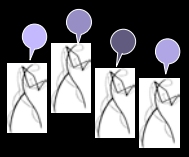
Learning-Driven Linguistic Evolution

Lisa Pearl
Cognitive Sciences, UC Irvine
February 10, 2009
Seminar on Social Dynamics

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 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.

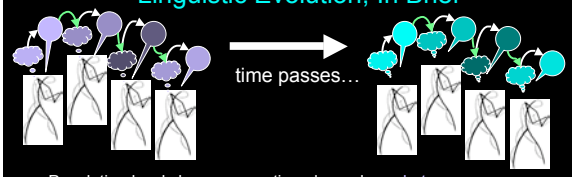
Example: Goblins steal children.
Form: "Subject Verb Object"
Meaning ≈ Stealer(Goblins) & Stolen(Children)

Focus today: the form of the language

Linguistic Evolution, In Brief

Speakers adjust their linguistic knowledge of the language form based on the data encountered from other population members.

Linguistic Evolution, In Brief



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

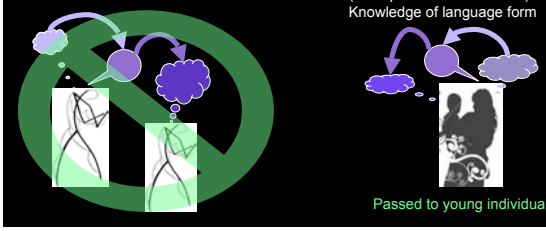
Passed to individuals of all ages

Integrating Linguistic Information

Premise from linguistics: Not all linguistic knowledge is created equal

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

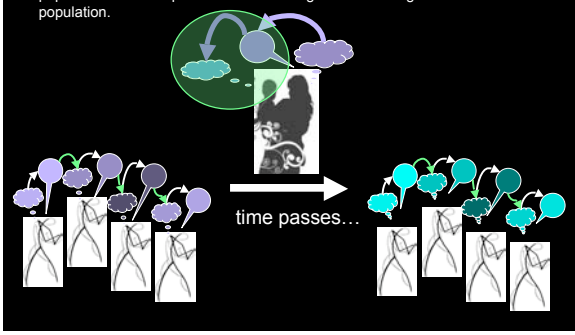
(example: word order rules)
Knowledge of language form



Passed to young individuals

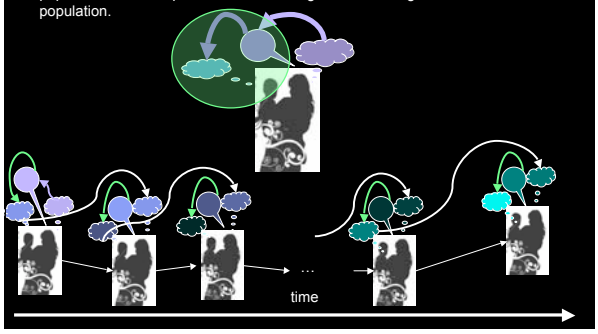
Change to knowledge that is alterable only early on

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.



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.



Road Map

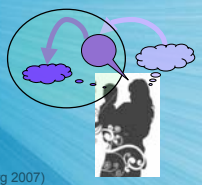
I. Individual Language Learning
   - The Nature of Linguistic Knowledge
   - Individual Learning Framework

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

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) Verb Subject Object
(move Subject to front) Subject Verb Object [~German]

The Nature of Linguistic Knowledge

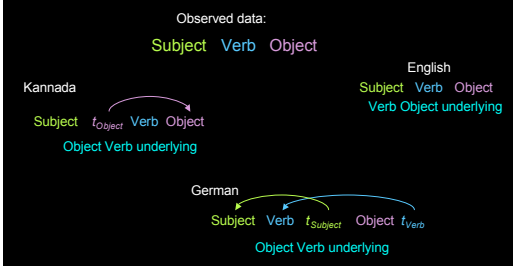
Premise from linguistics: The form of language (which is the observable word order, also known as syntax) is not necessarily generated by putting one word next to another in order. Instead, speakers have an underlying knowledge of the syntax of the language, which they use to generate the observable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

Observable Form: Subject Verb Object

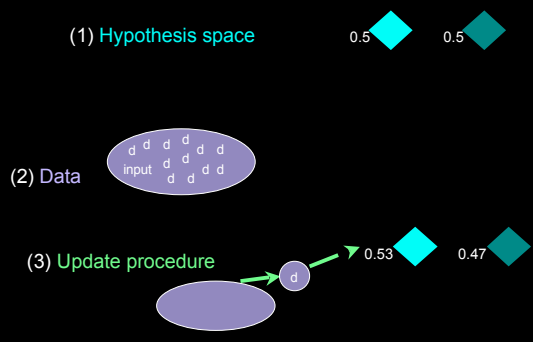
Another way: Subject + Object + Verb = Subject Object Verb
(move Object after Verb) Subject Verb Object
[~Kannada]

The Nature of Linguistic Knowledge

Problem for the learner: What is the generative system that creates the observed form of the language?
Ex: Is the basic word order Object Verb or Verb Object?



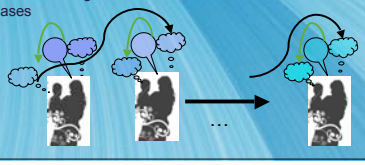
The individual learning framework: 3 components



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



Old English

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

Observed data:
Subject Verb Object

English ≠ Old English
Subject Verb Object

German ≈ Old English
Subject Verb $t_{Subject}$ Object t_{Verb}

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.

OV P_{OV} = ?? VO P_{VO} = ??

Observable Old English fact: shift from mostly OV order to mostly VO order within a fairly short period of time (historically speaking).

Old English

Changing basic word order in Old English:
Object-Verb (OV) vs. Verb-Object (VO) order

OV P_{OV} = ?? VO P_{VO} = ??

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

Old English

Changing basic word order in Old English:
Object-Verb (OV) vs. Verb-Object (VO) order

OV P_{OV} = ??

VO P_{VO} = ??

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

German/Old English

Observable order: Verb Object

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

Object Verb underlying

Speaker generates utterance

Underlying Distribution vs. Observable Distribution

Subject Verb Object

?

Observable order: Verb Obje

?

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

Object Verb underlying

Subject Verb Object

Verb Object underlying

?

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

Verb Object underlying

Learner interprets utterance

Underlying Distribution vs. Observable Distribution

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

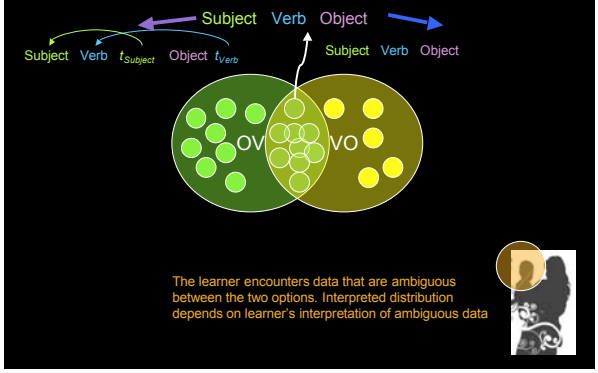
Subject Verb Object

OV

VO

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

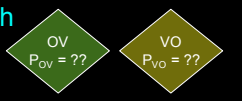
Underlying Distribution vs. Observable Distribution



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

Old English

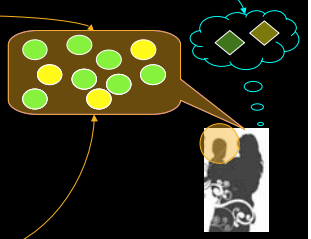
Changing basic word order in Old English:
Object-Verb (OV) vs. Verb-Object (VO) order



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.

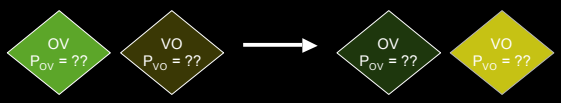


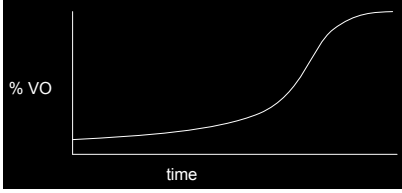
Old English

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

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

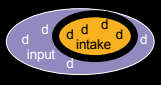
~1200 A.D.: VO-biased





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

Modeling Individuals: Learning Biases

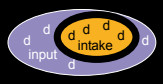


Biases human learners may have or need

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

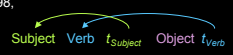
Interpretation Bias: Use only data that are more accessible (perhaps for language processing reasons) (Lightfoot 1991).

Modeling Individuals: Learning Biases



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

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

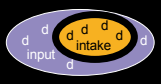


Interpretation Bias: Use only data that are more accessible (perhaps for language processing reasons) (Lightfoot 1991).

In some adult systems, the Verbs can move so that the observable word order does not necessarily reflect the basic word order.

When the Verb moves in some of these systems, it moves to the second position of the sentence.

Modeling Individuals: Learning Biases



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.

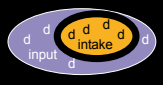
Interpretation Bias: Use only data perceived as most informative (Fodor 1998, Lightfoot 1999, Dresher 1999).

OV unambiguous data:
[…]$_{XP}$ … Object TensedVerb …
…TensedVerb … Object Verb-Marker …

VO unambiguous data:
[…]$_{XP}$ […]$_{XP}$ … TensedVerb Object …
…TensedVerb … Verb-Marker Object …

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

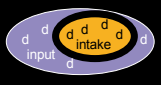
Modeling Individuals: Learning Biases



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

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

Modeling Individuals: Learning Biases



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

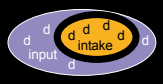
Lightfoot 1991: Data in structurally simple clauses (degree-0 clauses) should be used. Data in other clauses (degree-1 or more) should be ignored.

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

Degree = level of embedding

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

Modeling Individuals: Learning Biases



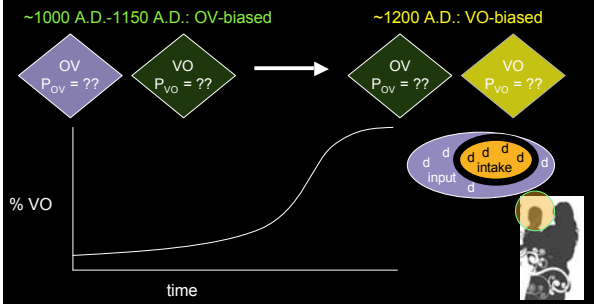
Interpretation Bias: Use only data perceived as unambiguous (Fodor 1998, Lightfoot 1999, Dresher 1999).

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

Data intake = degree-0 unambiguous data

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.

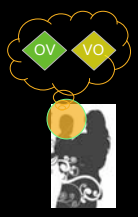


Modeling Individuals: Knowledge & Learning

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)



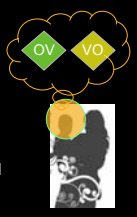
Modeling Individuals: Knowledge & Learning

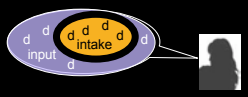
Individual learner tracks p_{VO} = probability of using VO
(probability of using OV = 1 - p_{VO})

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

Initial p_{VO} = 0.5 (unbiased)

Data comes from other members of population, filtered through interpretation biases.





Modeling Individuals: Knowledge & Learning

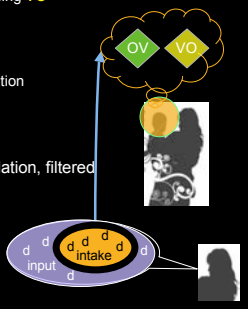
Individual learner tracks p_{VO} = probability of using VO
(probability of using OV = 1 - p_{VO})

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

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

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

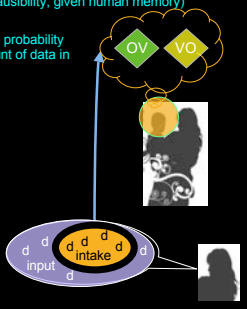
Important: Online update procedure (psychological plausibility, given human memory)

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 an individual encounters before setting p_{VO} (2000) = length of learning period



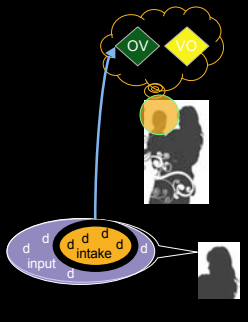
Individual-Level Learning Algorithm

(1) Initial p_{VO} = 0.5.

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

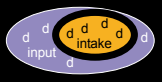
(3) If the data point is degree-0 and unambiguous, use update functions to shift hypothesis probabilities.

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



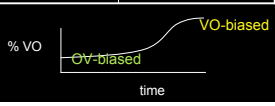
Biased Data Intake Distributions in Old English

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.

	OV Advantage in **Unamb D0**	
1000 A.D.	19.5%	OV-biased
1000-1150 A.D.	2.8%	
1200 A.D.	-2.7%	VO-biased



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.

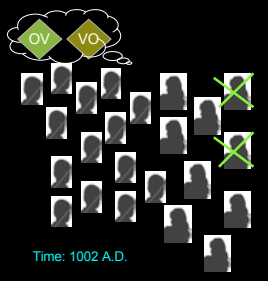
Population-Level Model

(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.

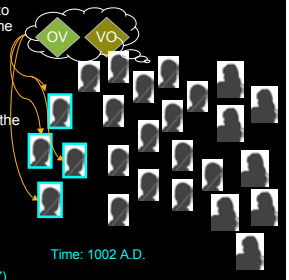
(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(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.

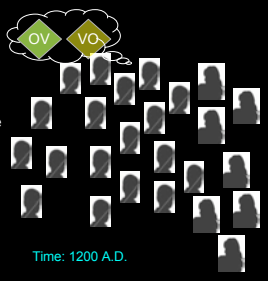
(2) Initialize the members of the population to the average p_{VO} at 1000 A.D. Set the time to 1000 A.D.

(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}.

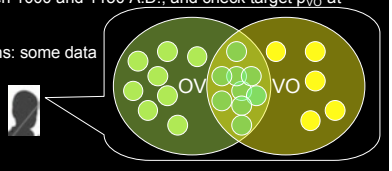
(6) Repeat steps (3-5) until the year 1200 A.D.



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.

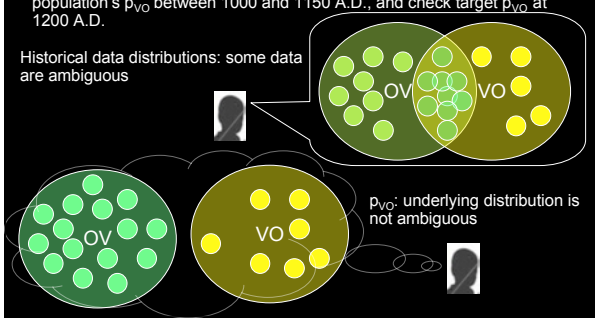
Historical data distributions: some data are ambiguous



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

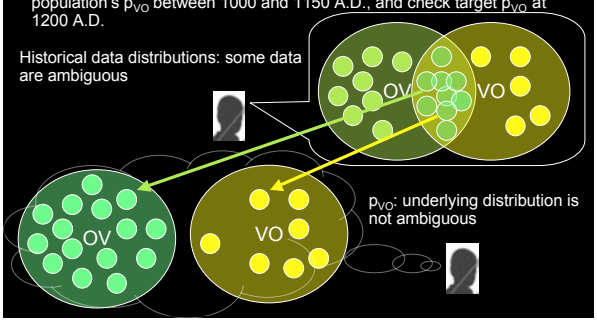


p_{VO}: underlying distribution is not ambiguous

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



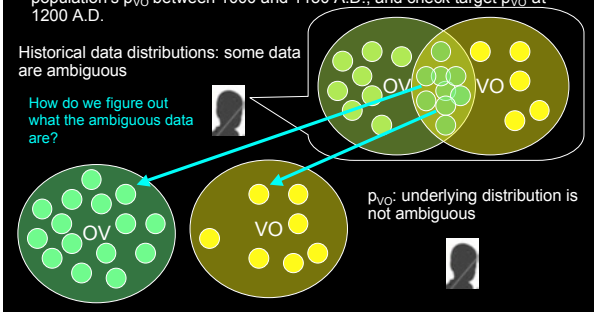
p_{VO}: underlying distribution is not ambiguous

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

	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.

Empirical Grounding Issues: What exactly is the underlying distribution?

(YCOE and PPCME2 Corpora)
% Advantage

	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%

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.

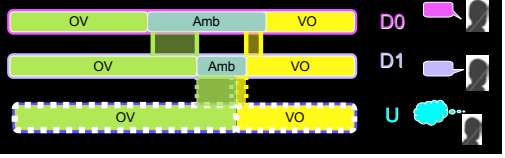
Idea: Ambiguous data distorts underlying distribution.
Observation: degree-1 distribution less distorted from underlying distribution.

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.

Idea: Ambiguous data distorts underlying distribution.
Observation: 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.

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

known quantities

γ = underlying p_{VO}
d0 = total degree - 0 data, **dl** = 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

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

derived quantities

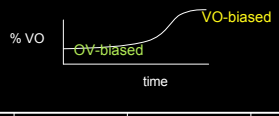
$$\gamma = \frac{-(\mathbf{d0})(\mathbf{d0} + \mathbf{u1d1'} - \mathbf{Ld1tod0} * (\mathbf{ad1'} + \mathbf{u1d1'}))}{2(\mathbf{Ld1tod0} + 1)(\mathbf{d0}^2)}$$

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

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.

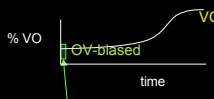


	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747
		OV-biased	VO-biased

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.

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

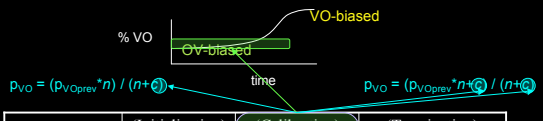


	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747
		OV-biased	VO-biased

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.

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

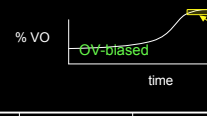


	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747
		OV-biased	VO-biased

Empirical Grounding Issues: What exactly is the underlying distribution?

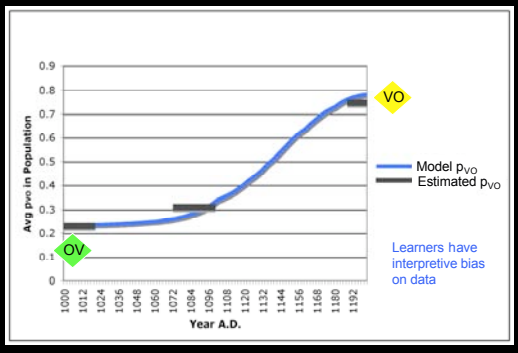
Observations:
(1) Degree-1 data less ambiguous than degree-0 data.
(2) Advantage is magnified in degree-1.

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



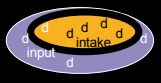
	(Initialization) 1000 A.D.	(Calibration) 1000-1150 A.D.	(Termination) 1200 A.D.
Average p_{VO}	0.234	0.310	0.747
		OV-biased	VO-biased

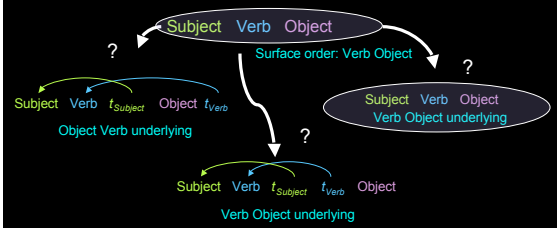
Linguistic Evolution: Change at the Historically-Attested Rate



Linguistic Evolution: Different Individual-Level Learning

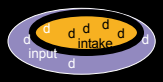
Learner uses ambiguous data. Strategy for learning: assume surface word order is actual order. (Fodor 1998)





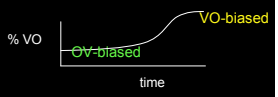
Linguistic Evolution: Different Individual-Level Learning

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



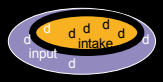
Advantage in intake determines learner's ending distribution between OV and VO order.

Need this trajectory



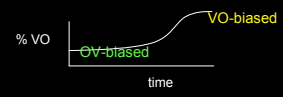
Linguistic Evolution: Different Individual-Level Learning

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.

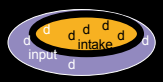
Need this trajectory



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

Linguistic Evolution: Different Individual-Level Learning

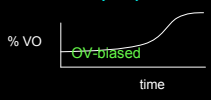
Learner uses ambiguous data. Strategy for learning: assume surface word order is actual order. (Fodor 1998)



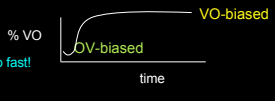
Advantage in intake determines learner's ending distribution between OV and VO order.

Need this trajectory

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



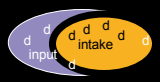
Problem: VO-biased all the way through, even at 1000 A.D.



Change is too fast!

Linguistic Evolution: Different Individual-Level Learning

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



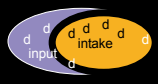
(YCOE and PPCME2 Corpora)
% Advantage

	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%

Very strongly OV-biased before 1150 A.D.

Linguistic Evolution: Different Individual-Level Learning

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



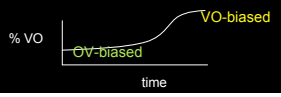
(YCOE and PPCME2 Corpora)
% Advantage

	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%

Very strongly OV-biased before 1150 A.D.

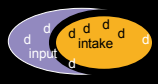
Need this trajectory

But population must become VO-biased.



Linguistic Evolution: Different Individual-Level Learning

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



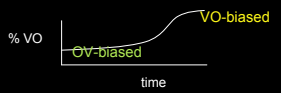
(YCOE and PPCME2 Corpora)
% Advantage

	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%

Very strongly OV-biased before 1150 A.D.

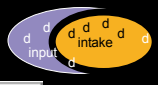
Need this trajectory

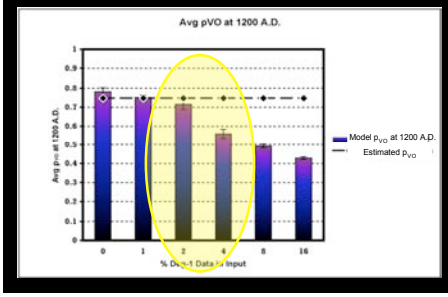
Can a population learning from degree-1 data make the change to VO-biased?



Linguistic Evolution: Different Individual-Level Learning

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

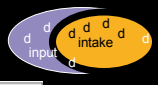


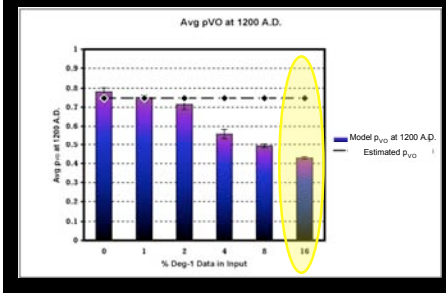


Modeled population can change at the right rate only if input contains less than 4% degree-1 data - otherwise, change is too slow for learners not using a degree-0 bias.

Linguistic Evolution: Different Individual-Level Learning

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



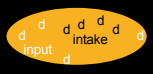


Estimates from modern English child-directed speech: Input consists of ~16% degree-1 data.

Prognosis: Change would be too slow without a degree-0 bias for individual learners.

Linguistic Evolution: Different Individual-Level Learning

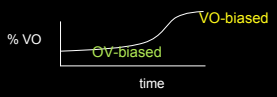
Learner uses degree-0 and degree-1 data, and learns from ambiguous data.



(YCOE and PPCME2 Corpora)
% Advantage

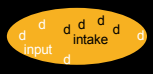
	OV Advantage in D0	OV Advantage in D1
1000 A.D.	-21.0%	28.1%

Need this trajectory



Linguistic Evolution: Different Individual-Level Learning

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

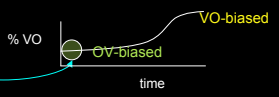


(YCOE and PPCME2 Corpora)
% Advantage

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

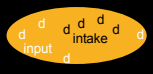
Need this trajectory

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



Linguistic Evolution: Different Individual-Level Learning

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

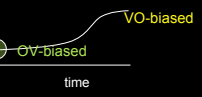


(YCOE and PPCME2 Corpora)
% Advantage

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

Need this trajectory

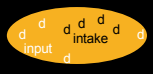
Population must remain OV-biased at 1000 A.D.



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.

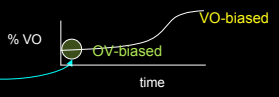


(YCOE and PPCME2 Corpora)
% Advantage

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

Need this trajectory

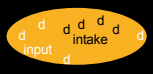
Population must remain OV-biased at 1000 A.D.



Requirement for OV advantage at 1000 A.D.: 43% of input is degree-1 data

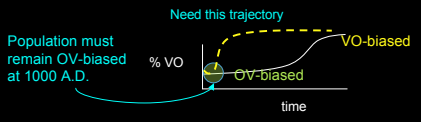
Linguistic Evolution: Different Individual-Level Learning

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



(YCOE and PPCME2 Corpora)
% Advantage

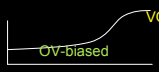
	OV Advantage in D0	OV Advantage in D1
1000 A.D.	-21.0%	28.1%



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

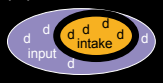
Linguistic Evolution: Summary

There may be some cases where linguistic evolution is driven by (young) individual-level learning. Suggested example: Old English word order.



Individual-level learning: can involve interpretive 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 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))?

(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.

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

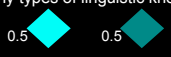
Thank You

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

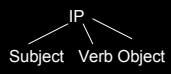
Individual Framework Applicability

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

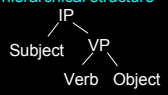


Discrete Representation: How much structure is posited for language?

A = linear structure



B = hierarchical structure



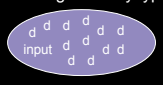
Discrete Representation : Is the basic word order Object Verb or Verb Object?

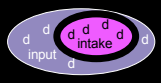
A = Object Verb

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.)

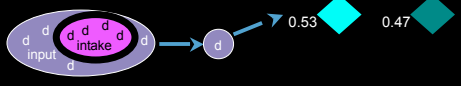
Selective Learning Bias: Use only data perceived as most informative (Fodor 1998, Lightfoot 1999, Dresher 1999).

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

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

Framework Applicability

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



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

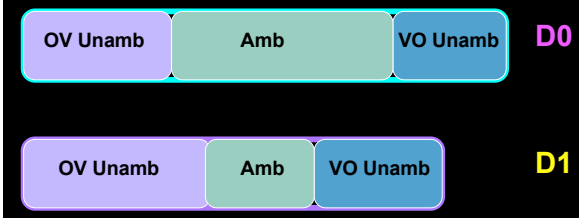
$$\text{Max}(\text{Prob}(p_{VO} \mid u)) = \text{Max}\left(\frac{\text{Prob}(u \mid p_{VO}) * \text{Prob}(p_{VO})}{\text{Prob}(u)}\right)$$

$$p_{OV} = p_{OV} + \gamma(1 - p_{VO})$$
$$p_{VO} = 1 - p_{OV}$$

Estimating Historical p_{VO}

Known quantities: Unambiguous and ambiguous data in d0 and d1

Estimating Historical p_{VO}

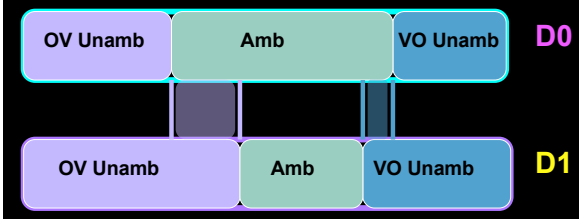


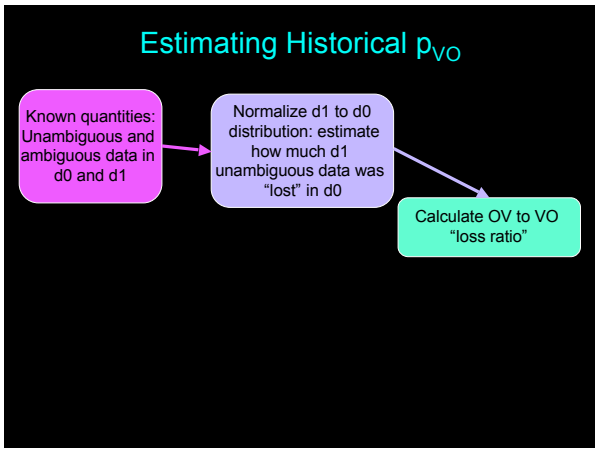
Estimating Historical p_{VO}

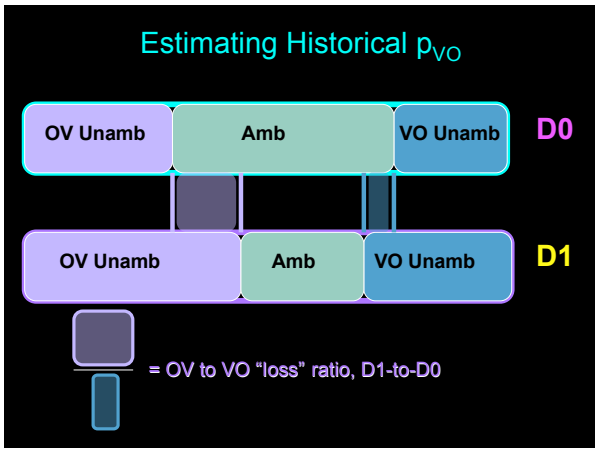
Known quantities: Unambiguous and ambiguous data in d0 and d1

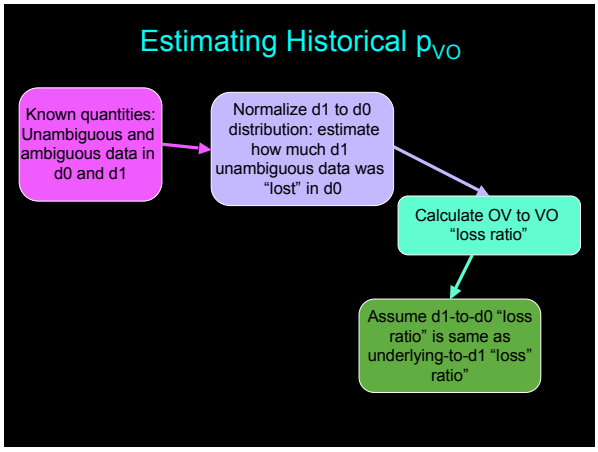
Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was "lost" in d0

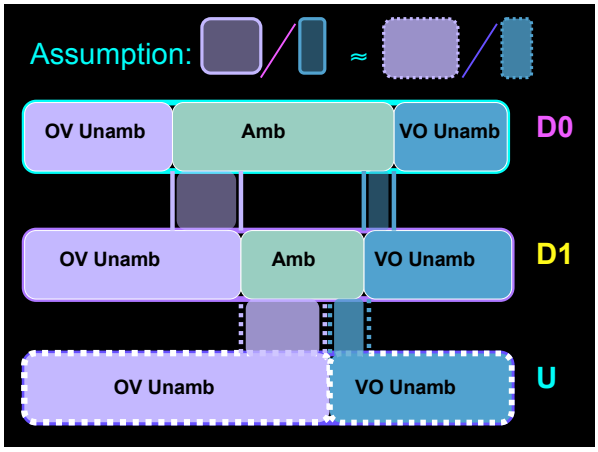
Estimating Historical p_{VO}

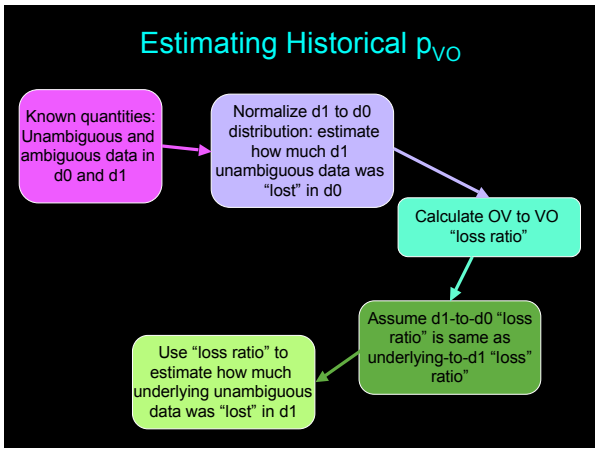


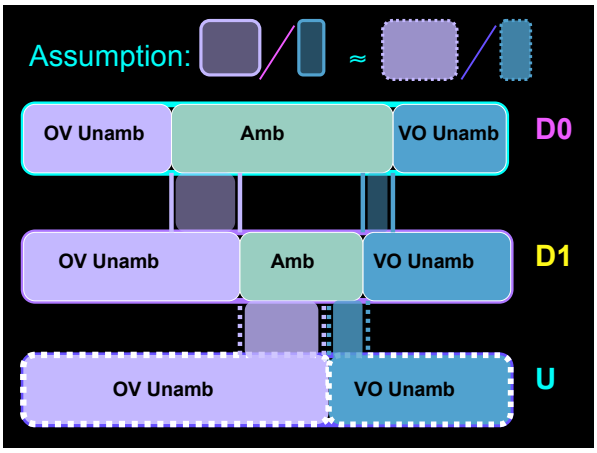

Estimating Historical p_{VO}
Known quantities: Unambiguous and ambiguous data in d0 and d1
Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was "lost" in d0
Calculate OV to VO "loss ratio"

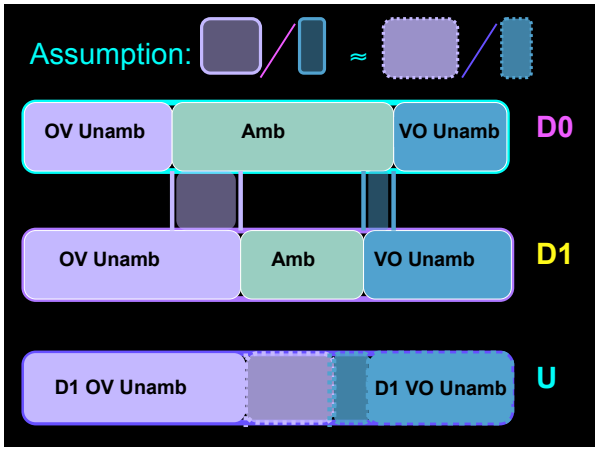

Estimating Historical p_{VO}
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
= OV to VO "loss" ratio, D1-to-D0

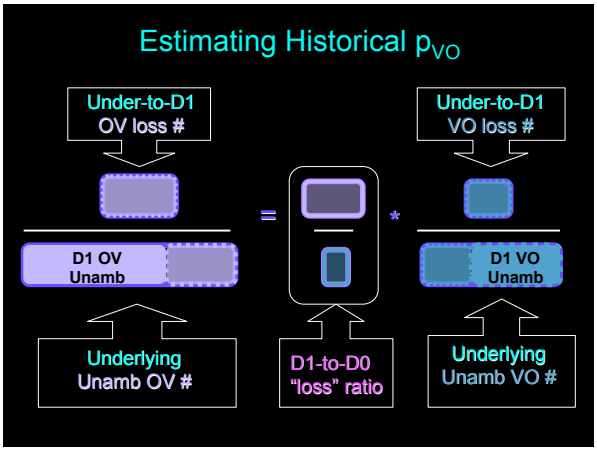

Estimating Historical p_{VO}
Known quantities: Unambiguous and ambiguous data in d0 and d1
Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was "lost" in d0
Calculate OV to VO "loss ratio"
Assume d1-to-d0 "loss ratio" is same as underlying-to-d1 "loss" ratio"


Assumption:
≈
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
OV Unamb
VO Unamb
U


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:
≈
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
OV Unamb
VO Unamb
U


Assumption:
≈
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
D1 OV Unamb
D1 VO Unamb
U


Estimating Historical p_{VO}
Under-to-D1 OV loss #
Under-to-D1 VO loss #
=
*
D1 OV Unamb
D1 VO Unamb
Underlying Unamb OV #
D1-to-D0 "loss" ratio
Underlying Unamb VO #

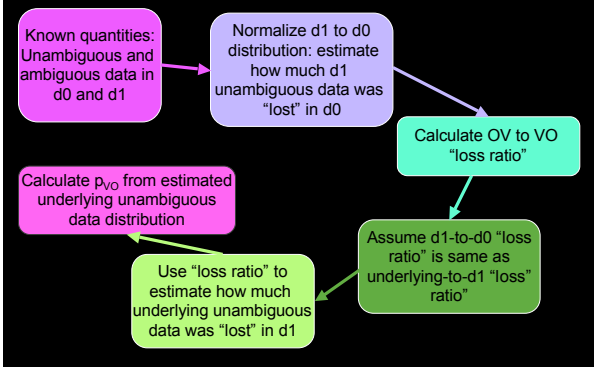
Estimating Historical p_{VO}

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

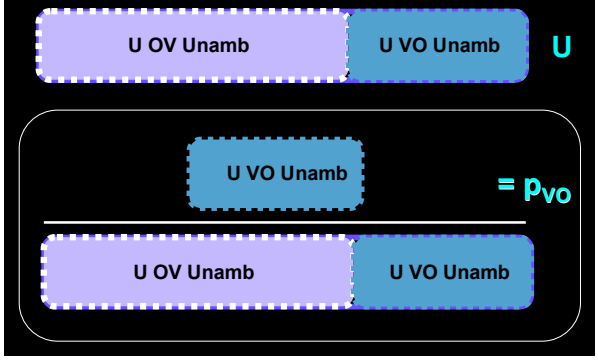
γ = underlying p_{VO}
d0 = total degree - 0 data, **d1** = total degree - 1 data
u1d1' = normalized unambiguous OV degree - 1 data
u2d1' = normalized unambiguous VO degree - 1 data
Ld1tod0 = loss ratio (OV/VO) from degree - 1 to degree - 0 distribution
ad1' = normalized ambiguous degree - 1 data

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

Estimating Historical p_{VO}



Estimating Historical p_{VO}



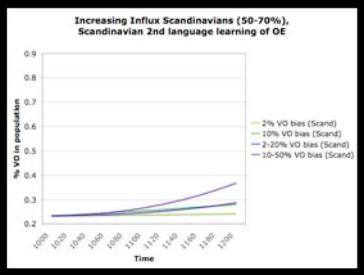
Potential Causes of Language Change

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.

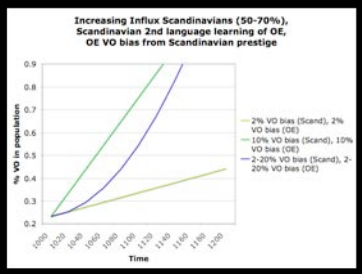
Evidence of individual probabilistic usage in Old English
- Historical records likely not the result of subpopulations of speakers who use only one order

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

$$\text{Max}(\text{Prob}(p_{VO} \mid u)) = \text{Max}(\frac{\text{Prob}(u \mid p_{VO}) * \text{Prob}(p_{VO})}{\text{Prob}(u)})$$

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

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

$$\text{Max}(\text{Prob}(p_{VO} \mid u)) = \text{Max}(\frac{\text{Prob}(u \mid p_{VO}) * \text{Prob}(p_{VO})}{\text{Prob}(u)})$$

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

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

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

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

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

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

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

Replace 1 in numerator and denominator with
$c = p_{VO_{prev}} * m$ if VO, $c = (1 - p_{VO_{prev}}) * m$ if OV
$3.0 \le m \le 5.0$

$$p_{VO} = \frac{p_{VO_{prev}} * n + c}{n + c}$$

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

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.)

(2) Verb-Marker is immediately post-Object

we$_{Subj}$ sculen$_{TensedVerb}$ [ure yfele þeawes]$_{Obj}$ forlæten$_{Verb\text{-}Marker}$
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

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

& [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.)

(2) Verb-Marker is immediately pre-Object

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.)

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

we$_{Subj}$ sculen$_{TensedVerb}$ [ure yfele þeawes]$_{Obj}$ forlæten$_{Verb-Marker}$
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$_{Negative}$ geseah$_{TensedVerb}$ ic$_{Subj}$ næfre$_{Adverbial}$ [ða burh]$_{Obj}$
NEG saw I never the city
'Never did I see the city.'
(Ælfric, *Homilies.* I.572.3, between 900 and 1000 A.D.)

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