Linguistic Evolution, In Brief

Linguistic knowledge is transmitted in a population via interaction with other speakers in the population.
Linguistic Evolution, In Brief

The information speakers transmit (observable data) is based on their own linguistic knowledge.
Linguistic Evolution, In Brief

Speakers adjust their linguistic knowledge based on the observable (and encountered) data from other population members.
Population-level changes over time depend on what information speakers pass to subsequent generations and how that information is integrated into an individual’s linguistic knowledge.
Integrating Linguistic Information

Not all linguistic knowledge is created equal

Some knowledge can be altered throughout an individual’s life

(example: vocabulary)
Integrating Linguistic Information

Not all linguistic knowledge is created equal

Some knowledge can be altered only during the early stages of an individual’s life

(example: word order rules)
Implication: The way in which young learners integrate linguistic information (along with the data available) determines the linguistic composition of the population and the speed at which the linguistic knowledge evolves within the population.
Change to knowledge that is alterable early

Implication: The way in which young learners integrate linguistic information (along with the data available) determines the linguistic composition of the population and the speed at which the linguistic knowledge evolves within the population.
I. Individual Language Learning
   The Nature of Linguistic Knowledge
   Individual Learning Framework

II. Linguistic Evolution: Case Study
   Old English Word Order
   Modeling Individuals  (Pearl & Weinberg 2007)
   Modeling Populations
   Issues in Empirical Grounding
   Selective Learning Biases
I. Individual Language Learning
   The Nature of Linguistic Knowledge
   Individual Learning Framework

II. Linguistic Evolution: Case Study
   Old English Word Order
   Modeling Individuals (Pearl & Weinberg 2007)
   Modeling Populations
   Issues in Empirical Grounding
   Selective Learning Biases
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering
  Ex: What are the contrastive sounds of a language?
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering
Ex: What are the contrastive sounds of a language?
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

Categorization/Clustering

Ex: What are the contrastive sounds of a language?
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Extraction
  Ex: Where are words in fluent speech?
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Extraction
Ex: Where are words in fluent speech?

húwz əfriːd əv ðə bɪɡ bæˈd əl ɪf
who’s afraid of the big bad wolf
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Extraction
Ex: Where are words in fluent speech?

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

húwz əfréjd əv ðə bɪg bæʼd wəʼlf
whoʼs afraid of the big bad wolf

blink~blinked   confide~confided

drink~drank
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

Extraction
Ex: Where are words in fluent speech?

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

húwz əfˈreɪd əv ˈðæ bɪɡ bæˈd əf ˈwɛrˈl̩f
whoˈs afrəɪd əv ðə bɪɡ bæˈd wəˈlf

blink~blinked
blɪŋk blɪŋkt
confide~confided
kənˈfaɪd kənˈfaɪdəd

drink~drank
drɪŋk drejŋk
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

- syntax = word order rules
- Learning problem: many ways to generate observable data
The Nature of Linguistic Knowledge

Different aspects: more and less transparent from data

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

syntax = word order rules
Learning problem: many ways to generate observable data

Observable data: word order Subject Verb Object
Generative system: syntax

Kannada
Subject t_{Object} Verb Object
Object Verb underlying

English
Subject Verb Object
Verb Object underlying

German
Subject Verb t_{Subject} Object t_{Verb}
Object Verb underlying
The individual learning framework: 3 components

(1) Hypothesis space

(2) Data

(3) Update procedure
I. Individual Language Learning
   The Nature of Linguistic Knowledge
   Individual Learning Framework

II. Linguistic Evolution: Case Study
   Old English Word Order
   Modeling Individuals  (Pearl & Weinberg 2007)
   Modeling Populations
   Issues in Empirical Grounding
   Selective Learning Biases
Old English

Changing Basic Word Order Rule in Old English: **Object-Verb (OV)** vs. **Verb-Object (VO)** order

**Individual Knowledge** (underlying probability in speaker’s mind): probability distribution between OV and VO orders
Changing Basic Word Order Rule in Old English: **Object-Verb (OV)** vs. **Verb-Object (VO)** order

**Individual Knowledge** (underlying probability in speaker’s mind): probability distribution between OV and VO orders

**Individual Usage** (observable data for learner): probability distribution between OV and VO orders (not necessarily same one as individual knowledge distribution, from learner’s perspective)

**Why not?**
Underlying Distribution vs. Observable Distribution

German/Old English

Subject Verb Object

Surface order: Verb Object

Speaker generates utterance
Underlying Distribution vs. Observable Distribution

Learner interprets utterance
Every utterance generated by speaker is either OV or VO order in the underlying distribution.
The learner encounters data that is ambiguous between the two options. Distribution depends on learner’s interpretation of ambiguous data.
Old English

Changing Basic Word Order Rule in Old English: **Object-Verb (OV)** vs. **Verb-Object (VO)** order

Individual Knowledge (underlying probability in speaker’s mind): probability distribution between OV and VO orders

Individual Usage (observable data for learner): probability distribution between OV and VO orders (not necessarily same one as individual knowledge distribution, from learner’s perspective)

Why not?
Old English

Changing Basic Word Order Rule in Old English: Object-Verb (OV) vs. Verb-Object (VO) order

Individual Knowledge (underlying probability in speaker’s mind): probability distribution between OV and VO orders

Individual Usage (observable data for learner): probability distribution between OV and VO orders (not necessarily same one as individual knowledge distribution, from learner’s perspective)

Due to learner interpretation bias
Old English

Estimates of average individual usage from historical corpora:
YCOE Corpus 2003; PPCME2 Corpus 2000

~1000 A.D.-1150 A.D.: OV-biased

~1200 A.D.: VO-biased

To get this rate of change, young individual learners at each time step must change their probability distribution the exact right amount from the previous population members’ distribution.
**Interpretation Bias**: Use only data perceived as most informative (Fodor 1998, Lightfoot 1999, Drescher 1999).

**Interpretation Bias**: Use only data that is more accessible (perhaps for language processing reasons) (Lightfoot 1991).
Modeling Individuals: Learning Biases

Interpretation Bias: Use only data perceived as most informative: **unambiguous data** (Fodor 1998, Lightfoot 1999, Dresher 1999).

Interpretation Bias: Use only data that is more accessible (perhaps for language processing reasons) (Lightfoot 1991).

Learner has heuristics for identifying unambiguous **OV/VO** data, based on partial knowledge of possible adult system rules (Fodor 1998, Lightfoot 1999, Dresher 1999).

Knowledge of tensed verb movement to 2nd phrasal position of sentence

**OV** unambiguous data:

\[
\text{[...]}_{\text{XP}} \quad \text{Object} \quad \text{TensedVerb} \quad \text{...}
\]

\[
\text{...TensedVerb} \quad \text{Object} \quad \text{Verb-Marker} \quad \text{...}
\]

**VO** unambiguous data:

\[
\text{[...]}_{\text{XP}} \quad [\text{...}]_{\text{XP}} \quad \text{TensedVerb} \quad \text{Object} \quad \text{...}
\]

\[
\text{...TensedVerb} \quad \text{Verb-Marker} \quad \text{Object} \quad \text{...}
\]
**Interpretation Bias**: Use only data perceived as most informative: 

**Interpretation Bias**: Use only data that is more accessible (perhaps for language processing reasons) (Lightfoot 1991).

OV unambiguous data:

\[ \ldots \text{XP} \ldots \text{Object} \ \text{TensedVerb} \ldots \text{TensedVerb} \ldots \text{Object} \ \text{Verb-Marker} \ldots \]

VO unambiguous data:

\[ \ldots \text{XP} \ [\ldots]_{\text{XP}} \ [\ldots]_{\text{XP}} \ \text{TensedVerb} \ \text{Object} \ldots \text{TensedVerb} \ldots \text{Verb-Marker} \ \text{Object} \ldots \]
Modeling Individuals: Learning Biases

**Interpretation Bias**: Use only data perceived as most informative: unambiguous data (Fodor 1998, Lightfoot 1999, Dresher 1999).

**Interpretation Bias**: Use only structurally simple (degree-0) data (Lightfoot 1991).

Jack told his mother that the giant was easy to fool.

[----Degree-0--------]

[----------Degree-1----------]
The point of interpretation biases: Unambiguous degree-0 data distribution may differ the right amount from population’s underlying distribution to change at the right rate.

~1000 A.D.-1150 A.D.: OV-biased

~1200 A.D.: VO-biased

% VO

time
Modeling Individuals: Knowledge & Learning

Individual learner tracks $p_{VO} = \text{probability of using VO}$
probability of using $OV = 1 - p_{VO}$

Old English: $0.0 \leq p_{VO} \leq 1.0$
Ex: $0.3 = 30\% \text{ VO, 70}\% \text{ OV}$ during generation

Initial $p_{VO} = 0.5$ (unbiased)
Modeling Individuals: Knowledge & Learning

Individual learner tracks $p_{VO} = \text{probability of using VO}$
probability of using $OV = 1 - p_{VO}$

Old English: $0.0 \leq p_{VO} \leq 1.0$
Ex: $0.3 = 30\% \text{ VO, 70}\% \text{ OV}$ during generation

Initial $p_{VO} = 0.5$ (unbiased)

Data from old members of population, filtered through selective learning biases.
Modeling Individuals: Knowledge & Learning

Individual learner tracks $p_{VO} = \text{probability of using } VO$

probability of using $OV = 1 - p_{VO}$

Old English: $0.0 \leq p_{VO} \leq 1.0$

Ex: $0.3 = 30\% \ VO, 70\% \ OV$ during generation

Initial $p_{VO} = 0.5$ (unbiased)

Data from old members of population, filtered through selective learning biases.

Individual update: Bayesian updating for binomial distribution (Chew 1971), adapted
Zoom-In on Updating Procedure

Max(Prob(p\textsubscript{VO} | u)) = \text{Max}\left(\frac{\text{Prob}(u | p\textsubscript{VO}) \ast \text{Prob}(p\textsubscript{VO})}{\text{Prob}(u)}\right)

\text{Prob}(p\textsubscript{VO} | u) = \frac{\text{p}\text{VO} \ast \binom{n}{r} \ast \text{p}\text{VO}^{r} \ast (1 - \text{p}\text{VO})^{n-r}}{\text{Prob}(u)}\text{ (for each point } r, 0 \leq r \leq n)\)

\frac{d}{dp\text{VO}}\left(\frac{\text{p}\text{VO} \ast \binom{n}{r} \ast \text{p}\text{VO}^{r} \ast (1 - \text{p}\text{VO})^{n-r}}{\text{Prob}(u)}\right) = 0\)

\frac{d}{dp\text{VO}}\left(\frac{\text{p}\text{VO} \ast \binom{n}{r} \ast \text{p}\text{VO}^{r} \ast (1 - \text{p}\text{VO})^{n-r}}{\text{Prob}(u)}\right) = 0\text{ (P}(u)\text{ is constant with respect to p}\text{VO)}\)

\text{p\textsubscript{VO}} = \frac{r + 1}{n + 1}, \text{ } r = \text{p\textsubscript{VO}\textsubscript{prev}} \ast n

Replace 1 in numerator and denominator with \text{c} = \text{p\textsubscript{VO}\textsubscript{prev}} \ast m\text{ if VO, } \text{c} = (1 - \text{p\textsubscript{VO}\textsubscript{prev}}) \ast m\text{ if OV}\n
3.0 \leq m \leq 5.0
Zoom-In on Updating Procedure

If \(OV\) data point
\[
p_{VO} = \left( p_{VO_{prev}} \times n \right) / (n+c)
\]

If \(VO\) data point
\[
p_{VO} = \left( p_{VO_{prev}} \times n+c \right) / (n+c)
\]

Important: Online update procedure
(psychological plausibility, given human memory)

Involves previous
probability & expected
amount of data in
learning period

Model parameters:

\(c\) represents learner’s confidence in data point
(calibrated from data)

\(n\) represents quantity of intake (2000)
Individual-Level Learning Algorithm

(1) Initial $p_{VO} = 0.5$.

(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 learning period is over, as determined by $n$. 
$p_{VO}$ 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).

So the bias in the degree-0 unambiguous data distribution controls an individual’s final $p_{VO}$ in this model.

<table>
<thead>
<tr>
<th></th>
<th>OV Advantage in Unamb D0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>19.5%</td>
</tr>
<tr>
<td>1000-1150 A.D.</td>
<td>2.8%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>-2.7%</td>
</tr>
</tbody>
</table>
Population-Level Model

(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 size estimated from population statistics of the time period (Koenigsberger & Briggs 1987)

Average $p_{VO}$ estimated from YCOE 2003 & PPCME2 2000

Time: 1000 A.D.
Population-Level Model

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.

Time: 1002 A.D.
Population-Level Model

(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.

Time: 1002 A.D.
Population-Level Model

(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 $p_{VO}$.

Population growth rate estimated from population statistics of the time period (Koenigsberger & Briggs 1987)

Time: 1002 A.D.
Population-Level Model

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 $p_{VO}$.

Time: 1002 A.D.
Population-Level Model

(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 $p_{\text{VO}}$ 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 $p_{\text{VO}}$.

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

Time: 1200 A.D.
Empirical Grounding Issues: What exactly is the underlying distribution?

Historical data used to initialize population’s $p_{VO}$ at 1000 A.D., calibrate population’s $p_{VO}$ between 1000 and 1150 A.D., and check target $p_{VO}$ at 1200 A.D.

Historical data distributions: some data are ambiguous
Empirical Grounding Issues: 
What exactly is the underlying distribution?

Historical data used to initialize population’s $p_{VO}$ at 1000 A.D., calibrate population’s $p_{VO}$ between 1000 and 1150 A.D., and check target $p_{VO}$ at 1200 A.D.

Historical data distributions: some data are ambiguous

$p_{VO}$: underlying distribution is not ambiguous
Empirical Grounding Issues:
What exactly is the underlying distribution?

Historical data used to initialize population’s $p_{VO}$ at 1000 A.D., calibrate population’s $p_{VO}$ between 1000 and 1150 A.D., and check target $p_{VO}$ at 1200 A.D.

Historical data distributions: some data are ambiguous

$p_{VO}$: underlying distribution is not ambiguous
Empirical Grounding Issues: What exactly is the underlying distribution?

Historical data used to initialize population’s $p_{VO}$ at 1000 A.D., calibrate population’s $p_{VO}$ between 1000 and 1150 A.D., and check target $p_{VO}$ at 1200 A.D.

Historical data distributions: some data are ambiguous

How do we figure out what the ambiguous data are?

$p_{VO}$: underlying distribution is not ambiguous
Empirical Grounding Issues: What exactly is the underlying distribution?

(YCOE and PPCME2 Corpora)
% Ambiguous Utterances

<table>
<thead>
<tr>
<th></th>
<th>Degree-0 % Ambiguous</th>
<th>Degree-1 % Ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>76%</td>
<td>28%</td>
</tr>
<tr>
<td>1000 - 1150 A.D.</td>
<td>80%</td>
<td>25%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>71%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Observations:
(1) Degree-1 data less ambiguous than degree-0 data.
Empirical Grounding Issues: What exactly is the underlying distribution?

(YCOE and PPCME2 Corpora)

% Advantage

<table>
<thead>
<tr>
<th></th>
<th>OV Advantage in Unamb D0</th>
<th>OV Advantage in Unamb D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>19.5%</td>
<td>41.7%</td>
</tr>
<tr>
<td>1000-1150 A.D.</td>
<td>2.8%</td>
<td>28.7%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>-2.7%</td>
<td>-45.2%</td>
</tr>
</tbody>
</table>

Observations:
(1) Degree-1 data less ambiguous than degree-0 data.
(2) Advantage is magnified in degree-1.
Empirical Grounding Issues:
What exactly is the underlying distribution?

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

Assumption: Ambiguous data distorts underlying distribution.
Assumption: degree-1 distribution less distorted from underlying distribution.
Empirical Grounding Issues: What exactly is the underlying distribution?

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

Assumption: Ambiguous data distorts underlying distribution.
Assumption: degree-1 distribution less distorted from underlying distribution.

Plan of Action: 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.
Empirical Grounding Issues: What exactly is the underlying distribution?

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

Assumption: Ambiguous data distorts underlying distribution
Assumption: degree-1 distribution less distorted from underlying distribution.

\[ \frac{\gamma \cdot d_0 - u_1 d_1'}{\gamma \cdot d_0} = L_{d_1 \text{to} d_0} \cdot \frac{a_1' - (\gamma \cdot d_0 - u_1 d_1')}{u_2 d_1' + a_1' - (\gamma \cdot d_0 - u_1 d_1')} \]

\[ \gamma = \frac{-(d_0)(d_0 + u_1 d_1' - L_{d_1 \text{to} d_0} \cdot (a_1' + u_1 d_1'))}{2(L_{d_1 \text{to} d_0} + 1)(d_0^2)} \]

\[ \pm \sqrt{\frac{((d_0)(d_0 + u_1 d_1' - L_{d_1 \text{to} d_0} \cdot (a_1' + u_1 d_1'))^2 - 4(L_{d_1 \text{to} d_0} + 1)(d_0^2)(-1)(d_0 \cdot u_1 d_1'))}{2(L_{d_1 \text{to} d_0} + 1)(d_0^2)}} \]

\[ d_0 = \text{total degree-0 data}, \quad d_1 = \text{total degree-1 data} \]
\[ u_1 d_1' = \text{normalized unambiguous OV degree-1 data} \]
\[ u_2 d_1' = \text{normalized unambiguous VO degree-1 data} \]
\[ L_{d_1 \text{to} d_0} = \text{loss ratio (OV/VO) from degree-1 to degree-0 distribution} \]
\[ a_1' = \text{normalized ambiguous degree-1 data} \]
Empirical Grounding Issues: What exactly is the underlying distribution?

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

Assumption: Ambiguous data distorts underlying distribution
Assumption: degree-1 distribution less distorted from underlying distribution.

<table>
<thead>
<tr>
<th></th>
<th>(Initialization)</th>
<th>(Calibration)</th>
<th>(Termination)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000 A.D.</td>
<td>1000-1150 A.D.</td>
<td>1200 A.D.</td>
</tr>
<tr>
<td>Average $p_{VO}$</td>
<td>0.234</td>
<td>0.310</td>
<td>0.747</td>
</tr>
</tbody>
</table>

OV-biased

VO-biased
Empirical Grounding Issues: What exactly is the underlying distribution?

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

Assumption: Ambiguous data distorts underlying distribution
Assumption: degree-1 distribution less distorted from underlying distribution.

<table>
<thead>
<tr>
<th></th>
<th>(Initialization)</th>
<th>(Calibration)</th>
<th>(Termination)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>0.234</td>
<td>0.310</td>
<td>0.747</td>
</tr>
<tr>
<td>1000-1150 A.D.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1200 A.D.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Average $p_{VO}$
Observations:

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

Assumption: Ambiguous data distorts underlying distribution
Assumption: degree-1 distribution less distorted from underlying distribution.

\[ p_{VO} = \frac{(p_{VO_{prev}} \times n)}{(n+c)} \]

<table>
<thead>
<tr>
<th></th>
<th>(Initialization)</th>
<th>(Calibration)</th>
<th>(Termination)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1000 A.D.</td>
<td>1000-1150 A.D.</td>
<td>1200 A.D.</td>
</tr>
<tr>
<td>Average ( p_{VO} )</td>
<td>0.234</td>
<td>0.310</td>
<td>0.747</td>
</tr>
</tbody>
</table>

OV-biased

VO-biased

\[ p_{VO} = \frac{(p_{VO_{prev}} \times n+c)}{(n+c)} \]
Empirical Grounding Issues: What exactly is the underlying distribution?

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

Assumption: Ambiguous data distorts underlying distribution
Assumption: degree-1 distribution less distorted from underlying distribution.

<table>
<thead>
<tr>
<th></th>
<th>(Initialization) 1000 A.D.</th>
<th>(Calibration) 1000-1150 A.D.</th>
<th>(Termination) 1200 A.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_{VO} )</td>
<td>0.234</td>
<td>0.310</td>
<td>0.747</td>
</tr>
</tbody>
</table>
Linguistic Evolution: Change at the Historically-Attested Rate

Learners have selective learning bias on data
Linguistic Evolution: Different Individual-Level Learning

Learner uses ambiguous data. Strategy for learning: assume surface order is actual order. (Fodor 1998)
Linguistic Evolution: Different Individual-Level Learning

Learner uses ambiguous data. Strategy for learning: assume surface order is actual order. (Fodor 1998)

Advantage in intake determines learner’s ending distribution between OV and VO order.
Linguistic Evolution: Different Individual-Level Learning

Learner uses ambiguous data. Strategy for learning: assume surface order is actual order. (Fodor 1998)

Advantage in intake determines learner’s ending distribution between OV and VO order.

<table>
<thead>
<tr>
<th>Degree-0 OV Advantage</th>
<th>% VO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>-21.0%</td>
</tr>
<tr>
<td>1000 - 1150 A.D.</td>
<td>-26.9%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>-21.8%</td>
</tr>
</tbody>
</table>

Need this trajectory
Linguistic Evolution: Different Individual-Level Learning

Learner uses ambiguous data. Strategy for learning: assume surface order is actual order. (Fodor 1998)

Advantage in intake determines learner’s ending distribution between OV and VO order.

<table>
<thead>
<tr>
<th></th>
<th>Degree-0 OV Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>-21.0%</td>
</tr>
<tr>
<td>1000 - 1150 A.D.</td>
<td>-26.9%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>-21.8%</td>
</tr>
</tbody>
</table>

Need this trajectory

Problem: VO-biased all the way through, even at 1000 A.D.

Change is too fast!
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.

(YCOE and PPCME2 Corpora)

<table>
<thead>
<tr>
<th>% Advantage</th>
<th>OV Advantage in Unamb D0</th>
<th>OV Advantage in Unamb D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>19.5%</td>
<td>41.7%</td>
</tr>
<tr>
<td>1000-1150 A.D.</td>
<td>2.8%</td>
<td>28.7%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>-2.7%</td>
<td>-45.2%</td>
</tr>
</tbody>
</table>

Very strongly OV-biased before 1150 A.D.
**Linguistic Evolution: Different Individual-Level Learning**

Learner uses degree-0 and degree-1 unambiguous data.

(YCOE and PPCME2 Corpora)

<table>
<thead>
<tr>
<th>Year</th>
<th>OV Advantage in Unamb (D_0)</th>
<th>OV Advantage in Unamb (D_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>19.5%</td>
<td>41.7%</td>
</tr>
<tr>
<td>1000-1150 A.D.</td>
<td>2.8%</td>
<td>28.7%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>-2.7%</td>
<td>-45.2%</td>
</tr>
</tbody>
</table>

% Advantage

- Advantage very strongly OV-biased before 1150 A.D.

But population must become VO-biased.

Need this trajectory

% VO

![Diagram showing the trajectory of % VO over time]

Very strongly OV-biased before 1150 A.D.
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.

(YCOE and PPCME2 Corpora)

% Advantage

<table>
<thead>
<tr>
<th></th>
<th>OV Advantage in Unamb D0</th>
<th>OV Advantage in Unamb D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>19.5%</td>
<td>41.7%</td>
</tr>
<tr>
<td>1000-1150 A.D.</td>
<td>2.8%</td>
<td>28.7%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>-2.7%</td>
<td>-45.2%</td>
</tr>
</tbody>
</table>

Very strongly OV-biased before 1150 A.D.

Can a population learning from degree-1 data make the change to VO-biased?
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.

Modeled population can change at the right rate only if input contains less than 4% degree-1 data - otherwise, change is too slow for learners not using a degree-0 bias.
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 unambiguous data.

Estimates from modern English child-directed speech: Input consists of ~16% degree-1 data.

Prognosis: Change would be too slow without a degree-0 bias for individual learners.
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

(YCOE and PPCME2 Corpora)

% Advantage

<table>
<thead>
<tr>
<th></th>
<th>OV Advantage in D0</th>
<th>OV Advantage in D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>-21.0%</td>
<td>28.1%</td>
</tr>
</tbody>
</table>

Need this trajectory

% VO

OV-biased

VO-biased

time
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

(YCOE and PPCME2 Corpora)

% Advantage

<table>
<thead>
<tr>
<th></th>
<th>OV Advantage in D0</th>
<th>OV Advantage in D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>-21.0%</td>
<td>28.1%</td>
</tr>
</tbody>
</table>

Population must remain OV-biased at 1000 A.D.

Need this trajectory

% VO

OV-biased

VO-biased

time
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

(YCOE and PPCME2 Corpora)

% Advantage

<table>
<thead>
<tr>
<th></th>
<th>OV Advantage in D0</th>
<th>OV Advantage in D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>-21.0%</td>
<td>28.1%</td>
</tr>
</tbody>
</table>

Population must remain OV-biased at 1000 A.D.

To do this, advantage in intake must be for OV order at 1000 A.D. Otherwise, population changes too quickly to VO-biased distribution.
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

(YCOE and PPCME2 Corpora)

<table>
<thead>
<tr>
<th>% Advantage</th>
<th>OV Advantage in D0</th>
<th>OV Advantage in D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>-21.0%</td>
<td>28.1%</td>
</tr>
</tbody>
</table>

Need this trajectory

Population must remain OV-biased at 1000 A.D.

Requirement for OV advantage at 1000 A.D.: 43% of input is degree-1 data
Linguistic Evolution: Different Individual-Level Learning

Learner uses degree-0 and degree-1 data, and learns from ambiguous data.

(YCOE and PPCME2 Corpora)

<table>
<thead>
<tr>
<th></th>
<th>OV Advantage in D0</th>
<th>OV Advantage in D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>-21.0%</td>
<td>28.1%</td>
</tr>
</tbody>
</table>

Population must remain OV-biased at 1000 A.D.

Need this trajectory

% VO

-28.1%

1000 A.D.

Requirement for OV advantage at 1000 A.D.: 43% of input is degree-1 data …but estimates show only ~16% of it is. Change will be too fast.
Linguistic Evolution: Summary

Some cases where linguistic evolution is driven by individual-level learning. Suggested example: Old English word order.

Individual-level learning: can involve selective learning biases, with strong effects on rate of linguistic change within a population.

Individual-Level Selective Learning:

1. unambiguous data
2. degree-0 data

Additional point: linguistic evolution can inform us about the nature of individual learning.
Linguistic Evolution: Open Questions

(1) If we add complexity to the population model, do we still need these individual-level selective learning biases?

Weight data points in individual intake using various factors:
(a) **spatial location of speaker** with respect to learner
(b) **social status of speaker**
(c) **speaker’s relation to learner** (family, friend, stranger)
(d) **context of data point** (social context, linguistic context)

(2) Are these learning biases necessary if we look at other language changes where individual-level learning is thought to be the main factor driving change at the population-level?
(1) Correct population-level behavior can result from correct individual-level learning behavior in some cases (small discrepancies compounded over time).
(1) Correct population-level behavior can result from correct individual-level learning behavior in some cases (small discrepancies compounded over time).

(2) In the case study examined here, linguistic evolution occurs at the correct rate only when learners employ selective learning biases that cause them to use only a subset of the available data.
(1) Correct population-level behavior can result from correct individual-level learning behavior in some cases (small discrepancies compounded over time).

(2) In the case study examined here, linguistic evolution occurs at the correct rate only when learners employ selective learning biases that cause them to use only a subset of the available data.

(3) Models of linguistic evolution can be empirically grounded and then more easily manipulated to fit the available data (less parameters of variation).

- **Individual-level**: learning period, data distribution, linguistic representation, probabilistic learning
- **Population-level**: population size, population growth rate, time period of change, rate of change
Thank You

Amy Weinberg
Colin Phillips

Norbert Hornstein
Philip Resnik

the Cognitive Neuroscience of Language Lab
at the University of Maryland
Pennsylvania Linguistics Colloquium
The Northwestern Institute on Complex Systems
Individual Framework Applicability

Benefit: Can combine discrete representations, selective learning biases, and probabilistic learning for many types of linguistic knowledge.

Discrete Representation: How much structure is posited for language?

A = linear structure

```
IP
 /   \
Subject Verb Object
```

B = hierarchical structure

```
IP
 /   \nSubject VP
     / \n      Verb Object
```

Discrete Representation: Is the basic word order Object Verb or Verb Object?

A = Object Verb

B = Verb Object
Framework Applicability

Benefit: Can combine discrete representations, selective learning biases, and probabilistic learning for many different problems.

Learning Bias: Use all available data. (Good for probabilistic learner - no data sparseness problem.)

Selective Learning Bias: Use only data perceived as most informative (Fodor 1998, Lightfoot 1999, Drescher 1999).

Selective Learning Bias: Use only data that is more accessible (perhaps for language processing reasons) (Lightfoot 1991).

Selective Learning Bias: Use only data that is perceived as more systematic (Yang 2005).
Framework Applicability

Benefit: Can combine discrete representations, selective learning biases, and probabilistic learning for many different problems.

This can be instantiated as Bayesian updating, a Linear reward-penalty scheme, or any other probabilistic learning procedure.

Max(\(\text{Prob}(p_{\text{vo}} | u)\)) = Max(\(\frac{\text{Prob}(u | p_{\text{vo}}) \times \text{Prob}(p_{\text{vo}})}{\text{Prob}(u)}\))

\(p_{\text{ov}} = p_{\text{ov}} + \gamma(1 - p_{\text{vo}})\)

\(p_{\text{vo}} = 1 - p_{\text{ov}}\)
Estimating Historical $\rho_{VO}$

Known quantities:
Unambiguous and ambiguous data in $d0$ and $d1$
Estimating Historical $\rho_{VO}$
Estimating Historical $\rho_{VO}$

Known quantities: Unambiguous and ambiguous data in $d_0$ and $d_1$

Normalize $d_1$ to $d_0$ distribution: estimate how much $d_1$ unambiguous data was “lost” in $d_0$
Estimating Historical $p_{VO}$
Estimating Historical $p_{VO}$

Known quantities: Unambiguous and ambiguous data in $d_0$ and $d_1$

Normalize $d_1$ to $d_0$ distribution: estimate how much $d_1$ unambiguous data was "lost" in $d_0$

Calculate OV to VO "loss ratio"
Estimating Historical $p_{VO}$

$\text{OV Unamb} \to \text{Amb} \to \text{VO Unamb}$

$\text{OV Unamb} \to \text{Amb} \to \text{VO Unamb}$

= OV to VO “loss” ratio, D1-to-D0
Estimating Historical $p_{VO}$

Known quantities: Unambiguous and ambiguous data in $d_0$ and $d_1$

Normalize $d_1$ to $d_0$ distribution: estimate how much $d_1$ unambiguous data was “lost” in $d_0$

Calculate OV to VO “loss ratio”

Assume $d_1$-to-$d_0$ “loss ratio” is same as underlying-to-$d_1$ “loss” ratio
Estimating Historical $p_{VO}$

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”

Assume d1-to-d0 “loss ratio” is same as underlying-to-d1 “loss” ratio

Use “loss ratio” to estimate how much underlying unambiguous data was “lost” in d1
Assumption: $\approx$
Estimating Historical $\rho_{VO}$

Under-to-D1 OV loss #

D1 OV Unamb

Underlying Unamb OV #

D1-to-D0 "loss" ratio

Under-to-D1 VO loss #

D1 VO Unamb

Underlying Unamb VO #
Estimating Historical $p_{VO}$

\[
\frac{\gamma \cdot d_0 - u_1d_1'}{\gamma \cdot d_0} = L_{d_1tod_0} \cdot \frac{a_1' - (\gamma \cdot d_0 - u_1d_1')}{u_2d_1' + a_1' - (\gamma \cdot d_0 - u_1d_1')}
\]

\[
\gamma = \text{underlying } p_{VO}
\]

\[
d_0 = \text{total degree -0 data, } d_1 = \text{total degree -1 data}
\]

\[
u_1d_1' = \text{normalized unambiguous OV degree -1 data}
\]

\[
u_2d_1' = \text{normalized unambiguous VO degree -1 data}
\]

\[
L_{d_1tod_0} = \text{loss ratio (OV/VO) from degree -1 to degree -0 distribution}
\]

\[
a_1' = \text{normalized ambiguous degree -1 data}
\]
Estimating Historical $p_{VO}$

- **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”

- Assume $d1$-to-$d0$ “loss ratio” is same as underlying-to-$d1$ “loss” ratio

- Use “loss ratio” to estimate how much underlying unambiguous data was “lost” in $d1$

- Calculate $p_{VO}$ from estimated underlying unambiguous data distribution
Estimating Historical $p_{VO}$
Potential 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

\[ \text{Max}(\text{Prob}(p_{vo} \mid u)) = \text{Max}(\frac{\text{Prob}(u \mid p_{vo}) \ast \text{Prob}(p_{vo})}{\text{Prob}(u)}) \]

Bayes’ Rule, find maximum of a posteriori (MAP) probability
Manning & Schütze (1999)
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

Max(Prob(p_{VO} | u)) = \max\left(\frac{\text{Prob}(u | p_{VO}) \times \text{Prob}(p_{VO})}{\text{Prob}(u)}\right)

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

\text{Prob}(p_{VO}) = \text{probability of seeing } r \text{ out of } n \text{ data points that are unambiguous for VO, for } 0 \leq r \leq n
\quad = \binom{n}{r} \times p_{VO}^r \times (1 - p_{VO})^{n-r}
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

\[
\text{Max}(\text{Prob}(p_{\text{VO}} | u)) = \text{Max}( \frac{p_{\text{VO}} \cdot \binom{n}{r} \cdot p_{\text{VO}}^r \cdot (1 - p_{\text{VO}})^{n-r}}{\text{Prob}(u)}) \quad \text{(for each point } r, 0 \leq r \leq n) \]

\[
\frac{d}{dp_{\text{VO}}} \left( \frac{p_{\text{VO}} \cdot \binom{n}{r} \cdot p_{\text{VO}}^r \cdot (1 - p_{\text{VO}})^{n-r}}{\text{Prob}(u)} \right) = 0
\]

\[
\frac{d}{dp_{\text{VO}}} \left( \frac{p_{\text{VO}} \cdot \binom{n}{r} \cdot p_{\text{VO}}^r \cdot (1 - p_{\text{VO}})^{n-r}}{\text{Prob}(u)} \right) = 0 \quad \text{(P}(u) \text{ is constant with respect to } p_{\text{VO}}) \]

\[p_{\text{VO}} = \frac{r + 1}{n + 1}\]
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

\[
p_{\text{VO}} = \frac{r + 1}{n + 1}, \quad r = p_{\text{VO} \text{prev}} \times n
\]

Replace 1 in numerator and denominator with
\[c = p_{\text{VO} \text{prev}} \times m \text{ if VO, } c = (1 - p_{\text{VO} \text{prev}}) \times m \text{ if OV}
\]
\[3.0 \leq m \leq 5.0\]

\[
p_{\text{VO}} = \frac{p_{\text{VO} \text{prev}} \times n + c}{n + c}
\]
Other Ways to Interpret Ambiguous Data

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)