Constrained Probabilistic Learning
for Complex Linguistic Systems

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

Human Language Learning

Theoretical work:
object of acquisition

(x x) x
H L H
EM pha sis

Human Language Learning

Theoretical work:
object of acquisition

(x x) x
H L H
EM pha sis

Experimental work:
time course of acquisition
& data

Human Language Learning

Theoretical work:
object of acquisition

(x x) x
H L H
EM pha sis

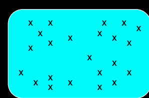
Experimental work:
time course of acquisition
& data

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

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

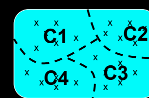
Categorization/Clustering
Ex: What are the contrastive sounds of a language?



The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

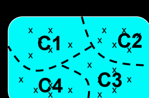
Categorization/Clustering
Ex: What are the contrastive sounds of a language?



The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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



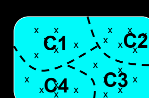
Extraction
Ex: Where are words in fluent speech?

húwzəfréjdəvðəbÍgbæ'dwə'lf

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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



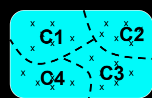
Extraction
Ex: Where are words in fluent speech?

húwz əfréjd əv ðə bÍg bæ'd wə'lf
who's afraid of the big bad wolf

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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



Extraction
Ex: Where are words in fluent speech?

húwz əfréjd əv ðə bÍg bæ'd wə'lf
who's afraid of the big bad wolf

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

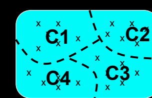
blink~blinked confide~confided

drink~drank

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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



Extraction
Ex: Where are words in fluent speech?

húwz əfréjd əv ðə bÍg bæ'd wə'lf
who's afraid of the big bad wolf

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

blink~blinked confide~confided
blɪŋk blɪŋkt kənfajd kənfajdəd

drink~drank
drɪŋk drejŋk

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Observable data: word order Subject Verb Object

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Observable data: word order Subject Verb Object

English
Subject Verb Object

Kannada
Subject t_{Object} Verb Object

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

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Observable data: stress contour EMphasis

The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Observable data: stress contour EMphasis

(S S) S
EM pha sis

(H L) H
EM pha sis

(H L L)
EM pha sis

(S S S)
EM pha sis

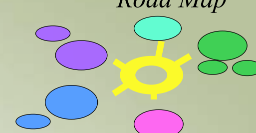
Road Map

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

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

Where next? Implications & Extensions

Road Map



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

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

Where next? Implications & Extensions

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

Which syllable of a larger unit is stressed?

Are all syllables included?

Are syllables differentiated?

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

Which syllable of a larger unit is stressed?

Are all syllables included?

Are syllables differentiated?

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

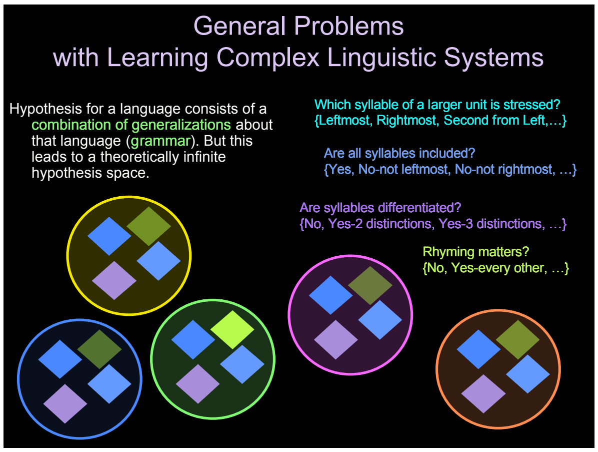
Moreover, data are often ambiguous, even if parameters of variation are known.

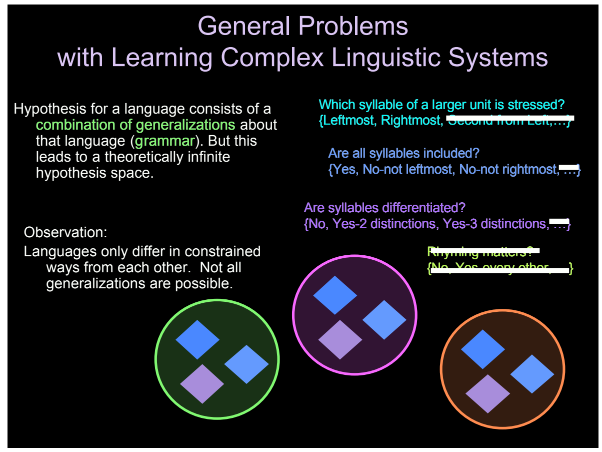
(S S S)
EM pha sis

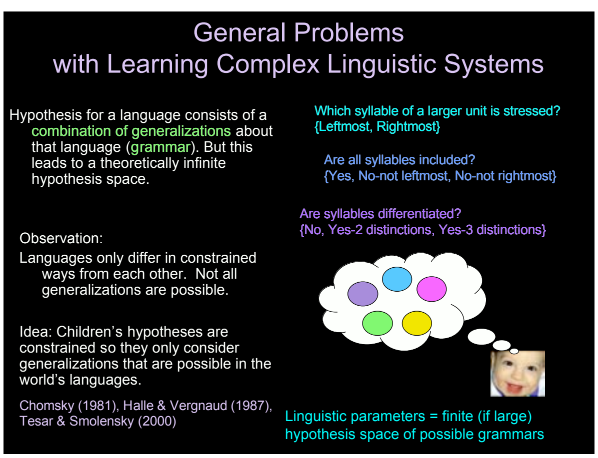
Levels of abstract structure

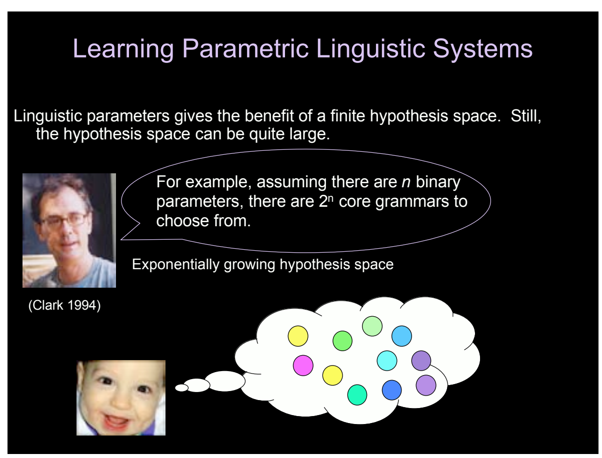
General Problems with Learning Complex Linguistic Systems

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


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


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


General Problems
with Learning Complex Linguistic Systems
Hypothesis for a language consists of a combination of generalizations about that language (grammar). But this leads to a theoretically infinite hypothesis space.
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: 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 gives the benefit of a finite hypothesis space. Still, the hypothesis space can be quite large.
For example, assuming there are n binary parameters, there are 2n core grammars to choose from.
Exponentially growing hypothesis space
(Clark 1994)

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. Only syllables in metrical feet can be stressed.

Stress assigned within metrical feet — (EM pha) sis

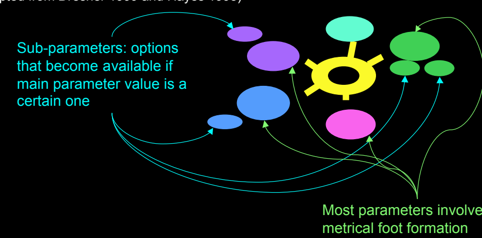
The way this is done also varies from language to language.

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)



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

CVV	CV	CCVC
lu	di	crous

krəs
crous
Syllable
onset — kr
rime
nucleus — ə
coda — s

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?

Yes: system has no extrametricality (Em-None)

(	...	)
L	L	H
VC	VC	VV
af	ter	**noon**

A Brief Tour of Parametric Metrical Phonology

Are all syllables included in metrical feet?

Yes: system has no extrametricality (Em-None)

(	...	)
L	L	H
VC	VC	VV
af	ter	**noon**

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?

Yes: system has no extrametricality (Em-None)

(	...	)
L	L	H
VC	VC	VV
af	ter	**noon**

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)

(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

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

narrowing of hypothesis space

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

narrowing of hypothesis space

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

narrowing of hypothesis space

(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) ↔ (H) (L L) (H)

compare

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)

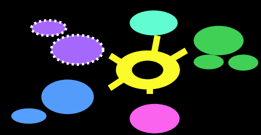
(H) (L L) (H)

Rightmost:

Stress the rightmost syllable (Ft Hd Right)

(H) (L L) (H)

Generating a Stress Contour



Process speaker uses to generate stress contour

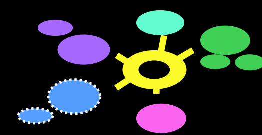
Are syllables differentiated?

Yes.

VC syllables are Heavy.

H	L	H
VC	CV	CVC
em	pha	sis

Generating a Stress Contour



Process speaker uses to generate stress contour

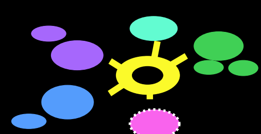
Are any syllables extrametrical?

Yes.

Rightmost syllable is not included in metrical foot.

(	...	)
H	L	H
VC	CV	CVC
em	pha	sis

Generating a Stress Contour



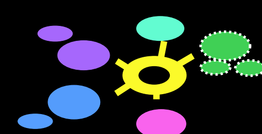
Process speaker uses to generate stress contour

Which direction are feet constructed from?

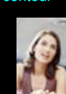
From the right.

H	L)	H
VC	CV	CVC
em	pha	sis

Generating a Stress Contour



Process speaker uses to generate stress contour

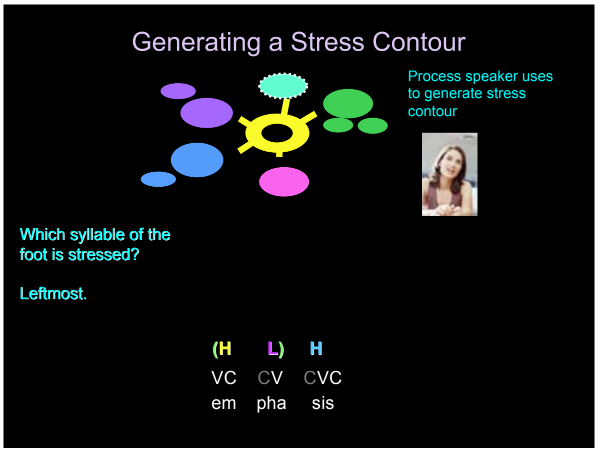


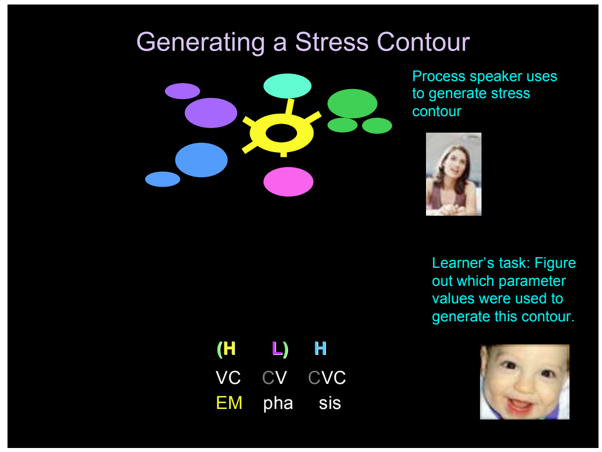
Are feet unrestricted?

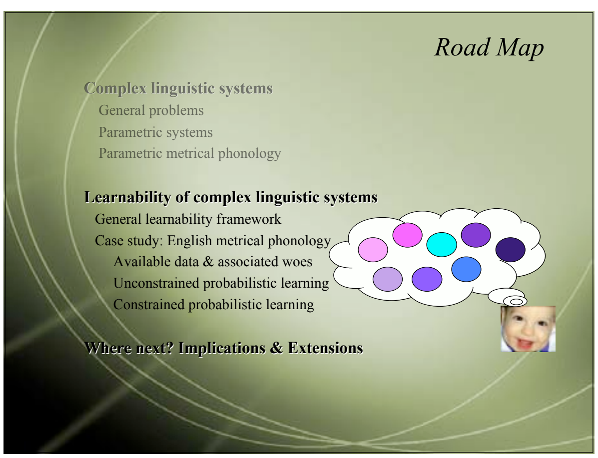
No.

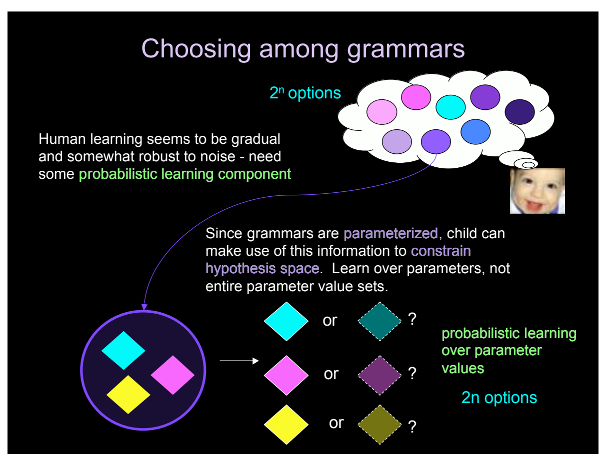
2 syllables per foot.

(H	L)	H
VC	CV	CVC
em	pha	sis


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


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


Road Map
Complex linguistic systems
General problems
Parametric systems
Parametric metrical phonology
Learnability of complex linguistic systems
General learnability framework
Case study: English metrical phonology
Available data & associated woes
Unconstrained probabilistic learning
Constrained probabilistic learning
Where next? Implications & Extensions


Choosing among grammars
2n options
Human learning seems to be gradual and somewhat robust to noise - need some probabilistic learning component
Since grammars are parameterized, child can make use of this information to constrain hypothesis space. Learn over parameters, not entire parameter value sets.
or ?
or ?
or ?
probabilistic learning over parameter values
2n options

A caveat about learning parameters separately

or ?
or ?
or ?

Parameters are system components that combine together to generate output.

Choice of one parameter may influence choice of subsequent parameters.

A caveat about learning parameters separately

1 or ?
or ?
or ?

Parameters are system components that combine together to generate output.

Choice of one parameter may influence choice of subsequent parameters.

A caveat about learning parameters separately

or ?
or ?
1 or ?

Parameters are system components that combine together to generate output.

Choice of one parameter may influence choice of subsequent parameters.

A caveat about learning parameters separately

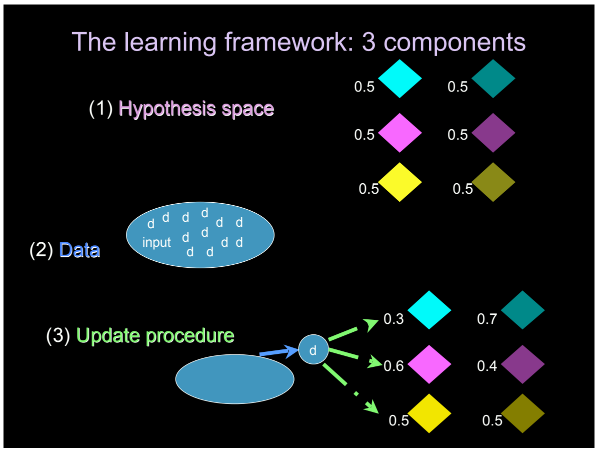
or ?
or ?
or ?

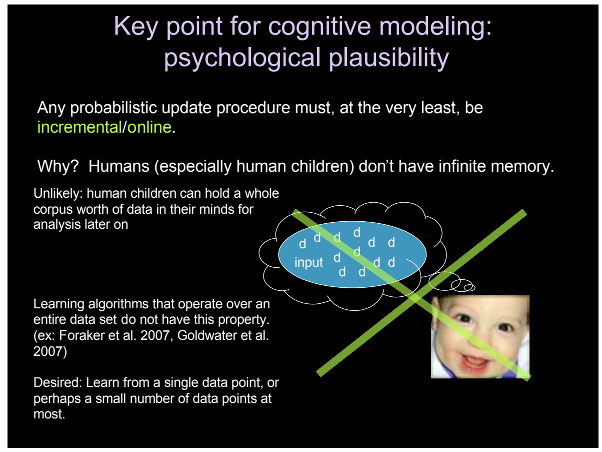
Parameters are system components that combine together to generate output.

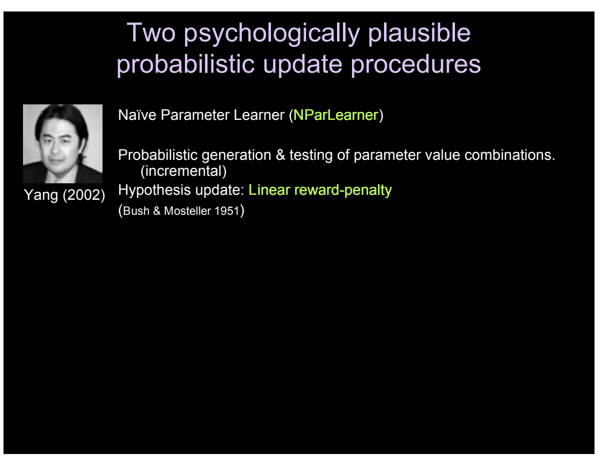
Choice of one parameter may influence choice of subsequent parameters.

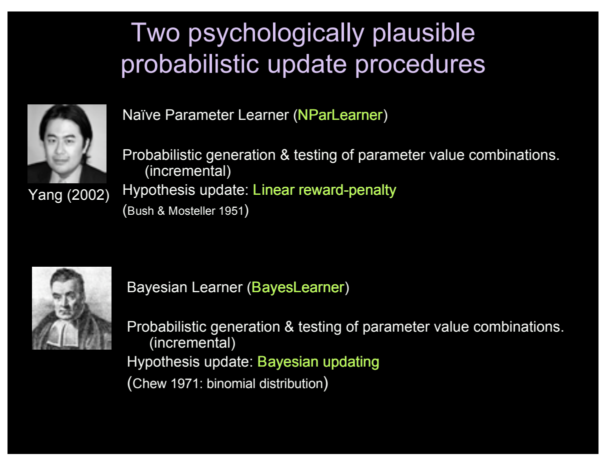
Point: The order in which parameters are set may determine if they are set correctly from the data.

Dresher 1999


The learning framework: 3 components
(1) Hypothesis space
0.5 0.5
0.5 0.5
0.5 0.5
(2) Data
input d d d d d d d d d d
(3) Update procedure
d
0.3 0.7
0.6 0.4
0.5 0.5


Key point for cognitive modeling: psychological plausibility
Any probabilistic update procedure must, at the very least, be incremental/online.
Why? Humans (especially human children) don't have infinite memory.
Unlikely: human children can hold a whole corpus worth of data in their minds for analysis later on
input d d d d d d d d d d
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.


Two psychologically plausible probabilistic update procedures
Yang (2002)
Naïve Parameter Learner (NParLearner)
Probabilistic generation & testing of parameter value combinations. (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 parameter value combinations. (incremental)
Hypothesis update: Linear reward-penalty
(Bush & Mosteller 1951)
Bayesian Learner (BayesLearner)
Probabilistic generation & testing of parameter value combinations. (incremental)
Hypothesis update: Bayesian updating
(Chew 1971: binomial distribution)

Case study: English metrical phonology

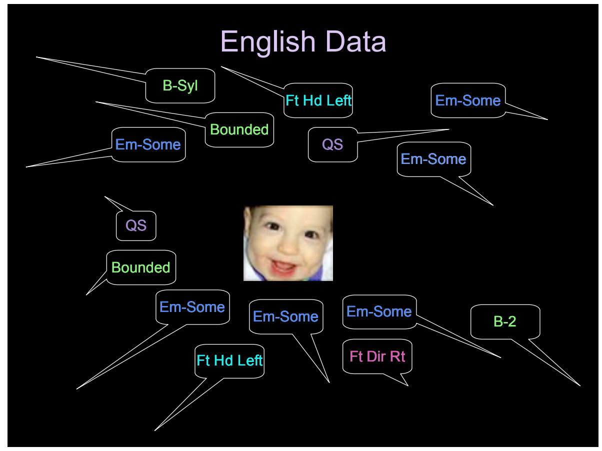
Adult English system values:

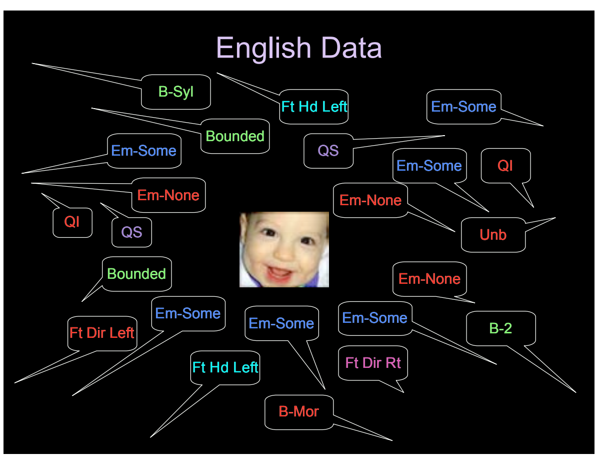
QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hd Left

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





Case study: English metrical phonology

Adult English system values:

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

Non-trivial language: English (full of exceptions)

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

Hard - therefore interesting!

Exceptions:

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

Probabilistic learning for English

Probabilistic generation and testing of parameter values (Yang 2002)

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

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 parameter values (Yang 2002)

For each data point encountered, the learner probabilistically generates a set of parameter values (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 parameter values (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, the grammar successfully "parses" the data point. All participating parameter values are rewarded.

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

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

reward all

Probabilistic learning for English

Probabilistic generation and testing of parameter values (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

reward all

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

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

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

punish all

Probabilistic learning for English

Probabilistic generation and testing of parameter values (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

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

punish all

Probabilistic learning for English

Probabilistic generation and testing of parameter values (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 parameter values (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 parameter values (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 parameter values (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

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, $0.01 \leq \gamma \leq 0.05$	1.2%
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

Probabilistic learning for English: Modifications

Probabilistic generation and testing of parameter values (Yang 2002)

Update parameter value probabilities

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

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

Invoke update procedure (Linear Reward-Penalty or Bayesian Updating) when batch limit *b* is reached. Then, reset parameter's batch counters.

Probabilistic learning for English: Modifications

Probabilistic generation and testing of parameter values (Yang 2002)

Update parameter value probabilities + Batch Learning

NParLearner (Yang 2002): Linear Reward-Penalty

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

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 *b*.

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, $0.01 \leq \gamma \leq 0.05$	1.2%
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, $0.01 \leq \gamma \leq 0.05$	1.2%
BayesLearner	0.0%
NParLearner + Batch, $0.01 \leq \gamma \leq 0.05$, $2 \leq b \leq 10$	0.8%
BayesLearner + Batch, $2 \leq b \leq 10$	1.0%

Probabilistic learning for English: Modifications

Probabilistic generation and testing of parameter values (Yang 2002)

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

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

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

Probabilistic learning for English: Modifications

Probabilistic generation and testing of parameter values (Yang 2002)

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

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

Build this bias into a model: set probability of QS = Ft Hd Left = 1.0.
These will always be chosen during generation.

QS…QSVCL or QSVCH?
…
Ft Hd Left

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

Update parameter value probabilities + Batch Learning

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, $0.01 \leq \gamma \leq 0.05$	1.2%
BayesLearner	0.0%
NParLearner + Batch, $0.01 \leq \gamma \leq 0.05$, $2 \leq b \leq 10$	0.8%
BayesLearner + Batch, $2 \leq b \leq 10$	1.0%

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, $0.01 \leq \gamma \leq 0.05$	1.2%
BayesLearner	0.0%
NParLearner + Batch, $0.01 \leq \gamma \leq 0.05$, $2 \leq b \leq 10$	0.8%
BayesLearner + Batch, $2 \leq b \leq 10$	1.0%
NParLearner + Batch + Bias, $0.01 \leq \gamma \leq 0.05$, $2 \leq b \leq 10$	5.0%
BayesLearner + Batch + Bias, $2 \leq b \leq 10$	1.0%

Probabilistic learning for English

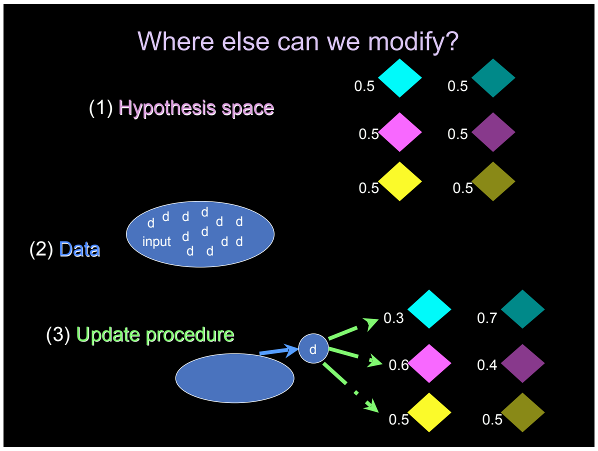
Goal: Converge on English values after learning period is over

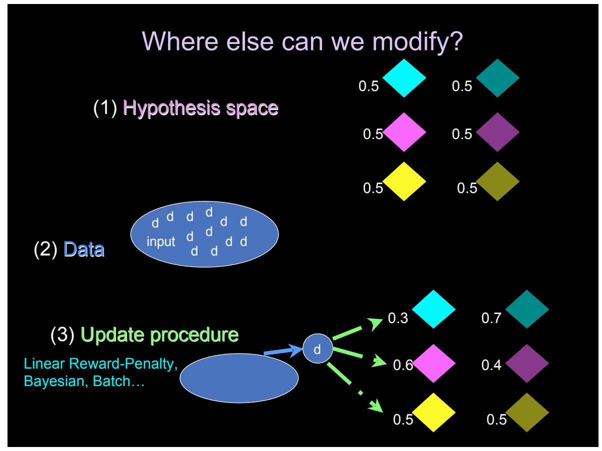
Learning Period Length: 1,666,667 words (based on estimates of words heard in a 6 month period, using Akhtar et al. (2004)).

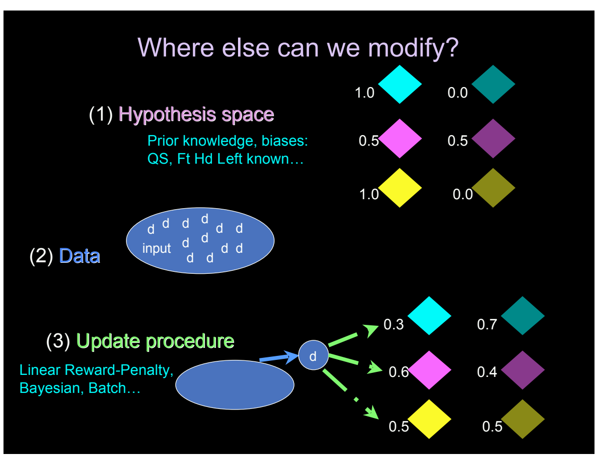
QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hd Left

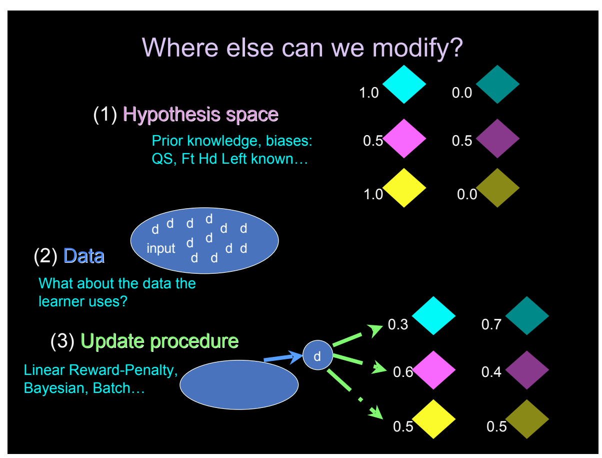
Model	Success rate (1000 runs)
NParLearner, $0.01 \leq \gamma \leq 0.05$	1.2%
BayesLearner	0.0%
NParLearner + Batch, $0.01 \leq \gamma \leq 0.05$, $2 \leq b \leq 10$	0.8%
BayesLearner + Batch, $2 \leq b \leq 10$	1.0%
NParLearner + Batch + Bias, $0.01 \leq \gamma \leq 0.05$, $2 \leq b \leq 10$	5.0%
BayesLearner + Batch + Bias, $2 \leq b \leq 10$	1.0%

The best isn't so great


Where else 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


Where else 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, Batch…
d
0.3 0.7
0.6 0.4
0.5 0.5


Where else can we modify?
(1) Hypothesis space
Prior knowledge, biases:
QS, Ft Hd Left known…
1.0 0.0
0.5 0.5
1.0 0.0
(2) Data
d d d d d d
input d d d d d d
(3) Update procedure
Linear Reward-Penalty,
Bayesian, Batch…
d
0.3 0.7
0.6 0.4
0.5 0.5


Where else can we modify?
(1) Hypothesis space
Prior knowledge, biases:
QS, Ft Hd Left known…
1.0 0.0
0.5 0.5
1.0 0.0
(2) Data
What about the data the
learner uses?
d d d d d d
input d d d d d d
(3) Update procedure
Linear Reward-Penalty,
Bayesian, Batch…
d
0.3 0.7
0.6 0.4
0.5 0.5

Data Intake Filtering
"Selective Learning"

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

d d d d d d input d d d d

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

One instantiation of "really good" = highly informative.

d d d d d d input intake d d

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

Where else can we modify?

(1) Hypothesis space

Prior knowledge, biases: QS, Ft Hd Left known…

1.0 0.0
0.5 0.5
1.0 0.0

(2) Data

d d d d d d input d d d d

What about the data the learner uses?

(3) Update procedure

Linear Reward-Penalty, Bayesian, Batch…

d

0.3 0.7
0.6 0.4
0.5 0.5

Where else can we modify?

(1) Hypothesis space

Prior knowledge, biases: QS, Ft Hd Left known…

1.0 0.0
0.5 0.5
1.0 0.0

(2) Data

d d d d d d input d d d d

d d d d d input intake d d

Data intake filter

(3) Update procedure

Linear Reward-Penalty, Bayesian, Batch…

d

0.3 0.7
0.6 0.4
0.5 0.5

Practical matters:
Feasibility of unambiguous data

Clark 1994

Existence?

"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"

AFterNOON

What's the same here, other than the output?

(S S) (S)
af ter noon

(L L) (H)
af ter noon

(L) (L H)
af ter noon

Identification?

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

Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

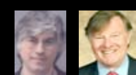
Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

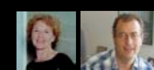
Identification?

Identifying unambiguous data:

Cues (Dresher, 1999; Lightfoot, 1999)



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



Both operate over a single data point at a time: compatible with incremental learning

Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

Identification?

Identifying unambiguous data:

Cues (Dresher 1999; Lightfoot 1999): heuristic pattern-matching to observable form of the data. Cues are available for each parameter value, known already by the learner.

$\boxed{\text{S...S}}$ af ter noon → Em-None

Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

Identification?

Identifying unambiguous data:

Cues (Dresher 1999; Lightfoot 1999): heuristic pattern-matching to observable form of the data. Cues are available for each parameter value, known already by the learner.

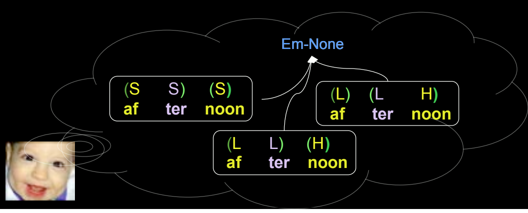
QS: 2 syllable word with 2 stresses	VV VV
Em-Right: Rightmost syllable is Heavy and unstressed	… H
Unb: 3+ unstressed S/L syllables in a row	…S S S… … L L L L
Ft Hd Left: Leftmost foot has stress on leftmost syllable	S S S… H L L …

Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

Identification?

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

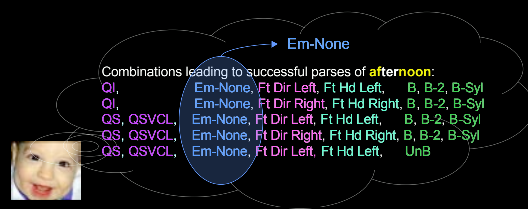


Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

Identification?

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



Probabilistic learning from unambiguous data

(Pearl 2008)

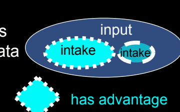
Each parameter has 2 values.

Probabilistic learning from unambiguous data

(Pearl 2008)

Each parameter has 2 values.

Advantage in data: How much more unambiguous data there is for one value over the other in the data distribution.



Assumption (Yang 2002):
The value with the greater advantage will be the one a probabilistic learner will converge on over time.

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

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

QS more probable

Em-None more 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		

Moving Targets & Unambiguous Data: What Happens After Parameter-Setting

Em-None more 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		

Moving Targets & Unambiguous Data: What Happens After Parameter-Setting

Em-Some more probable

QS		Extrametricality	
		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		

Getting to English

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

QS
?

Current knowledge of the system (parameters set) influences the perception of unambiguous data (subsequent parameters set).

Dresher 1999

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

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

Probabilistic learning from unambiguous data

(Pearl 2008)

Dresher 1999

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

Probabilistic learning from unambiguous data

(Pearl 2008)

Dresher 1999

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

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

Cues

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

(b) Em-Right
before Bounded-Syl

(c) Bounded-2
before Bounded-Syl

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

Parsing

Group 1:
QS, Ft Hd Left, Bounded

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

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

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

Feasibility & Sufficiency of the Unambiguous Data Filter

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

Clark 1994

Existence?

"It is unlikely that any example … would show the effect of only a single parameter value; rather, each example is the result of the interaction of several different principles and parameters"

Feasibility & Sufficiency of the Unambiguous Data Filter

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

√ Existence √ Identification

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

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

Road Map

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

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

Where next? Implications & Extensions

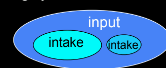
Where we are now

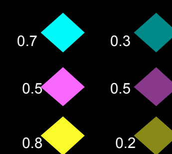
Modeling: aimed at understanding how children learn language, generating child behavior by using psychologically plausible methods

Learning complex systems: difficult.
Success comes from integrating biases into probabilistic learning models.

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

Bias on data:
interpretive bias to use highly informative data





Where we can go: 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)

Where we can go

(1) Interpretive bias:
- How successful on other difficult learning cases (noisy data sets, other complex systems)?
- How reasonable are cues/parsing for identifying unambiguous data? (Ask me!)
- Are there other methods of implementing interpretative biases that lead to successful learning (productive data: Yang 2005)?
- How necessary is an interpretive bias? Are there cleverer probabilistic learning methods than can succeed (Fodor & Sakas 2004, Bayesian strategies)?

+ biases?

Where we can go

(1) Interpretive bias:

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

How reasonable are cues/parsing for identifying unambiguous data? (Ask me!)

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

How necessary is an interpretive bias? Are there cleverer probabilistic learning methods than can succeed (Fodor & Sakas 2004, Bayesian strategies)?

+ biases?

(2) Hypothesis space bias:

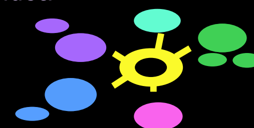
Is it possible to infer the correct parameters of variation given less structured information a priori (e.g. larger units than syllables are required)? [Model Selection]

Are other instantiations of hypothesis space restrictions learnable from realistic data (constraints (Tesar & Smolensky 2000))?

+ fewer/other 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.



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

What we can do with computational modeling:

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

(b) generate experimentally testable predictions about learning

Thank You

Amy Weinberg
Bill Idsardi
Bill Sakas

Jeff Lidz
Charles Yang
Janet Fodor

The audiences at

UC Irvine Machine Learning Group
University of California, Los Angeles Linguistics Department
University of Southern California Linguistics Department
BUCLD 32
UC Irvine Language Learning Group
UC Irvine Department of Cognitive Sciences
CUNY Psycholinguistics 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):

(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 — S S… → lu di crous → Ft Hd Left

Can find information in sub-part of data point — S S… → lu di… → Ft Hd Left

Can tolerate exceptions — S S… → lu di crous → Ft Hd Left

Cues vs. Parsing: Comparison

Cues:

Are heuristic — S S… → (x) (x x) / S S S / af ter noon ← Ft Hd Rt

Require additional knowledge

S S… → Ft Hd Left

May rely on default values

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

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

No default values used

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:

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

Cannot find information in
sub-part of data point

Cannot tolerate exceptions

????

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		+

Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

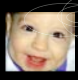
Identification?

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

The effect of learning parameters

Em-None

Combinations leading to successful parses of **afternoon**:
QI, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl
QI, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl
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



If Bounded is known..

Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

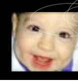
Identification?

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

The effect of learning parameters

Em-None, B, B-2, B-Syl

Combinations leading to successful parses of **afternoon**:
QI, Em-None, Ft Dir Left, Ft Hd Left, B, B-2, B-Syl
QI, Em-None, Ft Dir Right, Ft Hd Right, B, B-2, B-Syl
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



If Bounded is known..

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.

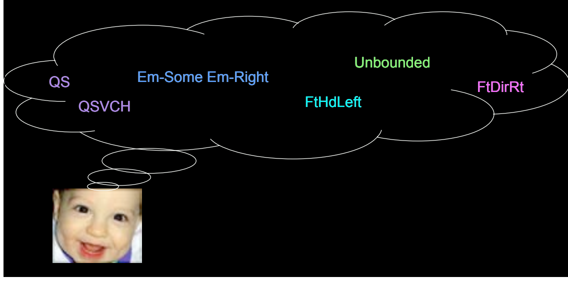
(b) Choose a currently unset parameter to set. The value chosen for this parameter is the value that has a higher probability in the data the learner perceives as unambiguous.

(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.



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

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 probabilistic learner manages to set the parameters in this order, the learner is guaranteed converge on English.

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

Order #23 looks good!

Order Constraints

Good: Order constraints exist that will allow the learner to converge on the adult system, provided the learner knows these constraints.

Better: These order constraints can be derived from properties of the learning system, rather than being stipulated, or they're already known through other means.

Knowing Through Other Means

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

Jusczyk, Cutler, & Redanz (1993): 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

Bounded-Syl as default (default values)

(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

Bounded-Syl as default (default values)

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

(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

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?