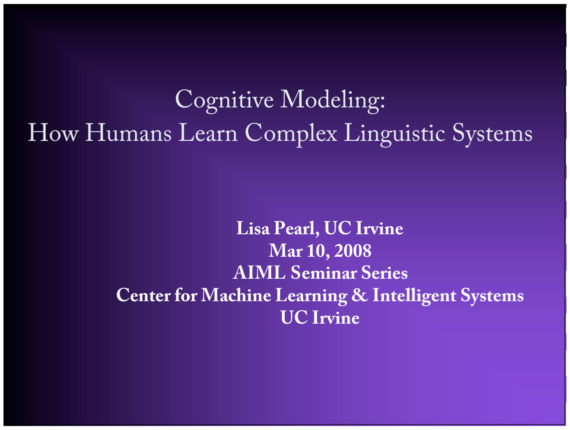
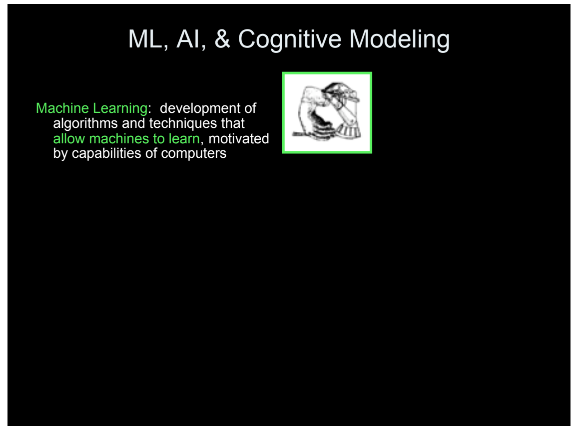
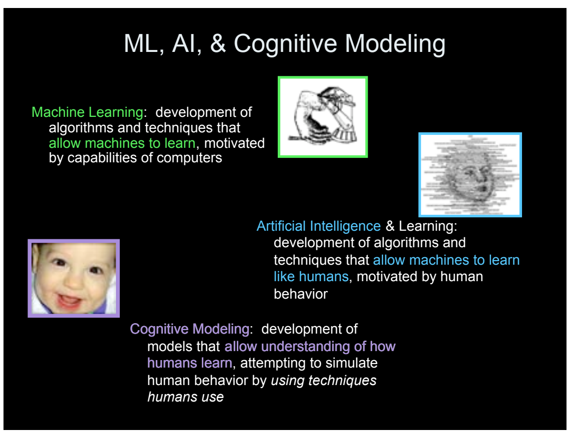
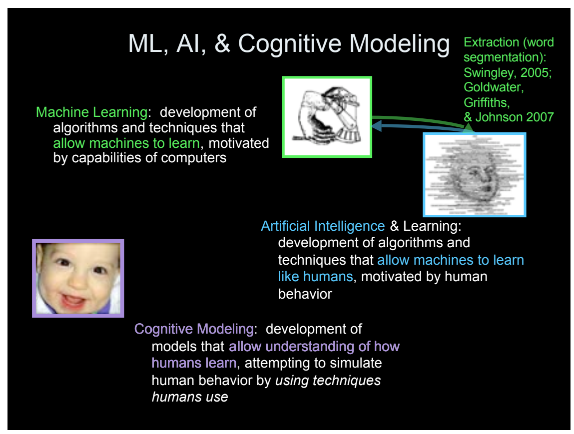
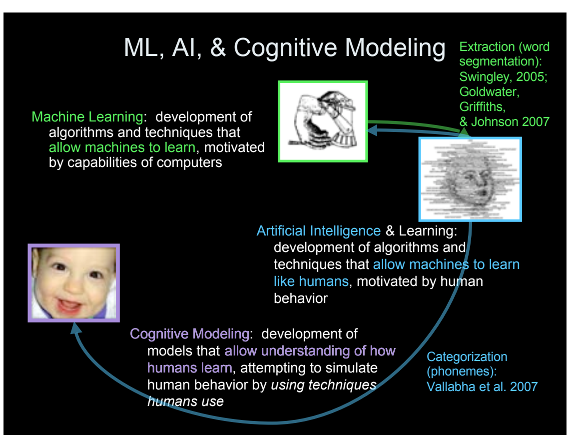
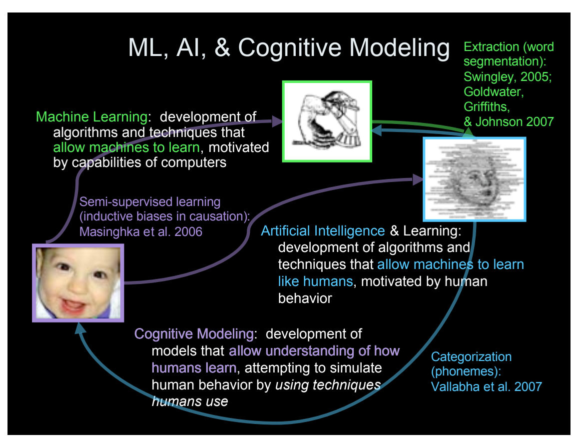

Cognitive Modeling:
How Humans Learn Complex Linguistic Systems
Lisa Pearl, UC Irvine
Mar 10, 2008
AIML Seminar Series
Center for Machine Learning & Intelligent Systems
UC Irvine


ML, AI, & Cognitive Modeling
Machine Learning: development of algorithms and techniques that allow machines to learn, motivated by capabilities of computers


ML, AI, & Cognitive Modeling
Machine Learning: development of algorithms and techniques that allow machines to learn, motivated by capabilities of computers
Artificial Intelligence & Learning: development of algorithms and techniques that allow machines to learn like humans, motivated by human behavior
Cognitive Modeling: development of models that allow understanding of how humans learn, attempting to simulate human behavior by using techniques humans use


ML, AI, & Cognitive Modeling
Extraction (word segmentation): Swingley, 2005; Goldwater, Griffiths, & Johnson 2007
Machine Learning: development of algorithms and techniques that allow machines to learn, motivated by capabilities of computers
Artificial Intelligence & Learning: development of algorithms and techniques that allow machines to learn like humans, motivated by human behavior
Cognitive Modeling: development of models that allow understanding of how humans learn, attempting to simulate human behavior by using techniques humans use

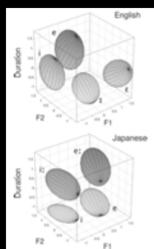

ML, AI, & Cognitive Modeling
Extraction (word segmentation): Swingley, 2005; Goldwater, Griffiths, & Johnson 2007
Machine Learning: development of algorithms and techniques that allow machines to learn, motivated by capabilities of computers
Artificial Intelligence & Learning: development of algorithms and techniques that allow machines to learn like humans, motivated by human behavior
Cognitive Modeling: development of models that allow understanding of how humans learn, attempting to simulate human behavior by using techniques humans use
Categorization (phonemes): Vallabha et al. 2007


ML, AI, & Cognitive Modeling
Extraction (word segmentation): Swingley, 2005; Goldwater, Griffiths, & Johnson 2007
Machine Learning: development of algorithms and techniques that allow machines to learn, motivated by capabilities of computers
Semi-supervised learning (inductive biases in causation): Masinghka et al. 2006
Artificial Intelligence & Learning: development of algorithms and techniques that allow machines to learn like humans, motivated by human behavior
Cognitive Modeling: development of models that allow understanding of how humans learn, attempting to simulate human behavior by using techniques humans use
Categorization (phonemes): Vallabha et al. 2007

Cognitive Modeling of Language

Different problems: more and less easily discernible from data

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



Vowel categories in English & Japanese

Hypothesis space: 3 dimensions of variation

English
  relevant dimensions: 1 and 2

Japanese
  relevant dimensions: 2 and 3

Vallabha et al. 2007

Cognitive Modeling of Language

Different problems: more and less easily discernible from data

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

Extraction
  Ex: Where are words in fluent speech?

Who's afraid of the big bad wolf?

huw zə frej dəv ðə bɪg bæd wəlf
who 'sa frai dof the big bad wolf

huw zə frejd əv ðə bɪgbædwəlf
who 'sa fraid of the bigbadwolf

húwzə fréjdəvðə bɪ́g bæ'd wə'lf
who'sa fraidofthe big bad wolf

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

Assumption from experimental work:
Relevant unit of word segmentation
for infants is the syllable

Swingley 2005

Gambell & Yang 2006

Cognitive Modeling of Language

Different problems: more and less easily discernible from data

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

Extraction
  Ex: Where are words in fluent speech?

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

regularity

blink~blinked blɪŋk blɪŋkt	ping~pinged pɪŋ pɪŋd	confide~confided kənfajd kənfajdəd

irregularity

drink~drank drɪŋk drejŋk	sing-sang sɪŋ sejŋ	hide-hid hajd hɪd
think~thought θɪŋk θɔt		

Cognitive Modeling of Language

Different problems: more and less easily discernible from data

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

Extraction
  Ex: Where are words in fluent speech?

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

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
Generative system: syntax

Cognitive Modeling of Language

Different problems: more and less easily discernible from data

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

Extraction
  Ex: Where are words in fluent speech?

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

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
Generative system: syntax

English
Subject Verb Object

Kannada
Subject t_{Object} Verb Object

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

Cognitive Modeling of Language

Different problems: more and less easily discernible from data

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

Extraction
  Ex: Where are words in fluent speech?

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

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

Observable data: stress contour   EMphasis
Generative system: metrical phonology

Cognitive Modeling of Language

Different problems: more and less easily discernible from data

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

Extraction
Ex: Where are words in fluent speech?

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

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

Observable data: stress contour EMphasis
Generative system: metrical phonology

(S S) S
EM pha sis

(H L) H
EM pha sis

(S S S)
EM pha sis

(H L L)
EM pha sis

Cognitive Modeling of Language

Different problems: more and less easily discernible from data

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

Extraction
Ex: Where are words in fluent speech?

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

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

Today's focus

Road Map

Introduction to 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

Introduction to 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

Levels of abstract structure

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

(S S S)
EM pha sis

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

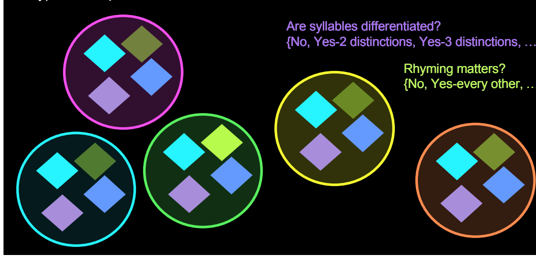
Hypothesis for a language consists of a combination of generalizations about that language (grammar). But this leads to a theoretically infinite hypothesis space.

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.

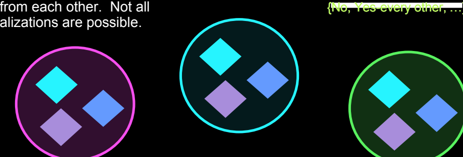
Which syllable of a larger unit is stressed?
{Leftmost, Rightmost, ~~Second from Left,…}~~

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.

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

Chomsky (1981), Halle & Vergnaud (1987)

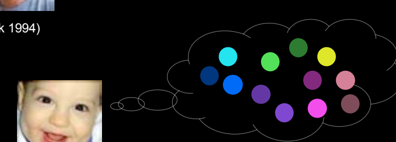
Learning Parametric Linguistic Systems

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

(Clark 1994)

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

Exponentially growing hypothesis space



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.

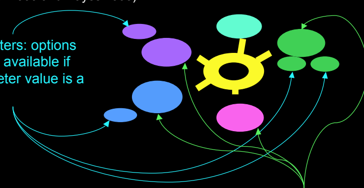
system parameters of variation - to be determined by learner from available data

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)

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



Most parameters involve metrical foot formation

All combine to generate stress contour output

A Brief Tour of Parametric Metrical Phonology

Are syllables differentiated?

No: system is quantity-insensitive (QI)

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

krɔs
crous
Syllable
onset — kr
rime
nucleus — ɔ
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 — narrowing of hypothesis space
Only allowed number of divisions: 2
Heavy vs. Light

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?

(...)
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 →   ← Ft Dir Right

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

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)
(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

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)

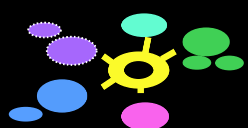
(H) (L L) (H)

(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

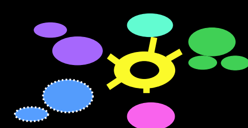
Are syllables differentiated?

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

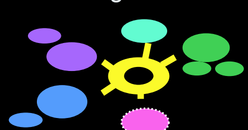
Are any syllables extrametrical?

Yes.

Rightmost syllable is not included in metrical foot.

(	…)	
H	L	H
VC	CV	CVC
em	pha	sis

Generating a Stress Contour



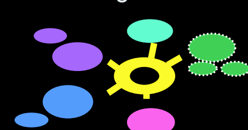
Process speaker uses to generate stress contour

Which direction are feet constructed from?

From the right.

H	L)	H
VC	CV	CVC
em	pha	sis

Generating a Stress Contour



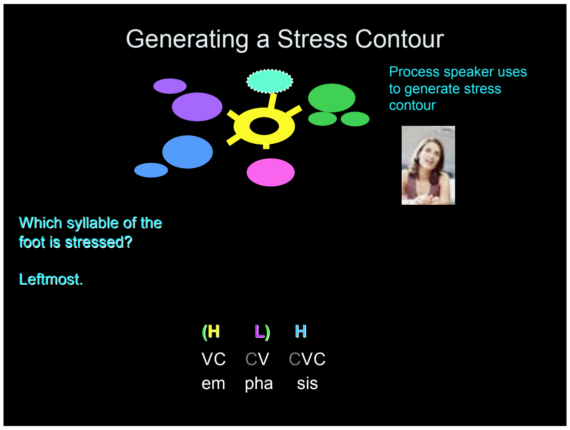
Process speaker uses to generate stress contour

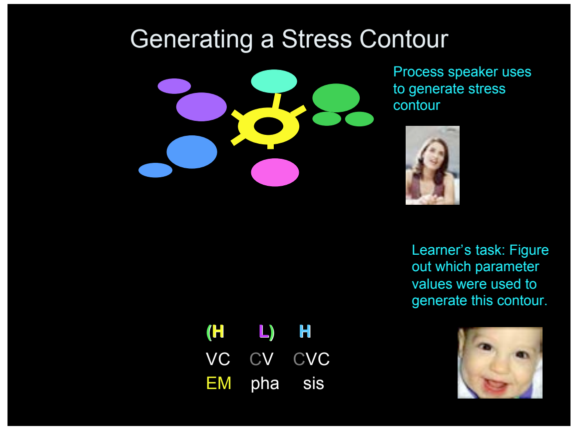
Are feet unrestricted?

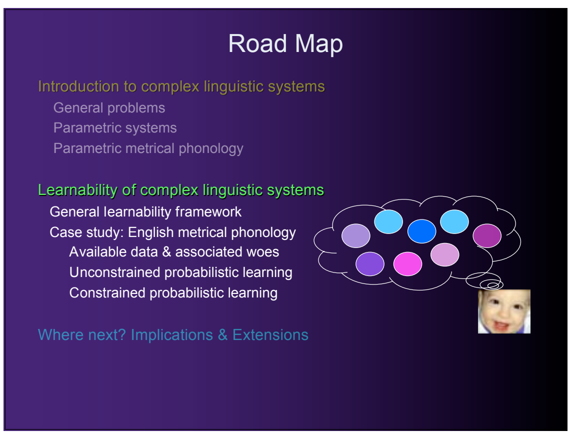
No.

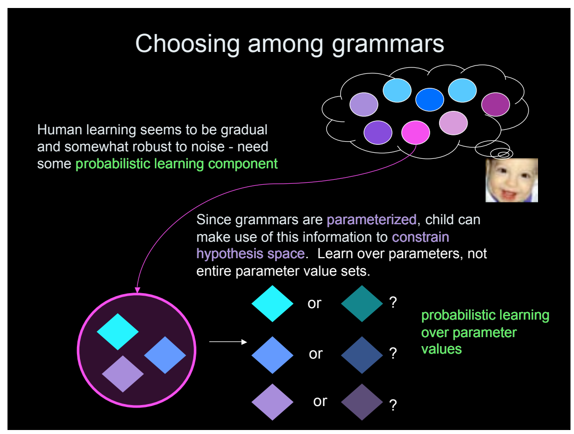
2 syllables per foot.

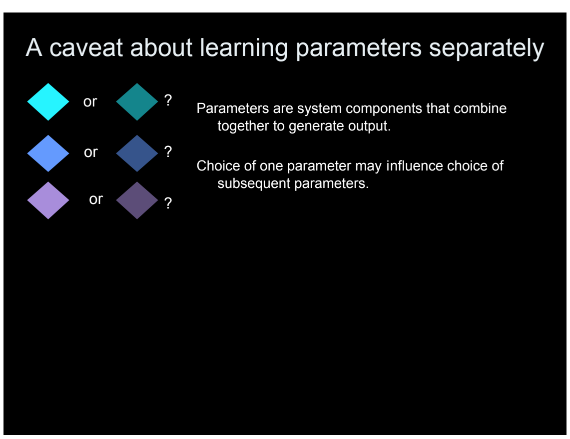
(H	L)	H
VC	CV	CVC
em	pha	sis

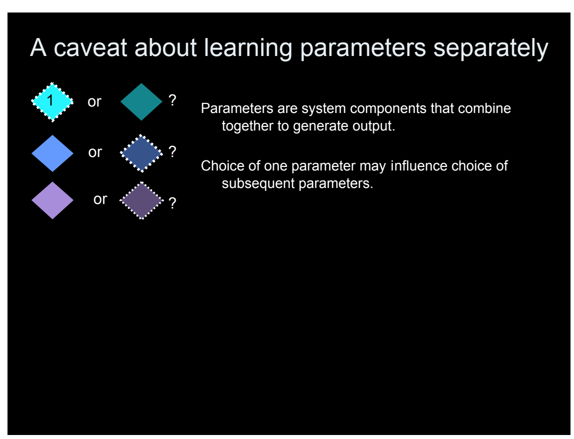

Generating a Stress Contour
Process speaker uses to generate stress contour
Which syllable of the foot is stressed?
Leftmost.
(H L) H
VC CV CVC
em pha sis

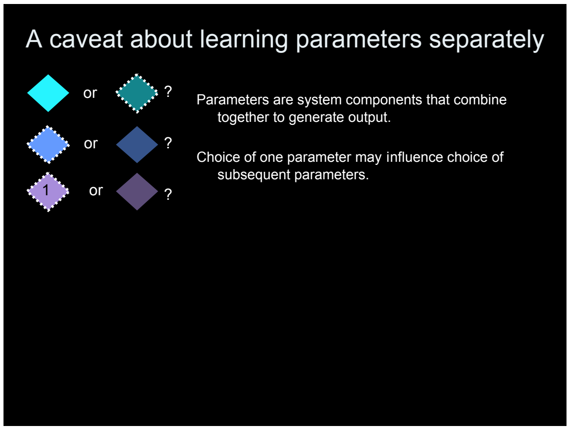

Generating a Stress Contour
Process speaker uses to generate stress contour
Learner's task: Figure out which parameter values were used to generate this contour.
(H L) H
VC CV CVC
EM pha sis

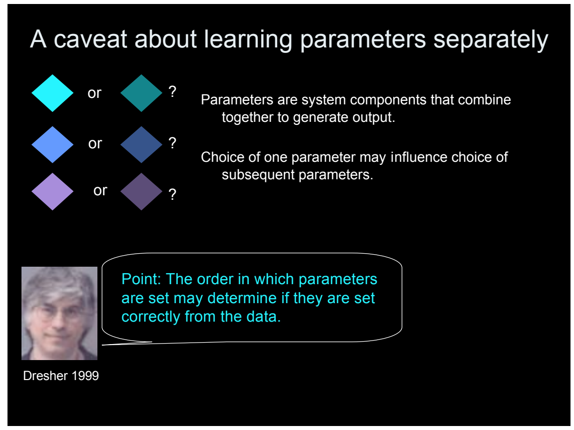

Road Map
Introduction to complex linguistic systems
General problems
Parametric systems
Parametric metrical phonology
Learnability of complex linguistic systems
General learnability framework
Case study: English metrical phonology
Available data & associated woes
Unconstrained probabilistic learning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

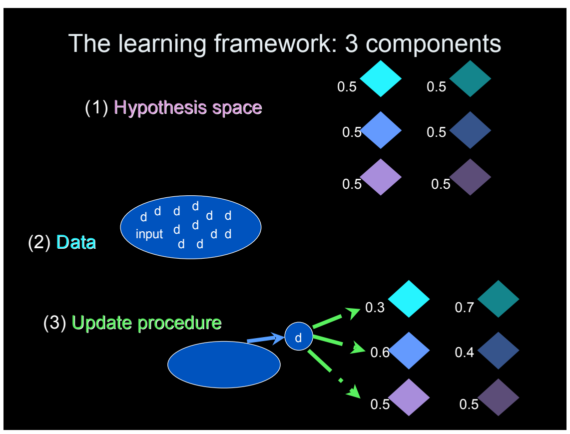

Choosing among grammars
Human learning seems to be gradual and somewhat robust to noise - need some probabilistic learning component
Since grammars are parameterized, child can make use of this information to constrain hypothesis space. Learn over parameters, not entire parameter value sets.
or ?
or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

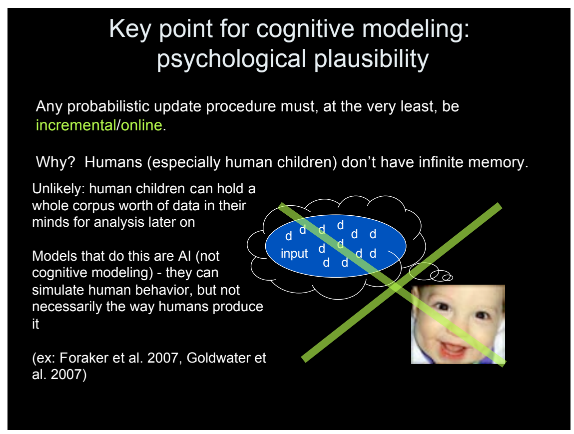

A caveat about learning parameters separately
or ?
or ?
or ?
Parameters are system components that combine together to generate output.
Choice of one parameter may influence choice of subsequent parameters.

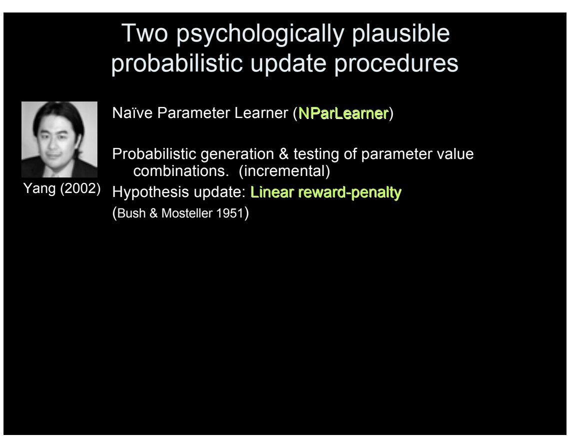

A caveat about learning parameters separately
1
or ?
or ?
or ?
Parameters are system components that combine together to generate output.
Choice of one parameter may influence choice of subsequent parameters.

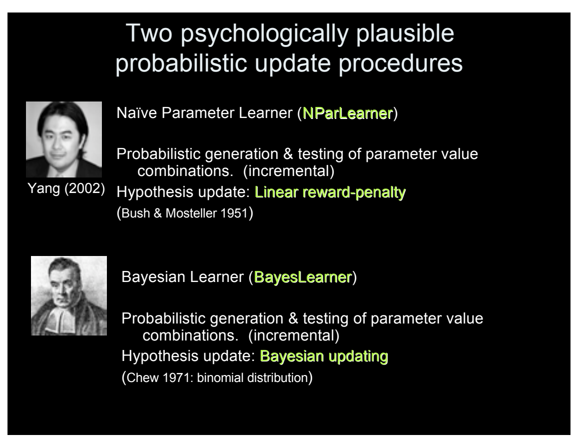

A caveat about learning parameters separately
or ?
or ?
1 or ?
Parameters are system components that combine together to generate output.
Choice of one parameter may influence choice of subsequent parameters.


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 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
Models that do this are AI (not cognitive modeling) - they can simulate human behavior, but not necessarily the way humans produce it
(ex: Foraker et al. 2007, Goldwater et al. 2007)
input d d d d d d d d d d d


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


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

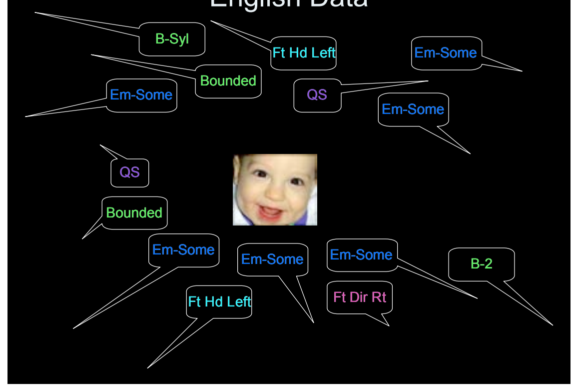
QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hd Left

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

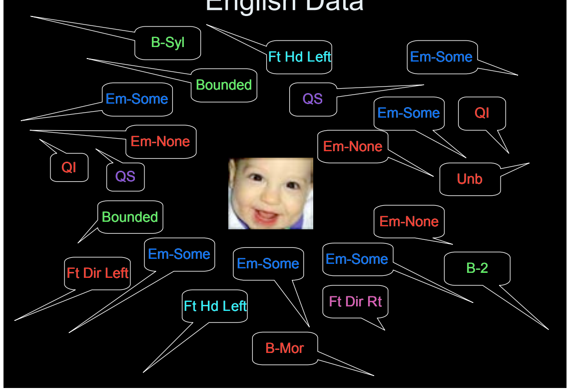
Total Words: 540505 Mean Length of Utterance: 3.5

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)

English Data



English Data



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

$p_{v1} = p_{v1} + (1 - p_{v1})$ $\quad$ $p_{v1} = (1-\gamma)p_{v1}$

$p_{v2} = 1 - p_{v1}$ $\quad$ $p_{v2} = 1 - p_{v1}$

reward v1 $\quad$ punish 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

$p_{v1} = p_{v1} + \gamma(1 - p_{v1})$ $\quad$ $p_{v1} = (1-\gamma)p_{v1}$

$p_{v2} = 1 - p_{v1}$ $\quad$ $p_{v2} = 1 - p_{v1}$

reward v1 $\quad$ punish v1

BayesLearner: Bayesian update of binomial distribution (Chew 1971)

Parameters α, β:

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

here: $\alpha = \beta = 0.5$

Parameter value v

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

reward: success + 1 $\quad$ 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
QSVCL = 0.6
Em-Some = 0.1

QS = 0.7
QSVCH = 0.4
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: 1,160,000 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: 1,160,000 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

$p_{v1} = p_{v1} + \gamma(1 - p_{v1})$ $\quad$ $p_{v1} = (1-\gamma)p_{v1}$

$p_{v2} = 1 - p_{v1}$ $\quad$ $p_{v2} = 1 - p_{v1}$

reward v1 $\quad$ punish 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 v

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

reward: success + 1 $\quad$ punish: success + 0

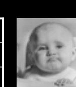
Probabilistic learning for English

Goal: Converge on English values after learning period is over

Learning Period Length: 1,160,000 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: 1,160,000 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 (Jusczyk, Cutler, & Redanz (1993) and QS (Turk, Jusczyk, & Gerken (1995).

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

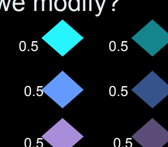
QS, QSVCH, Em-Some, Em-Right, Ft Dir Right, Bounded, Bounded-2, Bounded-Syllabic, Ft Hd Left

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%

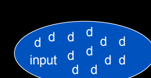
The best isn't so great

Where else can we modify?

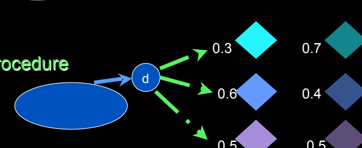
(1) Hypothesis space

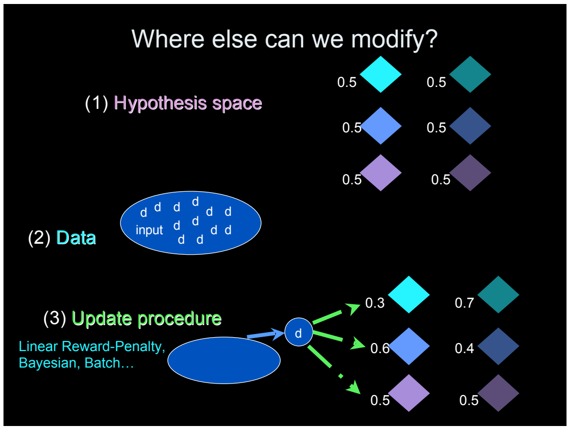


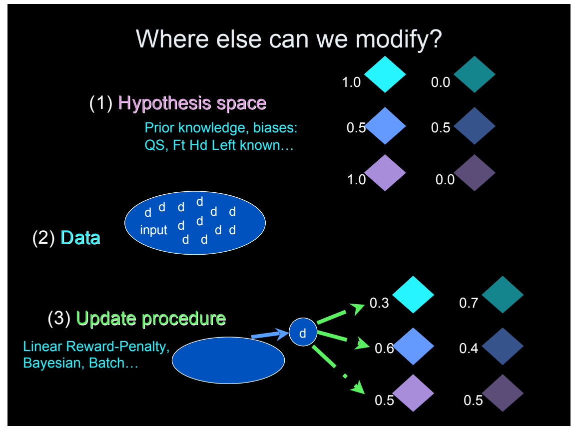
(2) Data

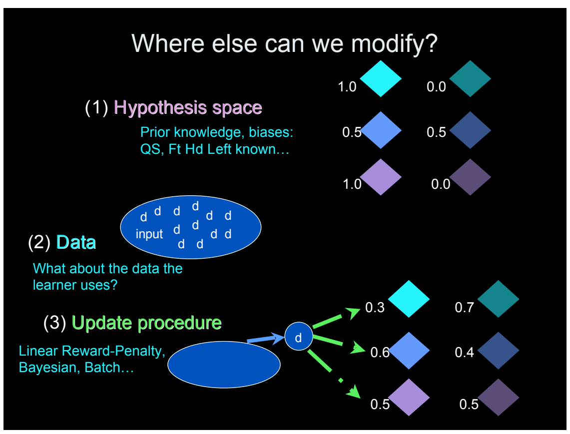


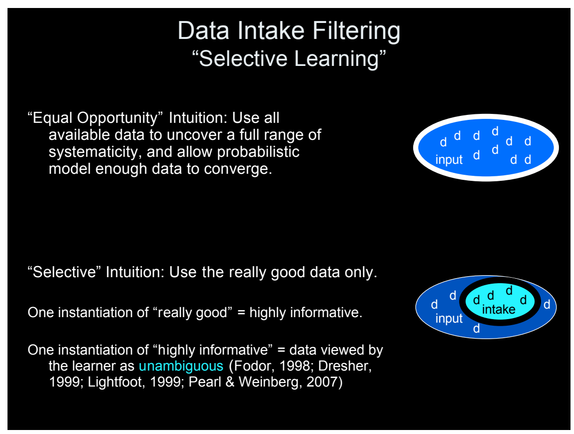
(3) Update procedure

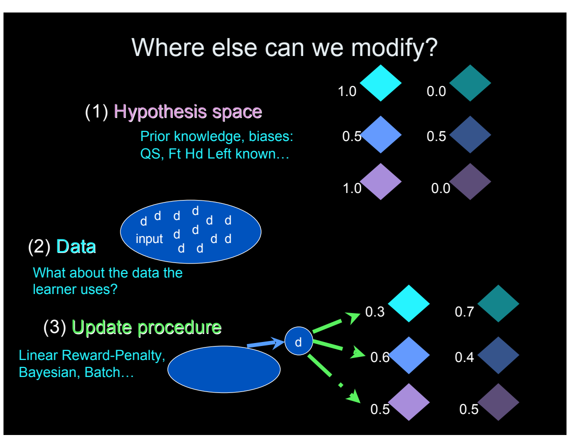


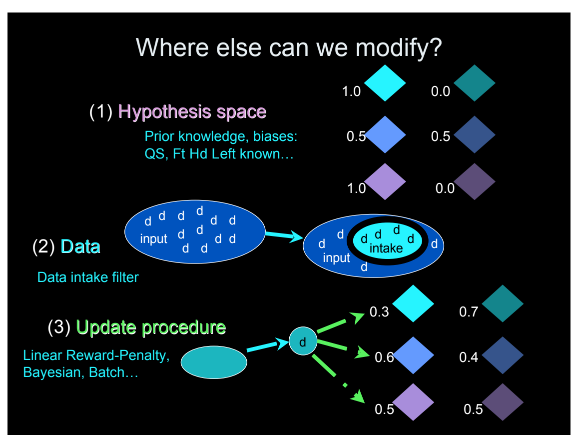

Where else can we modify?
(1) Hypothesis space
0.5
0.5
0.5
0.5
0.5
0.5
(2) Data
input
d d d d d d d d d d
(3) Update procedure
Linear Reward-Penalty,
Bayesian, Batch…
d
0.3
0.7
0.6
0.4
0.5
0.5


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
input
d d d d 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?
input
d d d d 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.
input
d d d d d d d d d
"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)
input
intake
d d d d d d d


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?
input
d d d d 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
Data intake filter
input
d d d d d d d d d d
input
intake
d 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

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

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.

S…S af ter noon → Em-None

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

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

Practical matters: Feasibility of unambiguous data

Existence? Depends on data set (empirically determined).

Identification?

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

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

S…S af ter noon → Em-None

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

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

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.

input

intake intake

has advantage

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.

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

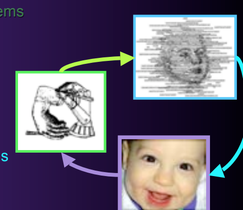
The parameters are freely ordered w.r.t. each other within each group.

Road Map

Introduction to 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

Cognitive modeling: aimed at understanding how humans solve problems, generating human behavior by using psychologically plausible methods

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

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

0.7 0.3

0.5 0.5

0.8 0.2

Bias on data:
interpretive bias to use highly informative data

input

intake intake

Where we can go

(1) Interpretive bias:

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

Are there other methods of implementing interpretative biases that lead to successful learning?

How necessary is an interpretive bias? Are there cleverer probabilistic learning methods than can succeed?

+ biases?

Where we can go

(1) Interpretive bias:

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

Are there other methods of implementing interpretative biases that lead to successful learning?

How necessary is an interpretive bias? Are there cleverer probabilistic learning methods than can succeed?

+ 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]

+ fewer biases?

Where we can go

(1) Interpretive bias:

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

Are there other methods of implementing interpretative biases that lead to successful learning?

How necessary is an interpretive bias? Are there cleverer probabilistic learning methods than can succeed?

+ biases?

(2) Hypothesis space bias:

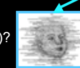
Is it possible to infer the correct parameters of variation given less structured information a priori (e.g. larger units than syllables are required)? [Model Selection]

+ fewer biases?

necessary biases

(3) Informing AI/ML:

Can we import the necessary biases for learning complex systems into language applications (ex: speech generation)?

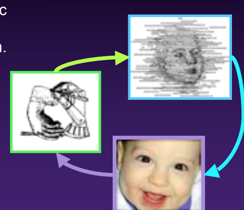


The big idea

Complex linguistic systems may well require something beyond probabilistic methods in order to be learned, and learned as well as humans learn them.

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

What we can do: take insights from cognitive modeling and apply them to problems in artificial intelligence and machine learning, & vice versa



Thank You

Amy Weinberg
Bill Idsardi
Bill Sakas
Jeff Lidz
Charles Yang
Janet Fodor

The audiences at

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