Language change, probabilistic models of

Kie Zuraw

University of California, Los Angeles

UCLA Department of Linguistics

3125 Campbell Hall

Box 951543

Los Angeles, CA 90095-1543

Abstract

Language change is gradual and takes place at the level of the population, not just at the level of the individual. Probabilistic models of language change have attempted to account for the spread of a new variant as a result of learners' responses to a variable environment, speakers' variable behavior, and properties of the population. Aspects of change accounted for in this way include the bias of learners towards one variant, the spread of a variant from one linguistic environment to another, and the rate of spread over time.

1. Properties of a probabilistic model of language change

In many areas of linguistics it makes sense to study the mental grammar of the individual in isolation. Language change, however, cannot be understood solely as something that happens within an individual (even if individuals' grammars do change over their lifetimes), but rather must be understood as something that happens in a speech community, through the accumulation of interactions between individuals over time. Each of these interactions has a probabilistic character: which individuals will interact, which variant the speaker will produce, and how the listener's mental grammar and lexicon will be affected.

Because language change is not instantaneous at the population level, any well developed model of language change must be probabilistic at least in the sense of specifying how an individual behaves when surrounded by variation and how that behavior in turn affects the rest of the population. In the simplest model, a learner could simply adopt the grammar of the first other speaker it encounters (assuming that the encounter provides enough information for the learner to identify that grammar). In that case, the model is probabilistic only in determining what type of speaker the learner encounters. But more realistic models require learning algorithms that can handle the learner's variable environment..

It can be difficult (or misleading) to try to understand how such a model will behave without being mathematically explicit about how the model works, and examining its behavior under varying conditions, whether analytically—by studying

properties of equations—or through computer simulation when direct analysis is infeasible. Being explicit requires simplifying reality somewhat, and depending on researchers' interests, models are typically more realistic in one area than in others.

A model of language must specify what the internal state of an individual looks like. First, does an individual have a single grammar that produces no variation, a single grammar that produces probabilistic variation, or multiple grammars each associated with a probability of use? ("Grammar" here is used loosely to include not only a language's phonology, morphology, and syntax, but also possibly its lexicon.) Second, how is a grammar represented? Possibilities that have been proposed include a vector of parameter settings, a connectionist network, a cloud of exemplars, or a probabilistic ranking of Optimality-Theoretic constraints.

It is also vital to specify how an individual's grammar is arrived at during initial acquisition, and how it may change during adulthood; the choice of learning or update algorithm of course depends on the nature of the grammar. Once a learning algorithm is specified, it is still necessary to consider whether there is a critical period after which learning ceases, and if so how much data the learner is exposed to during that period; if there is no critical period, the model must specify how plastic the grammar remains as the individual ages.

Moving to the level of the population, the model must specify who interacts with whom. Do children learn only (or preferentially) from their parents? Do all individuals interact equally with all members of the population, or is the population structured, whether geographically or socially, so that some interactions are more likely than others?

Does the listener in an interaction treat all data the same way, or is an utterance's influence a function of the speaker's (and possibly the listener's) age, the strength of the social connection between the two individuals, or even the social status of the speaker? Does the speaker's choice of utterance similarly depend on characteristics of the listener? Properties of the population itself must also be decided, such as size, nature of geographic and social structure (if any), birth and death rates, and whether these differ as function of grammar, geographic location, or status in the social network (see Milroy 1987 on the role of social networks in variation and change).

2. Probabilistic grammars

It has been well documented that when a linguistic change is in progress, productions vary, at rates that increasingly favor the innovative variant, not just across individuals but also within individuals (see Labov 1994, 2001; Kroch 2001 and references therein). Weinreich, Labov, and Herzog (1968) proposed to model this with *variable rules*, phonological rules that apply not obligatorily, but with a certain probability, which can change from generation to generation (see also Paolillo 2001). Little work has been done on modeling acquisition in this framework.

2.1.1 Probabilistic parameter settings

In syntax, Kroch (1989 and subsequent work by Kroch and colleagues; see Kroch 2001 for references) has attached probabilities to parameter settings. While change is in progress, individuals have nonzero probabilities associated with both settings of some parameters.

Learning algorithms have been proposed for parameter setting in variable environments, but these algorithms generally do not result in a variable grammar. Niyogi and Berwick (1995), for example, consider the behaviors of variants of Gibson and Wexler's (1994) Triggering Learning Algorithm when the learning environment contains speakers with different grammars; but even in the face of contradictory data, the learner chooses a single grammar. In Briscoe's approach (2000 and references therein), the learner also arrives at a single vector of parameter settings, but during learning maintains probabilities for those settings, based on both their prior probabilities (universally given) and their probabilities in light of the learning data; these probabilities determine how easily a parameter setting can be changed in response to further learning data.

2.1.2 Competing grammars

Rather than having variation within a single grammar, it is also possible to imagine that speakers can maintain multiple, independent grammars, each having some probability of use. (As a model of diglossic or bilingual competence, such multiple grammars would need to be associated with additional information about their appropriateness in different social settings.) Yang (2000) models acquisition of competing grammars in a reward-penalty scheme. When the learner encounters an utterance, a grammar is selected at random according to its current probability. If the grammar can parse the utterance, its probability is increased; otherwise, its probability is decreased.

Competing full grammars permit a wider range of variation types than grammars with probabilistic rules, parameter settings, or constraint rankings. For example, in a grammar with parameters A and B, each having possible settings 0 and 1, a speaker who

maintains separate grammars could use the grammars (0,0) (A = 0 and B = 0) and (1,1), each at probability 0.5. But if the speaker associates probabilities not with grammars but with individual parameter settings, this will be impossible: if (0,0) and (1,1) are both used, then the parameter settings A = 0, A = 1, B = 0, and B = 1 all must have nonzero probabilities, so the speaker will sometimes use (0,1) and (1,0) as well. The situation is similar for variable rules and constraint rankings. It should be noted that the term "competing grammars" is often applied to cases of a single grammar with variable parameter settings.

2.1.3   Probabilistic constraint ranking

Probabilistic ranking of Optimality-Theoretic (OT) constraints has also been applied to both phonological and syntactic change: in this version of OT, a grammar consists not of a linear ranking of constraints, but of a probability distribution over possible constraint rankings. One such model is Boersma's (1998) stochastic OT, in which each constraint is associated with a ranking value; at the time of generating an utterance, random noise is added to each ranking value, and the resulting perturbed values are used to rank the constraints linearly. In this theory, learning is accomplished by Boersma's Gradual Learning Algorithm.

2.1.4   Adaptive rules

An alternative to probabilistic grammars during language change is the theory of adaptive rules (Andersen 1973), in which each individual has a non-variable grammar. But, if that grammar represents an innovative departure from the current standard, the speaker

develops a layer of rules to make her utterances conform to community norms. Such a speaker's adult utterances are just like a conservative speaker's; the difference is that the innovative speaker is "tolerant" of the same innovation in children. These children, facing less opposition, would then develop their adaptive rules later in life—meanwhile producing utterances with the innovation, which might serve as learning data to younger learners—and in turn be even more tolerant of innovation. This process continues until the adaptive rules are used only in a special, conservative style associated with the old, and eventually the adaptive rule is eliminated entirely. This theory rests on the assumption that a child learns not just from adults' utterances, but also from adults' reactions to the child's own speech; most models do not share this assumption.

3. Probabilistic categories

The probabilistic models of grammar described above rely on discrete categories, such as nouns vs. verbs, or [+high] vs. [–high] vowels. In some models of language change, however, category membership is itself probabilistic.

3.1.1 Exemplar-based models

Pierrehumbert (2002) has explored certain types of phonological change within an exemplar-based framework. In exemplar theory, tokens of encountered speech are mapped onto a similarity space with dimensions such as duration and formant values, and possibly time. Not every token is necessarily stored separately, because the similarity space is granular, or discretized into "bins" called exemplars. When a new token is encountered, the strength of the exemplar it belongs to is augmented (this could be done

either incrementally or by giving an exemplar maximum strength whenever a token is mapped to it). Countervailingly, exemplar strength decays over time. The learner forms categories—corresponding, for example, to members of the ambient language's vowel inventory—by identifying clusters of exemplars. Each category, then, is made up of many exemplars of varying strengths. In perception, a category label is attributed to an incoming stimulus according to the number of exemplars from each category that are similar to it, with a weighting in favor of stronger exemplars. In order to produce an instance of a category, an exemplar is chosen at random, but again with weighting in favor of stronger exemplars, and neighbors of the chosen exemplar contribute to the resulting production in proportion to their strengths; random noise may also be added. Lexical entries, moreover, contain weights that are applied to exemplars in production, so that one word may "prefer" to be realized with a vowel of long duration, while another prefers short duration.

Pierrehumbert models Joan Bybee's discovery that lenitory phonological changes (those involving gestural reduction) progress more quickly in high-frequency words. For example, English optional schwa reduction/deletion is more advanced in high-frequency words such as *ev_e_ry*, *cam_e_ra*, *mem_o_ry*, *fam_il_y* than in low-frequency words such as *mamm_a_ry*, *artill_e_ry*, *hom_il_y* (see Bybee 2001 and references therein). Bybee proposes that the reason for the phenomenon is that lenition is imposed every time a word is used and the lexical entry (of the listener and/or speaker) is updated in response. Since this happens more often in high-frequency words, they change more rapidly. Crucial to this explanation is allowing lexical entries to have a high degree of phonetic concreteness: in

the English schwa example, the lexical entry cannot simply be a string of phonemes, either containing or not containing a medial vowel. Rather, it must somehow contain information about the typical duration range (including zero) of the potential schwa vowel, such as through lexical weighting of schwa exemplars. Pierrehumbert's simulations, using repeated rounds of production and perception in the exemplar model with an externally imposed constant tendency representing lenition, show that change does indeed progress more rapidly in frequent words.

Pierrehumbert proposes that a phonetic change can eventually change lexical entries at the phonological level. For example, it is plausible that words like *family*, whose phonology allows schwa deletion (sonorants surround the medial vowel) are currently nonetheless represented lexically with a medial schwa, just as are words like *attitude*. But if English schwa reduction continues to progress, words like *family* will come to weight short exemplars of schwa so heavily that their schwa productions cluster around a very short (or even zero) duration, whereas the schwa productions of words like *attitude* will cluster around a longer duration. Since, in Pierrehumbert's model, learners infer phoneme-like categories from clustering, in subsequent generations the lexical entries for the two types of words will refer to different category strings.

3.1.2   Connectionist models

Connectionist models of competence have also been applied to language change, by such authors as Tabor (1994) and Johansson (1997).

Tabor's model of the grammar/lexicon is associative, with no explicit parameter settings or grammatical categories. The model is a connectionist network with three

layers of nodes: input, output, and hidden. Input nodes represent lexical items, and output nodes represent word behaviors, such as taking a certain type of complement. Activation of an input node flows to strongly connected hidden-layer nodes, which in turn activate strongly connected output-layer nodes; there is no direct connection between input and output nodes. Fuzzy grammatical categories, which can be conceptualized as patterns of input-node-to-hidden-node connection weight, emerge from similar distribution. Two words that would both be classified in a traditional grammar as nouns, for example, have relatively similar distributions and therefore similarly weighted connections to the hidden units. The use of this network to model language change is discussed in 4 and 5 below.

4. Constant rate effects

When a change, such as from SOV (subject-object-verb) to SVO word order, occurs in multiple syntactic environments, such as main and subordinate clauses, we can imagine different patterns of change. If for some reason the change originates in main clauses, SVO could replace SOV fairly quickly there, and more slowly in subordinate clauses, perhaps because the change is only gradually being generalized to environments that are less similar to the original environment of the change.

An alternative, advocated by Anthony Kroch and colleagues, is that when the syntactic theory dictates that variation in multiple environments be controlled by the same parameter (such as head-complement order), historical replacement of one variant by the other should proceed at the same rate in all environments, though it may, for poorly understood reasons, begin earlier in some environments than others. This is the

constant rate hypothesis, and it has been argued to be borne out by numerous case studies of syntactic change (see Kroch 2001 and references therein)

In order to test the constant rate hypothesis, "rate of change" must be defined. Kroch uses the logistic function, $P = \dfrac{1}{1 + e^{-k-st}}$, to model change, where P is the probability of use of the new variant, t is time, and k and s are constant parameters that determine when the change begins and how quickly it progresses (see 7.1 below for derivation of this function). Algebra transforms this equation into $\ln \dfrac{P}{1-P} = k + st$; we see that the logistic transform, or logit, $\ln \dfrac{P}{1-P}$, is equal to $k + st$, a linear function of t with slope (steepness) s and intercept (starting point) k. Kroch takes the slope s to be the rate of change; thus s is expected to be constant across environments.

Tabor (1994), however (see 3.1.2), views constant-rate change as merely a special case of *frequency linkage*. When a word or construction's distribution changes, the input-output relationship cannot be directly changed in response, since there are no input-to-output connections; only the input-to-hidden and hidden-to-output connection weights can change. When the hidden-to-output weights change because one word or construction shifts in distribution, any word or construction with a similar pattern of input-to-hidden connection weights is necessarily "dragged along" in the change to some extent. That is, words and constructions undergo similar changes in distribution to the degree that their previous distributions were similar. Constant-rate cases, for Tabor, are simply extreme cases in which two constructions' initial distributions are highly similar.

5.   Learning biases

Why are speakers biased to adopt some linguistic innovations, so that they spread, rather than faithfully imitating whatever they encounter equally, so that new variants fail to spread? Attempts to model such bias have mostly taken one of two approaches. The first approach simply assumes that the new variant is preferred by speakers, perhaps because it carries social prestige of some kind. This approach is inspired by the sociolinguistic research into changes in progress (most phonological) of William Labov and colleagues (see Labov 1994, 2001 and references therein; a truly Labovian model would require more than merely tagging one variant as preferred.)

The second approach, which has been more extensively developed, explores cases in which one variant might have a structural advantage over another, usually because it generates fewer ambiguous utterances. (See Clark 2004 for another approach—a processing filter that skews the data actually usable for learning.) A commonly used example is word order. Niyogi and Berwick (1995), for example, model the change from Old French verb-second (V2) word order to Modern French subject-verb-object (SVO) order. In their simulation, a grammar is represented as a vector of (non-probabilistic) binary parameter settings, and language acquisition consists of determining those settings based on input utterances. Acquisition would be easy if every utterance uniquely identified the grammar it came from, but there are many utterances that could be produced by more than one grammar. For example (to use English lexical items), the utterance *dog bites man* is consistent with either a V2 or an SVO grammar, whereas *suddenly bites dog man* is not consistent with SVO and *dog suddenly bites man* is not

consistent with V2. We can imagine various strategies that a learner might use in response to an ambiguous utterance such as *dog bites man*: ignore the utterance, increment the probability of every grammar consistent with the utterance, or randomly choose one grammar consistent with the utterance and augment its probability. A learner need not even recognize the ambiguity: under a Triggering-Learning approach, the learner changes grammars only if the current grammar fails to parse an incoming utterance; an ambiguous sentence will cause no change if the learner's current grammar is consistent with it.

Under all these strategies, a grammar is at a disadvantage if it produces a higher proportion of ambiguous sentences than a competing grammar does. For example, consider a population in which 20% use innovative L_1 and 80% use conservative L_2. If 80% of L_1's utterances are unambiguous, compared with 10% of L_2's, then for every 100 utterances, 16 are triggers for L_1 and only 8 are triggers for L_2 (76 are ambiguous). Although L_1 is rarer in the population, from the learner's point of view there are more utterances coming from L_1 than from L_2. L_1 will therefore inexorably encroach on L_2.

Working in the framework of Stochastic Optimality Theory, Jäger (2003) has also modeled learning bias as a source of language change. In Jäger's variant of Optimality Theory, learning data are form-meaning pairs. The learner checks whether its current grammar (probabilistically) chooses the observed meaning as optimal given the observed form, but also whether its current grammar chooses the observed form given the observed meaning. If either type of mismatch occurs, the learner adjusts constraint ranking values accordingly, as in Boersma's (1998) Gradual Learning Algorithm. Jäger simulates the

behavior of this system through repeated rounds of production and learning, and finds that when one type of form is penalized by the constraint inventory more than another (regardless of ranking), the "better" form becomes associated to the more frequent meaning. This is because if the system begins unbiased, it will tend to use the "better" form for both meanings. If the learner's grammar picks the wrong meaning as optimal, it must demote any constraint that disfavors the observed form-meaning pair. The more frequent meaning is of course encountered more often, so the constraint against associating that meaning to the better form gets demoted more vigorously, and the frequent meaning increasingly is produced with the better form. Jäger applies this model to case-marking systems dependent on animacy; corpus data on real languages reveals pragmatically motivated asymmetries, such as that sentences in which an animate actor acts on an inanimate patient are common. Jäger is able to produce the result that zero case-marking (which he takes to be favored) tends to become associated with the common combinations of animacy and thematic role.

Tabor (1994) approaches biases from a different angle, focusing on possible trajectories from one grammar to another. Tabor asks why English *be going to*, on its journey from meaning solely 'be traveling towards' in Old English to acting as a future auxiliary today, passed through the stages it did. *Be going to* took on new traits one after the other, rather than simultaneously increasing the frequency of all those traits: first it began to appear with verb-phrase complements (rather than exclusively location complements) but with an intentional meaning and with motion plausibly involved (as in modern *we're going to fight*), then with no motion (*he's going to say something*), then

with lack of intention (*he's going to be fired*), then with non-sentient subjects (*the chair is going to break*), and finally with dummy subjects (*it's going to rain*). In Tabor's connectionist model (see 3.1.2 above), categories are represented in a continuous space, so if an item such as *be going to* moves from one category to another (by changing its input-to-hidden weights), it may pass through other categories on the way. The connection weights in the network determine what trajectories through this space are possible, and Tabor shows how a plausible model of verb categories would cause *be going to* to pass through the category of equi verbs (which take verb-phrase complements but require sentient subjects) on its way from motion verb to raising verb.

6.   Probabilistic aspects of reanalysis

Reanalysis has been argued to result from frequency shifts that cause a structure to become associated with a context that admits or encourages the new analysis (e.g., Lightfoot 1991). Others have argued, based on case studies, that reanalysis precedes the structural change (Frisch 1994, Pintzuk 1995); in those cases, it is unclear what triggers the reanalysis in the first place.

Fontana (1993) explores in detail a case in which one syntactic change affects a construction's environment and thus causes another syntactic change. Object pronouns in Old Spanish appeared in second position (enclitic on the first word of the clause), but in Modern Castilian Spanish they are verbal proclitics. Loss of topicalization in Old Spanish led to an increase in clauses beginning with the object pronoun; assuming that object pronouns were not able to stand on their own phonologically, these sentence-initial pronouns would have been procliticized to the following word. Another change, loss of I-

to-C movement, meant that the following word was often a verb. As the pronouns appeared as proclitics on verbs more and more often, they were reanalyzed to be permitted in that position only, whether sentence-initial or not.

Although the details of the analysis depend on issues in syntactic theory, if the basic outline is right, the Spanish situation can be seen as one in which an increasing number of forms become ambiguous, being consistent with either the existing grammar or an innovative one. The learning algorithm should then predict under what conditions, if any, the innovative grammar can take over (see §5 above).

7. Deriving S-shaped change

It has long been observed that many historical developments follow an S-shaped pattern of change (see Denison 2002 for discussion). At first, a new variant is rarely used, and its use increases slowly; in the middle period of the change, the new variant gains ground more and more quickly until the two variants' frequencies are about equal, after which point the change slows down; in the late period of the change, the older variant disappears more and more slowly and may even persist, at very low frequencies, indefinitely. An S-curve is illustrated in Figure 1.

<Figure 1 near here>

7.1 Contacts in the population

Many explanations involving probabilistic interactions among individuals in the speech community have been proposed for this pattern of change. One possibility, mentioned by Labov (1994, p. 66), is that the S shape of change simply reflects the time-course of

social contacts. If we assume, for simplicity's sake, that each individual converts to the new variant on first encountering it, a change that originates in one member of a population will still take time to spread. At the first time-step, only those who interact directly with the innovator "catch" the new variant, and its spread is slow. But as the number of speakers with the innovative variant increases, so does the number of speakers in contact with them at every time-step, so the spread accelerates, until more than half of the speech community uses the new variant, at which point the spread slows because there are fewer and fewer speakers left who do not already use the new variant.

To see this mathematically, suppose that there are n speakers in the population. Let x represent the number of speakers with the new variant at any time; $n - x$ is the number of speakers with the old variant. The number of possible pairwise interactions between speakers that would result in conversion of a speaker from the old variant to the new is simply the number of pairs composed of one new-variant speaker and one old-variant speaker, or $x(n - x)$. But, at a given time-step, not all those interactions will take place. Let p, a number between 0 and 1, represent, for each pair of speakers, the probability that they meet at any one time-step. At each time-step, then, the increase in speakers with the new variant is $px(n - x)$. When x is small this quantity is small—change is slow. For example, with $n = 100$ and $p = 0.01$, when $x = 1$, the number of speakers converted to the new grammar is $p(n\text{-}1) = 0.99$, just under one speaker. When x reaches $n/2 = 50$, the rate of change reaches its maximum, $pn^2/4 = 25$. When the change is almost complete, $x = 99$, the rate of change is again $p(n - 1) = 0.99$.

If we examine the proportion u (= x/n) of the population that has the innovation,

the rate of change at every timestep can be written $au(1-u)$, where a (=pn^2) plays a role

similar to p's in determining how frequent contact is. To obtain u as a function of time we

must solve the differential equation $\dfrac{du}{dt} = au(1-u)$. The solution is of the form $u =$

$\dfrac{1}{1+e^{-C-at}}$, where C is a constant. This is the logistic function from §4 above. Setting C so

that the population starts with 1% of members having the new variant, we obtain the plot

for $a = 0.5$ shown in Figure 2.

<Figure 2 near here>

An S-shaped pattern of spread is still derived if speakers' adoption of the new

variant is probabilistic (e.g., each contact with the new variant only increments a

speaker's use of it), or if the speech community is organized into a social network, with

contacts between speakers being governed by social bonds instead of evenly distributed.

As discussed above in 5, many authors have constructed models in which the

tendency of one linguistic variant to spread emerges from learning, rather than simply

being imposed, and there have been simulation studies (see Niyogi 2004 and references

therein) investigating under what learning conditions the new variant spreads according

to an S shape.

7.2 *A caution about rates of change*

Some authors have cautioned that, because of stylistic conservatism, change may appear, in the written record, to be slower than it really was. Shi (1989) argues that although the rise of the aspectual particle *le* in Mandarin Chinese appears, from the written record, to take about 1000 years, it was actually much more abrupt. Shi argues is that even after the rise of *le* was completed in the spoken language, the written language contained a mix of contemporary and classical styles even within documents, so that the rate of *le* use appears artificially low. To obtain the vernacular rate of *le* use from the written record requires factoring out the rate of classical-style use. To do this, Shi tracks the frequency of the classical copula/interjective *ye*. In true classical texts, there are 8 occurrences of *ye* per 1000 characters, so Shi infers that if a text has $n \leq 8$ occurrences of *ye* per 1000 characters, the probability that a character in that text should be attributed to the classical style is $n/8$. When the number of *le*s is plotted per 1000 putatively non-classical characters instead of per 1000 raw characters, the change is completed in at most 200 years. When change is so abrupt, it becomes difficult to determine whether the S shape is a good fit.

8. Model dynamics

Niyogi (2004) demonstrates some unexpected properties of population-level models of language change that would be difficult to intuit. Let us consider one example that uses a cue-based learning algorithm. Cue-based learning (Dresher & Kaye 1990) relies on the idea that certain sentences generated by a grammar could not come from any other grammar, and thus serve as unambiguous *cues* to the learner as to which grammar it

should choose. (On a smaller scale, a cue may also be partial, demonstrating unambiguously the setting of some parameter but being irrelevant to other aspects of the grammar.) For simplicity, assume that the learner faces a binary choice between two grammars, L_1 and L_2. Cue-based learning typically establishes one choice as the default (say L_2) by specifying that unless some minimum rate of L_1 cues is encountered, the learner will choose L_2. Varying the probability p that a speaker of L_1 produces a cue, we can see how the proportion of the population that eventually acquires L_1 (given infinite time) is affected. Niyogi finds that for small values of p, the population moves towards total L_2 use. In dynamical-systems terms, 0% L_1 use is the only stable fixed point for these low values of p—all other states will eventually move to 0% L_1 and stay there. If we examine increasing values of p, at some critical point (determined by the number of utterances in the critical period and the rate of cues that the learner requires before choosing L_1) there emerges a second stable fixed point of 100% L_1 use, as well as an unstable fixed point somewhere between 0% and 100%. That is, at higher p, a population that has little L_1 use will still move to 0% L_1 (since there are few utterances produced from L_1), but a population with more than some minimum number of L_1 speakers will now move towards 100% L_1 use and stay there; additionally, there is some intermediate proportion of L_1 use that can persist, but if the population diverges slightly from that value, it will be pulled towards 0% or 100% L_1 use. This is an example of a *bifurcation*: a small change in a parameter (here, p) results in a sudden change in the behavior of the system. Niyogi suggests that the potential for bifurcations helps address the actuation problem of Weinreich et al. 1968 (i.e., why does a language change start?): a small

change in the system, perhaps brought about by external forces such as population movements or cultural changes that change the frequency with which certain types of propositions are expressed, could lead to dramatic linguistic change. The calculations that Niyogi performs to arrive at the bifurcation result demonstrate the importance of making the model fully explicit—otherwise, its behavior cannot be seriously studied.

9. Further reading

Andersen, H. (1973). 'Abductive and deductive change' *Language* 49, 765-793.

Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

Briscoe, Edward J. (2000). 'Evolutionary perspectives on diachronic syntax' In Pintzuk, S., Tsoulas, G. & and Warner, A. (eds.) *Diachronic syntax: Models and mechanisms*. Oxford: Oxford University Press. 75-108.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Clark, B. (2004). *A Stochastic Optimality Theory approach to syntactic change*. Stanford University dissertation.

Denison, D. (2002). 'Log(ist)ic and simplistic S-curves' In Hickey, R. (ed.) *Motives for language change*. Cambridge: Cambridge University Press.

Dresher, E. & Kaye, J. (1990). 'A computational learning model for metrical phonology' *Cognition* 34, 137-195.

Fontana, J. M. (1993). *Phrase structure and the syntax of clitics in the history of Spanish*. University of Pennsylvania dissertation.

Frisch, S. (1994). 'Reanalysis precedes syntactic change: Evidence from Middle English' *Studies in the Linguistic Sciences* 24, 187-201.

Gibson, E. & Wexler, K. (1994). 'Triggers' *Linguistic Inquiry* 25, 407-454.

Hinskens, F., van Hout, R. &Wetzels, W.L. (eds.) (1997). *Variation, change and phonological theory*. Amsterdam: Benjamins.

Jäger, G. (2003). 'Simulating language change with Functional OT' In Kirby, S. (ed.) *Proceedings of language evolution and computation workshop/course at ESSLLI*. Vienna. 52-61.

Johansson, C. (1997). *A view from language: Growth of language in individuals and populations*. Lund: Lund University Press.

Kroch, A. (1989). 'Reflexes of grammar in patterns of language change' *Language Variation and Change* 1: 199-244.

Kroch, A. (2001). 'Syntactic change' In Baltin, M. & Collins, C. (eds.) *The Handbook of Contemporary Syntactic Theory*. Malden, MA: Blackwell. 699-729.

Labov, W. (1994). *Principles of linguistic change, Vol. 1: Internal factors*. Cambridge, MA: Blackwell.

Labov, W. (2001). *Principles of linguistic change, Vol. 2: Social factors*. Malden, MA: Blackwell.

Lightfoot, D. (1991). *How to set parameters: Evidence from language change*. Cambridge, MA: MIT Press.

Milroy, L. (1987). *Language and social networks* (2nd ed.). Oxford: Blackwell.

Niyogi, P. (forthcoming). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Niyogi, P. & Berwick, R. (1995). 'The logical problem of language change.' MIT Artificial Intelligence Laboratory Memo No. 1516.

Paolillo, J. C. (2001). *Analyzing linguistic variation: Statistical models and methods*. Chicago: University of Chicago Press.

Pierrehumbert, J. (2002). 'Word-specific phonetics' In Gussenhoven, C & Warner, N. (eds.) *Laboratory phonology VII*. Berlin: Mouton de Gruyter. 101-140.

Pintzuk, S. (1995). 'Variation and change in Old English clause structure' *Language Variation and Change* 7, 229-260.

Shi, Z. (1989). 'The grammaticalization of the particle *le* in Mandarin Chinese' *Language Variation and Change* 1, 99-114.

Tabor, W. (1994). *Syntactic innovation: a connectionist model*. Stanford University dissertation.

Weinreich, U., William L., and Marvin H. (1968). 'Empirical foundations for a theory of language change.' In Lehmann, W. & Malkiel, Y. (eds.), *Directions for historical linguistics*. Austin: University of Texas Press. 97-195.

Yang, Charles (2000). 'Internal and external forces in language change' *Language Variation and Change* 12, 231-250.

Figure captions

Figure 1: An S-shaped curve

Figure 2: Logistic function