At the Interface of Computational Learning Theory and Human Language Learning

Lisa Pearl
University of Maryland

UC Irvine: Mar 12, 2007

Human Language Learning

Theoretical work:
object of acquisition

IP
XP … VP …
Object Verb

Experimental work:
time course of acquisition

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

The Learning Problem

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

Syntactic System
Observable form: word order
Interference: movement rules

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

The Mechanism of Language Learning: Parameters

Premise: learner considers finite range of hypotheses (parameters)

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

The Mechanism of Language Learning: Extracting Systematicity

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

The Mechanism of Language Learning: Extracting Systematicity

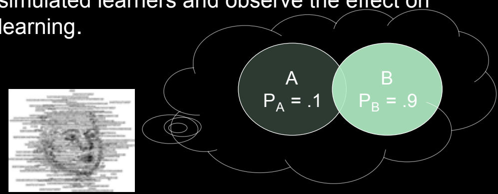
"It is unlikely that any example … would show the effect of only a single parameter value; rather, each example is the result of the interaction of several different principles and parameters" - Clark (1994)

Potential solution: the learner focuses in on an informative subset of the data.

Potential issue: data sparseness

Computational Modeling of Data Intake Filtering

Why? Can easily (and ethically) restrict data intake to simulated learners and observe the effect on learning.



Recent computational modeling surge: Yang, 2000; Sakas & Fodor, 2001; Yang, 2002; Pearl, 2005; Pearl & Weinberg, 2007

The Mechanism of Language Learning: Questions

The Mechanism of Language Learning: Questions

Hypothesis space formation

- What are the hypotheses under consideration?
- Where do these hypotheses come from?

The Mechanism of Language Learning: Questions

Hypothesis space formation

- What are the hypotheses under consideration?
- Where do these hypotheses come from?

Data Intake

- What data are used for learning?
- How are different data weighted by the learner?

The Mechanism of Language Learning: Questions

Hypothesis space formation

- What are the hypotheses under consideration?
- Where do these hypotheses come from?

Data Intake

- What data are used for learning?
- How are different data weighted by the learner?

Finding Systematicity

- Where/how is systematicity found, especially in the face of noise (exceptions, ambiguity)?

The Mechanism of Language Learning: Questions

Hypothesis space formation

- What are the hypotheses under consideration?
- Where do these hypotheses come from?

Data Intake

- What data are used for learning?
- How are different data weighted by the learner?

Finding Systematicity

- Where/how is systematicity found, especially in the face of noise (exceptions, ambiguity)?

Road Map

Learning Framework Overview

Computational Work:

- Data intake filtering and systematicity in metrical phonology (synchronic)
- Data intake filtering in syntax (diachronic)

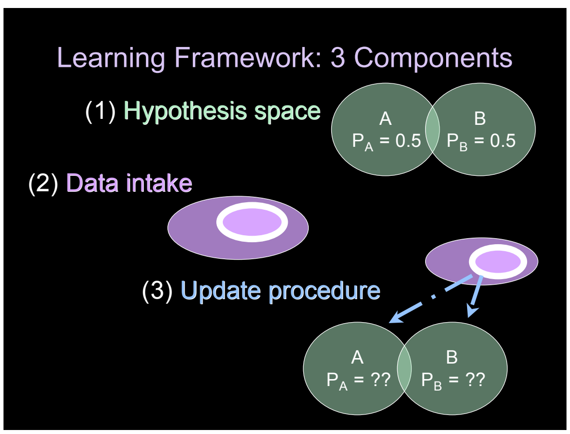
Road Map

Learning Framework Overview

Computational Work:

- Data intake filtering and systematicity in metrical phonology (synchronic)
- Data intake filtering in syntax (diachronic)

Important Feature: Case studies grounded in empirical data
- real data distributions
- searching realistic data space for evidence of underlying system



Benefits of Learning Framework

Components:
(1) hypothesis space (2) data intake (3) update procedure

Application to a wide range of learning problems, provided these three components are defined

Ex: hypothesis space defined in terms of parameter values (Yang, 2002) or in terms of how much structure is posited for the language (Perfors, Tenenbaum, & Regier, 2006)

Can combine discrete representations (hypothesis space) with probabilistic components (update procedure)

The Hypothesis Space & The Update Procedure

Hypothesis Space: theoretical and experimental work on what hypotheses children entertain (ex: Lidz, Waxman, & Freedman, 2003; Thornton & Crain, 1999; Hamburger & Crain, 1984)

Update Procedure: recent experimental work on probabilistic learning as feasible in adults (Tenenbaum, 2000; Thompson & Newport, 2007) and infants (Newport & Aslin, 2004; Gerken, 2006).

The Hypothesis Space & The Update Procedure

Hypothesis Space: theoretical and experimental work on what hypotheses children entertain (ex: Lidz, Waxman, & Freedman, 2003; Thornton & Crain, 1999; Hamburger & Crain, 1984)

Update Procedure: recent experimental work on probabilistic learning as feasible in adults (Tenenbaum, 2000; Thompson & Newport, 2007) and infants (Newport & Aslin, 2004; Gerken, 2006).

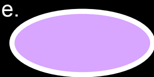
Bayesian updating

Infers likelihood of given hypothesis, given data.

Amount of probability shifted depends on layout of hypothesis space.

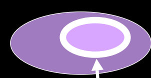
Investigating Data Intake Filtering

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



input

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



subset of input

Modeling Case Studies of Data Intake Filters

Case One: Synchronic Metrical Phonology
- Hypothesis Space: parameters
- Update Procedure: Bayesian updating
- Difficult Features: multiple interactive parameters; noisy input

Case Two: Diachronic Syntax
- Hypothesis Space: parameters
- Update Procedure: Bayesian updating
- Difficult Feature: adult target state is a probability distribution

Data Intake Filtering: The Big Questions

(1) Is it feasible to filter?
Can we filter and get success?

(2) Is it necessary to filter?
Must we filter to get success?

Data Intake Filtering: The Big Questions

(1) Is it feasible to filter?
Can we filter and get success?

(2) Is it necessary to filter?
Must we filter to get success?

Road Map

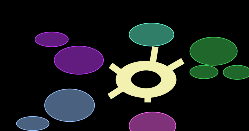
Learning Framework Overview

Computational Work: Case Studies
- Data intake filtering and systematicity in metrical phonology (synchronic)
  - Finding unambiguous data in a complex system: cues vs. parsing
  - Metrical phonology overview: interacting parameters
  - Cues vs. parsing in metrical phonology
  - English metrical phonology
  - Logical problem of language acquisition
  - Filter feasibility & constraints on parameter-setting orders
- Data intake filtering in syntax (diachronic)

Filter Feasibility

How feasible is an unambiguous data filter in a complex system?
- Data sparseness: are there unambiguous data? (Clark 1992)
- How could a learner identify such data?

Metrical phonology (9 interacting parameters)



Interactive Parameters

The order in which parameters are set may determine if they are set correctly (Dresher, 1999): parameter-setting influences perception of unambiguous data.

Identifying unambiguous data:

Cues (Dresher, 1999; Lightfoot, 1999)

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

Cues vs. Parsing: Overview

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

XP Obj Verb … → OV

Parsing tries to analyze a data point with "all possible parameter value combinations", conducting an "exhaustive search of all parametric possibilities." (Fodor, 1998)

Subj Verb Obj → OV, +V2 / VO, +V2 / VO, -V2

Cues vs. Parsing: Comparison

	Cues	Parsing
Easy identification of unambiguous data	+	
Can find information in datum sub-part	+	
Can tolerate exceptions	+	
Is not heuristic		+
Does not require additional knowledge		+
Does not use default values		+

Cues vs. Parsing in a Probabilistic Framework

"Both models ... cannot capture the variation in and the gradualness of language development…when a parameter *is* set, it is set in an all-or-none fashion." - Yang (2002)

Benefit of using learning framework to sidestep this problem - separable components used in combination:

(1) cues/parsing to identify unambiguous data

(2) probabilistic framework of gradual updating based on unambiguous data

Road Map

Learning Framework Overview

Computational Work: Case Studies

Data intake filtering and systematicity in metrical phonology (synchronic)

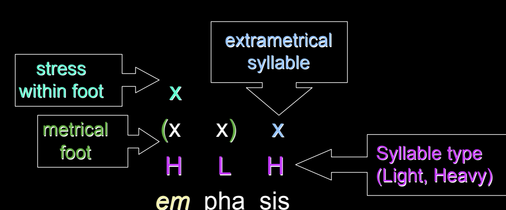
- Finding unambiguous data in a complex system: cues vs. parsing
- Metrical phonology overview: interacting parameters
- Cues vs. parsing in metrical phonology
- English metrical phonology
- Logical problem of language acquisition
- Filter feasibility & constraints on parameter-setting orders

Data intake filtering in syntax (diachronic)

Metrical Phonology

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

sample metrical phonology structure



Why Parameters?

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

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

Why Parameters?

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

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

(1) If word-by-word association, expect piece-meal change over time at the individual word level. Instead, historical linguists posit changes to underlying *systems* to best explain the observed data.

Why Parameters?

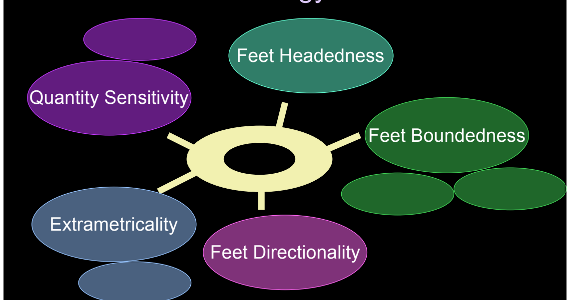
Why posit parameters instead of just associating stress contours with words?

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

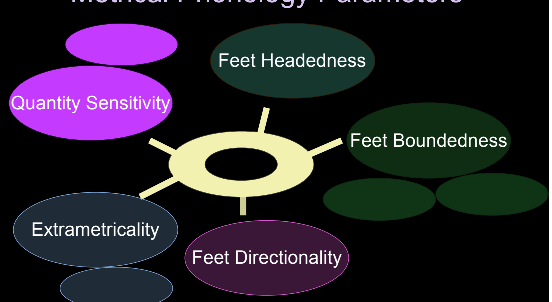
(1) If word-by-word association, expect piece-meal change over time at the individual word level. Instead, historical linguists posit changes to underlying *systems* to best explain the observed data.

(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.

Metrical Phonology Parameters



Metrical Phonology Parameters



Quantity Sensitivity: QI

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

S	S	S
VV	V	VC
CVV	CV	CCVC
lu	di	crous

Quantity Sensitivity: QS

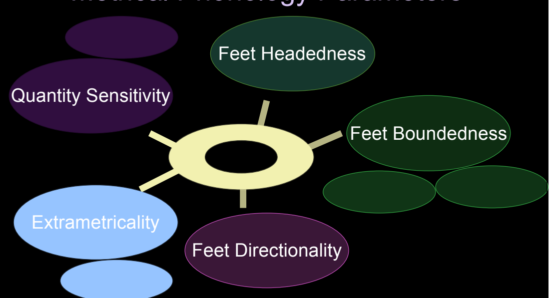
Quantity-**S**ensitive (**QS**):
Syllables are separated into **L**ight and **H**eavy
V are always L, VV are always H
VC-**L**ight (**QSVCL**) = **VC** syllable is **L**
VC-**H**eavy (**QSVCH**) = **VC** syllable is **H**

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

Quantity Sensitivity: Stress

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

Metrical Phonology Parameters



Extrametricality, Metrical Feet, and Stress

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

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

Extrametricality: None

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

metrical foot ⇒ **(L** **L)** **(H)**
VC VC VV
af ter **noon**

Extrametricality: Some

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

extrametrical syllable ⇒ **L** **(H** **L)** ⇐ metrical foot
V VC V
a **gen** da

Extrametricality: Some

Extrametricality-**Some (Em-Some): One edge syllable not in foot**

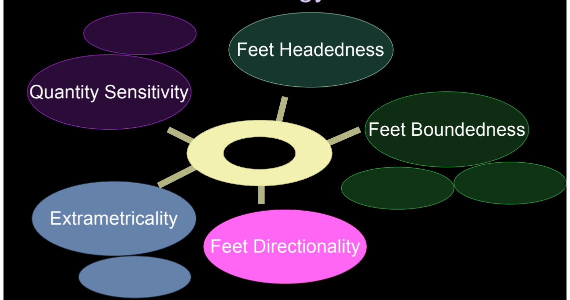
Extrametricality-**Right (Em-Right)**: **Rightmost syllable** not in foot - ***cannot* have stress**

metrical foot → (H L) H ← extrametrical syllable

VV V VC

lu di crous

Metrical Phonology Parameters



Feet Directionality

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

Feet Direction Left: start from left edge

H L H

Feet Direction Right: start from right edge

H L H

Feet Directionality

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

Feet Direction Left: start from left edge

(H L) H

Feet Direction Right: start from right edge

H L H

Feet Directionality

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

Feet Direction Left: start from left edge

(H L) (H)

Feet Direction Right: start from right edge

H L H

Feet Directionality

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

Feet Direction Left: start from left edge

(H L) (H)

Feet Direction Right: start from right edge

H (L H)

Feet Directionality

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

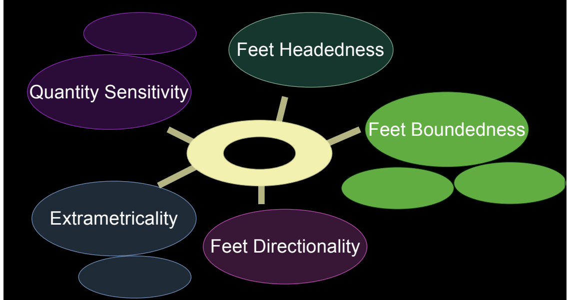
Feet Direction Left: start from left edge

(H L) (H)

Feet Direction Right: start from right edge

(H) (L H)

Metrical Phonology Parameters



Boundedness: Unbounded Feet

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

Boundedness: Unbounded Feet

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

start from left → L L L H L

Boundedness: Unbounded Feet

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

start from left → (L L L) H L

Boundedness: Unbounded Feet

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

start from left → (L L L)(H L)

Boundedness: Unbounded Feet

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

start from left ⇒ (L L L)(H L)

L L L H L ⇐ start from right

Boundedness: Unbounded Feet

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

start from left ⇒ (L L L)(H L)

L L L H (L) ⇐ start from right

Boundedness: Unbounded Feet

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

start from left ⇒ (L L L)(H L)

(L L L H)(L) ⇐ start from right

Boundedness: Unbounded Feet

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

start from left ⇒ (L L L)(H L)

(L L L H)(L) ⇐ start from right

start from left ⇒ L L L L L

Boundedness: Unbounded Feet

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

start from left ⇒ (L L L)(H L)

(L L L H)(L) ⇐ start from right

start from left ⇒ (L L L L L)

Boundedness: Unbounded Feet

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

start from left ⇒ (L L L)(H L)

(L L L H)(L) ⇐ start from right

start from left ⇒ (L L L L L)

(L L L L L) ⇐ start from right

Boundedness: Bounded Feet

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

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

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

Boundedness: Bounded Feet

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

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

start from left ⇒ x x x x x

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

Boundedness: Bounded Feet

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

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

start from left ⇒ (x x)(x x)(x)

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

Boundedness: Bounded Feet

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

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

start from left ⇒ (x x)(x x)(x)

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

start from left ⇒ x x x x x

Boundedness: Bounded Feet

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

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

start from left ⇒ (x x)(x x)(x)

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

start from left ⇒ (x x x)(x x)

Boundedness: Bounded Feet

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

Bounded-**Mor**aic: counting unit is **mora**
H = **2** moras, **L** = **1** mora

Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is **syllable**

start from left → L H L L H

bounded-2 →

Bounded-Moraic: counting unit is **mora**

H = 2 moras, L = 1 mora

Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is **syllable**

start from left → (L H)(L L)(H)

bounded-2 →

Bounded-Moraic: counting unit is **mora**

H = 2 moras, L = 1 mora

Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is **syllable**

start from left → (L H)(L L)(H)

bounded-2 → H H L L H

Bounded-Moraic: counting unit is **mora**

H = 2 moras, L = 1 mora

Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is **syllable**

start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

Bounded-Moraic: counting unit is **mora**

H = 2 moras, L = 1 mora

Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is **syllable**

start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

S S S S S

Bounded-Moraic: counting unit is **mora**

H = 2 moras, L = 1 mora

Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is **syllable**

start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

(S S)(S S)(S)

Bounded-Moraic: counting unit is **mora**

H = 2 moras, L = 1 mora

Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is syllable

start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

(S S)(S S)(S)

Bounded-Moraic: counting unit is mora

H = 2 moras, L = 1 mora

start from left → x x x x x x x x

bounded-2 → H H L L H

Boundedness: Bounded Feet

Bounded-Syllabic: counting unit is syllable

start from left → (L H)(L L)(H)

bounded-2 → (H H)(L L)(H)

(S S)(S S)(S)

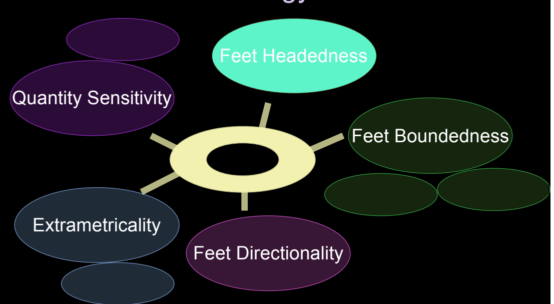
Bounded-Moraic: counting unit is mora

H = 2 moras, L = 1 mora

start from left → (x x)(x x) (x x) (x x)

bounded-2 → (H)(H) (L L)(H)

Metrical Phonology Parameters



Feet Headedness

Feet Headedness: which syllable of metrical foot gets stress

Feet Head Left: leftmost syllable in foot gets stress

(H) (L H)

Feet Head Right: rightmost syllable in foot gets stress

(H) (L H)

Feet Headedness

Feet Headedness: which syllable of metrical foot gets stress

Feet Head Left: leftmost syllable in foot gets stress

(H) (L H)

Feet Head Right: rightmost syllable in foot gets stress

(H) (L H)

Feet Headedness

Feet Headedness: which syllable of metrical foot gets stress

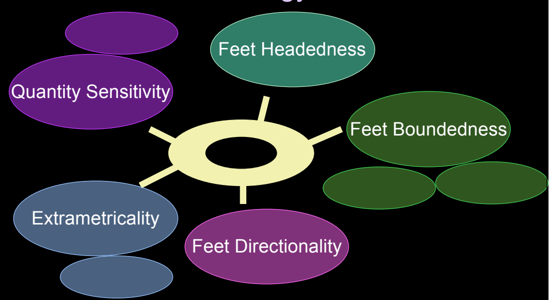
Feet Head Left: leftmost syllable in foot gets stress

(H) (L H)

Feet Head Right: rightmost syllable in foot gets stress

(H) (L H)

Metrical Phonology Parameters



Road Map

Learning Framework Overview

Computational Work: Case Studies

Data intake filtering and systematicity in metrical phonology (synchronic)

- Finding unambiguous data in a complex system: cues vs. parsing
- Metrical phonology overview: interacting parameters
- Cues vs. parsing in metrical phonology
- English metrical phonology
- Logical problem of language acquisition
- Filter feasibility & constraints on parameter-setting orders

Data intake filtering in syntax (diachronic)

Cues for Metrical Phonology Parameters

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

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

Parsing with Metrical Phonology Parameters

parse data with all available values of all parameters
(values cease to be available when one value is chosen as the correct one for the language - the other value(s) is(are) then unavailable)

If only one value for a parameter leads to a successful parse of the datum (e.g. "Extrametrical None"), that datum is considered unambiguous for that parameter value.

Parsing with Metrical Phonology Parameters

Sample Datum: VC VC VV ('afternoon')

Parsing with Metrical Phonology Parameters

Sample Datum: VC VC VV ('afternoon')

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

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

Parsing with Metrical Phonology Parameters

Sample Datum: VC VC VV ('afternoon')

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

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

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

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

Parsing with Metrical Phonology Parameters

Sample Datum: VC VC VV ('afternoon')

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

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

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

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

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

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

Parsing with Metrical Phonology Parameters

Values leading to successful parses of datum:
(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)

Datum is unambiguous for Em-None.

Parsing with Metrical Phonology Parameters

Values leading to successful parses of datum:
(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)

Datum is unambiguous for Em-None.

If QI already set, datum is unambiguous for Em-None, B, B-2, and B-Syl.

Road Map

Learning Framework Overview

Computational Work: Case Studies
- Data intake filtering and systematicity in metrical phonology (synchronic)
  - Finding unambiguous data in a complex system: cues vs. parsing
  - Metrical phonology overview: interacting parameters
  - Cues vs. parsing in metrical phonology
  - English metrical phonology
  - Logical problem of language acquisition
  - Filter feasibility & constraints on parameter-setting orders
- Data intake filtering in syntax (diachronic)

Finding Unambiguous Data: English Metrical Phonology

Non-trivial system: metrical phonology

Non-trivial language: English (full of exceptions)
- exceptions: data unambiguous for the *incorrect* value in the adult system

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

Logical problem of language acquisition: Are there any viable parameter-setting orders using unambiguous data (found with cues or parsing)?

Empirical Grounding in Realistic Data: Estimating English Data Distributions

Caretaker speech to children between the ages of 6 months and 2 years (CHILDES: MacWhinney, 2000)

Total Words: 540505
Mean Length of Utterance: 3.5

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

Road Map

Learning Framework Overview

Computational Work: Case Studies

- Data intake filtering and systematicity in metrical phonology (synchronic)
  - Finding unambiguous data in a complex system: cues vs. parsing
  - Metrical phonology overview: interacting parameters
  - Cues vs. parsing in metrical phonology
  - English metrical phonology
  - Logical problem of language acquisition
  - Filter feasibility & constraints on parameter-setting orders
- Data intake filtering in syntax (diachronic)

Viable Parameter-Setting Orders:

Encapsulating the Knowledge for Acquisition Success

Viable orders are derived for each method (cues and parsing) via an exhaustive walk through all possible parameter-setting orders.

Worst Case: No orders lead to correct system

Slightly Better Case: Viable orders available, but fairly random

Better Case: Viable orders available, can be captured by small number of ***order constraints***

Best Case: All orders lead to correct system

Identifying Viable Parameter-Setting Orders

(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.

(d) Compare final set of values to English set of values. If they match, this is a viable parameter-setting order.

(e) Repeat (a-d) for all parameter-setting orders.

Road Map

Learning Framework Overview

Computational Work: Case Studies

- Data intake filtering and systematicity in metrical phonology (synchronic)
  - Finding unambiguous data in a complex system: cues vs. parsing
  - Metrical phonology overview: interacting parameters
  - Cues vs. parsing in metrical phonology
  - English metrical phonology
  - Logical problem of language acquisition
  - Filter feasibility & constraints on parameter-setting orders
- Data intake filtering in syntax (diachronic)

Cues: Parameter-Setting Orders

Cues: Sample viable orders

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

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

Cues: Sample failed orders

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

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

Parsing: Parameter-Setting Orders

Parsing: Sample viable orders

(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

Parsing: Sample failed orders

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

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

Cues vs. Parsing: Order Constraints

Cues

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

(b) Em-Right
before Bounded-Syl

(c) Bounded-2
before Bounded-Syl

The rest of the parameters are freely ordered with respect to each other.

Parsing

Group 1:
QS, Ft Head Left, Bounded

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

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

The parameters are freely ordered with respect to each other within each group.

Take Home Message: Feasibility of the Unambiguous Data Filter

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

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

Take Home Message: Feasibility of the Unambiguous Data Filter

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

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

(1) Unambiguous data can be identified in sufficient quantities to extract the correct systematicity.

(2) This filter is robust across a realistic (highly ambiguous, exception-filled) data set.

Cues vs. Parsing Again

Is there any (additional) reason to prefer one method of identifying unambiguous data over the other?

Cues

VV VV L H H

… L L L L

H L L … …S S S S…

S S S…

Parsing

(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)

Deriving 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.

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 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)

(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)

(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)

(b) Em-Right
before Bounded-Syl

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

(c) Bounded-2
before Bounded-Syl

Deriving Constraints: Cues

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

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

(b) Em-Right
before Bounded-Syl

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

(c) Bounded-2
before Bounded-Syl

Bounded-Syl as default (default values)

Deriving Constraints: Cues

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

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

(b) Em-Right before Bounded-Syl

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

(c) Bounded-2 before Bounded-Syl

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)

Deriving Constraints: Parsing

Group 1:
QS, Ft Head 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 Head 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 Head Left, Bounded

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)

Cues vs. Parsing for Unambiguous Data

The order constraints a learner would need to succeed can be derived in a principled manner for cues but must be mostly stipulated for parsing.

Open Questions

(1) Can we combine the strengths of cues and parsing?

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 datum

Viable combination of cues & parsing:
parsing of datum 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 parameteric system) is that the syllable is extrametrical.

Combining Cues and Parsing

Viable combination of cues & parsing:
parsing of datum subpart = derivation of cues?

Would partial parsing
(a) derive cues that lead to successful acquisition?
(b) be a more realistic representation of the learning mechanism?

Open Questions

(1) Can we combine the strengths of cues and parsing?

(2) Are order constraints *not* derivable from the learning system consistent cross-linguistically?

Non-derivable Constraints

Parsing Constraints

Group 1:
QS, Ft Head 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?

Open Questions

(1) Can we combine the strengths of cues and parsing?

(2) Are order constraints *not* derivable from the learning system consistent cross-linguistically?

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

Experimental Predictions for English

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 Head Left, Bounded

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

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

Open Questions

(1) Can we combine the strengths of cues and parsing?

(2) Are order constraints *not* derivable from the learning system consistent cross-linguistically?

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

(4) Is the unambiguous data filter successful for other languages besides English? Other complex linguistic domains?

Data Intake Filtering: The Big Questions

(1) Is it feasible to filter?
Can we filter and get success?

(2) Is it necessary to filter?
Must we filter to get success?

Data Intake Filtering: The Big Questions

(1) Is it feasible to filter?
Can we filter and get success?

(2) Is it necessary to filter?
Must we filter to get success?

Road Map

Learning Framework Overview

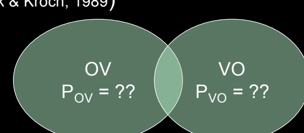
Computational Work: Case Studies

- Data intake filtering and systematicity in metrical phonology (synchronic)
- Data intake filtering in syntax (diachronic)
  - Old English description & proposed filters
  - Using language change to explore language learning
  - Old English data
  - Modeled learners and populations
  - Estimating ground truth
  - Sufficiency & necessity of filtering

Diachronic Investigation: Old English

Learning: Old English Object Verb (OV) vs. Verb Object order (VO)

Target State: probabilistic distribution between OV and VO hypotheses (YCOE Corpus, 2003; PPCME2 Corpus, 2000; similar models: Yang, 2002; Pintzuk, 2002; Kroch & Taylor, 1997; Bock & Kroch, 1989)



Old English Filters

Filter 1: Use data perceived as **unambiguous** (Dresher, 1999; Lightfoot, 1999; Fodor, 1998)

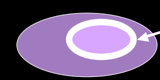
Filter 2: Use structurally "simple" data - matrix clause or "**degree-0**" data (Lightfoot, 1991)

Jack told his mother that the giant was easy to fool.
[----Degree-0-------]
[-------------Degree-1----------]

Problems

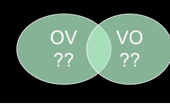
Potential problem: **data sparseness**

degree-0 unambiguous data set is significantly smaller than entire input set



Modeling problem:

How do we know if the final probabilistic state of the simulated learners is correct? What is our metric of success?



Road Map

Learning Framework Overview

Computational Work: Case Studies

- Data intake filtering and systematicity in metrical phonology (synchronic)
- Data intake filtering in syntax (diachronic)
  - Old English description & proposed filters
  - Using language change to explore language learning
  - Old English data
  - Modeled learners and populations
  - Estimating ground truth
  - Sufficiency & necessity of filtering

Modeling Solution

Using **language change** to test language learning

Old English, 1000 A.D. to 1200 A.D.: shift from a strongly OV-biased distribution to a strongly VO-biased distribution (YCOE Corpus, 2003; PPCME2 Corpus, 2000)

Old English shift proposed to be the result of ***imperfect learning*** of precisely the right amount **at the individual-level** (Lightfoot, 1991)

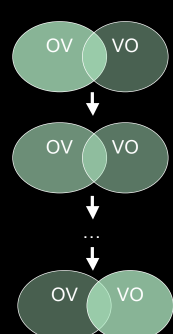
Imperfect Learning = Language Change

Individuals: the learner's final probability distribution is different from the adult's by a certain amount

These individuals: source of data for future individuals

Future individuals: converge on a probability distribution that is different.

Population-level: the population as a whole shifts at a certain rate, based on the amount individual learners differ from the rest of the population.



Language Learning Success

If we instantiate a certain learning model for individuals of a population and the population changes at the correct rate, we conclude:

(1) individuals misconverged precisely the right amount

(2) the learning model that allows this amount of misconvergence is correct

Road Map

Learning Framework Overview

Computational Work: Case Studies

- Data intake filtering and systematicity in metrical phonology (synchronic)
- Data intake filtering in syntax (diachronic)
  - Old English description & proposed filters
  - Using language change to explore language learning
  - Old English data
  - Modeled learners and populations
  - Estimating ground truth
  - Sufficiency & Necessity of Filtering

Old English OV and VO

OV-biased: between 1000 and 1150 A.D.

VO-biased: by 1200 A.D.

Old English OV and VO

OV-biased: between 1000 and 1150 A.D.

he$_{Subj}$ Gode$_{Obj}$ þancode$_{TensedVerb}$
he God thanked
'He thanked God'
(*Beowulf,* 625, ~1100 A.D.)

VO-biased: by 1200 A.D.

Old English OV and VO

OV-biased: between 1000 and 1150 A.D.

he$_{Subj}$ Gode$_{Obj}$ þancode$_{TensedVerb}$
he God thanked
'He thanked God'
(*Beowulf,* 625, ~1100 A.D.)

VO-biased: by 1200 A.D.

& [mid his stefne]$_{PP}$ he$_{Subj}$ awecð$_{TensedVerb}$ deade$_{Obj}$ [to life]$_{PP}$
& with his stem he awakened the-dead to life
'And with his stem, he awakened the dead to life.'
(*James the Greater,* 30.31, ~1150 A.D.)

Ambiguous Data

Subject TensedVerb Object is ambiguous (most common data type)

OV, +V2

heo$_{Subj}$ clænsað$_{TensedVerb}$ t_{Subj} [þa sawle þæs rædendan]$_{Obj}$ $t_{TensedVerb}$
they purified the souls [the advising]-Gen

VO, -V2

heo$_{Subj}$ clænsað$_{TensedVerb}$ [þa sawle þæs rædendan]$_{Obj}$
they purified the souls [the-advising]-Gen

'They purified the souls of the advising ones.'
(*Alcuin's De Virtutibus et Vitiis*, 83.59, ~1150 A.D.)

Perceived Unambiguous Data: Examples

Unambiguous OV

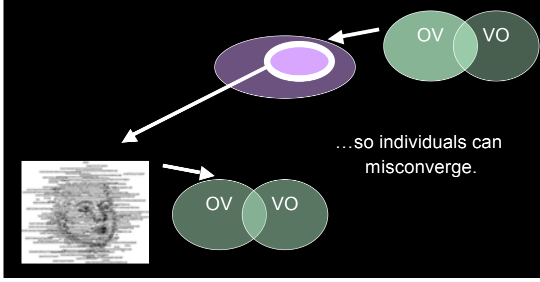
he$_{Subj}$ hyne$_{Obj}$ gebidde$_{TensedVerb}$
He him may-pray
'He may pray (to) him'
(*Ælfric's Letter to Wulfsige,* 87.107, ~1075 A.D.)

Unambiguous VO

þa$_{Adv}$ ahof$_{TensedVerb}$ Paulus$_{Subj}$ up$_{Verb-Marker}$ [his heafod]$_{Obj}$
then lifted Paul up his head
'Then Paul lifted his head up.'
(*Blickling Homilies,* 187.35, between 900 and 1000 A.D.)

The Effect of Filtering

Unambiguous degree-0 data distribution may differ from adult distribution used to generate data



…so individuals can misconverge.

Road Map

Learning Framework Overview

Computational Work: Case Studies

- Data intake filtering and systematicity in metrical phonology (synchronic)
- Data intake filtering in syntax (diachronic)
  - Old English description & proposed filters
  - Using language change to explore language learning
  - Old English data
  - Modeled learners and populations
  - Estimating ground truth
  - Sufficiency & necessity of filtering

The Model: Individual-Level

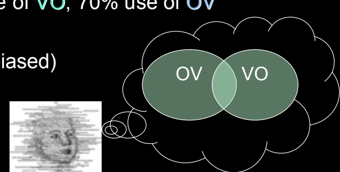
Individual learner tracks p_{VO} = probability of using VO

probability of using OV = 1 - p_{VO}

Old English: 0.0 <= p_{VO} <= 1.0

Ex: 0.3 = 30% use of VO, 70% use of OV

Initial p_{VO} = 0.5 (unbiased)



The Model: Individual-Level

Update using adaptation of Bayesian Updating (Manning & Schütze, 1999) for hypothesis space with 2 hypotheses

$$\mathrm{Max}(\mathrm{Prob}(p_{VO} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{Prob}(u \mid p_{VO}) * \mathrm{Prob}(p_{VO})}{\mathrm{Prob}(u)}\right)$$

$$\mathrm{Prob}(p_{VO} \mid u) = \frac{p_{VO} * \binom{n}{r} * p_{VO}^{\,r} * (1 - p_{VO})^{n-r}}{\mathrm{Prob}(u)} \text{(for each point } r,\ 0 \leq r \leq n)$$

$$\frac{d}{dp_{VO}}\left(\frac{p_{VO} * \binom{n}{r} * p_{VO}^{\,r} * (1 - p_{VO})^{n-r}}{\mathrm{Prob}(u)}\right) = 0$$

$$\frac{d}{dp_{VO}}\left(\frac{p_{VO} * \binom{n}{r} * p_{VO}^{\,r} * (1 - p_{VO})^{n-r}}{\cancel{\mathrm{Prob}(u)}}\right) = 0 \quad \text{(P}(u)\text{ is constant with respect to } p_{VO})$$

$$p_{VO} = \frac{r+1}{n+1},\ r = p_{VOprev} * n$$

Replace 1 in numerator and denominator with $c = p_{VOprev} * m$ if VO, $c = (1 - p_{VOprev}) * m$ if OV

$3.0 \leq m \leq 5.0$

The Model: Individual-Level

Update using adaptation of Bayesian Updating (Manning & Schütze, 1999) for hypothesis space with 2 hypotheses

If OV data point
$p_{VO} = (p_{VOprev} * n) / (n + c)$

If VO data point
$p_{VO} = (p_{VOprev} * n + c) / (n + c)$

c represents learner's confidence in input (calibrated), *n* represents quantity of intake (2000)

Individual-Level Learning Algorithm

Individual-Level Learning Algorithm

(1) Set initial p_{VO} to 0.5.

Individual-Level Learning Algorithm

(1) Set initial p_{VO} to 0.5.

(2) Encounter data point from an “average” member of the population.

(3) If the data point is degree-0 and unambiguous, use update functions to shift hypothesis probabilities.

Individual-Level Learning Algorithm

(1) Set initial p_{VO} to 0.5.

(2) Encounter data point from an “average” member of the population.

(3) If the data point is degree-0 and unambiguous, use update functions to shift hypothesis probabilities.

(4) Repeat (2-3) until the fluctuation period is over, as determined by *n*.

Biased Data Intake Distributions

p_{VO} shifts away from 0.5 when there is more of one data type in the intake than the other (advantage (Yang, 2000) of one data type)

Biased Data Intake Distributions

p_{VO} shifts away from 0.5 when there is more of one data type in the intake than the other (advantage (Yang, 2000) of one data type)

	OV Advantage in Unamb D0	OV Advantage in Unamb D1
1000 A.D.	19.5%	41.7%
1000-1150 A.D.	2.8%	28.7%
1200 A.D.	-2.7%	-45.2%

Population-Level Algorithm

Population-Level Algorithm

(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

Population-Level Algorithm

(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(3) Move forward 2 years.

(4) Members age 59-60 die off. The rest of the population ages 2 years.

Population-Level Algorithm

(1) Set the age range of the population from 0 to 60 years old and create 18,000 population members.

(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(3) Move forward 2 years.

(4) Members age 59-60 die off. The rest of the population ages 2 years.

(5) New members are born. These new members use the individual acquisition algorithm to set their p_{VO}.

(6) Repeat steps (3-5) until the year 1200 A.D.

Road Map

Learning Framework Overview

Computational Work: Case Studies

Data intake filtering and systematicity in metrical phonology (synchronic)

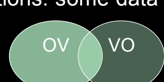
Data intake filtering in syntax (diachronic)

- Old English description & proposed filters
- Using language change to explore language learning
- Old English data
- Modeled learners and populations
- Estimating ground truth
- Sufficiency & necessity of filtering

Estimating Historical p_{VO}

Historical data used to initialize population at 1000 A.D., calibrate population between 1000 and 1150 A.D., and check target state at 1200 A.D.

Historical data distributions: some data are ambiguous



p_{VO}: underlying distribution used to produce data, so no ambiguous data

Estimating Historical p_{VO}

(YCOE and PPCME2 Corpora)
% Ambiguous Utterances

	Degree-0 % Ambiguous	Degree-1 % Ambiguous
1000 A.D.	76%	28%
1000 - 1150 A.D.	80%	25%
1200 A.D.	71%	10%

Observations:
(1) Degree-1 data less ambiguous than degree-0 data.
(2) Advantage is magnified in degree-1.

Assumption: degree-1 distribution less distorted from underlying distribution.

Estimating Historical p_{VO}

Use the difference in distortion between the degree-0 and degree-1 unambiguous data distributions to estimate the difference in distortion between the degree-1 distribution and the underlying unambiguous data distribution in a speaker's mind.

$$\frac{\gamma * \mathbf{d0} - \mathbf{u1d1'}}{\gamma * \mathbf{d0}} = \mathbf{Ld1tod0} * \frac{\mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}{\mathbf{u2d1'} + \mathbf{ad1'} - (\gamma * \mathbf{d0} - \mathbf{u1d1'})}$$

γ = underlying p_{VO}
d0 = total degree-0 data, **d1** = total degree-1 data
u1d1' = normalized unambiguous OV degree-1 data
u2d1' = normalized unambiguous VO degree-1 data
Ld1tod0 = loss ratio (OV/VO) from degree-1 to degree-0 distribution
ad1' = normalized ambiguous degree-1 data

$$\gamma = \frac{-(\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$
$$+/- \frac{\sqrt{((\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'})))^2 - 4(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)((-1)(\mathbf{d0} * \mathbf{u1d1'}))}}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$

Estimating Historical p_{VO}

	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747

Road Map

Learning Framework Overview

Computational Work: Case Studies

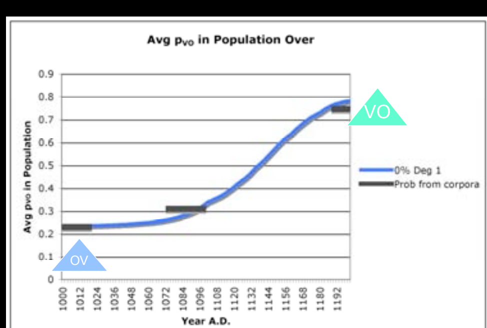
- Data intake filtering and systematicity in metrical phonology (synchronic)
- Data intake filtering in syntax (diachronic)
  - Old English description & proposed filters
  - Using language change to explore language learning
  - Old English data
  - Modeled learners and populations
  - Estimating ground truth
  - Sufficiency & necessity of filtering

Questions to Answer

(1) ***sufficiency***: Can an Old English population whose learners filter their intake down to the **degree-0 unambiguous data** shift at the correct rate?

(2) ***necessity***: If the proposed intake filtering is sufficient to cause an Old English population to change at the correct rate, is it in fact necessary? **Are the filters responsible?**

Sufficiency of Filters



Necessity of Filters: Remove Unambiguous Filter

Learner can use ambiguous data. Strategy: assume base-generation (surface order is actual order).

(Fodor, 1998)

Example: *Subject TensedVerb Object* = VO

Necessity of Filters: Remove Unambiguous Filter

Learner can use ambiguous data. Strategy: assume base-generation (surface order is actual order).

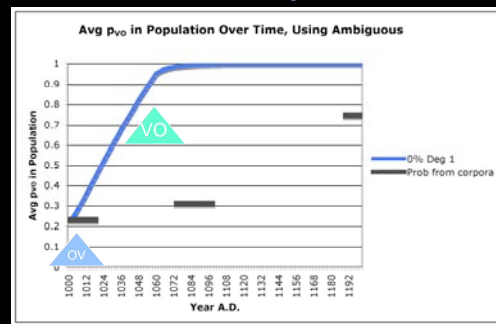
(Fodor, 1998)

Example: *Subject TensedVerb Object* = VO

	Degree-0 OV Advantage
1000 A.D.	-21.0%
1000 - 1150 A.D.	-26.9%
1200 A.D.	-21.8%

VO order has advantage, even at 1000 A.D.!

Necessity of Filters: Remove Unambiguous Filter



Necessity of Filters: Removing Degree-0 Filter

Learner can use unambiguous data in both degree-0 and degree-1 clauses.

Necessity of Filters: Removing Degree-0 Filter

Learner can use unambiguous data in both degree-0 and degree-1 clauses.

	OV Advantage in Unamb D0	OV Advantage in Unamb D1
1000 A.D.	19.5%	41.7%
1000-1150 A.D.	2.8%	28.7%
1200 A.D.	-2.7%	-45.2%

Necessity of Filters: Removing Degree-0 Filter

Learner can use unambiguous data in both degree-0 and degree-1 clauses.

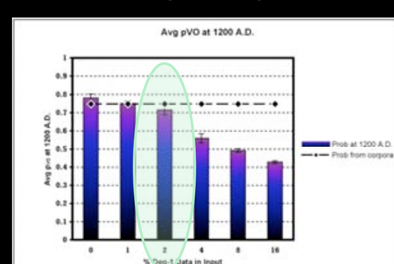
	OV Advantage in Unamb D0	OV Advantage in Unamb D1
1000 A.D.	19.5%	41.7%
1000-1150 A.D.	2.8%	28.7%
1200 A.D.	-2.7%	-45.2%

Degree-1 data is strongly OV-biased.

What is the threshold of permissible % of degree-1 data so the population can still be strongly VO-biased by 1200 A.D.?

How does this compare to the amount available to children?

Necessity of Filters: Allowing in Degree-1 Data



Permissible Threshold: <4% degree-1 data in intake.

Necessity of Filters: Removing Degree-0 Filter

Permissible threshold: <4%

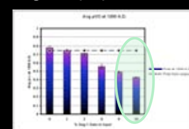
Estimated amount available to children (from corpora): ~16%

Necessity of Filters: Removing Degree-0 Filter

Permissible threshold: <4%

Estimated amount available to children (from corpora): ~16%

Conclusion: Filter required so that 16% degree-1 data does not cause Old English population to be too OV-biased



Necessity of Filters: Removing Both Filters

Dropping Unambiguous Data Filter: too much VO
(change is too fast)
Dropping Degree-0 Filter: too much OV
(change is too slow)
Drop both?

Necessity of Filters: Removing Both Filters

Dropping Unambiguous Data Filter: too much VO
(change is too fast)
Dropping Degree-0 Filter: too much OV
(change is too slow)
Drop both?

	OV Advantage in D0	OV Advantage in D1
1000 A.D.	-21.0%	28.1%

Requires 43% of the intake to be degree-1 data just to get the intake to be OV-biased at 1000 A.D.

Old English Language Change Summary

Language change modeling results: existence proof for sufficiency & necessity of data intake filtering

(1) unambiguous data
(2) degree-0 data

Additional moral: interaction of language change modeling and language learning theory

Data Intake Investigation: Take Home Messages

(1) Learners can extract the correct systematicity by looking at a subset of the data.

(2) The Old English model is empirically grounded, with learners searching through realistic data distributions.

(3) These results could not be obtained through standard experimental techniques.

Open Questions

(1) Are these filters robust across different language changes?

(2) Are these filters robust across different population models? (Ex: using population models with data weighting based on spatial location or social status of speaker, or context)

Answering Questions & Asking More

Answering Questions & Asking More

Data Intake

Unambiguous data & degree-0 data filtering: feasibility, sufficiency, necessity

True for other learning situations and domains?

Should different data be weighted differently?

Answering Questions & Asking More

Data Intake

Unambiguous data & degree-0 data filtering: feasibility, sufficiency, necessity

True for other learning situations and domains?

Should different data be weighted differently?

Finding Systematicity & Hypothesis Space Formation

Systematicity found in noisy systems

Systematicity even for exceptions to the rule?

Where /when do new hypotheses and hypothesis spaces (e.g. for exceptions) form?

Take Home Messages

Take Home Messages

(1) Defining the hypothesis space and discovering the time course of acquisition isn't enough to explain language learning - we need a theory of the mechanism.

Take Home Messages

(1) Defining the hypothesis space and discovering the time course of acquisition isn't enough to explain language learning - we need a theory of the mechanism.

(2) Uncovering the right systematicity in a realistic data set is a difficult task, but (perhaps contrary to intuition) *not* impossible if the learner has a restricted data intake (Clark's assessment was too pessimistic).

Take Home Messages

(1) Defining the hypothesis space and discovering the time course of acquisition isn't enough to explain language learning - we need a theory of the mechanism.

(2) Uncovering the right systematicity in a realistic data set is a difficult task, but (perhaps contrary to intuition) *not* impossible if the learner has a restricted data intake (Clark's assessment was too pessimistic).

(3) Computational modeling can explore questions we can't address experimentally, in addition to generating predictions that we can explore with standard experimental techniques.

Thank You

Amy Weinberg	Jeff Lidz
Bill Idsardi	Charles Yang
Colin Phillips	Norbert Hornstein
Elizabeth Royston	Philip Resnik
Raven Alder	David Poeppel

the Cognitive Neuroscience of Language Lab
at the University of Maryland

Causes of Language Change

Old Norse influence before 1000 A.D.: VO-biased
- If sole cause of change, requires exponential influx of Old Norse speakers.

Old French at 1066 A.D.: embedded clauses predominantly OV-biased (Kibler, 1984)
- Matrix clauses often SVO (ambiguous)
- OV-bias would have hindered Old English change to VO-biased system.

Evidence of individual probabilistic usage in Old English
- Historical records likely not the result of subpopulations of speakers who use only one order

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

$$\mathrm{Max}(\mathrm{Prob}(p_{VO} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{Prob}(u \mid p_{VO}) * \mathrm{Prob}(p_{VO})}{\mathrm{Prob}(u)}\right)$$

Bayes' Rule, find maximum of a posteriori (MAP) probability
Manning & Schütze (1999)

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

$$\mathrm{Max}(\mathrm{Prob}(p_{VO} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{Prob}(u \mid p_{VO}) * \mathrm{Prob}(p_{VO})}{\mathrm{Prob}(u)}\right)$$

Prob(u | p_{VO}) = probability of seeing unambiguous data point u, given p_{VO}.
= p_{VO}

Prob(p_{VO}) = probability of seeing r out of n data points that are unambiguous for VO, for 0 <= r <= n
= $\binom{n}{r} * p_{VO}^{\,r} * (1 - p_{VO})^{n-r}$

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

$$\mathrm{Max}(\mathrm{Prob}(\mathrm{p_{VO}} \mid u)) = \mathrm{Max}\left(\frac{\mathrm{p_{VO}} * \binom{n}{r} * \mathrm{p_{VO}}^{r} * (1 - \mathrm{p_{VO}})^{n-r}}{\mathrm{Prob}(u)}\right) \quad \text{(for each point } r,\ 0 \le r \le n\text{)}$$

$$\frac{d}{d\mathrm{p_{VO}}}\left(\frac{\mathrm{p_{VO}} * \binom{n}{r} * \mathrm{p_{VO}}^{r} * (1 - \mathrm{p_{VO}})^{n-r}}{\mathrm{Prob}(u)}\right) = 0$$

$$\frac{d}{d\mathrm{p_{VO}}}\left(\frac{\mathrm{p_{VO}} * \binom{n}{r} * \mathrm{p_{VO}}^{r} * (1 - \mathrm{p_{VO}})^{n-r}}{\cancel{\mathrm{Prob}(u)}}\right) = 0 \quad \text{(P}(u)\text{ is constant with respect to p}_{\mathrm{VO}}\text{)}$$

$$\mathrm{p_{VO}} = \frac{r+1}{n+1}$$

Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

$$\mathrm{p_{VO}} = \frac{r+1}{n+1},\quad r = \mathrm{p_{VOprev}} * n$$

Replace 1 in numerator and denominator with

$c = \mathrm{p_{VOprev}} * m$ if VO, $c = (1 - \mathrm{p_{VOprev}}) * m$ if OV

$3.0 \le m \le 5.0$

$$\mathrm{p_{VO}} = \frac{\mathrm{p_{VOprev}} * n + c}{n + c}$$

Estimating Ground Truth

Estimating Ground Truth

Known quantities: Unambiguous and ambiguous data in d0 and d1

Estimating Ground Truth

Known quantities: Unambiguous and ambiguous data in d0 and d1

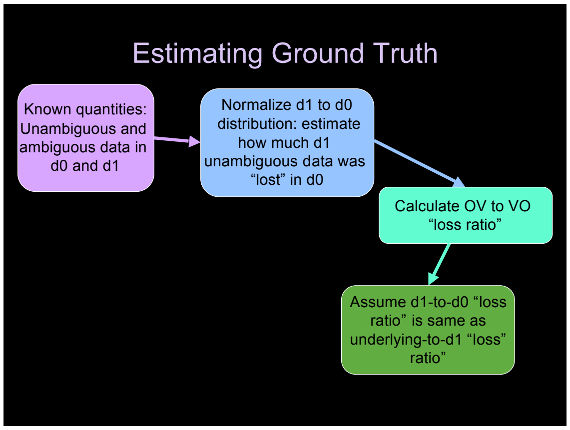
Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was “lost” in d0

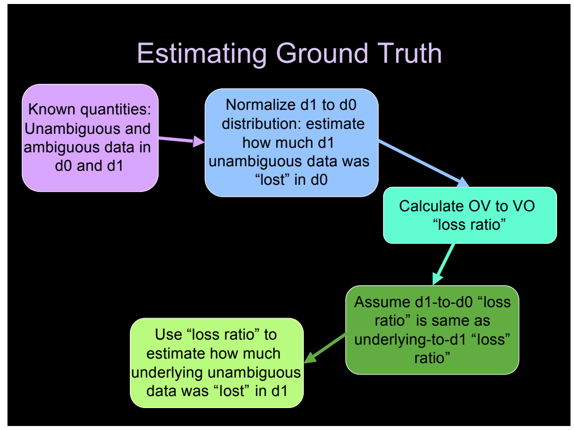
Estimating Ground Truth

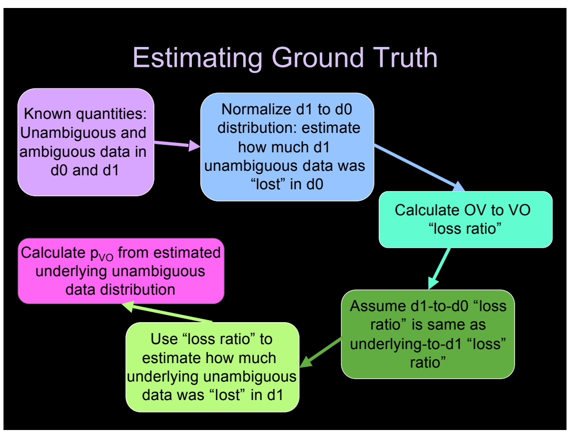
Known quantities: Unambiguous and ambiguous data in d0 and d1

Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was “lost” in d0

Calculate OV to VO “loss ratio”







Other Ways to Remove the Unambiguous Filter

Strategies for assessing ambiguous data
- (1) assume base-generation
  - attempted and failed
  - system-dependent (syntax)
- (2) weight based on level of ambiguity (Pearl & Lidz, in submission)
  - unambiguous = highest weight
  - moderately ambiguous = lower weight
  - fully ambiguous = lowest weight (ignore)
- (3) randomly assign to one hypothesis (Yang, 2002)

Making Parsing More Robust

Main problem with the instantiation considered: if can't parse the entire data point, can't extract information from it

Potential Solution: partial parsing
- Examples
  - sentences: clause by clause
  - words: syllables including word edge (#)

(**x** x
#**VV** VC … : Feet Headed Left, not Em-Left

- Benefits
  - may be able to derive cues rather than requiring them to be part of the learner's innate endowment (Dresher, 1999)

Perceived Unambiguous Data: OV

Unambiguous OV data

Perceived Unambiguous Data: OV

Unambiguous OV data

(1) Tensed Verb is immediately post-Object

he$_{\text{Subj}}$ hyne$_{\text{Obj}}$ gebidde$_{\text{TensedVerb}}$
He him may-pray
'He may pray (to) him'
(*Ælfric's Letter to Wulfsige,* 87.107, ~1075 A.D.)

Perceived Unambiguous Data: OV

Unambiguous OV data

(1) Tensed Verb is immediately post-Object

he$_{\text{Subj}}$ hyne$_{\text{Obj}}$ gebidde$_{\text{TensedVerb}}$
He him may-pray
'He may pray (to) him'
(*Ælfric's Letter to Wulfsige,* 87.107, ~1075 A.D.)

(2) Verb-Marker is immediately post-Object

we$_{\text{Subj}}$ sculen$_{\text{TensedVerb}}$ [ure yfele þeawes]$_{\text{Obj}}$ forlæten$_{\text{Verb-Marker}}$
we should our evil practices abandon
'We should abandon our evil practices.'
(*Alcuin's De Virtutibus et Vitiis,* 70.52, ~1150 A.D.)

Perceived Unambiguous Data: VO

Unambiguous VO data

Perceived Unambiguous Data: VO

Unambiguous VO data

(1) Tensed Verb is immediately pre-Object, 2+ phrases precede (due to interaction of V2 movement)

& [mid his stefne]$_{\text{PP}}$ he$_{\text{Subj}}$ awecð$_{\text{TensedVerb}}$ deade$_{\text{Obj}}$ [to life]$_{\text{PP}}$
& with his stem he awakened the-dead to life
'And with his stem, he awakened the dead to life.'
(*James the Greater,* 30.31, ~1150 A.D.)

Perceived Unambiguous Data: VO

Unambiguous VO data

(1) Tensed Verb is immediately pre-Object, 2+ phrases precede (due to interaction of V2 movement)

& [mid his stefne]$_{\text{PP}}$ he$_{\text{Subj}}$ awecð$_{\text{TensedVerb}}$ deade$_{\text{Obj}}$ [to life]$_{\text{PP}}$
& with his stem he awakened the-dead to life
'And with his stem, he awakened the dead to life.'
(*James the Greater,* 30.31, ~1150 A.D.)

(2) Verb-Marker is immediately pre-Object

þa$_{\textbf{Adv}}$ ahof$_{\text{TensedVerb}}$ Paulus$_{\text{Subj}}$ up$_{\text{Verb-Marker}}$[his heafod]$_{\text{Obj}}$
then lifted Paul up his head
'Then Paul lifted his head up.'
(*Blickling Homilies,* 187.35, between 900 and 1000 A.D.)

Verb-Markers

Sub-piece of the verbal complex that is semantically associated with a Verb, used to determine original position of Verb

Examples: particle ('up', 'out'), a non-tensed complement to tensed Verbs, a closed-class adverbial ('never'), or a negative ('not') (Lightfoot, 1991).

þa$_{\textbf{Adv}}$ ahof$_{\text{TensedVerb}}$ Paulus$_{\text{Subj}}$ up$_{\text{Verb-Marker}}$ [his heafod]$_{\text{Obj}}$
then lifted Paul up his head
'Then Paul lifted his head up.'

Verb-Markers

Sub-piece of the verbal complex that is semantically associated with a Verb, used to determine original position of Verb

Examples: particle (‘up’, ‘out’), a non-tensed complement to tensed Verbs, a closed-class adverbial (‘never’), or a negative (‘not’) (Lightfoot, 1991).

þa$_{\textbf{Adv}}$ ahof$_{\text{TensedVerb}}$ Paulus$_{\text{Subj}}$ up$_{\text{Verb-Marker}}$ [his heafod]$_{\text{Obj}}$
then lifted Paul up his head
‘Then Paul lifted his head up.’

we$_{\text{Subj}}$ sculen$_{\text{TensedVerb}}$ [ure yfele þeawes]$_{\text{Obj}}$ forlæten$_{\text{Verb-Marker}}$
we should our evil practices abandon
‘We should abandon our evil practices.’

Unreliable Verb-Markers

Sometimes the Verb-Marker would not remain adjacent to the Object.

ne$_{\text{Negative}}$ geseah$_{\text{TensedVerb}}$ ic$_{\text{Subj}}$ næfre$_{\text{Adverbial}}$ [ða burh]$_{\text{Obj}}$
NEG saw I never the city
‘Never did I see the city.’
(Ælfric, *Homilies.* I.572.3, between 900 and 1000 A.D.)

Unreliable Verb-Markers

Sometimes the Verb-Marker would not remain adjacent to the Object.

ne$_{\text{Negative}}$ geseah$_{\text{TensedVerb}}$ ic$_{\text{Subj}}$ næfre$_{\text{Adverbial}}$ [ða burh]$_{\text{Obj}}$
NEG saw I never the city
‘Never did I see the city.’
(Ælfric, *Homilies.* I.572.3, between 900 and 1000 A.D.)