Language Change as a Source of Word Order Correlations*

Brady Clark, Matthew Goldrick, and Kenneth Konopka
Northwestern University

Abstract

Typological work has demonstrated that there are constraints on word order variation. For example, auxiliary verbs tend to precede content verbs in VO languages (Dryer 1992). Further, typologically recurrent structural preferences are reflected in language change. In this paper, we present agent-based modeling work that suggests that Filtered Learning Models (e.g. Kirby 1999) can capture the emergence of word order correlations over time. We identify limitations of the Filtered Learning Model of Kirby (1999), and demonstrate that an extended (filtered) version of the Variational Learning Model presented in Yang (2002) overcomes these limitations while preserving the insight that constraints on word order variation are emergent in a population through repeated cycles of language acquisition and use.

1 Structural preferences in typology and change

1.1 Introduction

One main way in which natural languages differ is in their word order. For example, auxiliary verbs (Aux) can precede or follow content verbs in both verb-object (VO) and object-verb (OV) languages.[1] All of the logically possible combinations of orderings of auxiliary verbs, content verbs, and objects, given in (1), are attested stable grammatical states.

(1) a. OV&VAux (e.g. Slave, Siroi; Dryer 2006)
b. OV&AuxV (e.g. Seme, Sorbian; Matthew Dryer, p.c.)
c. VO&VAux (e.g. Akan, Gumuz; Matthew Dryer, p.c.)
d. VO&AuxV (e.g. English)

*For valuable comments on this work, we would like to thank Gerhard Jäger, Janet Pierrehumbert, the audience at the Blankensee Colloquium 2005, and an anonymous reviewer.

[1]We adopt Dryer's (1992:100) use of AUXILIARY VERB here: tense/aspect words that are specifically verbal. For English, this category includes *will*, *have*, and progressive *be*, but not the passive auxiliary *be* and modal auxiliaries such as *can* and *should*. A CONTENT VERB is the verb with which the auxiliary verb combines.

Despite extraordinary cross-linguistic variation, though, certain word orders are more frequent than others. In the following two sections we provide evidence from typology and change that there are robust preferences for certain word order patterns.

1.2 Evidence from typology

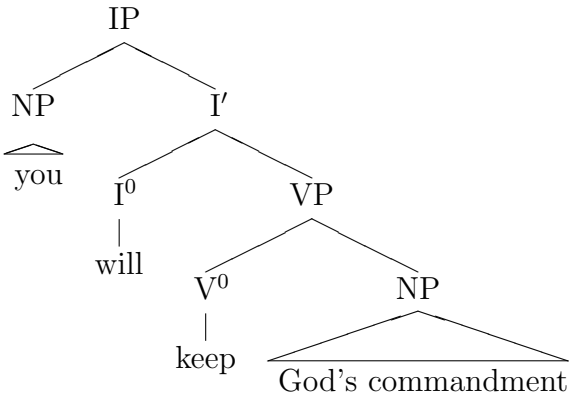
Typological work has demonstrated that there are robust word order correlations cross–linguistically (Greenberg 1966; Hawkins 1983; Dryer 1992). For example, there is a strong tendency for auxiliaries to precede the content verb in VO languages (i.e. VO&AuxV), while auxiliaries tend to follow in OV languages (i.e. OV&VAux) (Dryer 1992:100). This tendency is illustrated in Table 1.[2]

Table 1: Order of content verb and auxiliary verb (Dryer 1992: 100)

	Africa	Eurasia	SEAsia&Oc	Aus–NewGui	NAmer	SAmer	Total
OV&VAux	[5]	[12]	[2]	[8]	[1]	[8]	36
OV&AuxV	3	0	0	0	0	0	3
VO&VAux	1	1	0	[1]	0	1	4
VO&AuxV	[15]	[5]	[3]	0	[4]	1	28

Present-day English is VO&AuxV. English clauses with both an auxiliary and a content verb have a consistently right-branching structure, illustrated in (2), where auxiliaries sit in the head of IP:[3]

(2)

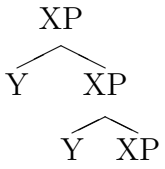


[2]The form of the data in Table 1 is discussed in detail in Dryer (1992). The numbers represent the number of genera that contain languages of the given type in the geographic area listed. A genus is a genetic group roughly comparable in time depth to the subfamilies of Indo–European.

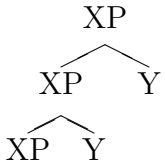
[3]In (2), the category I describes auxiliary verbs, the category N describes nouns, and the category V describes verbs.

In general, there is a typological tendency for languages to converge on one of two ideals (Dryer 1992): right–branching languages (where phrasal categories such as VP follow non–phrasal categories such as I^0, e.g. English), as in (3a), or left–branching languages (where phrasal categories precede non–phrasal categories, e.g. Japanese), as in (3b).

(3) a. **Right–branching**:



b. **Left–branching**:

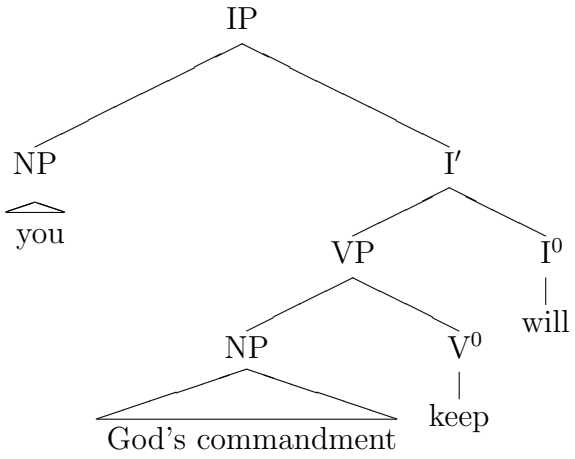


1.3 Evidence from change

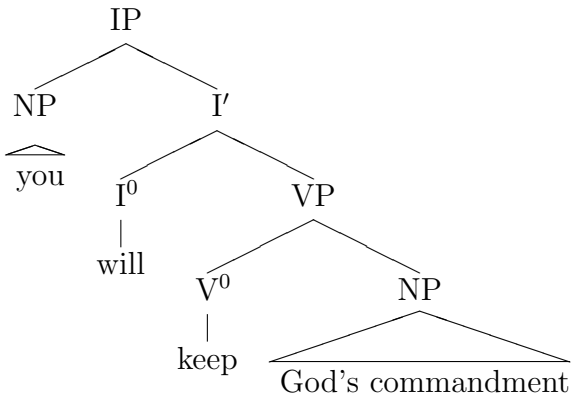
The preference for consistent branching observed in typology can also be seen diachronically, e.g. in the history of English. Late Old English (925-1150) and early Middle English (1150-1325) subordinate clauses displayed both intertextually and intratextually (at least) three structures (Pintzuk 1999; Kroch and Taylor 2001; Clark 2004), given in (4a–c). Note that the BRACE construction in (4c) has inconsistent branching: the non–phrasal category I^0 is a left–sister of the phrasal category VP, while the non–phrasal category V^0 is a right–sister of a phrasal category YP. In contrast, the ALL-FINAL and ALL–MEDIAL constructions in (4a) and (4b) have consistent branching.[4]

[4]We are simplifying a bit here. Clark (2004) argues for a verbal cluster analysis of the all-final construction.

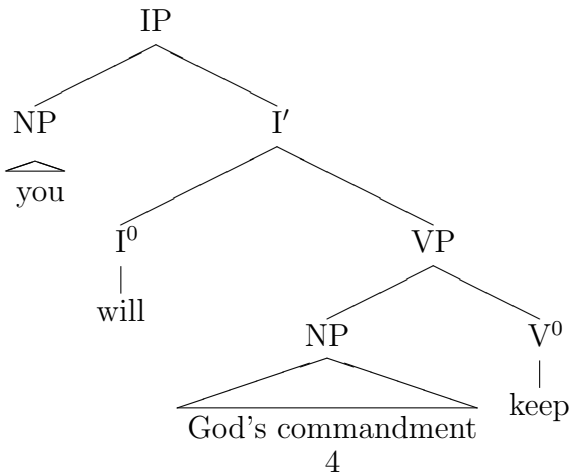
(4) a. **All–final (OV&VAux, *you God's commandment keep will*):**



b. **All–medial (VO&AuxV, *you will keep God's commandment*):**



c. **Brace (OV&AuxV, *you will God's commandment keep*):**



Roughly speaking, of the three variants in (4a–c), the all–final structure in (4a) was most frequent within Old English subordinate clauses. In early Middle English the all–medial structure was most frequent. While the inconsistent brace construction was available at a low frequency at both of these stages of the language, the frequency of the brace construction gradually declined over the course of Middle English. (5)–(7) give examples of all three variants in late Old English. (8)–(10) give evidence from early Middle English.

(5) **All–final**
him þær se gionga cyning þæs oferfæreldes **forwiernan mehte**
him there the young king the crossing prevent could
'... the young king could prevent him from crossing there'
(c800-900, Orosius 44.19-20, [SOURCE: Pintzuk 1996, 245])

(6) **All–medial**
he **wolde adræfan** ut anne æþeling
he would drive out a prince
'... he would drive out a prince...'
(c1000-1100, ChronB(T) 82.18-19, [SOURCE: Pintzuk 1999, 104])

(7) **Brace**
he **mæg** þa synfullan sawle þurh his gife **geliffæstan**
he may the sinful soul through his gift endow-with-life
'He can endow the sinful soul with life through his grace'
(c900-1000, Ælfric's Homilies I, 33.496.30, [SOURCE: Fischer 2000, 143])

(8) **All–final**
ȝef ȝe þus godes heste **halden wulleð**
if you thus God's commandment keep will
'if you will thus keep God's commandment'
(c1225, Ancrene Riwle, II.141.1889, [SOURCE: Kroch and Taylor 2001, 141, PPCME2])

(9) **All–medial**
oðet he **habbe iȝetted** ou al þet he wulleð
until he has granted you all that you desire
'until he has granted you all that you desire'
(c1225, Ancrene Riwle, II.68.229, [SOURCE: Kroch and Taylor 2001, 145, PPCME2])

(10) **Brace**
ðanne hie **willeð** here ibede to godde **bidden**
when they will their prayer to God pray
'when will they pray their prayer to God'
(c1200, Vices and Virtues I, 143.1773, [SOURCE: Kroch and Taylor 2001, 154, PPCME2])

Table 2 illustrates the relative frequency of the three variants in (4a–c) within two texts, the late Old English text Chronicle A (Scribe 1) and the early Middle English text *Festis Marie*. The estimated frequency distributions of structures (4a-c) are given in Figure 2. The numbers are inferred from the frequency information in Pintzuk (1999) (for Chronicle A, Scribe 1) and Allen (2000) (for *Festis Marie*). Crucially, all three variants are present intratextually. As we argue below, our model of language acquisition and use must be able to capture intratextual variability of this sort (Kroch 2001; Yang 2002; Clark 2004).

Table 2: Estimated frequencies of structures (4a-c) in Chron A, Scribe 1 (OE) and *Festis Marie* (early ME)

structures	% for Chron A, Scribe 1	% for *Festis Marie*
(4a) $[_{IP}$ XP $[_{I'}$ $[_{VP}$ YP V] I]]	61	8
(4b) $[_{IP}$ XP $[_{I'}$ I $[_{VP}$ V YP]]]	1	67
(4c) $[_{IP}$ XP $[_{I'}$ I $[_{VP}$ YP V]]]	38	25

In sum, in the history of English we see a gradual convergence on the consistent right–branching structure in (4b), where phrasal categories such as direct objects follow non–phrasal categories such as non–finite verbs. This change is arguably a reflection of the typological preference for consistent branching discussed above. The typologically rare brace order in (4c) was available at each stage of early English, but was never the preferred option, neither within nor across speakers.

2 Accounting for structural preferences

The previous section discussed the following two observations:

i. There is extraordinary cross-linguistic word order variation. For example, all of the logically possible combinations of orderings of auxiliary verbs, content verbs, and objects are attested stable grammatical states.

ii. Certain word order patterns (e.g. VO&AuxV, OV&VAux) are more frequent than others (e.g. VO&VAux, OV&AuxV) and this reflects a preference for consistently branching structures.

There are several types of explanations for why languages tend to converge on consistently right–branching or consistently left–branching structures. We focus on two related types of explanations here: purely syntactic explanations and cultural evolution explanations.

2.1 Purely syntactic accounts

Starting with Greenberg (1966), there is a long tradition of purely syntactic explanations for cross–linguistic word order correlations (e.g. the fact that OV languages tend to be VAux), see, e.g., Greenberg (1966), Lehmann (1973), Vennemann (1973), Hawkins (1983), Svenonius (2000), and Biberauer and Roberts (2005). Syntactic accounts attempt to explain word order correlations solely in terms of constraints on phrase structure relations, e.g. between heads and dependents, or between phrasal categories and non–phrasal categories. For example, Greenberg (1966) suggests that word order correlations reflect a tendency to consistently order heads with respect to their dependents/modifiers. Greenberg's syntactic explanation for word order correlations was the germ for later syntactic explanations, e.g. Hawkins' (1983) principle of Cross–Category Harmony.

An underlying assumption of the syntactic approach is that consistent languages involve simpler grammars while inconsistent grammars involve more complex grammars, and that language learners disprefer complex grammars:

> A disharmonic language requires more category–particular rules and thus the grammar of such a language is more complex. (Mallison and Blake 1981:417)

For example, in recent work Biberauer and Roberts (2005) tie word order correlations to a "least–effort" strategy applied to parameter setting. Biberauer and Roberts (2005:38) suggest that language acquirers "will, given evidence for a particular setting of one of a series of isomorphic parameters, set all the isomorphic parameters [e.g. for T and ν] the same way", unless overridden by primary linguistic data. Consequently, grammars with simpler structural representations (e.g. consistently branching structures) are preferred over grammars with more complex representations (e.g. inconsistently branching structures). The crucial property of this and earlier syntactic explanations is that they seek to explain word order correlations purely in terms of a language–specific predisposition for simpler structures. This predisposition for simpler structures is assumed to be part of the genetic endowment of the language learner. As pointed out by Brighton, Kirby, and Smith (2005), this type of account of typological generalizations depends on "the assumption that properties of the cognitive mechanisms supporting language map *directly* onto the universal features of language we observe."

2.2 Cultural evolution accounts

In contrast to purely syntactic accounts, typological generalizations such as word order correlations can be explained in terms of non–genetic, cultural evolution, i.e. language

change (Kirby 1999; Jäger and van Rooij 2005). In cultural evolution accounts, typological generalizations are emergent in a population from repeated cycles of language use and acquisition. This type of account makes certain key assumptions about how language change progresses. First, in order for a linguistic form to spread after it has been introduced into a population (e.g. as a consequence of language contact), language users must be able to learn and use the new form. Second, language users must have a BIAS (incentive) to do so, i.e. the new form has some social or structural advantage over the old form. There are several reasons why there might be a bias for a particular linguistic form. These reasons include (Jäger and van Rooij 2005):

- Learnability: Some forms are easier to learn than others.
- Processing: Some forms are less costly in processing.
- Use: Some forms are more useful in actual conversation.

A central claim of cultural evolution accounts is that the selective pressure of biases for particular linguistic forms results in the emergence of typological generalizations over many generations. Proponents of cultural evolution accounts are typically concerned with explaining typological generalizations via evidence of fit between structure and language use (Kirby 1999:10). For example, Hawkins (1994) proposes that word order correlations are ultimately a reflection of parsing complexity. Kirby (1999) shows how parsing principles such as those proposed by Hawkins can have a selective effect on the forms that make up the learning experience.

Functionalist explanations for typological generalizations such as Hawkins (1994) have been criticized on methodological grounds, e.g. that they are constructed after the fact "in the sense that there tends to be an *ad hoc* search for functions that match the universals to be explained" (Kirby 1999:13). One tool that can be used to circumvent this criticism is computer simulations of language use and acquisition. Computer simulations enable us to model cultural evolutionary explanations for language universals and explore the consequences of varying side conditions (Jäger and van Rooij 2005).

2.3 Overview of this paper

Our goal in this paper is to provide a cultural evolution account for the two observations presented at the beginning of this section. We use agent-based modeling to demonstrate that these properties of word order variation emerge in a population of BIASED VARIATIONAL learners. Along the way, we will contrast several different models of language use and acquisition, highlighting results that would not have been discovered by looking at solely one model. We first examine properties of FILTERED LEARNING MODELS (Kirby 1999). As demonstrated by Kirby (1999), this class of model can explain the emergence of typological generalizations such as word order correlations. However, the Filtered Learning Model presented by Kirby (1999) makes questionable assumptions about language learners. The second class of models we analyze are what we call VARIATIONAL LEARNING MODELS (Yang 2002; Clark 2004). This class of model makes reasonable

assumptions about language learners but fails to capture the emergence of typological generalizations. We discuss computer simulations that suggest that Filtered Learning Models can capture the emergence of typological generalizations, independent of assumptions about the language learner. Further, we demonstrate that a filtered version of the Variational Learning Model proposed by Yang (1999, 2000, 2002) overcomes the limitations of Kirby's (1999) Filtered Learning Model, while simultaneously capturing the emergence of typological generalizations.

3 Filtered Learning Models

3.1 Filtering

Filtered Learning Models (Kirby 1999; Briscoe 2001) introduce biases toward certain linguistic structures, e.g. consistently branching structures (Kirby 1999). In this class of model, these biases act as filters on the language data that speakers produce and acquirers perceive. As a consequence, the input that acquirers use to establish their language model[5] are adjusted in favor of the preferred structures. Correspondingly, preferred structures increase in frequency over time. (11) presents our assumptions about the transmission process from speakers to acquirers.[6] Filtering can happen at Step 1 and/or Step 2 in (11).

(11) **Steps of the transmission process:**

1. The language model of the speaker is used to produce utterances.
2. The acquirer perceives utterances.
3. The acquirer uses perceived utterances to establish their language model.

In the Filtered Learning Model presented by Kirby (1999), only successfully parsed (i.e. perceived) observations affect learning. The parser will occasionally fail, and, consequently, acts as a filter on the raw language data, i.e. filtering happens at Step 2 in (11). The set of utterances that is used by acquirers to establish their language model is a subset of the raw language data. Thus, we can capture the observation that more parsable (learnable) variants increase in frequency over time. In contrast, rather than claiming that differential parsability causes differential learnability, Hawkins (1994:83-95) argues that parsing influences GENERATION, and that more parsable variants will be used more frequently than less parsable ones, i.e. filtering happens at Step 1 in (11).

[5] We use *language model* here as a neutral term for a language user's mental linguistic competence. A language model could include multiple grammars.

[6] (11) is meant to encompass PURELY VERTICAL, OBLIQUE, and HORIZONTAL transmission (Cavalli-Sforza and Feldman 1981; Niyogi 2002). Purely vertical transmission involves transmission from parents to children. Oblique transmission involves transmission where members of the parental generation other than the parents affect acquisition. Horizontal transmission involves transmission where members of the same generation influence the acquirer.

Briscoe (1998) shows that either Hawkins' model or Kirby's, or a combination thereof, accounts for language change in favor of more parsable variants. The implementation of the Filtered Learning Model discussed in this section is completely agnostic about whether parsing influences acquisition or generation and about the source(s) of biases for particular linguistic forms.

3.2 Kirby's (1999) Filtered Learning Model

In this section, we discuss the individual components of the Filtered Learning Model. The discussion is modeled after Kirby (1999:42–47). Imagine a language with the brace construction as the basic order for clauses with a nonfinite verb and an auxiliary verb. Such languages do exist, as indicated in Table 1. For example, Koopman (1984) argues that the West African language Vata is an INFL–medial OV (OV&AuxV, brace) language. If the consistent all–medial structure is introduced into the language (by language contact, by expressiveness), then the relative well–formedness of the all–medial structure and the brace structure predicts that the all–medial structure should win over time. The acquisition process in the Filtered Learning Model is described in (12) (adapted from Kirby 1999, 45):

(12) **Acquisition process with filtering:**

1. consider the number of brace constructions and all–medial constructions in the input;
2. convert these numbers into probabilities reflecting the chance of each variant being chosen at random from the sample to trigger acquisition;
3. scale those probabilities (e.g. by using the Early Immediate Constituents metric for the variants; see Hawkins 1994 and Kirby 1999) so that the probability of the all–medial construction being used for acquisition is raised and the probability of the brace construction being used is lowered.

The equation in (13) accounts for the way in which the filtered subset of utterances that is used by acquirers to establish their language model is selected.[7] $p(f)$ (e.g. p(all–medial)) is the probability of the construction f occuring in the trigger experience. n_f is the number of tokens of the construction f in the language data. α corresponds to the learning bias for the preferred structure.

[7](13) is equivalent to Kirby's (1999, 46) equation, given in (i):

(i) $p(\text{all-medial}) = \frac{.89.n_{\text{all}-\text{medial}}}{.89.n_{\text{all}-\text{medial}} + .61.n_{\text{brace}}}$

Assuming a two-word NP, the values .89 and .61 correspond to the aggregate Immediate Constituent-to-word ratios for the all-medial construction and brace construction, respectively (see Hawkins 1994 and Kirby 1999). For this case, $(1 - \alpha)$ in (13) would equal $\frac{.61}{.89}$.

(13) $$p(\text{all-medial}) = \frac{n_{\text{all-medial}}}{n_{\text{all-medial}} + (1-\alpha)n_{\text{brace}}}$$

In order to explore the consequences of the Filtered Learning Model, we constructed computer simulations within Swarm, a software package for multi–agent simulation of complex systems.[8] The simulations include the following components (Kirby 1999:43–44):

(14) a. Utterances: Features of sentences, e.g. OV, VO.

b. Arena of use: An unstructured pool of utterances.

c. Grammars: List of possible utterances, e.g. [OV].

The implementation of the Filtered Learning Model described in Kirby (1999) makes certain key assumptions about speakers and language learners that differentiate it from the Variational Learning Model we describe in Section 4. As noted in the introduction, during periods of word order change, linguistic behavior is variable both at the level of the community and at the level of the individual (see Table 2). For Kirby (1999:37), the frequency of use of a particular word order during language change is taken to be a reflection of the use of that order by a particular speech community. Individual language learners, though, do not learn the frequency of use of a particular word order. For example, individual linguistic competence can only be OV and VO, not both (Kirby 1999:45):

> "... it is possible to have different frequencies for different orders without compromising a theory of 'all–or–nothing' competence." (Kirby 1999:37)

In this section, we discuss computer simulations of Kirby's Filtered Learning Model, showing that this model can capture the emergence of typological generalizations. We then discuss certain architectural and empirical limitations of Kirby's model.

In addition to the components described in (14), Kirby's Filtered Learning Model includes the components in (15).

(15) a. Speakers: A speech community which is made up of a set of speakers each of which consists of a single grammar. These grammars produce utterances for input to the arena of use.

b. Acquirers: There are speakers who have not been assigned grammars. They take as learning data utterances from the arena of use.

Two dynamic processes— production and acquisition— govern the interaction of the components in (14) and (15) (Kirby 1999:44). These processes are described in (16). A key feature of the acquisition process is the assumption that the trigger experience

[8] `www.swarm.org`

is mapped directly to a unique grammar. In this way, Kirby's model of acquisition is an example of the TRANSFORMATIONAL LEARNING APPROACH to language acquisition: the state of the acquirer undergoes direct changes as an old hypothesis is replaced by a new one. As Yang (2002:Ch. 2) points out, this approach is formally insufficient and incompatible with what is known about child language acquisition. We return to this point in the conclusion to this section.

(16) a. Production: Speakers add utterances to the arena of use in line with their grammars.

b. Acquisition: Acquirers develop a grammar (and become speakers) by

1. taking a random subset of utterances from the arena of use,
2. modifying the subset through the process of filtering described in (12),
3. choosing an utterance— the *trigger*— from the modified subset, and
4. mapping the trigger directly to a grammar

3.3 Population-level characteristics of Kirby's (1999) model

Figure 1 shows the time course of change for simulation runs of Kirby's Filtered Learning Model, at varying levels of the learning bias (henceforth, α) for the preferred variant. The initial frequency of the preferred structure was held constant at 20% for each run. Likewise, the random subset of utterances that acquirers take from the arena of use (the sample rate) was held constant at 10%. Each simulation was run for 100 iterations, after which the arena of use consisted entirely of the preferred variant when $\alpha > 0$. A key feature of this graph is that when $\alpha > 0$ (i.e. any positive bias) the slopes of change resemble the S–curve (Weinreich et al. 1968; Kroch 1989). When $\alpha = 0$ (no bias), random walk behavior is observed.

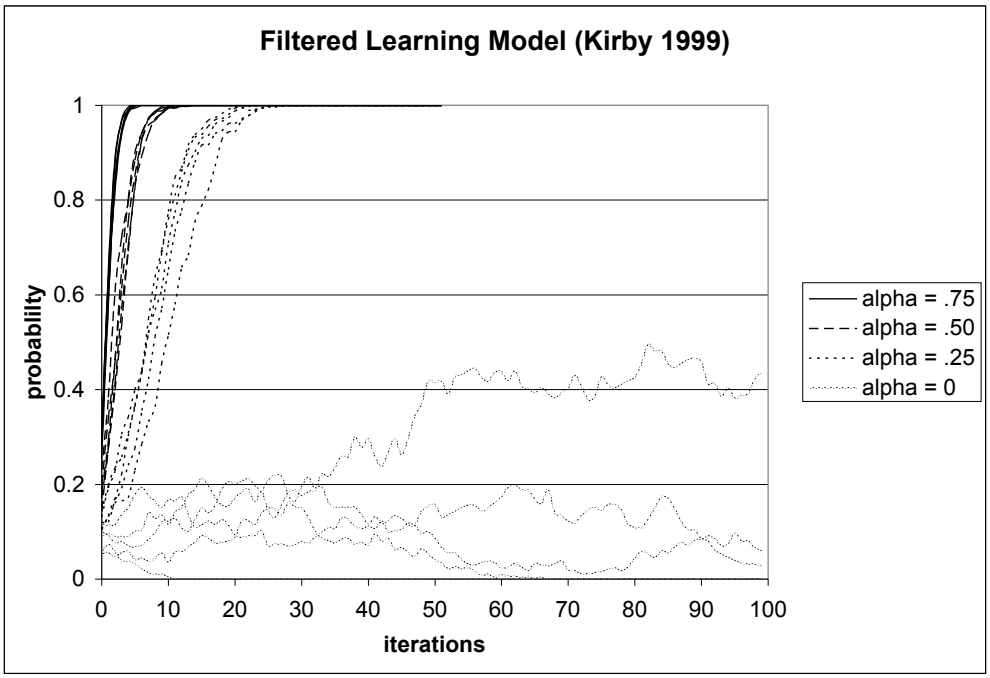


Figure 1: Filtered Learning Model (Kirby 1999) with varying levels of bias

These simulation results suggest that languages can adapt to an asymmetric functional pressure through a process of non–genetic cultural evolution (i.e. language change). If this type of model is on the right track, we should reject accounts that depend entirely on a direct mapping between the cognitive mechanisms supporting language and typological generalizations such as the purely syntactic accounts described in Section 2.1 (Kirby 1999:135).

3.4 Individual-level characteristics of Kirby's (1999) model

The Filtered Learning Model presented in Kirby (1999) is inadequate for several reasons hinted at earlier. Kirby assumes a TRANSFORMATIONAL LEARNING APPROACH[9] to language acquisition in which the state of the acquirer undergoes direct changes as an old hypothesis is replaced by a new one. This approach has been shown to be formally insufficient, see Yang (2002:18–20) for a summary. The transformational learning approach has also been demonstrated to be incompatible with what is known about

[9]This term is borrowed by Yang (2002:15) from evolutionary biology (Lewontin 1985).

children's linguistic development. In all transformational learning models, including the one assumed by Kirby, the learner is always identified with a single, unique grammar. This predicts, among other things, that abrupt changes in the use of linguistic expressions should be observed as the acquirer shifts from grammar to grammar. There is no evidence of this developmentally. Rather, language development is gradual (Yang 2002:22).

Figure 2 presents a simulation run illustrating the gradual spread of a preferred grammar through a speech community. In contrast to Figure 1, which showed the time course of change for the entire population, Figure 2 shows the distribution of variants at the level of individual speakers. At each stage of the change we see INTERSPEAKER, but not INTRASPEAKER, variation.

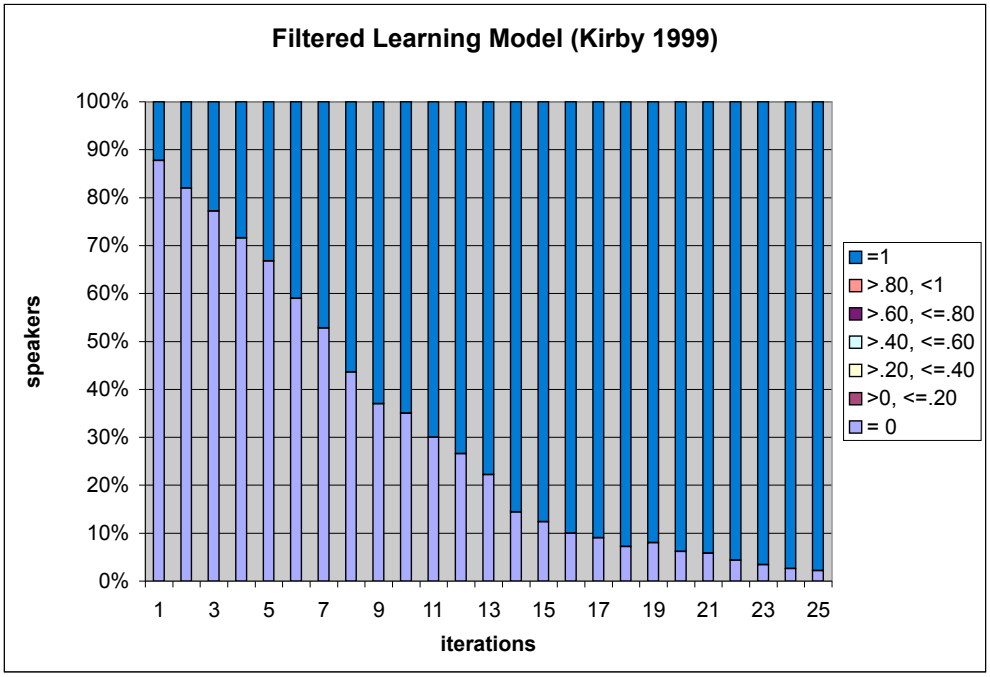


Figure 2: Grammars in the speech community during a period of change

The historical linguistics and variationist literature strongly suggest that the trajectory predicted by Figure 2 is completely unattested: both intraspeaker and interspeaker variation are always observed during periods of change (Weinreich et al. 1968; Kroch 2001; Clark 2004). Work in the variationist tradition (starting with Weinreich et al. 1968) has provided ample evidence that linguistic competence accommodates and gen-

erates variation. Intratextual variability was illustrated in Table 2, repeated here as Table 3. For the authors of the two texts described in Table 3, at least three orderings are available for subordinate clauses with a subject, an object, a (pre-)auxiliary finite verb, and nonfinite verb.

Table 3: Estimated frequencies of structures (4a-c) in Chron A, Scribe 1 (OE) and *Festis Marie* (early ME)

structures	% for Chron A, Scribe 1	% for *Festis Marie*
(4a) $[_{IP}$ XP $[_{I'}$ $[_{VP}$ YP V] I]]	61	8
(4b) $[_{IP}$ XP $[_{I'}$ I $[_{VP}$ V YP]]]	1	67
(4c) $[_{IP}$ XP $[_{I'}$ I $[_{VP}$ YP V]]]	38	25

Data such as that in Table 3, along with the generalizations about child language development noted above, suggests that our model should make it is possible to have different frequencies for different structures without adopting an empirically false notion of 'all–or–nothing' competence. Rather, we need a model of linguistic competence that accommodates and generates variation.

Lastly, a key underlying assumption of Kirby's Filtered Learning Model is that language acquisition is probabilistic. In Kirby's model, a language acquirer's grammar is determined by the probability of the grammar in the filtered subset of utterances taken from the arena of use. For example, if the frequency of the preferred variant in the filtered subset of utterances is 90%, there is a 10% chance that the learner will acquire the less preferred structure. Kirby's assumption that language acquisition is probabilistic is contrary to what is known about language acquisition. Research on language acquisition has shown that children are highly competent and robust learners: "it seems unlikely that, given similar experience, children would attain languages that differ substantially" (Yang 2000:237).

In the next section, we describe a model of language acquisition and use which overcomes the limitations of Kirby's Filtered Learning Model. We show how this model can be extended to capture the emergence of typological generalizations such as word order correlations.

4 Filtered Variational Learning Models

4.1 Variational Learning Models

In this section, we focus on approaches that capture intraspeaker variable linguistic behavior in terms of models of linguistic competence that accommodate and generate variation (Clark 2004; Yang 2002). Following Yang (2002), we call models of this sort VARIATIONAL LEARNING MODELS. The key property of Variational Learning Models

are their guiding assumption that learning involves "coexisting hypotheses in competition and gradual selection" (Yang 2002:34). Consequently, Variational Learning Models avoid the architectural and empirical problems with the Filtered Learning Model presented by Kirby (1999). To illustrate the variational learning approach, we focus on the model presented by Yang (1999, 2000, 2002). Clark (2004) discusses a related Variational Learning Model within the framework of Stochastic Optimality Theory.

In the Variational Learning Model presented by Yang (1999: 431, 2002: 26–30), intraspeaker variable linguistic behavior is modeled in terms of a population of grammars. Each grammar G_i is associated with a weight p_i, $0 \leq p_i \leq 1$ and $\sum p_i = 1$. Each of these weights denote the probability with which the learner can access the associated grammar. In a learning environment E, the weight $p_i(E, t)$ is determined by the learning function, E, and t (the time since the onset of language acquisition).

Learning is modeled in terms of the changing weights of grammars in response to the sentences incrementally presented to the acquirer. Suppose that there are N grammars in the population. Write $G_i \rightarrow s$ if a grammar G can analyze sentence s. Write γ for the learning rate.[10] Write p_i for $p_i(E, t)$ at time t , and p_i' for $p(E, t + 1)$ at time $t + 1$, where each time instance corresponds to the presentation of an input sentence. Learning takes place as in (17), the Linear reward–penalty scheme (Bush and Mosteller 1951, 1958). In this paper, we are only going to look at a simple two grammar case.

(17) Given an input sentence s, the learner selects a grammar G_i with probability p_i:

a. if $G_i \rightarrow s$ then $\begin{cases} p_i' = p_i + \gamma(1 - p_i) \\ p_j' = (1 - \gamma)p_j \text{ if } j \neq i \end{cases}$

b. if $G_i \not\rightarrow s$ then $\begin{cases} p_i' = (1 - \gamma)p_i \\ p_j' = \frac{\gamma}{N-1} + (1 - \gamma)p_j \text{ if} j \neq i \end{cases}$

Yang (1999, 2002) shows that, in the general case, when learning stops, grammars more compatible with the learning data are better represented in the population than other, less compatible grammars. As a consequence, in a heterogeneous learning environment where no single grammar can analyze every input sentence, the acquirer converges on a stable combination of grammars. This consequence of the Variational Learning Model is very relevant for language change: the linguistic competence of speakers during periods of language change is modeled as the combination of multiple grammars. This result is supported by much work in historical syntax, see, e.g., Pintzuk (1999), Kroch (2001), and Clark (2004).

[10] In what follows, the value of γ is low (.01), i.e. the learner does not alter the weight of grammars too radically in response to input sentences (Yang 2002:32).

4.2 Incorporating filtering in Variational Learning Models

As Yang (2002:54) notes, there is no bias in grammar evaluation in the Variational Learning Model: "all grammars are there to begin with, and input–grammar compatibility is the only criterion for rewarding/punishing grammars." Consequently, while the Variational Learning Model overcomes the Filtered Learning Model's failure to capture intraspeaker variation, it cannot account for typological generalizations such as those observed by Greenberg (1966), Hawkins (1983), and Dryer (1992). The simulations described in Section 3 suggest that languages can adapt glossogenetically to an asymmetric functional pressure through a process of non–genetic cultural evolution (i.e. language change).

To illustrate the problem, consider a situation in which the all–medial and brace construction are in competition, and the brace construction is more frequent than the all–medial construction. The Filtered Learning Model presented by Kirby predicts that the consistent, all–medial construction should win out over time. In contrast, for Yang's Variational Learning Model, we expect relative random fluctuations in the relative frequency of the brace construction and the all–medial construction (i.e. stable variation), barring external changes such as language contact.

To capture the emergence of typological generalizations, we can extend the Variational Learning Model presented in Yang (2002) so that certain structures are preferentially selected for by the learner. We do this by adding a filtering function F to the Variational Learning Model.[11] Figure 3 illustrates the application of the filtering function F to the weight $p_{\text{preferred}}$ associated with the preferred grammar at different levels of bias (α). The filtering function F has two key properties. First, F drives probabilities to extrema. Consequently, one form will always drive the other form out of use. Second, there is an asymmetric drive towards extrema. Correspondingly, there is pressure to adopt the preferred form over time. In this way, typological asymmetries such as word order correlations are captured.

[11] The filtering function F is given in (i):

i. $F(p) = \dfrac{1}{1 + e^{-z'(p,\alpha)}}$, where $z'(p, \alpha) = 5((10(\alpha + .1))2p - 1)$

A reviewer asks why we did not just use the the sigmoid function in (ii) and scale α appropriately. The sigmoid function in (ii) captures the observation that one form will always drive the other form out of use and that there is pressure to adopt the preferred form over time. However, the function in (ii) does not map from {0,1} to the range {0,1}. Rather, it has a range of -1 to 1. While less elegant than (ii), the function in (i) correctly maps to the range {0,1}, i.e. $z'(p)$ $(= 5(2p - 1))$ compresses the function F to {0,1}.

ii. $F'(p) = \dfrac{1}{1 + e^{-\alpha \times p}}$

The reviewer also states that "the sigmoid function [in (i)] gives no principled way to generalize to multiple grammars". So long as the grammar space is defined by binary parameters, biases can be attached to every parameter. As a consequence, we get an asymmetric drive towards extrema in a grammar space containing more than two grammars. This type of approach is adopted by Kirby (1999:48-49).

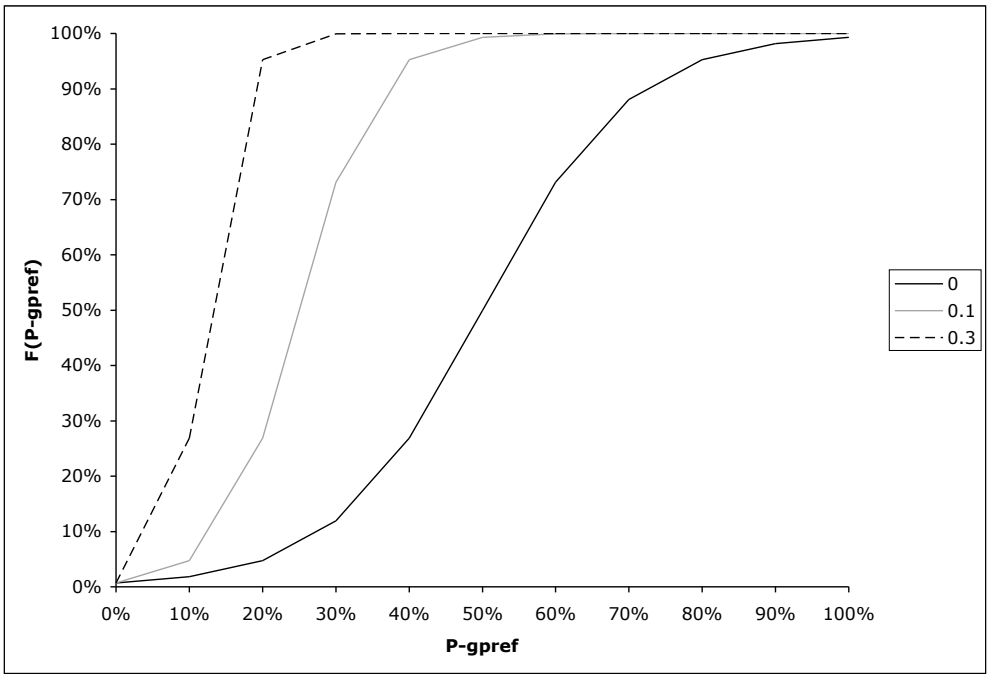


Figure 3: Filtering with different levels of bias

In the Variational Learning Model, language change is defined in terms of the diffusion of grammars in successive generations of language learners. In order to explore the consequences of the Variational Learning Model for language change, we ran computer simulations within NetLogo, a cross-platform multi–agent programmable modeling environment for the simulation of complex systems.[12]

The simulations discussed in this section share the components in (18) with the Filtered Learning Model presented in Kirby (1999:43–44):

(18) a. Utterances: Features of sentences, e.g. OV, VO.

b. Arena of use: An unstructured pool of utterances.

c. Grammars: List of possible utterances, e.g. [OV].

As noted above, Variational Learning Models crucially differ from the Filtered Learning Model presented by Kirby (1999) in the assumptions they make about language

[12]http://ccl.northwestern.edu/netlogo/

users. (19) describes speakers and acquirers in our computer simulations of the Variational Learning Model.

(19) a. Speakers: A speech community which is made up of a set of speakers each of which consists of two grammars G_1 and G_2. Each grammar is associated with a weight, p_1 and p_2. These grammars produce utterances for input to the arena of use.

b. Acquirers: There are speakers whose grammars have been assigned neutral weights, $p_1 = .5$ and $p_2 = .5$. They take as learning data utterances from the arena of use.

As with Kirby's Filtered Learning Model, two dynamic processes, production (described in (20)) and acquisition (described in (21)), govern the interaction of the components in (18) and (19). As noted above, the learning algorithm in Yang's Variational Learning Model is the Linear Reward–Penalty scheme given in (17), rather than the probabilistic learning algorithm assumed by Kirby's Filtered Learning Model. We make the simplifying assumption here that there are no overlapping generations. Note that the acquisition process encapsulates multiple interactions with different speakers. In the simulations described below, the speaker's linguistic knowledge is stable after acquisition.

(20) Production: Speakers probabilistically add utterances to the arena of use in line with the weights p_1 and p_2 associated with G_1 and G_2.

(21) Acquisition:

1. interact with a speaker;
2. receive the preferred form with probability $F(p_{\text{preferred}})$ (where F is the filtering function and $p_{\text{preferred}}$ is the weight the speaker associates with grammar $G_{\text{preferred}}$), else less preferred form;
3. apply the Linear Reward–Penalty learning algorithm in (17).

The amount by which the probability of receiving the preferred structure from a speaker is raised is dependent on α (the degree of bias), as shown in Figure 3.

4.3 Results

Recall the guiding observations that we wish to account for:

i. There is extraordinary cross-linguistic word order variation. For example, all of the logically possible combinations of orderings of auxiliary verbs, content verbs, and objects are attested stable grammatical states.

ii. Certain word order patterns (e.g. VO&AuxV, OV&VAux) are more frequent than others (e.g. VO&VAux, OV&AuxV) and this reflects a preference for consistently branching structures.

The left side of Figure 4 shows the time course of change for multiple simulation runs of the filtered Variational Learning Model at a single level of α (= 0.0015). The initial random distribution of the preferred structure was varied from .25 to .75 in .05 increments. 10 simulations were performed for each initial distribution. Learners received 500 utterances as input (50 per speaker, 10 speakers total (1% of the population)). Each simulation was run for 20 iterations. The right side of Figure 4 shows that the proportion of outcomes that resulted in convergence to the preferred form (e.g. the all-medial construction) was larger than the proportion of outcomes that resulted in convergence to the dispreferred form (e.g. the brace construction). Crucially, the dispreferred form is a possible stable grammatical state; the population occasionally converged on the less preferred form. In this way, the filtered Variational Learning Model captures both typological asymmetries and multistability (i.e. the fact that both typologically preferred and dispreferred structures are possible stable states).[13]

[13]Note that the simulations reported here assume equal probability over a range of initial conditions; this assumption is critical for the conclusions drawn here (thanks to Gerhard Jäger for drawing our attention to this point). To the extent that the initial conditions do not cover a sufficiently wide range, the number of typologically stable states will decrease. For example, as shown in Figure 4, above an initial probability of .6 $G_{\text{less-preferred}}$ is never observed as a final state. (Conversely, below .4 $G_{\text{preferred}}$ is never observed as a final state.) Similarly, if the initial distributions are not equiprobable, the relative probability of final states may not reflect the biases of learners. For example, if initial states are more likely to be drawn from below .4, the influence of the bias towards the preferred grammar will be less apparent.

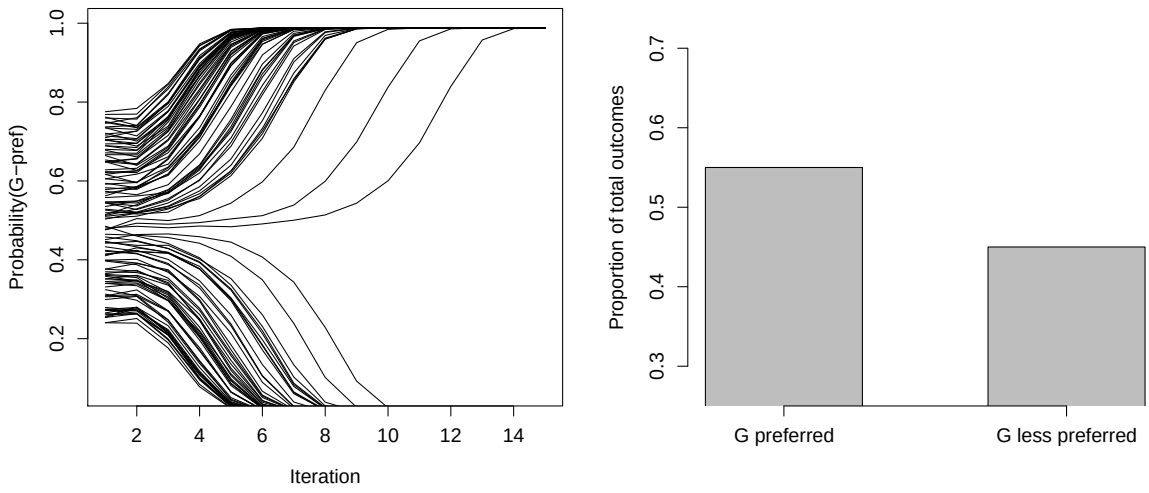


Figure 4: Filtered Variational Learning Model

Figure 5 shows that that the higher the bias (α), the more likely it is that the population will converge on the preferred grammar. The number of simulation runs in each column of the graph is 75, with the initial distribution of the preferred structure held constant at .50. The error bars in Figure 5 show the standard error of the mean.

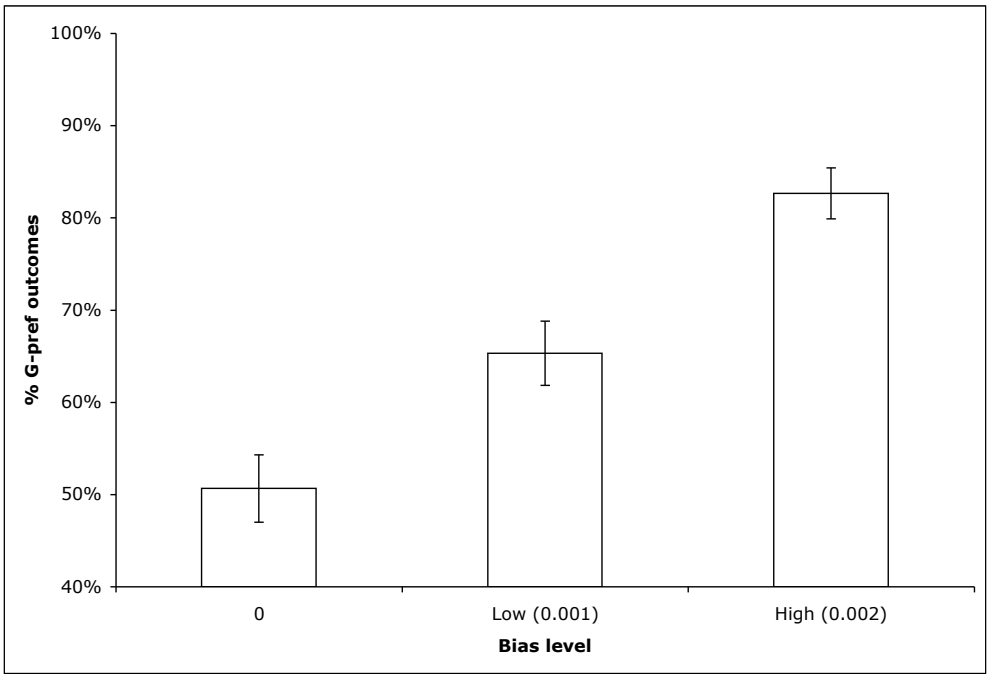


Figure 5: Filtered Variational Learning Model with varying levels of bias

Recall that a guiding assumption of the Filtered Learning Model presented in Kirby (1999) is that linguistic competence is 'all–or–nothing': speakers can be OV or VO, but not both. Figure 2 presented a simulation run illustrating the gradual spread of a preferred grammar through a speech community. At each stage of the change we saw interspeaker, but not intraspeaker, variation. In contrast, Variational Learning Models are able to capture both interspeaker and intraspeaker variation during periods of language change. Figure 6 illustrates a single simulation run in which a preferred structure (e.g. the all–medial construction) is spreading through a community of speakers. At each stage of the change, we can see that the community is composed of individuals whose linguistic competence accommodates and generates variation (Weinreich et al. 1968; Kroch 2001; Clark 2004). At Stages 0 and 1, the majority of the speech community nearly categorically produces utterances of the less preferred form. At Stages 2 and 3, the majority of the speech community variably produces both the preferred and the less preferred form. Lastly, at Stage 4, the majority of the population nearly categorically produces the preferred form. This result demonstrates that Variational Learning Models are more empirically adequate than the Filtered Learning Model presented in

Kirby (1999).

Further, the Variational Learning Model gives us a way to capture the process of GRADUAL LANGUAGE DEATH. Gradual language death is language loss due to "the gradual shift to the dominant language in a contact situation" (Wolfram 2002:766). During periods of gradual language death, "there is often a continuum of language proficiency that correlates with different generations of speakers" (ibid.). In Figure 6, monolingual speakers of the preferred grammar only take over the whole speech community after a period in which the majority of the speech community is bilingual. The Filtered Learning Model presented in Kirby (1999) is not able to capture this kind of change.

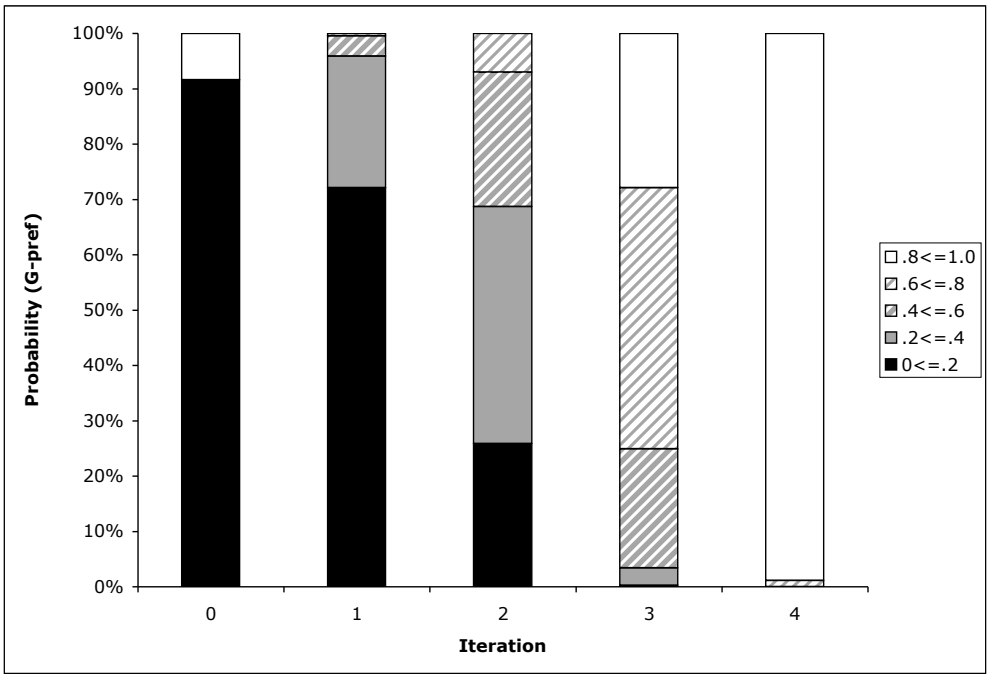


Figure 6: Grammars in the speech community during a period of change

5 Summary and discussion

In sum, in this paper we have presented computer simulations that suggest that Filtered Learning Models can account for the emergence of typological generalizations over time,

independent of particular assumptions about the learner. Filtered Variational Learning Models avoid the architectural and empirical limitations of the Filtered Learning Model presented in Kirby (1999). We have shown that filtered Variational Learning Models are able to capture the emergence of typological generalizations through a process of non–genetic cultural evolution, thus preserving the insights of Kirby (1999).

This paper reports on ongoing work. We are currently building on what was presented here, and extending it in two directions. First, one of the major benefits of computer simulations of language acquisition and use is the ability to explore the ramifications of different side conditions. In related work we have demonstrated that differing side conditions of the various models described here (e.g. sample rate) have important consequences (Konopka 2006). Second, we have assumed, with Kirby (1999:Ch.2), that speakers input into an unstructured arena of use, and that language learners sample from random points in the arena. In future work, we hope to incorporate models of local social structures (Milroy 2002) such as social networks into our simulations.

References

Allen, Cynthia. 2000. Obsolescence and sudden death in syntax: The decline of verb-final order in early Middle English. In Generative Theory and Corpus Studies: A Dialogue from 10 ICEHL, ed. by Ricardo Bermúdez-Otero, David Denison, Richard M. Hogg, and C.B. McCully, 3–25. Berlin: Mouton de Gruyter.

Biberauer, Theresa, and Ian Roberts. 2005. Changing EPP–parameters in the history of English: accounting for variation and change . English Language and Linguistics 9.5–46.

Brighton, Henry, Simon Kirby, and Kenny Smith. 2005. Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In Language origins: Perspectives on evolution, ed. by Maggie Tallerman, chapter 13, 291–309. Oxford: Oxford University Press.

Briscoe, Ted. 1998. Language as a complex adaptive system: Co-evolution of language and of the language acquisition device. In 8th Meeting of Computational Linguistics in the Netherlands, ed. by P. Coppen, H. van Halteren, and L. Teunissen, 75–105. Amsterdam: Rodopi.

——. 2001. Evolutionary Perspectives on Diachronic Syntax. In Diachronic Syntax: Models and Mechanisms, ed. by Susan Pintzuk, George Tsoulas, and Anthony Warner, 75–105. Oxford: Oxford University Press.

Bush, Robert R., and Frederick Mosteller. 1951. A Mathematical Model for Simple Learning. Psychological Review 68.313–323.

——, and ——. 1958. Stochastic Models for Learning. New York, NY: Wiley.

Cavalli-Sforza, L., and M. W. Feldman. 1981. Cultural Transmission and Change: A Quantitative Approach. Princeton, NJ: Princeton University Press.

Clark, Brady Z. 2004. *A Stochastic Optimality Theory Approach to Syntactic Change*. Stanford, CA: Stanford University dissertation.

Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. Language 68.81–138.

—— 2006. Word Order. In Clause Structure, Language Typology and Syntactic Description, Vol. 1, ed. by Timothy Shopen. Cambridge University Press.

Fischer, Olga, Ans van Kemenade, Willem Koopman, and Wim van der Wurff. 2000. The Syntax of Early English. Cambridge: Cambridge University Press.

Greenberg, Joseph H. 1966. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Universals of Language (Second Edition), ed. by Joseph H. Greenberg, 73–113. Cambridge, MA: MIT Press.

Hawkins, John A. 1983. Word Order Universals. New York, NY: Academic Press.

—— 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Jäger, Gerhard, and Robert van Rooij. 2005. Language Structure: Psychological and Social Constraints. University of Bielefeld and University of Amsterdam, ms.

Kirby, Simon. 1999. Function, Selection, and Innateness. Oxford: Oxford University Press.

Konopka, Kenneth, 2006. A computational model of language change. Northwestern University.

Koopman, Hilda. 1984. The Syntax of Verbs: From Verb Movement Rules in the Kru Languages to Universal Grammar. Dordrecht: Foris.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. Language Variation and Change 1.199–244.

——. 2001. Syntactic Change. In The Handbook of Contemporary Syntactic Theory, ed. by Mark Baltin and Chris Collins, 699–729. Oxford: Blackwell.

——, and Ann Taylor. 2001. Verb-Object Order in Early Middle English. In Diachronic Syntax: Models and Mechanisms, ed. by Susan Pintzuk, George Tsoulas, and Anthony Warner, 132–163. Oxford: Oxford University Press.

Lehmann, Winfred P. 1973. A structural principle of language and its implications. Language 49.42–66.

Lewontin, Richard. 1985. The Organism as the Subject and Object of Evolution. In The Dialectical Biologist, ed. by Richard Levins and Richard Lewontin, 85–106. Cambridge, MA: Harvard University Press.

Mallison, Graham, and Barry J. Blake. 1981. Language Typology: Cross–linguistic Studies in Syntax. Amsterdam: North–Holland.

Milroy, Lesley. 2002. Social Networks. In The Handbook of Language Variation and Change, ed. by J.K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 549–572. Malden, MA: Blackwell Publishing.

Niyogi, Partha. 2002. Theories of cultural evolution and their application to language change. In Linguistic Evolution through Language Acquisition, ed. by Ted Briscoe, 205–233. Cambridge: Cambridge University Press.

Pintzuk, Susan. 1996. Old English Verb-Complement Word Order and the Change from OV to VO. In York Papers in Linguistics, number 17. University of York.

——. 1999. Phrase Structures in Competition: Variation and Change in Old English Word Order. New York, NY: Garland Publishing, Inc.

Svenonius, Peter. 2000. Introduction. In The Derivation of VO and OV, ed. by Peter Svenonius, 1–26. Amsterdam: Benjamins.

Vennemann, Theo. 1973. Explanation in syntax. In Syntax and Semantics 2, ed. by John Kimball, 1–50. New York, NY: Seminar Press.

Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In Directions for historical linguistics: a symposium, ed. by W.P. Lehmann and Y. Malkiel, 95–188. Austin, TX: University of Texas Press.

Wolfram, Walt. 2002. Language Death and Dying. In The Handbook of Language Variation and Change, ed. by J.K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 764–787. Malden, MA: Blackwell Publishing.

Yang, Charles D. 2000. Internal and external forces in language change. Language Variation and Change 12.231–250.

—— 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

—— 1999. A selectionist theory of language development. In Proceedings of 37th Meeting of the Association for Computational Linguistics, 429–435, East Stroudsburg, PA. Association for Computational Linguistics.