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.

Linguistic Evolution, In Brief
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)

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.

Road Map

I. Individual Language Learning
   The Nature of Linguistic Knowledge
   Individual Learning Framework

II. Linguistic Evolution: Case Study
    Old English Word Order
    Modeling Individuals
    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?

Ex:
What are the contrastive sounds of a language?

Extraction
Ex: Where are words in fluent speech?

Ex:
Where are words in fluent speech?

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

blink~blinked
confide~confided
drink~drank

who's afraid of the big bad wolf
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
- Object Verb underlying
- Subject Verb Object underlying

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 as individual knowledge distribution, from learner’s perspective) Why not?
Underlying Distribution vs. Observable Distribution

Subject Verb Object

German/Old English

Subject Verb Object underlying

Speaker generates utterance

Subject Verb Object

Every utterance generated by speaker is either OV or VO order in the underlying distribution

Learner interprets utterance

The learner encounters data that is ambiguous between the two options. Distribution depends on learner’s interpretation of ambiguous data

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 as individual knowledge distribution, from learner’s perspective)

Why not?

Due to learner interpretation bias
% VO

Old English

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

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

<table>
<thead>
<tr>
<th>VO</th>
<th>OV</th>
</tr>
</thead>
<tbody>
<tr>
<td>P&lt;sub&gt;VO&lt;/sub&gt; = 77</td>
<td>P&lt;sub&gt;OV&lt;/sub&gt; = 77</td>
</tr>
</tbody>
</table>

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.

~1200 A.D.: VO-biased

<table>
<thead>
<tr>
<th>VO</th>
<th>OV</th>
</tr>
</thead>
<tbody>
<tr>
<td>P&lt;sub&gt;VO&lt;/sub&gt; = 77</td>
<td>P&lt;sub&gt;OV&lt;/sub&gt; = 77</td>
</tr>
</tbody>
</table>

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

VO unambiguous data:
- TensedVerb...
- Verb-Marker...
- Object...

OV unambiguous data:
- TensedVerb...
- Object...

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

Old English

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

VO unambiguous data:
- TensedVerb...
- Verb-Marker...
- Object...

OV unambiguous data:
- TensedVerb...
- Object...

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

Modeling Individuals: Learning Biases

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

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

\[ \text{P} = \text{tensed verb} \rightarrow \text{2nd phrase} \]

Modeling Individuals: Learning Biases

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

<table>
<thead>
<tr>
<th>VO</th>
<th>OV</th>
</tr>
</thead>
<tbody>
<tr>
<td>P&lt;sub&gt;VO&lt;/sub&gt; = 77</td>
<td>P&lt;sub&gt;OV&lt;/sub&gt; = 77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VO</th>
<th>OV</th>
</tr>
</thead>
<tbody>
<tr>
<td>P&lt;sub&gt;VO&lt;/sub&gt; = 77</td>
<td>P&lt;sub&gt;OV&lt;/sub&gt; = 77</td>
</tr>
</tbody>
</table>

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

VO unambiguous data:
- TensedVerb...
- Verb-Marker...
- Object...

OV unambiguous data:
- TensedVerb...
- Object...

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

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% 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

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

Model parameters:
- $c$ represents learner’s confidence in data point (calibrated from data)
- $n$ represents quantity of intake (2000)

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

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 $n$.

Biased Data Intake Distributions in Old English

$p_{VO}$ 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>Time Period</th>
<th>OV Advantage in Unamb OV</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>

$% VO_{OV}$ vs $% VO_{VO}$
(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 $p_{0.6}$ 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 $p_{0.6}$.
Empirical Grounding Issues: What exactly is the underlying distribution?

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

Historical data distributions: some data are ambiguous

Observations:
(1) Degree-1 data less ambiguous than degree-0 data.

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

<table>
<thead>
<tr>
<th>Year Range</th>
<th>OV Advantage in Unamb</th>
<th>OV Advantage in Unamb</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>-3.7%</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.

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.

Average $p_{VO} = \frac{(OV_{bias} - VO_{bias})}{(1 + (OV_{bias} - VO_{bias}))}$

<table>
<thead>
<tr>
<th>Time</th>
<th>Initialization</th>
<th>Calibration</th>
<th>Termination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average $p_{VO}$</td>
<td>0.234</td>
<td>0.310</td>
<td>0.747</td>
</tr>
</tbody>
</table>

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.

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.

Average $p_{VO} = \frac{(OV_{bias} - VO_{bias})}{(1 + (OV_{bias} - VO_{bias}))}$

<table>
<thead>
<tr>
<th>Time</th>
<th>Initialization</th>
<th>Calibration</th>
<th>Termination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average $p_{VO}$</td>
<td>0.234</td>
<td>0.310</td>
<td>0.747</td>
</tr>
</tbody>
</table>

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>Year</th>
<th>% VO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 A.D.</td>
<td>0.234</td>
</tr>
<tr>
<td>1000-1150 A.D.</td>
<td>0.310</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>0.747</td>
</tr>
</tbody>
</table>

Linguistic Evolution: Change at the Historically-Attested Rate
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.

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.

Problem: VO-biased all the way through, even at 1000 A.D.
Need this trajectory
Change is too fast!
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.

But population must become VO-biased.

Need this trajectory.

% VO

OV-biased

VO-biased

Need this trajectory.

Can a population learning from degree-1 data make the change to VO-biased?

Time

Estimated \( p_{\text{VO}} \)

Model \( p_{\text{VO}} \) at 1200 A.D.

Need this trajectory.

% VO

OV-biased

VO-biased

Need this trajectory.

Can a population learning from degree-1 data make the change to VO-biased?

Time

Estimated \( p_{\text{VO}} \)

Model \( p_{\text{VO}} \) at 1200 A.D.

Need this trajectory.

% VO

OV-biased

VO-biased

Need this trajectory.

Can a population learning from degree-1 data make the change to VO-biased?

Time

Estimated \( p_{\text{VO}} \)

Model \( p_{\text{VO}} \) at 1200 A.D.

Need this trajectory.

% VO

OV-biased

VO-biased

Need this trajectory.

Can a population learning from degree-1 data make the change to VO-biased?

Time

Estimated \( p_{\text{VO}} \)

Model \( p_{\text{VO}} \) at 1200 A.D.

Need this trajectory.

% VO

OV-biased

VO-biased

Need this trajectory.

Can a population learning from degree-1 data make the change to VO-biased?

Time

Estimated \( p_{\text{VO}} \)

Model \( p_{\text{VO}} \) at 1200 A.D.

Need this trajectory.

% VO

OV-biased

VO-biased

Need this trajectory.

Can a population learning from degree-1 data make the change to VO-biased?

Time

Estimated \( p_{\text{VO}} \)

Model \( p_{\text{VO}} \) at 1200 A.D.

Need this trajectory.

% VO

OV-biased

VO-biased

Need this trajectory.
**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>

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

Need this trajectory

% VO OV-biased VO-biased

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: 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 social complexity to the population model, do we still need these individual-level selective learning biases? Weight data points in individual intake using various factors:
   - spatial location of speaker with respect to learner
   - social status of speaker
   - speaker's relation to learner (family, friend, stranger)
   - 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?
Learning-Driven Linguistic Evolution: Take-Home Messages

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 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
  - B = Hierarchical structure

- **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, Dresher 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).

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

\[
\text{Max(Prob}(p_{VO} | u)) = \text{Max}(\text{Prob}(u | p_{VO}) \times \text{Prob}(p_{VO}) / \text{Prob}(u))
\]

\[
p_{VO} = p_{VO} + (1-p_{VO})
\]

\[
p_{VO} = 1 - p_{VO}
\]

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 \( d_0 \)
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"
Assumption: $\approx$

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 $p_{VO}$ from estimated underlying unambiguous data distribution.**

- Use "loss ratio" to estimate how much underlying unambiguous data was "lost" in $d_1$

- Assume $d_1$-to-$d_0$ "loss ratio" is same as underlying-to-$d_1$ "loss ratio"
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 (Köber, 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.

Scandinavian Influence, Perfect Learning

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

$$\text{Max}(\text{Prob}(p_{\text{VO}}|u)) = \text{Max}(\frac{\text{Prob}(u|p_{\text{VO}}) \cdot \text{Prob}(p_{\text{VO}})}{\text{Prob}(u)})$$

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(pVO | u)) = Max(Prob(u | pVO) * Prob(pVO) / Prob(u))

Prob(u | pVO) = probability of seeing unambiguous data point u, given pVO

Prob(pVO) = probability of seeing r out of n data points that are unambiguous for VO, for 0 <= r <= n

= \binom{n}{r} pVO^r (1- pVO)^{n-r}

pVO = \frac{r + 1}{n + 1}

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 & Litz, in submission)
- unambiguous = highest weight
- moderately ambiguous = lower weight
- fully ambiguous = lowest weight (ignore)

(3) randomly assign to one hypothesis (Yang 2002)
Perceived Unambiguous Data: OV

Unambiguous OV data
(1) Tensed Verb is immediately post-Object

\[
\begin{align*}
\text{he} & \quad \text{him} \quad \text{may-pray} \\
\text{He may pray (to) him}
\end{align*}
\]

(Ælfric's Letter to Wulfsige, 87.107, ~1075 A.D.)

(2) Verb-Marker is immediately post-Object

\[
\begin{align*}
\text{we} & \quad \text{sculen} \quad \text{should} \quad \text{our evil practices} \quad \text{abandon} \\
\text{We should abandon our evil practices.}
\end{align*}
\]

(Alcuin's De Virtutibus et Vitiis, 70.52, ~1150 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).

Unreliable Verb-Markers

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

\[
\begin{align*}
\text{NEG} & \quad \text{I never saw} \quad \text{the city} \\
\text{Never did I see the city.}
\end{align*}
\]

(Ælfric, Homilies. I.572.3, between 900 and 1000 A.D.)