Linguistic knowledge is transmitted in a population via interaction with other speakers in the population.

The information speakers transmit (observable data) is generated from their own linguistic knowledge, which involves (mostly unconscious) knowledge of the form of the language and (mostly conscious) knowledge of the meaning of the information.

Example: Goblins steal children.
Form: “Subject Verb Object”
Meaning = Stealer(Goblins) & Stolen(Children)

Focus today: the form of the language.
Population-level changes over time depend on what language forms speakers pass to subsequent generations and how those language forms are integrated into an individual’s linguistic knowledge of the language’s form.

Integrating Linguistic Information

Premise from linguistics: Not all linguistic knowledge is created equal

Some knowledge can be altered throughout an individual’s life

(example: vocabulary)

Knowledge of meaning

Passed to young individuals

Passed to individuals of all ages

Change to knowledge that is alterable only early on

Implication: The way in which young learners integrate this 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.

Passed to young individuals
Implication: The way in which young learners integrate this 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 (Pearl & Weinberg 2007)
   Modeling Populations
   Issues in Empirical Grounding
   Interpretation Biases

The Nature of Linguistic Knowledge

Premise from linguistics: The form of language (which is the observable word order, also known as syntax) is not necessarily generated by putting one word next to another in order. Instead, speakers have an underlying knowledge of the syntax of the language, which they use to generate the observable form. Sometimes, this generative system involves reordering words and phrases.

Observable Form: Subject  Verb  Object

One way to generate this form: Subject + Verb + Object [-English] = Subject  Verb  Object
The Nature of Linguistic Knowledge

Premise from linguistics: The form of language (which is the observable word order, also known as syntax) is not necessarily generated by putting one word next to another in order. Instead, speakers have an underlying knowledge of the syntax of the language, which they use to generate the observable form. Sometimes, this generative system involves reordering words and phrases.

Observable Form: Subject Verb Object

Another way: Subject + Object + Verb = Subject Object Verb
(move Verb to front)

Subject Verb Object [~German]

Verb Subject Object
(move Subject to front)

Subject Verb Object [~German]

Subject Verb

Objective Form: Subject Verb Object

Another way: Subject + Object + Verb = Subject Object Verb
(move Object after Verb)

Subject Verb Object

~Kannada

The individual learning framework: 3 components

1. Hypothesis space

2. Data

3. Update procedure

Observed data:
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

Old English

From linguistics: Old English generative system is similar to the German generative system.

Subject      Verb      Object
English = Old English

Subject      Verb      Object
German = Old English

Basic question for the Old English learner: Is the basic word order Object Verb or Verb Object?

Old English: Word Order

The not-so-basic answer (from Pintzuk 2002, and other historical linguists): Old English apparently used both kinds of word orders (Object Verb and Verb Object) for several hundred years.

An individual speaker had a probability distribution between these two orders. Learners therefore want to learn the appropriate probability distribution over these two orders, rather than simply which word order is correct.

<table>
<thead>
<tr>
<th></th>
<th>OV</th>
<th>VO</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_{OV}</td>
<td>??</td>
<td>??</td>
</tr>
</tbody>
</table>

Observable Old English fact: shift from mostly OV order to mostly VO order within a fairly short period of time (historically speaking).
Individual Knowledge (underlying probability in speaker’s mind): probability distribution between OV and VO orders.

Individual Usage (which is the observable data for the learner): probability distribution between OV and VO orders.

Important: From the learner’s perspective, this distribution is not necessarily the same as the individual knowledge distribution. Why not?

Underlying Distribution vs. Observable Distribution

Subject Verb Object
Observable order: Verb Object

Subject Verb Object underlying
Learner interprets utterance

Every utterance generated by speaker is either OV or VO order in the underlying distribution.
**Underlying Distribution vs. Observable Distribution**

The learner encounters data that are ambiguous between the two options. Interpreted distribution depends on learner's interpretation of ambiguous data.

**Old English**

Changing basic word order 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 (which is the observable data for the learner): probability distribution between OV and VO orders.

Important: From the learner's perspective, this distribution is not necessarily the same as the individual knowledge distribution, due to learner interpretation bias.

**Estimates of average individual usage from historical corpora:**

- ~1000 A.D.-1150 A.D.: OV-biased
- ~1200 A.D.: VO-biased

<table>
<thead>
<tr>
<th>Time</th>
<th>OV</th>
<th>VO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>??</td>
<td>??</td>
</tr>
<tr>
<td>1150</td>
<td>??</td>
<td>??</td>
</tr>
<tr>
<td>1200</td>
<td>??</td>
<td>??</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.

**Modeling Individuals: Learning Biases**

Biases human learners may have or need

- Interpretation Bias: Use only data perceived as most informative (Fodor 1998, Lightfoot 1999, Dresher 1999).
- Interpretation Bias: Use only data that are more accessible (perhaps for language processing reasons) (Lightfoot 1991).

% VO vs. time
Modeling Individuals: Learning Biases

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

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

Unambiguous data are the most informative data. Look for data that seem to be unambiguous for OV order, and for data that seem to be unambiguous for VO order. From linguistics: these data will take a specific form.

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

Lightfoot 1991: Data in structurally simple clauses (degree-0 clauses) should be used. Data in other clauses (degree-1 or more) should be ignored.  

Degree = level of embedding  

Lightfoot 1991: Data in structurally simple clauses (degree-0 clauses) should be used. Data in other clauses (degree-1 or more) should be ignored.  

Degree = level of embedding  

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

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

[-------------Degree-1----------]
Modeling Individuals: Learning Biases

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

Interpretation Bias: Use only data that are more accessible, which is in degree-0 clauses (Lightfoot 1991).

Data intake = degree-0 unambiguous data

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.

Modeling Individuals: Knowledge & Learning

Individual learner tracks \( p_{VO} \) = probability of using VO (probability of using OV = 1 - \( p_{VO} \)).

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

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

Data comes from other members of population, filtered through interpretation biases.
Modeling Individuals: Knowledge & Learning

Individual learner tracks $p_{VO}$ = 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 comes from other members of population, filtered through interpretation biases.

Individual update: Bayesian updating for binomial distribution (Chew 1971), adapted

Zoom-In on Updating Procedure

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

Involves previous probability is 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 an individual encounters before setting $p_{VO}$ (2000 = length of learning period)

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</th>
<th>OV Advantage in Unamb. DO</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>

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 $n$. 
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 growth rate estimated from population statistics of the time period (Koenigsberger & Briggs 1987)
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 $p_{VO}$.

(6) Repeat steps (3-5) until the year 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

$p_{VO}$: 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

Observations:

1. Degree-1 data less ambiguous than degree-0 data.

<table>
<thead>
<tr>
<th>Year</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>89%</td>
<td>25%</td>
</tr>
<tr>
<td>1200 A.D.</td>
<td>71%</td>
<td>16%</td>
</tr>
</tbody>
</table>

Observations:

1. Degree-1 data less ambiguous than degree-0 data.

Idea: Ambiguous data distorts underlying distribution. Observation: 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.

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 underlying unambiguous data distribution and the degree-1 distribution in a speaker’s mind.

\[
\text{Idea:} \quad \text{Ambiguous data distorts underlying distribution.}
\]

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

\[
\begin{align*}
\text{Degree-1 data} & \quad \text{less ambiguous than degree-0 data.} \\
\text{Advantage is} & \quad \text{magnified in degree-1.}
\end{align*}
\]

\[
\begin{align*}
\text{Plan of Action:} & \quad \text{Use the difference in distortion between the degree-0 and degree-1 unambiguous data distributions to estimate the difference in distortion between the underlying unambiguous data distribution and the degree-1 distribution in a speaker’s mind.}
\end{align*}
\]

\[
\text{Idea:} \quad \text{Ambiguous data distorts underlying distribution.}
\]

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

\[
\begin{align*}
\text{Degree-1 data} & \quad \text{less ambiguous than degree-0 data.} \\
\text{Advantage is} & \quad \text{magnified in degree-1.}
\end{align*}
\]

\[
\begin{align*}
\text{Plan of Action:} & \quad \text{Use the difference in distortion between the degree-0 and degree-1 unambiguous data distributions to estimate the difference in distortion between the underlying unambiguous data distribution and the degree-1 distribution in a speaker’s mind.}
\end{align*}
\]

\[
\text{Idea:} \quad \text{Ambiguous data distorts underlying distribution.}
\]

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

\[
\begin{align*}
\text{Degree-1 data} & \quad \text{less ambiguous than degree-0 data.} \\
\text{Advantage is} & \quad \text{magnified in degree-1.}
\end{align*}
\]

\[
\begin{align*}
\text{Plan of Action:} & \quad \text{Use the difference in distortion between the degree-0 and degree-1 unambiguous data distributions to estimate the difference in distortion between the underlying unambiguous data distribution and the degree-1 distribution in a speaker’s mind.}
\end{align*}
\]

\[
\text{Idea:} \quad \text{Ambiguous data distorts underlying distribution.}
\]

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

\[
\begin{align*}
\text{Degree-1 data} & \quad \text{less ambiguous than degree-0 data.} \\
\text{Advantage is} & \quad \text{magnified in degree-1.}
\end{align*}
\]

\[
\begin{align*}
\text{Plan of Action:} & \quad \text{Use the difference in distortion between the degree-0 and degree-1 unambiguous data distributions to estimate the difference in distortion between the underlying unambiguous data distribution and the degree-1 distribution in a speaker’s mind.}
\end{align*}
\]
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.

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

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

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.

Linguistic Evolution:
Change at the Historically-Attested Rate

Learners have interpretive bias on data

Linguistic Evolution:
Different Individual-Level Learning

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

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

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

Need this trajectory

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

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

Learner uses degree-0 and degree-1 unambiguous data

(YCOE and PPCME2 Corpora)

<table>
<thead>
<tr>
<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>
</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>

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>Time</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>

Need this trajectory
Very strongly OV-biased before 1150 A.D.

But population must become VO-biased.

% VO

Estimated $p_{VO}$ at 1200 A.D.
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.

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)

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

Need this trajectory

% VO

To do this, advantage in intake must be for OV order at 1000 A.D. Otherwise, population changes too quickly to VO-biased distribution.

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.

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: Open Questions

1. If we add social complexity to the population model, do we still need these individual-level 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)
   - Second language speaker influence (Scandinavian (VO) vs. Old English (OV))? (a)

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?

3. What if we relax the constraint that probability for word order usage can only be altered early on? Young individual's learning is not as crucial, but perhaps a model of this kind can still produce the same linguistic evolution.

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).
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 interpretive 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 length, data distribution, linguistic representation (generative system), incremental 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, UMaryland
The Workshop on Psychocomputational Models of Human Language Acquisition
Pennsylvania Linguistics Colloquium
The Northwestern Institute on Complex Systems
The Institute for Mathematical Behavioral Sciences, UC Irvine
Individual Framework Applicability

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

**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 1997).

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

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

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

\[
p_{vO} = p_{vO} + \gamma(1-p_{vO})
p_{vO} = 1 - p_{vO}
\]
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"

Assumption: $\approx$

$D_0$

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

$D_1$

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

$= OV$ to $VO$ "loss" ratio, $D_1$-to-$D_0$
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"

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"

Assumption:

$$D_0 \approx D_1$$

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" 
- Assume $d_1$-to-$d_0$ "loss ratio" is same as underlying-to-$d_1$ "loss ratio"
- Use "loss ratio" to estimate how much underlying unambiguous data was "lost" in $d_1$
- Calculate $p_{VO}$ from estimated underlying unambiguous data distribution

### UOV Unamb

- $d_0$: normal degree: 0 data
- $d_1$: total degree: 1 data
- $d_0'$: normalized unambiguous OV degree: 0 data
- $d_1'$: normalized unambiguous VO degree: 1 data
- $L_{d1to d0}$: loss ratio (OV/VO) from degree 1 to degree 0 distribution
- $a_{d1}'$: normalized ambiguous degree: 1 data

\[
\begin{align*}
\text{Estimating Historical } p_{VO} &= \text{UOV Unamb} \quad \text{U VO Unamb} \\
\text{U OV Unamb} &= \text{U VO Unamb} \\
\text{U OV Unamb} &= \text{U VO Unamb} \\
\end{align*}
\]
**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

**Scandinavian Influence, Perfect Learning**

Even with severe Scandinavian “second language learning” VO biases and an increasing influx of Scandinavians, the change still does not happen swiftly enough.

**Scandinavian Influence, Perfect Learning**

May be able to get change to occur quickly enough with an additional pressure of social prestige for the VO order that Scandinavian used. However, it’s not clear there’s evidence for this historically.
Deriving the Bayesian Update Equations for a Hypothesis Space with 2 Hypotheses

Max(ProbpVO | u) = Max(proba | pVO) * Prob(pVO) \over Prob(u) \) 

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

Max(ProbpVO | u) = Max(proba | pVO) * Prob(pVO) \over 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 = r + 1 \over n + 1 

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

\[ \frac{d}{dpVO} \left[ \frac{pVO * \binom{n}{r} * pVO^r * (1-pVO)^{n-r}}{Prob(u)} \right] = 0 \] 

(P(a) is constant with respect to pVO)

Replace 1 in numerator and denominator with 
c = pVOprev * m if VO, c = (1 - pVOprev) * m if OV 
3.0 <= m <= 5.0

\[ pVO = \frac{pVOprev * n + c}{n + c} \]
Other Ways to Interpret Ambiguous Data

Strategies for assessing ambiguous data
(1) assume base-generation (surface order is correct order)
   - 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)

Perceived Unambiguous Data: OV

Unambiguous OV data
(1) Tensed Verb is immediately post-Object
   hehe
   hyne
   gebedde
   gebidden
   ‘He may pray (to) him’
   (Ælfric’s Letter to Wulfsige, 87.107, ~1075 A.D.)

(2) Verb-Marker is immediately post-Object
   wesce
e
   sculcn
   [word yfel beases]
   [word verb marker]
   ‘We should abandon our evil practices’
   (Alcuin’s De Virilibus et Vilitis, 70.52, ~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)

A [mid his stefne] he awoke 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

Then Paul lifted his head up.
(Bickling 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).

Unreliable Verb-Markers

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

Never did I see the city.
(Ellis, Homilies. I.572.3, between 900 and 1000 A.D.)

This can lead to ambiguous data (if only one were present) or data that appear to be unambiguous for the VO order (mafræ) even though Old English at this period of time was strongly OV-biased.