Language in Populations: The Interaction Between Learning & Change

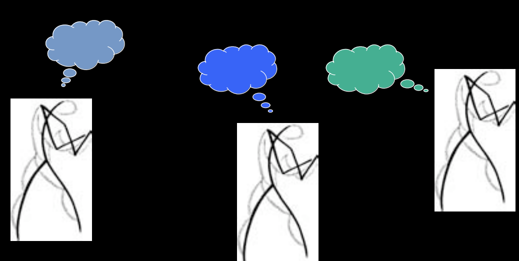
Lisa Pearl
University of Maryland
April 2, 2007
Northwestern Institute on Complex Systems

Language

Language Characteristics

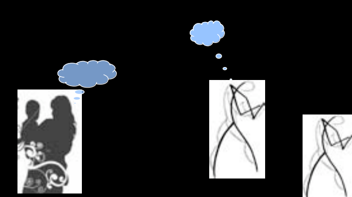
Language

Language Characteristics
(1) a trait individuals within a population have



Language

Language Characteristics
(1) a trait individuals within a population have
(2) transmitted via learning, rather than solely genetic

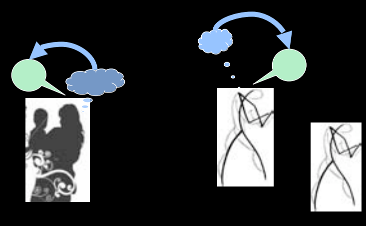


Language

Language Characteristics

(1) a trait individuals within a population have

(2) transmitted via learning, rather than solely genetic

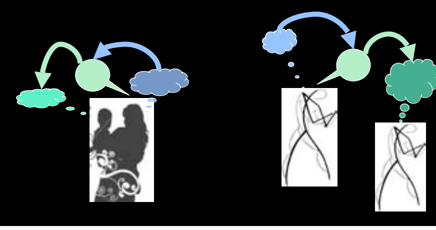


Language

Language Characteristics

(1) a trait individuals within a population have

(2) transmitted via learning, rather than solely genetic



Language

Language Characteristics

(1) a trait individuals within a population have

(2) transmitted via learning, rather than solely genetic

(3) some parts mutable only during learning period

Language

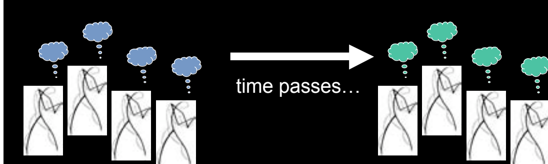
Language Characteristics

(1) a trait individuals within a population have

(2) transmitted via learning, rather than solely genetic

(3) some parts mutable only during learning period

(4) linguistic composition of population can change



Language Change in a Population

Two opposing linguistic structures (e.g. Object Verb and Verb Object order) can be used probabilistically by individuals in a population

Change in the probability of usage within the population proceeds at a certain rate

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

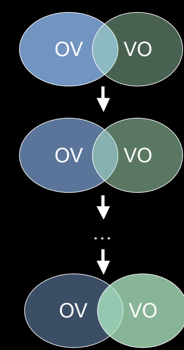
Imperfect Learning = Language Change

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

These individuals: source of data for future individuals

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

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



Modeling Correct Linguistic Behavior

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

(1) individuals misconverged precisely the right amount

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

Language Learning in Individuals: The Tricky Part

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

Language Learning in Individuals: The Tricky Part

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

Syntactic System
- Observable form: word order
- Interference: movement rules

Subject *Object* *Verb*

Language Learning in Individuals: The Tricky Part

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

Syntactic System
- Observable form: word order
- Interference: movement rules

Verb *Subject* *Object* t_{Verb}

Verb-Second (V2) movement

Language Learning in Individuals: The Tricky Part

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

Syntactic System
- Observable form: word order
- Interference: movement rules

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

Language Learning: Extracting Systematicity

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

Language Learning: Extracting Systematicity

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

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

Potential issue: data sparseness

Road Map

Individual Learning Framework Overview

Population Modeling: Syntactic Language Change

Road Map

Individual Learning Framework Overview

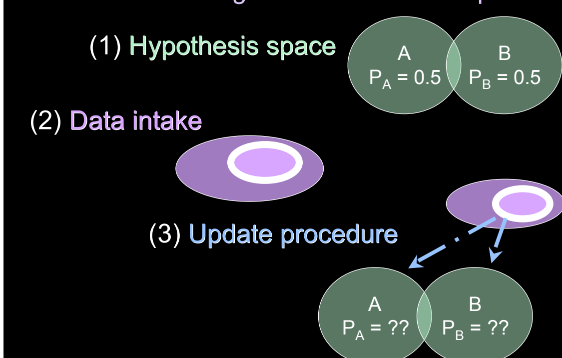
Population Modeling: Syntactic Language Change

Important Feature: grounded in empirical data

Individual-level: learning period, data distributions, discrete representations, probabilistic update procedure

Population-level: population size, population growth rate, time period of change, rate of change

Individual Learning Framework: 3 Components



Benefits of Learning Framework

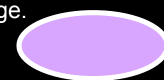
Components:

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

Can combine discrete representations (hypothesis space) with probabilistic components (update procedure): get gradualness and variation found in real language learning

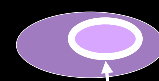
Individual Model: Data Intake Filtering

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



input

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



subset of input

Case Study: Model Specifics

Case Study: Model Specifics

Hypothesis Space: word order parameters

Object Verb (OV) vs. Verb Object (VO) order

Case Study: Model Specifics

Hypothesis Space: word order parameters
Object Verb (OV) vs. Verb Object (VO) order

Update Procedure: adapted Bayesian updating
shifts probabilities between opposing hypotheses

amount shifted depends on layout of hypothesis space

Case Study: Model Specifics

Hypothesis Space: word order parameters
Object Verb (OV) vs. Verb Object (VO) order

Update Procedure: adapted Bayesian updating
shifts probabilities between opposing hypotheses

amount shifted depends on layout of hypothesis space

Difficult Feature: adult target state is a probability distribution
target state is usually one hypothesis or the other

converging on the right probability is harder

Learning & Change: The Big Questions

(1) Is it feasible to filter?
Is there a data sparseness problem?

(2) Is it sufficient to filter?
Can we get the right population behavior if we filter?

(3) Is it necessary to filter?
Must we filter to get the right population behavior?

Road Map

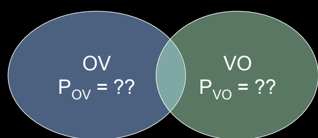
Individual Learning Framework Overview

Population Modeling: Syntactic Language Change
- Old English: description & proposed individual filters
- Old English data & feasibility of filtering
- Modeled learners and populations
- Estimating ground truth
- Sufficiency & necessity of filtering

Old English

Learning: Old English OV vs. VO order

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



Old English Filters

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

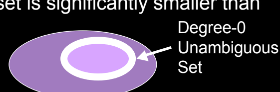
Filter 2: Use structurally “simple” data - matrix clause or “degree-0” data (Lightfoot, 1991)

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

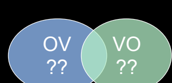
Problems

Potential problem for feasibility: **data sparseness**

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



Learners must use this data set to **misconverge** the **exact right amount at each point in time** so that the population changes at the correct rate



Road Map

Individual Learning Framework Overview

Population Modeling: Syntactic Language Change

- Old English: description & proposed individual filters
- Old English data & feasibility of filtering
- Modeled learners and populations
- Estimating ground truth
- Sufficiency & necessity of filtering

Old English OV and VO

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

VO-biased: by 1200 A.D.

Old English OV and VO

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

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

VO-biased: by 1200 A.D.

Old English OV and VO

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

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

VO-biased: by 1200 A.D.

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

Ambiguous Data

Subject ***TensedVerb Object*** is ambiguous
(most common data type)

OV, +V2
heo$_{Subj}$ clænsað$_{TensedVerb}$ t_{Subj} [þa sawle þæs rædendan]$_{Obj}$ $t_{TensedVerb}$
they purified the souls [the advising]-Gen

VO, -V2
heo$_{Subj}$ clænsað$_{TensedVerb}$ [þa sawle þæs rædendan]$_{Obj}$
they purified the souls [the-advising]-Gen

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

Perceived Unambiguous Data: Examples

Unambiguous OV

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

Unambiguous VO

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

Perceived Unambiguous Data: Making "Unambiguous" Feasible

Definitions of data perceived as unambiguous are ***heuristic*** and/or involve only partial knowledge of the adult linguistic system (Lightfoot 1999, Dresher 1999, Fodor 1998)

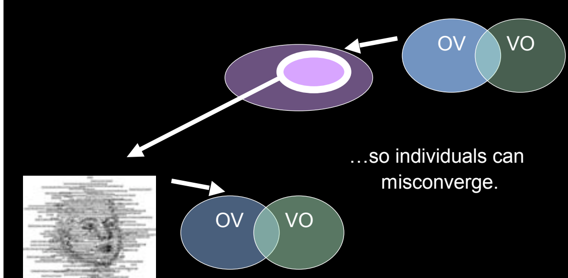
OV:
[…]$_{\text{XP}}$ … Object TensedVerb …
… Object Verb-Marker …

VO:
[…]$_{\text{XP}}$ […]$_{\text{XP}}$ … TensedVerb Object …
… Verb-Marker Object …

This allows the learner to identify *some* data points as unambiguous (even if they're actually not for someone with full knowledge of the adult linguistic system)

The Effect of Filtering

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



…so individuals can misconverge.

Road Map

Individual Learning Framework Overview

Population Modeling: Syntactic Language Change
- Old English: description & proposed individual filters
- Old English data & feasibility of filtering
- Modeled learners and populations
- Estimating ground truth
- Sufficiency & necessity of filtering

The Model: Individual-Level

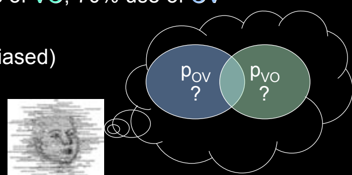
Individual learner tracks p_{VO} = probability of using VO

probability of using OV = 1 - p_{VO}

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

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

Initial p_{VO} = 0.5 (unbiased)



The Model: Individual-Level

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

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

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

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

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

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

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

$3.0 \le m \le 5.0$

The Model: Individual-Level

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

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

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

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

Individual-Level Learning Algorithm

Individual-Level Learning Algorithm

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

Individual-Level Learning Algorithm

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

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

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

Individual-Level Learning Algorithm

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

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

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

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

Biased Data Intake Distributions

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

Biased Data Intake Distributions

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

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

Population-Level Algorithm

Population-Level Algorithm

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

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

Population-Level Algorithm

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

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

(3) Move forward 2 years.

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

Population-Level Algorithm

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

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

(3) Move forward 2 years.

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

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

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

Road Map

Individual Learning Framework Overview

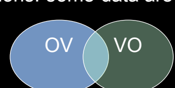
Population Modeling: Syntactic Language Change

- Old English: description & proposed individual filters
- Old English data & feasibility of filtering
- Modeled learners and populations
- Estimating ground truth
- Sufficiency & necessity of filtering

Estimating Historical p_{VO}

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

Historical data distributions: some data are ambiguous



p_{VO}: underlying distribution is not ambiguous

Estimating Historical p_{VO}

(YCOE and PPCME2 Corpora)
% Ambiguous Utterances

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

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

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

Estimating Historical p_{VO}

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

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

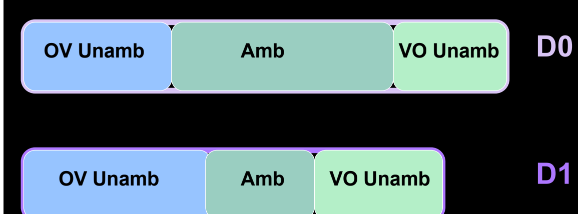
γ = underlying p_{VO}
d0 = total degree-0 data, **d1** = total degree-1 data
u1d1' = normalized unambiguous OV degree-1 data
u2d1' = normalized unambiguous VO degree-1 data
Ld1tod0 = loss ratio (OV/VO) from degree-1 to degree-0 distribution
ad1' = normalized ambiguous degree-1 data

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

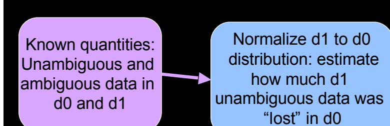
Estimating Historical p_{VO}

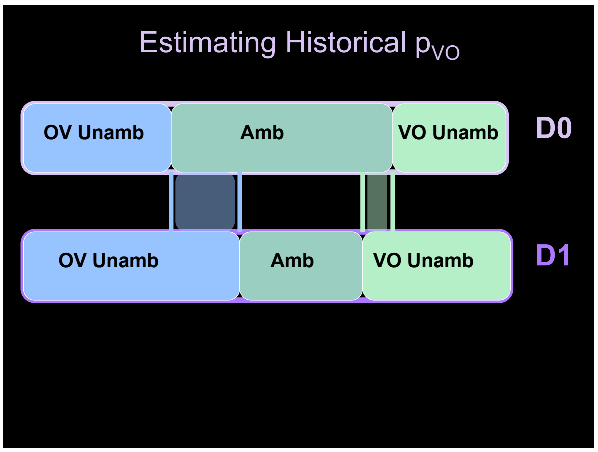
Known quantities: Unambiguous and ambiguous data in d0 and d1

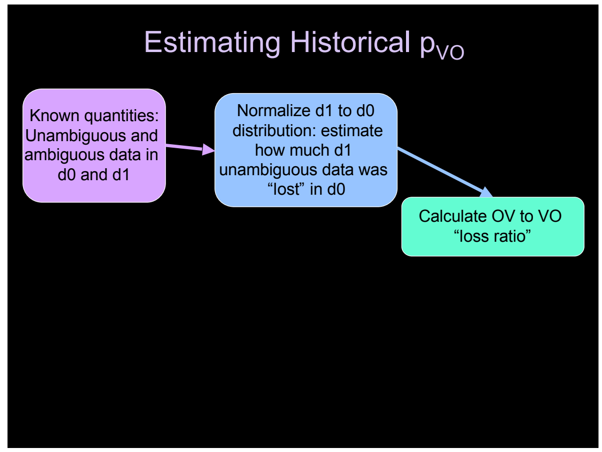
Estimating Historical p_{VO}

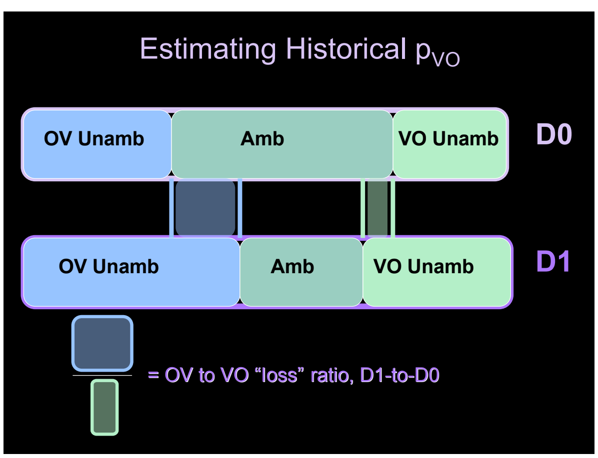


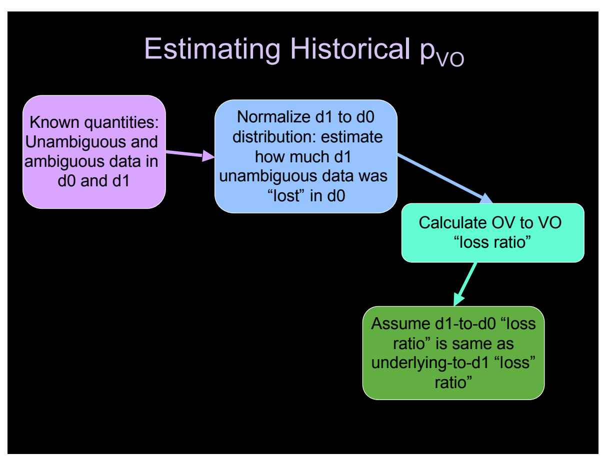
Estimating Historical p_{VO}

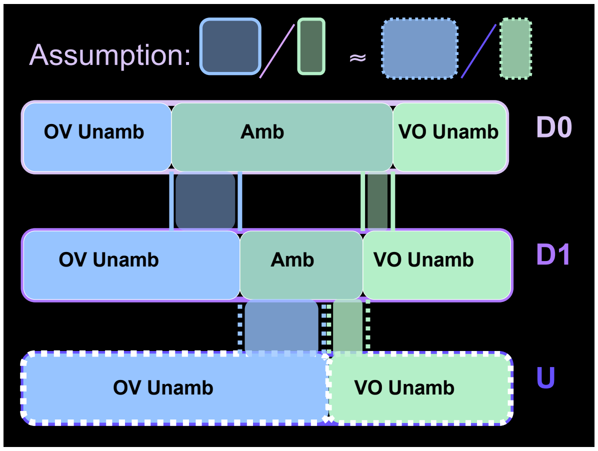


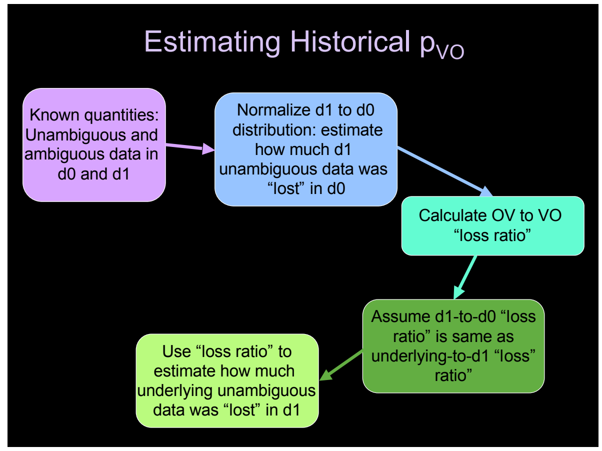

Estimating Historical p_{VO}
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1

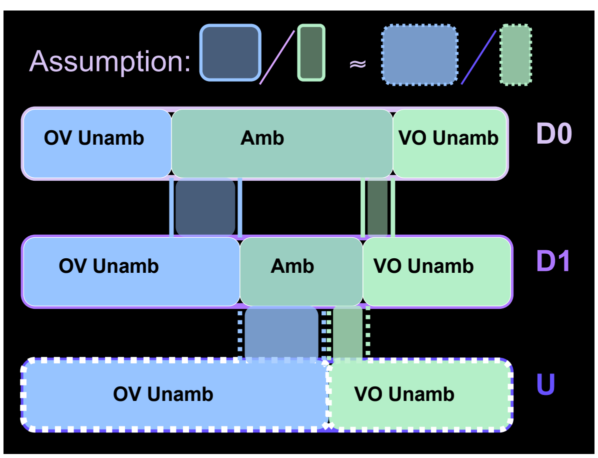

Estimating Historical p_{VO}
Known quantities: Unambiguous and ambiguous data in d0 and d1
Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was "lost" in d0
Calculate OV to VO "loss ratio"

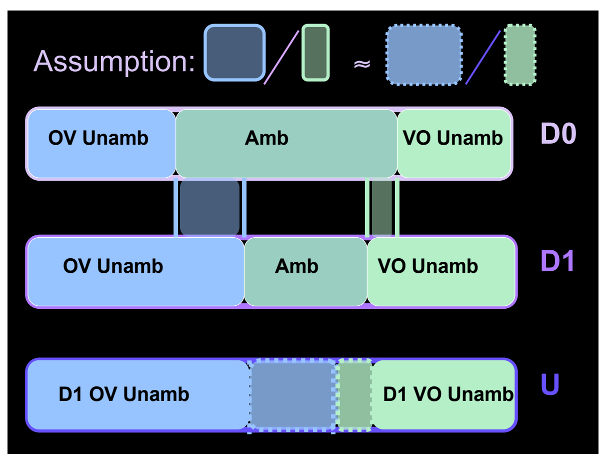

Estimating Historical p_{VO}
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
= OV to VO "loss" ratio, D1-to-D0

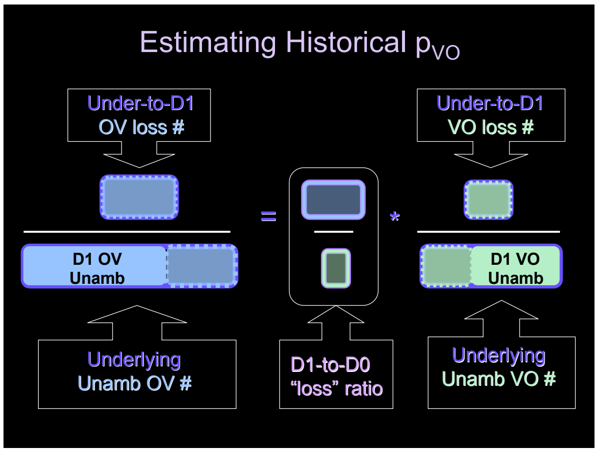

Estimating Historical p_{VO}
Known quantities: Unambiguous and ambiguous data in d0 and d1
Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was "lost" in d0
Calculate OV to VO "loss ratio"
Assume d1-to-d0 "loss ratio" is same as underlying-to-d1 "loss" ratio"

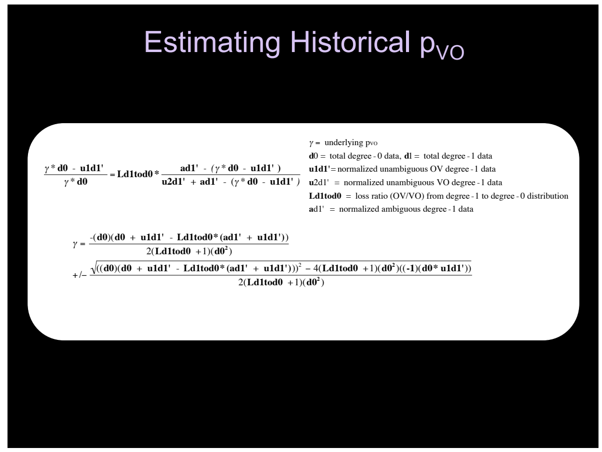

Assumption:
≈
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
OV Unamb
VO Unamb
U

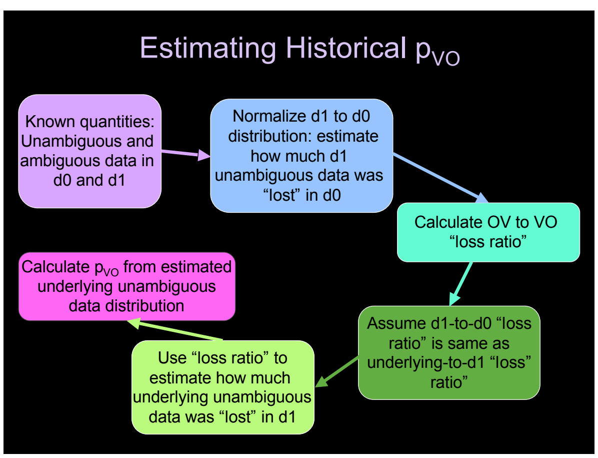

Estimating Historical p_{VO}
Known quantities: Unambiguous and ambiguous data in d0 and d1
Normalize d1 to d0 distribution: estimate how much d1 unambiguous data was "lost" in d0
Calculate OV to VO "loss ratio"
Assume d1-to-d0 "loss ratio" is same as underlying-to-d1 "loss" ratio"
Use "loss ratio" to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

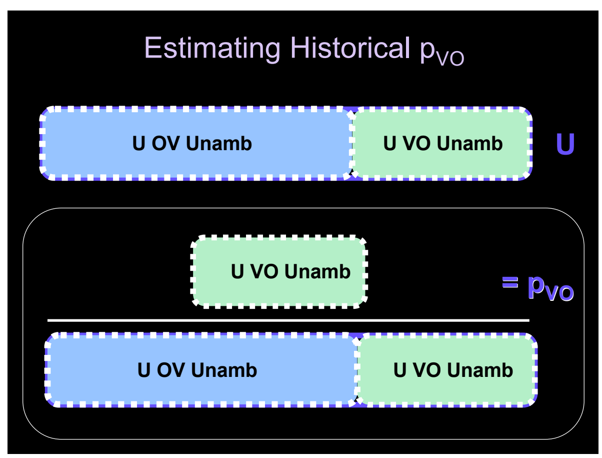

Assumption:
≈
OV Unamb
Amb
VO Unamb
D0
OV Unamb
Amb
VO Unamb
D1
OV Unamb
VO Unamb
U


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}
$\frac{\gamma * d0 - u1d1'}{\gamma * d0} = Ld1tod0 * \frac{ad1' - (\gamma * d0 - u1d1')}{u2d1' + ad1' - (\gamma * d0 - u1d1')}$
γ = underlying pVO
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
$\gamma = \frac{-(d0)(d0 + u1d1' - Ld1tod0 * (ad1' + u1d1'))}{2(Ld1tod0 + 1)(d0^2)}$
$+/- \frac{\sqrt{((d0)(d0 + u1d1' - Ld1tod0 * (ad1' + u1d1')))^2 - 4(Ld1tod0 + 1)(d0^2)((-1)(d0 * u1d1'))}}{2(Ld1tod0 + 1)(d0^2)}$


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


Estimating Historical p_{VO}
U OV Unamb
U VO Unamb
U
U VO Unamb
= p_{VO}
U OV Unamb
U VO Unamb

Estimating Historical p_{VO}

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

Road Map

Individual Learning Framework Overview

Population Modeling: Syntactic Language Change

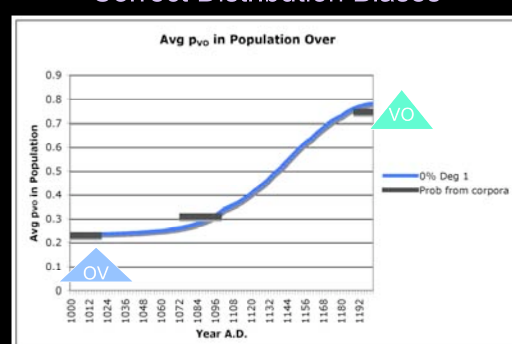
- Old English: description & proposed individual filters
- Old English data & feasibility of filtering
- Modeled learners and populations
- Estimating ground truth
- Sufficiency & necessity of filtering

Remaining Questions to Answer

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

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

Sufficiency of Filters: Correct Distribution Biases



Necessity of Filters: Remove Unambiguous Filter

Learner can use ambiguous data. Strategy: assume surface order is actual order. (Fodor, 1998)

Example: *Subject TensedVerb Object* = VO

Necessity of Filters: Remove Unambiguous Filter

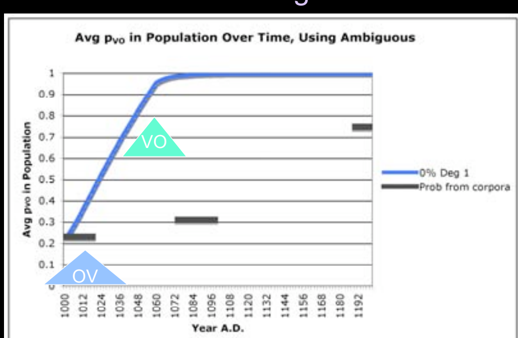
Learner can use ambiguous data. Strategy: assume surface order is actual order. (Fodor, 1998)

Example: *Subject TensedVerb Object* = VO

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

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

Necessity of Filters: Remove Unambiguous Filter



Necessity of Filters: Removing Degree-0 Filter

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

Necessity of Filters: Removing Degree-0 Filter

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

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

Necessity of Filters: Removing Degree-0 Filter

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

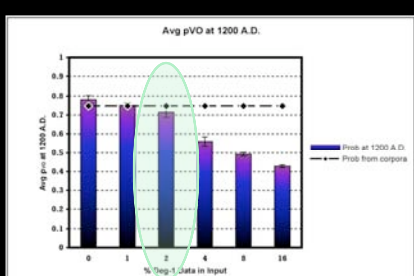
	OV Advantage in Unamb D0	OV Advantage in Unamb D1
1000 A.D.	19.5%	41.7%
1000-1150 A.D.	2.8%	28.7%
1200 A.D.	-2.7%	-45.2%

Degree-1 data is strongly OV-biased.

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

How does this compare to the amount available to children?

Necessity of Filters: Allowing in Degree-1 Data



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

Necessity of Filters: Removing Degree-0 Filter

Permissible threshold: <4%

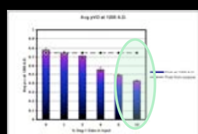
Estimated amount available to children (from corpora): ~16%

Necessity of Filters: Removing Degree-0 Filter

Permissible threshold: <4%

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

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



Necessity of Filters: Removing Both Filters

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

Necessity of Filters: Removing Both Filters

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

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

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

Old English Language Change Summary

Language change modeling results: existence proof for feasibility, sufficiency, and necessity of data intake filtering during individual learning

Individual-Level Filters:
(1) unambiguous data
(2) degree-0 data

There is an interaction of language change modeling and language learning theory. Each can be used to constrain the other.

Open Questions

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

Open Questions

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

Weight data points in individual intake using various factors:

(a) spatial location of speaker with respect to learner

(b) social status of speaker

(c) speaker's relation to learner (family, friend, stranger)

(d) context of data point (social context, linguistic context)

Open Questions

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

Weight data points in individual intake using various factors:

(a) spatial location of speaker with respect to learner

(b) social status of speaker

(c) speaker's relation to learner (family, friend, stranger)

(d) context of data point (social context, linguistic context)

(2) Are these filters 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?

Population Modeling: Take Home Messages

Population Modeling: Take Home Messages

(1) Models of language change can (and should) be empirically grounded.

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

Population Modeling: Take Home Messages

(1) Models of language change can (and should) be empirically grounded.

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

(2) Learners can extract the correct system by looking at a subset of the data.

Population Modeling: Take Home Messages

(1) Models of language change can (and should) be empirically grounded.

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

(2) Learners can extract the correct system by looking at a subset of the data.

(3) Correct population-level behavior can result from correct individual-level behavior (small misconvergences compounded over time).

Thank You

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

the Cognitive Neuroscience of Language Lab
at the University of Maryland

Causes of Language Change

Old Norse influence before 1000 A.D.: VO-biased

- If sole cause of change, requires exponential influx of Old Norse speakers.

Old French at 1066 A.D.: embedded clauses predominantly OV-biased (Kibler, 1984)

- Matrix clauses often SVO (ambiguous)
- OV-bias would have hindered Old English change to VO-biased system.

Evidence of individual probabilistic usage in Old English

- Historical records likely not the result of subpopulations of speakers who use only one order

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Replace 1 in numerator and denominator with

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

$3.0 \leq m \leq 5.0$

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

Other Ways to Remove the Unambiguous Filter

Strategies for assessing ambiguous data

(1) assume base-generation
- attempted and failed
- system-dependent (syntax)

(2) weight based on level of ambiguity (Pearl & Lidz, in submission)
- unambiguous = highest weight
- moderately ambiguous = lower weight
- fully ambiguous = lowest weight (ignore)

(3) randomly assign to one hypothesis (Yang, 2002)

Perceived Unambiguous Data: OV

Unambiguous OV data

Perceived Unambiguous Data: OV

Unambiguous OV data

(1) Tensed Verb is immediately post-Object

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

Perceived Unambiguous Data: OV

Unambiguous OV data

(1) Tensed Verb is immediately post-Object

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

(2) Verb-Marker is immediately post-Object

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

Perceived Unambiguous Data: VO

Unambiguous VO data

Perceived Unambiguous Data: VO

Unambiguous VO data

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

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

Perceived Unambiguous Data: VO

Unambiguous VO data

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

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

(2) Verb-Marker is immediately pre-Object

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

Verb-Markers

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

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

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

Verb-Markers

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

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

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

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

Unreliable Verb-Markers

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

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

Unreliable Verb-Markers

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

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