.

Parameter values v1 vs. v2

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

BayesLearner: Bayesian update of binomial distribution (Chew 1971)

Invoke when the batch counter for p_{v1} or p_{v2} equals *c*.

Note: total data seen + 1

Parameter value v1

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

reward: success + 1 punish: success + 0

Probabilistic learning for English

Goal: Converge on English values after learning period is over

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

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

Model	Success rate (1000 runs)
NParLearner, γ = .001, .0025, .01, .025	0.0%
BayesLearner	0.0%

Probabilistic learning for English

Goal: Converge on English values after learning period is over

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

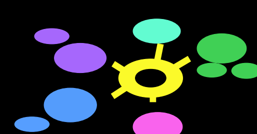
QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hd Left

Model	Success rate (1000 runs)
NParLearner, γ = .001, .0025, .01, .025	0.0%
BayesLearner	0.0%
NParLearner + Counting, γ = .001, .0025, .01, .025, c = 2, 5, 7, 10, 15, 20	0.033%
BayesLearner + Counting, c = 2, 5, 7, 10, 15, 20	0.0%

What gives?

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

156 viable grammars



English is *not* the optimal grammar

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

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

Unbiased probabilistic learning is more likely to find the optimal grammar

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

Unbiased probabilistic learning is more likely to find the optimal grammar

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

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

Unbiased probabilistic learning is more likely to find the optimal grammar

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

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

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

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

Probabilistic learning for English: Bias

Probabilistic generation and testing of grammars (Yang 2002)

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

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

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

Ft Hd Left
S S

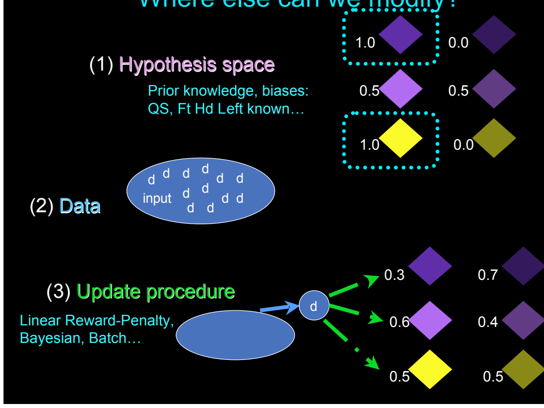
Ft Hd Rt
S S

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

Where else can we modify?



Probabilistic learning for English

Goal: Converge on English values after learning period is over

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

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

Model	Success rate (1000 runs)
NParLearner, γ = .001, .0025, .01, .025	0.000%
BayesLearner	0.100%
NParLearner + Counting, γ = .001, .0025, .01, .025, c = 2, 5, 7, 10, 15, 20	1.650%
BayesLearner + Counting, c = 2, 5, 7, 10, 15, 20	1.780%

The best isn't so great, even with this restricted hypothesis space

Why not?
Because English is *still not* the optimal grammar

Adult English system values:

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

Of the 60 available grammars, English is ranked

18th by token compatibility

18th by type compatibility

Other Biases for Children

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

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

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

Where we are now

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

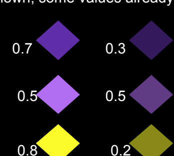
Learning complex systems: difficult.
Correct grammar is not the optimal grammar for child's input data without some kind of additional bias.

Where we are now

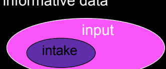
Modeling: aimed at understanding how children learn language, generating child behavior by using psychologically plausible methods

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

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



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



Where we can go

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

+ biases?

Where we can go

(1) Selective bias:

How successful on other difficult learning cases (noisy data sets, other complex systems)?

Are there other methods of implementing selective biases that lead to successful learning (productive data: Yang 2005)?

How necessary is a selective bias? Are there other biases that might cause the correct grammar to be the optimal grammar for the English data?

+ biases?

(2) Hypothesis space bias:

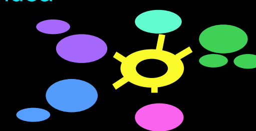
Will other hypothesis space instantiations allow the correct grammar to be the optimal grammar (constraints (Tesar & Smolensky 2000))? What learning mechanisms make the correct grammar learnable in these hypothesis spaces?

Is it possible to converge on the correct grammar given a less well-defined hypothesis space a priori (e.g. only knowing that units larger than syllables are required)?

+ other/fewer biases?

The big idea

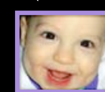
Complex linguistic systems may well require something beyond probabilistic methods in order to be learned as well as children learn them given the data children are given.



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

What we can do with computational modeling:

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



(b) generate experimentally testable predictions about learning (Pearl 2008: learning trajectory)