Probability in Language Change[1]

Kie Zuraw

0. Introduction

Why do languages change? If children are able to infer surrounding adults' grammatical systems from their utterances, and if adults adjust their lexicons and perhaps grammars to achieve better communication with their interlocutors, any linguistic innovations that might somehow arise should be quickly stamped out. This is a combination of Weinreich et al.'s (1968) "actuation problem" (how and why does a particular change occur at a particular time) and what we might call the "continuation problem": what sustains the momentum of a change, causing an innovation to increase in frequency, to spread from word to word, or to spread from speaker to speaker, rather than stalling or receding?

Any answer to the continuation problem that has been proposed must rely on a probabilistic model of the language faculty. If the rise in frequency of an innovation results from snowballing mislearnings, we require a model of how the learner responds to her or his variable environment. If the rise in frequency results from individuals' adopting the speech patterns of some social group, we require a probabilistic model of the speech community, in which individuals probabilistically and incrementally update their grammars and lexicons in

response to interlocutors' behavior. Moreover, when the rise in frequency of an innovation involves variation within individuals, as we can often see that it does in written records, we require a probabilistic model of language representation and/or use. Otherwise, we have no way of representing difference between a generation whose members use a new variant 20% of the time and a generation whose members use a new variant 40% of the time.

The fact that language change happens seems to demand a probabilistic view of the language faculty, in phonology, morphology, syntax, semantics, processing, acquisition, and the social use of language. The probabilistically oriented study of language change therefore relies on probabilistic models of all the areas of linguistics discussed in this volume.

This chapter surveys the role of probability in the study of language change. The first section describes the use of probabilistic tools in establishing language relatedness through vocabulary comparison, an important task when historical and textual records are lacking, and inferences about language change must be drawn from the ways in which related languages differ. The second section examines how the frequencies of linguistic traits change over time in the historical record, and how the timecourse of a change can shed light on its motivation and on the continuation problem. The third section discusses the role that the frequencies of lexical items and constructions play in their susceptibility to change, and what this can tell us about the synchronic effects of frequency. The

fourth section asks how language change is directly molded by probabilistic behavior on the part of its participants—speakers, hearers, and learners.

1. Probability as a tool for investigating language relatedness

An important task in historical linguistics is establishing which linguistic changes are possible or probable (the "constraints problem" of Weinreich et al. 1968). In many cases, we can look to synchronic variation to tell us which changes are in progress in a particular language (see Labov 1994). In rare cases, we have written records of change within a language. But the vast majority of language changes that have taken place in human history have left no trace either in synchronic variation or in the written record. The only way to discover them is through comparison of related languages: if we can reconstruct a proto-phoneme **p*, for example, that became *b* in some context in a daughter language, then we know that the change from *p* to *b* in that context is a possible one; if we find many such cases, then we know that the change is a common one. Moreover, once we have established by reconstruction that a change took place, we can use synchronic evidence to answer questions such as how regular the change was, and which types of exceptions were allowed to persist.

But how are we to know if two languages are related in the first place, so that an attempt at reconstruction makes any sense? A common method for establishing the relatedness of languages is to compare their vocabularies. A list

of words is collected for each language, based on a standard list of 100 or 200 meanings (e.g., the lists proposed in Swadesh 1952, 1955) that are expected to have a name in every language. If the languages are related, we expect to find either similarities between the sounds of words with identical or similar meanings (e.g., if the word for meaning *i* begins with a labial consonant in Language *A*, then the word for meaning *i* begins with a labial consonant in Language *B* too) or consistent correspondences between them (e.g., wherever we see a *t* in Language *A*, there is a *k* in the corresponding word of Language *B*). Because reflexes of a proto-phoneme can differ considerably in daughter languages, consistent correspondences are a more appropriate criterion for languages that are not known to be closely related.[2] The more frequent and consistent the correspondences are, the more likely it is that the two languages are connected, whether through descent from a common ancestor, or perhaps through borrowing.[3]

The mathematical challenge in using this method is, how sure can we be that the similarities or correspondences found are not merely due to chance? It turns out that the degree of similarity or correspondence necessary to establish relatedness is greater than we might intuit. Ringe (1992) gives a detailed and highly accessible demonstration of this fact using randomly generated and real wordlists. Ringe's method is flawed, as discussed below, but he makes an important point: even though a particular event may be very unlikely to occur by

chance, it may be an instantiation of a larger class of events one or more of which is relatively likely to occur.

Suppose, for example, that we hypothesize that two languages are related, and our criterion for relatedness is similarity (rather than regular correspondence). If we find that the word for *eye* begins with *t* in both languages, that is a piece of evidence in favor of the hypothesis, but how striking is it? How likely is it to have occurred by chance if the two languages were not related? The probability that language *A* and language *B*'s words for *eye* should both begin with *t* by chance is equal to the proportion of words in language *A* that begin with *t* (A_t) times the proportion of words in language *B* that begin with *t* (B_t). If, in each language, only 5% of words begin with *t*, then $A_tB_t = 0.0025$, a low probability. But this is a misleading result: the hypothesis being tested is not that both language's words for *eye* begin with *t*, but that the language's vocabularies are similar. The probability we should be interested in is the probability that at least one word-pair on the list would begin with the same sound by chance—this probability will depend on the phoneme distributions in the two languages, but will be much, much higher than 0.0025. For example, if each language has the same 20 phonemes, each occurring word-initially 5 times in a list of 100 meanings, the chance of obtaining at least one match is nearly 100%.

Because this issue has caused so much confusion in the literature (see Manaster Ramer and Hitchcock 1996 for an attempt to sort out one exchange), it

is worth belaboring. Manaster Ramer and Hitchcock call the confusion of a specific event with the class to which it belongs the "birthday fallacy": the chance that two randomly chosen people share the birthday of February 1st is small (1 in 365^2 = 133,225), but the chance that they merely share the same birthday is much greater (1 in 365).[4] Choosing a specific date when calculating the probability of a shared birthday is analogous to requiring a correspondence to involve a particular sound or pair of sounds, or occur in a particular word.

The same issue arises when we seek correspondences across multiple languages, as suggested by Greenberg and colleagues (Greenberg 1987, Greenberg and Ruhlen 1992). A correspondence seen in any two languages out of a group of, say, 15 languages is not as significant as a correspondence seen in a comparison between just two languages, because there are $\binom{15}{2}$ = 105 pairs in which such a correspondence could have occurred, rather than just one.[5] As Baxter and Manaster Ramer (1996) argue, it should be possible to determine the number of matches across a set of *n* languages that would be as significant a match in a two-language comparison, but the determination becomes much more complicated when, for example, each language participating in the correspondence is required to be from a different family (Manaster Ramer and Hitchcock 1996).

Considering just the simpler case of a comparison between two languages, what we would like to do is determine how different a contingency table, as the in (1) (from Ringe), is from what would be expected by chance. The contingency table represents how often each word-initial consonant (or Ø for vowel-initial words) in a list of 100 English words corresponds to each word-initial consonant in the German word with the same meaning. For example, there are 4 words that begin with *w* in English and with *v* in German. We could construct similar tables for any other type of correspondence we were interested in, such as medial consonants, or consonant-vowel sequences.

(1) (table is at end of manuscript)

If the two languages were not related, we would expect to see, on average, the values in (2), which has preserved the row and column totals of (1) but eliminated any row-column interaction. In a given case, the numbers will differ from those in (2) (minimally, they must be integers), so our question is, how unusual is it for a chance-generated table to deviate from (2) as strongly as (1) does?

(2) (table is at end of manuscript)

Ringe proposes calculating the probability, for each cell, that the number of observed matches or more would be seen by chance, by summing binomials. That is, he proposes that the probability of finding exactly one match of *x* in Language *A* with *y* in Language *B* in a list of 100 words is the product of three numbers: A_xB_y (the probability that a particular pair show the correspondence), $(1-A_xB_y)^{99}$ (the probability that the 99 other pairs do not), and 100 (the number of places in the list where a matching pair could be found). Similarly, the probability of finding exactly two such pairs would be $A_xB_y^2 \cdot (1-A_xB_y)^{98} \cdot 9900$. To find the probability of finding *n* or more matches, we would sum the probabilities of finding *n* through 100 matches:

(3) $$\sum_{i=n}^{100} (A_x B_y)^i \cdot (1 - A_x B_y)^{100-i} \cdot \binom{100}{i}$$

For $A_x = B_y = 0.05$ and $n = 3$, this sum is 0.20—in other words, at least one correspondence between *x* and *y* is a fairly likely event.

The problem with Ringe's method, as pointed out by Baxter and Manaster Ramer (1996), is that it wrongly assumes that the probability of seeing a correspondence in one word-pair is independent of whether the same correspondence occurs in another pair. A_x and B_y are based on frequencies within the chosen word list. Suppose that $A_t = B_t = 0.05$: there are 5 instances of initial *t* in each language's word-list. If the words for *eye* begin with *t* in both languages, then the chance that the words for *cheek* will also both begin with *t* is lowered,

because there is one fewer *t* left in the pool from which *cheek* can draw its initial consonant. The probability that *cheek* would begin with *t* in both languages is now not $\frac{5}{100} \times \frac{5}{100} = 0.0025$, but rather $\frac{4}{99} \times \frac{4}{99} = 0.0016$. The values to be summed are not binomials as shown in (3), but hypergeometrics, which are unwieldy for numbers as high as 100.

How, then, can we accurately determine whether a table like (1) is significantly different from the average table we would expect to see if the two languages were not at all related? Statistics such as χ^2 give a measure of how different an observed contingency table is from the expected table,[6] but in order to determine the significance of that difference—how likely it would be to arise by chance—we must rely on lookup tables that are inappropriate to the task. Lookup tables for the distribution of χ^2, for example, assume that the data points are independent, and that expected cell values are relatively high (expected frequencies of at least 5 in each cell)—much too high for a table with dozens of cells and only 100 instances to go around.

Kessler (2001) proposes an ingenious solution to the inapplicability of standard lookup tables, similar in spirit to Oswalt's (1970) shift test,[7] but much more robust. We want to know the distribution of values of χ^2, or some other measure of skewedness, if languages *A* and *B* are not related, so that we can see how unusual the observed value of χ^2 is. If *A* and *B* are not at all related, and if we

have excluded from the word-list words subject to sound-symbolism and onomatopoeia, then any lineup of *A*'s words with *B*'s should be equally likely—the particular lineup that occurs in reality is the result of mere chance. Thus, the universe of possible arrangements that should occur by chance, while preserving the individual phoneme distribution of each language, is well represented by keeping the order of *A*'s list constant and permuting *B*'s list in all possible ways. If we calculate χ^2 for each such permutation, we obtain the distribution of χ^2. We can compare the value of χ^2 obtained for the actual word-list to this distribution: if it is larger than 99% of the χ^2 values in the distribution, then we know that a contingency table as skewed as the one we obtained will occur only 1% of the time if the two languages are unrelated.

In practice, however, we cannot consider all permutations of *B*'s list. For a list of 100 words, there are astronomically many permutations: $100! \approx 9.3 \times 10^{157}$. This is too many to consider, even for a computer. Kessler's solution is to instead randomly generate some large number of permutations to get a close estimate of how often the resulting χ^2 values are greater or smaller than the observed one. The more permutations sampled, the more accurate the count; Kessler uses 10,000 permutations. The same method can be used for any other measure: Kessler considers R^2, the sum of the square of each cell entry (minus one if nonzero); various breakdowns by phonetic feature; and matching phoneme sequences rather than individual phonemes. For Kessler's example data, R^2 seems to work best.

The problem of determining whether a language resemblance is stronger than would be expected by chance is a tractable one, then, at least in simple cases such as correspondences between phonemes. As all the authors cited here agree, however, establishing relatedness is only a starting point. These statistical methods do not replace the work of establishing which words are cognates, determining the contextual determinants of sound changes that lead to inexact correspondences, or reconstructing proto-forms. They do, however, give us a tool with which to determine how striking an apparently striking connection really is, so that we can decide whether an attempt at reconstruction is warranted.

2. Changes in probabilities over time

Language change appears to take place gradually, with innovations being used at different rates in different parts of the speech community and in different linguistic or social contexts, and with an innovation's overall rate of use rising gradually, often over centuries (though see discussion of Shi 1989 below). Changes in observed probabilities in the historical record can give evidence for the nature of the linguistic system underlying variable linguistic behavior, the nature and proximal cause of a particular change, and the way in which changes take hold and spread.

2.1. Correlations in rate of change

Suppose that a language is observed to undergo a gradual change from an SOV (subject-object-verb) word order to an SVO order, that in texts from intermediate stages the innovative order is found more frequently in main clauses than in subordinate clauses, and that in the intermediate stages, variation is observed even within each individual writer. How should the linguistic system of an individual living during the middle stages be represented? If it is a grammar that encodes separately the probabilities of employing SOV or SVO in various contexts, then the innovative word order may spread at quite unrelated rates in main and subordinate clauses. If, however, the difference between SOV and SVO is controlled by a single parameter in the grammar—whose setting can be probabilistic to allow variation—and it is some orthogonal force (stylistic, perhaps) that prefers SOV in subordinate clauses, then although the frequency of use of the innovative order may differ according to clausal context, the rate of change of those contextual frequencies should all be the same, assuming that orthogonal forces remain constant. This is the Constant Rate Hypothesis, proposed by Kroch (1989): because changes occur at the level of abstract grammatical parameters, they spread at the same rate in every context, although the base frequencies of use in each context may differ for external reasons.

Kroch and colleagues have tested the Constant Rate Hypothesis by modeling S-shaped language changes with a logistic function. It has long been

observed (e.g., Osgood and Sebeok 1954, Weinreich, Labov, and Herzog 1968, Bailey 1973) that language change takes an S-shaped course: a new variant appears rarely for a long time, then quickly increases in frequency; finally, the rate of change slows as the frequency approaches its maximum (100% in the case of a total replacement of the earlier form). There are several mathematical functions that produce an S-like shape. Kroch chooses the logistic function because every logistic function has associated with it a slope, and therefore the slopes of frequency changes that should be linked, according to the Constant Rate Hypothesis, can be compared.

The logistic function takes the form in (4),[8] where *P*, interpreted here as the probability of seeing some variant in some context that it could potentially occupy, is a function of *t*, time.

(4) $$P = \frac{1}{1 + e^{-k-st}}$$

Simple algebra transforms (4) into (5), where now the logistic transform, or logit, $\ln \frac{P}{1-P}$, is a linear function of *t*, with a slope (steepness) *s* and an intercept (initial value) *k*.[9]

(5) $$\ln \frac{P}{1-P} = k + st$$

When frequency changes over time are plotted for the same innovation in different contexts, the logit for each context should have approximately the same slope under the Constant Rate Hypothesis, although they may have different intercepts. In (6) are illustrated two logistic functions whose logits have the same slope, but different intercepts, and one that has a different slope.

(6) (figure is at end of manuscript)

The Constant Rate Hypothesis can also be tested using the multivariate analysis performed by the VARBRUL program (see Mendoza-Denton et al., this volume). VARBRUL represents the logit as the sum of some contextual weights, representing the positive and negative effects of various features of the context, plus a base rate of use, in this case a linear function *f* of time:

(7) $$\ln \frac{P}{1-P} = f(t) + a_1 + a_2 + a_3 + \ldots$$ (Kroch 1989, p. 6)

If the values of the a_j do not change as *t* changes, then the contribution of the context is constant over time, and only the base rate of use of the innovative variant changes.

Kroch and colleagues have found evidence for the Constant Rate Hypothesis in several cases of language change. Kroch (1989) illustrates how the results of Noble (1985), Oliveira e Silva (1982), and Fontaine (1985) support the

Constant Rate Hypothesis in the replacement of possessive *have* by *have got* in British English, the rise of the definite article in Portuguese possessive noun phrases, and the loss of verb-second in French, respectively. Kroch (1989) also reanalyzes Ellegård's data of the rise of periphrastic *do* in English. Pintzuk (1995) finds that Old English I′-initial Infl roes in frequency at the same rate in main and subordinate clauses, and Santorini (1993) finds that the rise of a similar I′-initial Infl phenomenon in early Yiddish proceeded at the same rate with both simple and complex verbs, although in this case the intercepts of the logits are also similar, so we cannot be sure that the breakdown is into two truly different contexts.

These findings suggest that syntactic and morphosyntactic changes do indeed occur at some abstract level of the grammar, affecting all contexts equally, and subject only to independent influences on various contexts. Tabor (1994), using a very different model of grammar, essentially agrees, but views constant rate effects as a special case of frequency linkage effects—related changes proceeding at related, though not necessarily identical, rates.

Tabor's model of (morphosyntactic) language change uses a connectionist network to learn associations between words and the contexts in which they tend to occur and among words that tend to occur in similar contexts. Words, represented by input nodes, are connected to intermediate hidden-layer nodes; words that are strongly associated to the same hidden-layer nodes act as clusters

somewhat like traditional grammatical categories (e.g., Noun, Verb), although cluster membership is a gradient property, so that a single word may belong to different clusters to different degrees. Hidden-layer nodes are connected to each other, to represent sequential information, and to output nodes, representing behaviors in various syntactic constructions. How strongly a word is associated to some syntactic behavior is therefore mediated by the hidden units, and thereby by the behavior of cluster-mates.

If a network that has been trained on a corpus is exposed to an altered version of the original training data (representing an externally motivated shift in frequency), it adjusts its connection weights in response, but not only those aspects of the language that changed in the training data will be affected: aspects of the language that were strongly linked to the changed aspects will be affected also. In particular, if the frequency with which some word occurs in some context changes in the training data, the network will adjust the word's association strengths with the hidden units in response, thereby altering the word's indirect association to other words; as a consequence, other aspects of the word's behavior will also change, under the influence of the word's new cluster-mates.[10] At the same time, the network must adjust associations between the word's strongly associated hidden units and the output units, so that the behavior of other words that were strongly associated to the same hidden units will change too. This is where frequency linkage effects come from in Tabor's model: a change in one

aspect of the language drags along with it changes in other aspects of the language.

Frequency linkage of the constant-rate variety will be observed when two words or constructions have the same distribution (or nearly so) before the change begins. Tabor demonstrates with an abstract example: two nouns, N1 and N2, behave similarly along five binary contextual dimensions, C1 through C5 (e.g., if C1 is what possessive verb the nouns appear as the object of, they might appear as the object of *have* 96% of the time and as the object of *have got* 4% of the time). A third noun, N3, behaves like N1 and N2 along dimension C1, but shows different frequencies on the other four dimensions. A network is trained on a corpus with these properties, then re-trained on a corpus that is the same except that N2 undergoes a frequency change in C1, from choosing option A 4% of the time to choosing it 100% of the time; no examples of N1 and N3 are given for C1 in the new corpus. The point of the experiment is to observe how the frequencies N1 and N3's choosing option A in C1 change as the network approaches the desired frequency for N2 choosing option A in C1. The slope of the logit for how often N1 exhibits option A in C1 is almost identical to the slope of the logit for N2, but N3's slope is much shallower. Because N3 does not share N2's properties as well as N1 does, N3 is not "dragged along" as much as N1 is. Thus, for Tabor, constancy of rate is gradient; he would predict that SVO would spread at the same

rate in main and subordinate clauses to the extent that the behavior of main and subordinate clauses is otherwise similar.

It remains to be seen whether any convincing cases of demonstrably partial frequency linkage exist. Tabor argues that the rise of English periphrastic *do* is such a case, but Kroch (1989) proposes that certain syntactic assumptions can explain why the slopes for some of the contexts for *do* are unequal. If clear cases can be found, then we have evidence that language change is indeed abstract, occurring at the level of structures and categories, but that structures and categories can be fuzzy, and membership in them gradient.

2.2. Reanalysis and frequency change

Besides bearing on the abstractness of language change, rates of use over time can also shed light on the relationship between reanalysis and frequency. It seems clear in many cases of morphological, syntactic, and semantic change that some word or construction has been reanalyzed—that is, its behavior has changed in a radical way, indicating that it has joined a different grammatical category. For example, *be going to*, which once obligatorily indicated motion towards, is now used as an all-purpose future marker in English.

How and why does reanalysis occur? In the (otherwise very different) models of Lightfoot (1991) and Tabor (1994), reanalysis results from frequency shifts that encourage or even force learners to assign a new structural analysis to a

form because of the contexts in which it appears. Others argue that reanalysis is a prerequisite for syntactic change: only after two structural options become available can one rise in frequency. Santorini (1993) and Pintzuk (1995), for example, argue that in Yiddish and English respectively, the availability of an I′-initial position for Infl in both main and subordinate clauses occurs at the beginning of the rise in frequency of medial Infl, not at the end (the alternative analysis is that in the early stages, clause-medial Infl results from movement, and not until later is Infl reanalyzed as potentially I′-initial). In these two cases, the argument for the availability of I′-initial Infl at the early stages is mainly a syntactic one, but it also has a probabilistic element. Surface non-final Infl could be the result of base-generated I′-initial Infl or base-generated I′-final Infl, with rightward movement of other constituents. Santorini and Pintzuk both argue that the rate of such rightward movements observed in unambiguous contexts is too low to account for the relatively high rate of non-final Infl. Therefore, I′-initial Infl must have been used at least some of the time at the early stages of the change, before it became very frequent.

Frisch (1994) presents another case in which evidence that reanalysis precedes syntactic change comes from frequencies over time. In Middle English, *not* acted like a sentence-level adverb: it could appear preverbally or postverbally, much like modern *never*; it carried an emphatic meaning; and it alone was not

sufficient to indicate negation (*ne* was instead the usual marker of non-emphatic negation).

(8) Þat Jesuss nohht ne wollde Ben boren nowwhar i Þe land, ...
that Jesus not *neg* would be born nowhere in the land, ...
'That Jesus did not (at all) want to be born anywhere in the land, ..."
(Frisch 1994 p. 189, Ormulum I: 122)

It is standardly proposed (Kroch 1989, Pollock 1989, Shanklin 1990, Roberts 1993) that *not* was reanalyzed as a sentential negator, losing its preverbal position and emphatic meaning, because the phonological loss of the clitic *ne* eventually forced *not* to be so interpreted. Frisch demonstrates, however, that the loss of preverbal *not* was well underway before the loss of *ne* began.[11] Frisch argues, therefore, that a semantic reanalysis of *not* as non-emphatic, allowing it to occupy the Specifier position of NegP rather than a sentence-level adverb position, caused *ne* to become redundant and be lost. With *ne* gone, *not* was free to occupy either the Spec or the head of NegP.

An important assumption is that the rate of adverbial use of *not* in ambiguous cases can be extrapolated from the behavior of the unambiguous sentence adverb *never*. *Never* is preverbal 16% of the time, and during the first 70

years of Middle English, *not* has the same distribution. Frisch presents the following formula:

(9) number of preverbal *not* = 0.16 × total number of adverbial *not*

total number of adverbial *not* = number of preverbal *not* / 0.16

Assuming that the rate at which true sentence-level adverbs appear preverbally is constant at 16% throughout the period, Frisch obtains an estimate of how often *not* is used adverbially from 1150 to 1500. The key finding is that this percentage falls drastically before the percentage of negative sentences containing *ne* begins to drop much at all.

We have, then, cases in which that reanalysis appears to occur at the beginning of a frequency shift, rather than at the end. Does this contradict Tabor's claim that frequency shift leads to gradient reanalysis, in which a word begins to belong more and more to a different cluster, gradually taking on properties of that cluster? Perhaps not: In Tabor's model, reanalysis and frequency shifts can be gradual and mutually reinforcing. If Tabor's model were extended to include semantics, an increasing use of *not* in non-emphatic contexts (a typical case of semantic bleaching) could cause *not* to be gradually reanalyzed as a non-emphatic negator. The more strongly it was so re-categorized, the less often it would appear preverbally. Reanalysis would thus follow one frequency shift, and precipitate

another: frequency shifts affect probabilistic learners and in turn are affected by probabilistic speakers.

2.3. The timecourse of language change

As mentioned above, it has long been observed that language change proceeds along an S-shaped curve. Why should change begin and end slowly? If changes spread from speaker to speaker, the rate of spreading depends on the number of interactions between a speaker who has the new variant and one who has the old variant. There will be few such exchanges at first, because there are few speakers who have the new variant, and few such exchanges at the end, because there are few remaining speakers to whom the change has not yet spread (Bloomfield 1933). When there is variation within individuals, as there is in nearly all studies of historical texts, the picture is more complicated, because there are no speakers with 100% use of the new variant at first. We must assume that speakers can slightly increment their use of a variant, and that some force (such as group identification or learnability) encourages the change to continue in one direction. The remainder of this section discusses some attempts to derive S-shaped language change mathematically, with limited success.

But first, is a cautionary note, based on Shi's (1989) findings: Shi argues that a gradual, S-shaped change that appears to have taken place over 1000 years is actually an abrupt change that was completed in at most 200 years. The illusion

of gradualness comes from the persistence of classical style in modern texts. Shi tracks the rise of the aspectual particle *le* in Mandarin, which derives from the classical verb *liao* 'finish'. When the number of uses of *le* per 1000 characters is tracked for a corpus from the pre-10th to 20th centuries, the rate of use rises slowly from the 10th to 12th century, then quickly until the 17th century, and continues to rise slowly (though unevenly) to the present.

Shi finds, however, that *le* seems to be inhibited by classical verbs, and hypothesizes that avoidance of *le* in more recent writers is merely an attempt to emulate classical style. Shi uses occurrences of the sentence-final copula or interjective *ye* as an index of classicalness. Classical texts have approximately 8 occurrences of *ye* per 1,000 characters, so if there are *n* occurrences of *ye* per 1,000 characters in a text, there are approximately *n*/8 classical characters per actual character in the text; the rest can be considered vernacular. When the number of *le*s per 1,000 vernacular characters is plotted, the picture is very different from when raw character count was used: there is how a sharp rise in use of *le* from the 10th to the 12th century, and the rate of *le* use has not risen since. Shi's study points out an important potentially distorting effect of the unavoidable use of written records: even when a change is abrupt, the conservatism of written styles may cause it to appear gradual.

Assuming, however, that the S-shaped model is accurate (though it may appear artificially stretched in the written record), are there any models that can

derive it? Manning (this volume), points out that that stochastic Optimality Theory (Boersma 1998, Boersma & Hayes 2001) predicts S-shaped change if one constraint rises or falls at a constant rate through the grammar. In stochastic OT, surface forms are chosen according to their satisfaction of constraints whose rankings are normally distributed. Change is therefore slow when constraints' distributions overlap only at the outer edges, accelerates as the centers of the bell curves begin to overlap, and slows as the distributions again overlap only at the edges. The mechanism by which the grammar is transmitted from generation to generation in such a way that a change is in ranking persistent and linear is not known, however. Below are reviewed some attempts at achieving S-shaped change through modeling transmission of the grammar from adults to children over time.

Niyogi and Berwick (1995) present an abstract simulation of language change that does derive a logistic function for change, among other possibilities. In Niyogi and Berwick's model, different members of the population use different grammars, and learners must decide which grammar to adopt. (Admittedly, this is an unrealistic assumption, as it predicts no variation within individuals.) Each grammar is a series of *n* binary parameters, and the distribution of sentences produced by each grammar is uniform (all well-formed sentences are equally likely). Learners set parameters on the basis of examples, permanently and without tracking probabilities. Because the learner has limited opportunity to

adjust its grammar, mislearning is likely, especially if many utterances are ambiguous, making even homogeneous populations potentially unstable.

Learning proceeds as follows in Niyogi and Berwick's model: the learner draws at random two utterances by members of the surrounding population. If the second trigger utterance unambiguously supports one parameter setting, the learner chooses that setting. If only the first trigger is unambiguous, the learner chooses that setting. And if both triggers are ambiguous, the learner makes an unbiased choice at random. In other words, the critical period is just two utterances, and if they conflict, the more recent utterance prevails.

Niyogi and Berwick investigate by simulation the case of three parameters governing constituent order (yielding eight possible grammars) and find that the distribution of grammars sometimes changes according to a logistic function (S-shaped curve) that varies in steepness. But with some starting distributions and maturation times, the function is not logistic: rapid change can occur right away (the initial tail of the S is cut off), or the function may fall off towards the end rather than continuing to approach an asymptote.

Niyogi and Berwick apply their model to the change from Old French V2 to Modern French SVO, using the five binary parameters suggested by Clark and Roberts (1993) to yield 32 possible grammars. If learning time is limited, so that the younger generation does not have full opportunity to acquire the older generation's grammar, then in simulations even a population that begins

homogeneously V2 shifts away from V2, though the change is slow and does not proceed very far. But when even small numbers of SVO speakers were included in the initial population (perhaps representing foreign speakers), there was relatively rapid loss of V2.[12]

Niyogi and Berwick's model is deterministic if the population of agents is infinite (and generations do not overlap). Briscoe (2000) extends the investigation to cases in which the population is finite and small, and finds that the results are quite different. For example, if two competing grammars are initially equally distributed and produce equal proportions of ambiguous sentences, in the infinite-population model the two grammars should remain in balance: half the learners will adopt one, and half will adopt the other. In a finite population, however, the probability that exactly half the learners will adopt one grammar on any given trial is low (just as the probability is low that exactly half of a finite number of coin tosses will come up heads). Therefore, one grammar will probably gain ground over the other. As one grammar becomes much more common than the other, however, it becomes less and less likely that it can maintain its advantage. At the extreme, if a grammar is used by 100% of the population, as long as there are some unambiguous sentences, some learners will learn the other grammar. Even a moderate bias such as 75%-25% is untenable if there are a high proportion of ambiguous sentences: if 75% of the population uses grammar *A*, with 50% of sentences from each grammar being ambiguous, and there are 100 learners, the

probability that 75 or more of the learners will adopt *A* is only 0.07. Grammar *A* begins to lose ground, then, falling towards 50%, which we have already seen is itself an unstable state. The proportions of the two grammars will therefore oscillate endlessly.

Clearly, a realistic and complete model of how changes spread remains to be implemented.

3. The role of frequency in language change

We have seen the importance of changes in frequency over time. The individual frequencies of linguistic items also appear to play an important role in language change. Words' frequencies affect their susceptibility to phonological, morphological, and morphosyntactic change. This fact reinforces the findings elsewhere in this volume that not lexical items are treated alike, and that the strength of lexical entries is gradient. These differing strength values are important in the lexical (word-to-word) spread of linguistic innovations.

3.1. Frequency and phonological erosion

Bybee (1994) proposes that a usage-based model of phonology[13] can account for two relationships between word frequency and phonological change: frequent lexical items are the first to adopt automatic, phonetic rules, and the last to abandon nonphonetic rules. By "phonetic rules" Bybee means rules, like

American English flapping, that involve minimal articulatory or acoustic change. Nonphonetic rules include morphologically conditioned rules, like stress in Spanish verbs or English noun-verb pairs, and lexical generalizations, like the English *sing-sang-sung* and *ring-rang-rung* patterns.

An important assumption for Bybee in explaining the effect of frequency on susceptibility to phonetic changes is that lexical representations do not include only the idiosyncratic aspects of a word. Redundancies and phonetic detail are included, so that different words may be reliably associated with slightly different patterns of articulatory timing, and other sub-phonemic properties.

Phonetic rules tend to spread gradually through the lexicon, affecting frequent words to a greater extent. For example, in Hooper (1976), Bybee found that medial schwa deletion was most advanced in frequent words like *ev̲ery* (it is nearly obligatory), and less advanced in less frequent words like *artill̲ery* (it is nearly forbidden). In Bybee's usage-based model, this is because lexical entries are updated by speakers and/or listeners every time they are used. If schwa deletion has some probability of applying every time a word is used, then there is a related probability that the word's lexical entry will be updated to reflect the change. Because there is no reverse rule of "schwa restoration," once the strength of the schwa in a lexical entry is reduced, it cannot later increase—it can only stay where it is or reduce further. The more often a word is used, the more chances it

has to drift irreversibly towards schwa deletion. Thus, highly frequent words are the innovators in phonetic change.

Pierrehumbert (2000; see also this volume), in developing an exemplar-theory based model of production, derives this finding quantitatively. In exemplar theory, categories are represented mentally as clouds of remembered tokens (projected onto a similarity map) that are typically densest in the middle. Highly similar tokens are grouped into a single exemplar, whose strength is augmented when tokens are added to the group (and, countervailingly, decays over time). An incoming stimulus is classified according to the number of exemplars from each category that are similar to it, with a weighting in favor of stronger exemplars. Categories are produced by choosing an exemplar at random, but with a preference for stronger exemplars, and with some amount of noise added, so that the actual production may differ slightly from the exemplar chosen. Pierrehumbert shows that when exemplars are chosen in this way and the resulting tokens added to memory, the exemplar cloud gradually becomes more diffuse, but its center does not shift.

When a persistent bias (motivated by some external force) is added, however, drift does occur. If there is a tendency for productions to be slightly hypoarticulated with respect to the exemplar chosen for production (i.e., the articulatory gesture is reduced in magnitude), the center of the exemplar cloud gradually shifts towards hypoarticulation. For example, if an exemplar is chosen

whose articulatory effort along some dimension is 0.9, it may be produced with 0.89 effort instead. The 0.89 token is then added as an exemplar, and if it is chosen in a later production, it may be pronounced with 0.88 effort, and so on.

The shift increases as the number of productions of the category increases. This means that if individual words have their own exemplar clouds, then words that are used more often shift more rapidly, as predicted by Bybee. Pierrehumbert further shows how an infrequent category that is subject to lenition (or any other persistent bias) is absorbed into a frequent category that is not subject to lenition.

Bybee argues that frequent words are more subject to phonetic rules for an additional reason: phonetic rules tend to be lenition rules, involving reduced articulatory gestures. Frequent words are more likely to be used in prosodically unemphasized positions, which are associated with less articulatory effort. This is because a frequent word is likely to be used more than once in a discourse, and subsequent occurrences of a word in a discourse tend to be less emphasized prosodically than the first occurrence (Fowler and Housum 1987). In addition, frequent words or constructions are more likely to become semantically bleached (see Bybee 2000, discussed below), and thus less likely to be the carrier of important discourse information that is subject to prosodic emphasis.

Frequent words' lexical entries are thus doubly subject to a phonetic rule when that rule is lenitive: not only does the word's more frequent use give it more opportunities to undergo the change, but the word's tendency to occur in

repetitive or semantically bleached contexts disproportionately subjects it to lenition.

3.2. Frequency and nonphonetic rules

Highly frequent words are conservative, however, when it comes to nonphonetic rules like the English irregular past tenses (Hooper 1976) or English noun-verb stress shifts (Phillips 1998, 2001):[14] when the language begins to lose or gain a rule, they are the last words to change.[15] There are two reasons for this. The first reason is the competition between irregulars (residual archaic forms) and regulars (the innovative form). This competition proceeds differently in different models of regulars and irregulars, but in every case an irregular requires a strong lexical entry in order to resist regularizing. Under the dual-mechanism model of Pinker and Prince (1994), for example, listed irregular words and regular morphological rules compete in the brain: irregular, listed *sang* competes with regular, synthesized *sing+ed*. If the lexical entry of an irregular word is not strong enough, it may not be accessed in time or with enough certainty to win the competition, and the regular pronunciation will win. In a model such as Albright and Hayes' (2000) that encodes both regular and irregular patterns in the grammar, the competition is between very specific irregular rules and more general regular rules; in a connectionist model it is between patterns in associative memory (Rumelhart and McClelland 1986, Daugherty and Seidenberg 1994).

Frequent words' lexical entries are strong from frequent use and reinforcement, and thus will tend to beat out synthesized, regular pronunciations, whether the pressure for those pronunciations comes from the grammar or from elsewhere in the lexicon. Infrequent words' lexical entries, on the other hand, may not be strong enough to win reliably.

The second, related reason for the retention of nonproductive rules in frequent words concerns transmission from one generation to the next. Infrequent irregulars may fail to be transmitted to the next generation—if a word is too infrequent, the child may never encounter it—and the younger generation will apply regular rules to the word. An abstract simulation performed by Kirby (in press) confirms that this mechanism can have the observed effect. Although the population in Kirby's simulation begins with no lexicon at all, as a lexicon begins to develop, it is only the most frequent words that are able to retain an irregular form; words that are too infrequent to be reliably transmitted fall under a regular compositional rule.

3.3. Frequency and the undertransmission of morphosyntax

The instability of infrequent irregulars is one type of "undertransmission." Richards (1997) is a study of a more drastic type of undertransmission in the morphosyntactic realm that leads to a change not just in particular lexical items, but in the whole grammatical system. Richards compares the word order and

verbal morphology of current speakers of Lardil, an Australian language, to data collected by Hale from speakers in the 1960s. Lardil is being replaced in everyday use by English, but Richards argues that the changes observed in Lardil are due not to the linguistic influence of English, but to the scarcity of Lardil data available to learners. (Richards' arguments rest on syntactic sensitivities of the changes that would not be expected if the language were merely adopting English morphosyntax.)

The morphological difference between "Old Lardil" and "New Lardil" that Richards discusses is the frequent absence of inflection on objects in New Lardil (the syntactic difference is the resulting rigidification of word order). Richards' explanation is that in Old Lardil, certain phonological rules could delete object suffixes. New Lardil learners exposed to these apparently unsuffixed forms, and not exposed to enough overtly suffixed forms to learn that suffix deletion is phonologically conditioned, might conclude that overt suffixes alternate freely with null suffixes.

In analyzing the behavior of particular nouns and pronouns, Richards found that the pronoun on which inflection was most often produced had a highly irregular paradigm in Old Lardil. The pronoun on which inflection was least often produced had a more regular paradigm. Richards suggests that although regular morphophonological rules have been lost in New Lardil because of insufficient evidence, individual lexical entries exhibiting idiosyncratic inflection have been

retained when frequent enough. Similarly, Richards finds that the regular morphophonological rules of verb augmentation have been lost, but that certain (presumably) irregular verbs forms of Old Lardil have been retained. High frequency, then, can allow a word to retain various idiosyncratic properties in the face of a more general language change.

3.4. Frequency and grammaticalization

Frequency may also have an effect on which words or morphemes will undergo morphosyntactic change. Grammaticalization, the process by which content morphemes or morpheme sequences become function elements, tends to be correlated with an increase in frequency (see Traugott and Heine 1991, Hopper and Traugott 1993, for overviews and many case studies of grammaticalization). Is this increase merely the result of grammaticalization, as the morpheme becomes needed in more contexts, or could it also be a cause?

Bybee (2000) argues that it can. Bybee traces the evolution of English *can* from a content word meaning ‘have mental ability/knowledge’ to a function word meaning ‘possibility exists’. Following Haiman (1994), Bybee views grammaticalization as a form of ritualization, whereby repetition of a frequent act (in this case, the uttering of a word or construction) bleaches the act of its significance, reduces its (phonological) form, and allows the act to become associated to a wider range of meanings. Bybee shows how *cunnan*, the ancestor

of *can*, which first took only noun-phrase objects, began to take as object infinitives of verbs relating to intellectual states and activities, communication, and skills. Bybee argues that because *cunnan* with a noun-phrase object was already common, additional mental verbs began to be added to "bolster the meaning," creating seemingly redundant expressions like *cunnan ongitan*, 'know how to understand.' This use of *cunnan* further weakened it semantically—presumably, a learner who encounters the phrase *cunnan ongitan* is likely to attribute all the meaning to *ongitan* and treat *cunnan* as merely grammatical.

The token frequency of *can* increased greatly from Old to Middle English, partly as *can* came to be used with a larger number of verbs, partly as some of the *can+VERB* combinations became more frequent. The increase in both type and token frequency, Bybee argues, further bleached *can* semantically, and its verbal objects expanded to include emotional states, non-mental states, verbs that take as object another person, verbs indicating an action (rather than merely a skill). A few instances of inanimate subjects also began to occur. Eventually, as the 'possibility' meaning became more common, the use of inanimate subjects increased. Thus, increasing frequency and semantic bleaching reinforce each other.

Bybee further notes (citing crosslinguistic findings in Bybee et al. 1991, 1994) that grammaticalized morphemes tend to be shorter and more phonologically fused with surrounding material, for reasons discussed above:

frequent morphemes (including grammatical morphemes) are more susceptible to erosive lenition rules, which can cause loss and overlap of gestures. Bybee proposes that the units of lexical storage are not only morphemes or words, but also highly frequent phrases or sequences. When grammatical morphemes enter into high-frequency sequences such as *going to*, those sequences too are subject to erosion (*gonna*). As these sequences gain their own lexical representations, they can also develop idiosyncratic meanings and syntactic functions.

Tabor's (1994) connectionist model, described above, similarly views frequency as a driver of syntactic change. Tabor focuses not on the overall type or token frequency of a lexical item, but on the frequency with which it occurs in a particular context. Tabor performs a series of experiments, simulating real changes that occurred in English, in which a network is trained on a corpus, then trained on a frequency-altered version of that corpus, and a word or sequence of words consequently changes its categorical affiliation, exhibiting new behaviors that were previously ungrammatical (i.e., below some probability threshold). The cases simulated include the rise of periphrastic *do*, the development of *sort of*/*kind of* as a degree modifier, and the development of *be going to* as a future auxiliary.

Tabor, like Bybee, argues that the changes in frequency that often precede a reanalysis can be the cause of the reanalysis: the more a word appears in the same context as words from some other category, the more it is pushed to take on the characteristics of that category. For example, in the *sort of*/*kind of* case,

sentences like (10) would have been parsed only as (10a) until the 19th century ('It was a type of dense rock'), but can currently also be parsed as (10b) ('It was a somewhat dense rock'). Tabor argues that a high frequency for sentences like (10)—where *sort of*/*kind of* is followed by and adjective+noun and therefore appears in a position that the degree modifiers *quite* or *rather* also can appear in—caused *sort of*/*kind of* to become affiliated with the degree modifiers, and therefore become able to appear in unambiguously degree-modifying contexts, like (11).

(10) It was a sort/kind of dense rock (Tabor p. 137)

(a) It was [a [[sort/kind $_{N}$] [of [dense rock $_{NP}$] $_{PP}$] $_{N'}$] $_{NP}$]

(b) It was [a [[[sort/kind of $_{DegMod}$] [dense $_{Adj}$] $_{AdjP}$] rock $_{N'}$] $_{NP}$]

(11) We are sort/kind of hungry (Tabor p. 137)

Tabor finds a sharp rise in the late 18th century how often *sort of*/*kind of* is followed by an adjective (crucially, preceding the rise of sentences like (11)). The simulation shows that increasing the frequency of <a sort/kind of Adj N> noun phrases does lead unambiguously degree-modified utterances like (11) to rise above the threshold of grammaticality (i.e., become more frequent than a near-grammatical control sentence type that is never attested in the corpus and does not

become more likely over the course of the training) and continue to rise in frequency. Thus, again we see that reanalysis and frequency change are mutually reinforcing.

4. Language agents in a probabilistic environment

Speaker-hearer interactions, whether involving adults, children, or a combination, are the atoms of language change. What we call a language change is not a single event, but rather a high-level description of millions of individual interactions over time, with early interactions influencing later ones. If a participant, child or adult, comes away from an interaction with her grammar or lexicon slightly changed, then her altered behavior in a subsequent interaction may cause a change in the grammar of her interlocutor, and so on.

The mathematics of a model built up from many probabilistic interactions of agents can be unwieldy, however. Rather than trying to calculate directly how the system will behave, researchers often use computer simulation as an experimental tool. Artificial agents with the desired properties and behaviors are left to interact and change, and the results observed. Through such simulations, the effects of probabilistic learning and behavior on language change can be explored, and we can determine under what conditions a change will continue or accelerate, and under what conditions variation is stable.[16]

4.1. The adoption of new words

Zuraw (2000), in presenting a model of exceptions and regularities in the phonological grammar, shows that listeners' probabilistic updating of their lexicons can shape the integration of new words into a language. The puzzle I attempted to address is that even though there is a resistance to applying semiproductive phonology to new words, as words become integrated into the lexicon they begin to undergo semiproductive phonology at rates similar to those seen in the established lexicon.

In the model proposed, semiproductive phonology is encoded in a stochastic optimality-theoretic grammar (see Boersma 1998, Boersma and Hayes 2001) by low-ranking markedness constraints. Existing words' behavior is determined by high-ranking faithfulness constraints that require the preservation of idiosyncratic properties encoded in lexical entries. These constraints do not apply to new words, however, because those words lack lexical entries. The 'subterranean' constraints emerge, therefore, to determine the pronunciation of new words.

The example examined in the greatest depth is Tagalog nasal coalescence, which fuses a prefix-final nasal with a stem-initial obstruent:

(12)

stem		*nasal-coalesced*	
bákat	'mark, scar'	ma<u>m</u>ákat	'to leave a scar'

stem	*prefixed but not nasal-coalesced*	
bajáni	‘hero, helper’	mambajáni ‘to offer cooperation’

As shown in nasal coalescence appears to be distributed in the lexicon according to a pattern—voiceless obstruents are much more likely than voiced to undergo it, and obstruents with fronter places of articulation are somewhat more likely than those with backer places to undergo it—but the pronunciation of individual words is unpredictable and must be memorized.

(13) (figure is at end of manuscript)

I argue that words with nasal-coalescing prefixes (or at least some of them) have their own lexical entries, and thus high-ranking faithfulness constraints against coalescing, splitting, or inserting segments within a lexical entry ensure that they are pronounced correctly. An additional constraint, USELISTED, which prefers inputs to be a single lexical entry, ensures that if a lexical entry exists, it is used as the basis for evaluating faithfulness.

When no lexical entry exists, as when a prefixed form is created for the first time, USELISTED cannot be satisfied, and the faithfulness constraints do not apply, so it falls to low-ranked constraints to decide probabilistically whether nasal coalescence should apply. I further assume that the strength of a lexical

entry grows gradually as instances of the word are encountered, and that a lexical entry with strength of 0.5, for example, is available for use only half the time. Thus, in 50% of utterances, USELISTED and the faithfulness constraints will enforce the memorized pronunciation of such a half-strength word, but in the other 50% of utterances, the lower-ranked constraints will decide, because the lexical entry has not been accessed.

Boersma's (1998) Gradual Learning Algorithm is shown to be able to learn the distribution of nasal coalescence from exposure to the lexicon and encode that distribution in the ranking of subterranean constraints, preferring nasal coalescence on voiceless obstruents and dispreferring nasal coalescence on back obstruents (cross-linguistic motivations are suggested for both). The behavior of the resulting grammar in generating and assigning acceptability ratings to new morphologically complex words is shown to be a fair match to experimental results with speakers. The part of the model that I will describe here concerns the grammar and lexicon's effects on the adoption of new words by the speech community.

The grammar is biased against applying semiproductive phonology to new words (this is just the definition of semiproductivity in this model: an unfaithful mapping from input to output is productive to the extent that the ranking values in the grammar allow it to apply to new words). This is consistent with experiments in several languages' finding that speakers are reluctant to apply semiproductive

phonology to new words—although various aspects of the experimental design can increase apparent productivity, it is always less than what might be expected from looking at the lexicon (Bybee and Pardo 1981; Eddington 1996; Albright, Andrade, and Hayes 2000; Suzuki, Maye, and Ohno 2000). It has also been observed in many cases, however, that words eventually tend to conform to existing lexical patterns after they have been in the vocabulary for some time. In the Tagalog case, prefixed forms of Spanish loan-stems undergo nasal coalescence at rates similar to those seen in the native vocabulary, as shown in (14). This phenomenon seems counterintuitive, if we expect that the more frequent pronunciation early in a word's life (i.e., without nasal coalescence) should take over and become the conventionalized pronunciation as the word establishes a lexical entry in the minds of speakers.

(14) (figure is at end of manuscript)

I propose that the solution lies in probabilistic interactions between speakers and hearers, specifically in probabilistic reasoning on the part of the listener. Because morphologically complex words can be either drawn directly from their own lexical entries or formed synthetically by morpheme concatenation, the listener must decide whether a morphologically complex word that she hears was lexical or synthesized for her interlocutor, assuming that she wants to maintain a lexicon that is similar to her interlocutors'. For example, if a

listener hears *mambulo*, she must guess whether the speaker was using a lexicalized word *mambulo*, or merely concatenating the prefix *maN-* with the stem *bulo*. Factors that should enter into the calculation include how strong the listener's own lexical entry for the word is (if she has one at all), and how likely it is that a lexicalized or concatenated input, respectively, would produce the observed pronunciation. The listener can apply Bayes' Law:

(15) $P(synthesized \mid pronunciation)$

$$= \frac{P(pronunciation \mid synthesized) \cdot P(synthesized)}{P(pronunciation)}$$

$P(lexicalized \mid pronunciation)$

$$= \frac{P(pronunciation \mid lexicalized) \cdot P(lexicalized)}{P(pronunciation)}$$

The grammar influences that calculation, because the probabilities $P(pronunciation \mid synthesized)$ and $P(pronunciation \mid lexicalized)$ depend on the grammar. $P(pronunciation \mid lexicalized)$ is always close to one, because of the high-ranking faithfulness constraints. $P(pronunciation \mid synthesized)$ is higher for non-nasal-coalesced pronunciations than for nasal-coalesced pronunciations—recall that the grammar somewhat disfavors nasal coalescence on new words.

Therefore, there is a bias towards classifying non-nasal-coalesced words as synthesized and nasal-coalesced words as lexical. Intuitively, the low productivity of a phonological rule encourages speakers to interpret words that do display the rule as exceptional and therefore listed.

The lexicon also influences the calculation, by contributing to *P*(*synthesized*) and *P*(*lexicalized*). *P*(*synthesized*) depends on the construction's productivity, determined by how many morphologically and semantically eligible words participate in the construction. *P*(*lexicalized*) depends on the candidate word's similarity to existing words. Thus, pronunciations that are similar to existing, lexicalized words (e.g., nasal-coalesced voiceless front obstruents and non-nasal-coalesced voiced back obstruents) are more likely to be interpreted as lexical.

If a hearer does decide that a word was lexicalized for her interlocutor, she will create a weak lexical entry for it. The existence of this weak lexical entry means that when it is the hearer's turn to speak, she has some small probability of using it. The bias towards recording nasal-coalesced words as lexical, especially when they resemble existing nasal-coalesced words (and ignoring non-nasal-coalesced words as synthesized) results in stronger lexical entries for nasal-coalesced pronunciations, which in turn results in an increase in the number of nasal-coalesced productions, leading to further strengthening of lexical entries for nasal-coalesced pronunciations.

The model was implemented in a computer simulation with ten agents of varying ages who from time to time "die" and are replaced by agents with empty lexicons. To avoid undue influence from young speakers with immature lexicons, in each speaker-hearer interaction the hearer probabilistically decides, as a function of the speaker's age, whether to let her lexicon be affected by the speaker's utterance.[17] Different pronunciations for the same word (nasal-coalesced and not) do not directly compete, but if they are not reinforced, lexical entries decay.[18] Therefore, if two pronunciations remain prevalent, agents can have two strong pronunciations for the same word. This is a desirable result, because there are certain nasal-coalesced words whose pronunciation is variable within speakers. But if one pronunciation becomes much more common than the other, the lexical entry for the uncommon pronunciation will gradually decay.

The result of the simulations, shown in (16), was that new words were incorporated into the lexicon in a pattern similar to that seen among Spanish stems (though somewhat more extreme—this could be because some of the Spanish-derived words have not been in use long enough for their pronunciations to reach their final state). That is, for voiceless obstruents the final rate of nasal coalescence is nearly 100%, for front, voiced obstruents it is around 50%, and for back, voiced obstruents it is close to 0%.

(16) (figure is at end of manuscript)

Probabilistic reasoning by adults, then, could explain the maintenance of lexical regularities over historical time. Such reasoning requires speakers to have a probabilistic grammar, so that there is variation in the treatment of new words, and it requires listeners to have access, whether direct or indirect, to the statistical characteristics of the lexicon.

4.2. Learners' response to the probabilistic environment

If language change is a shift in the distribution of competing variants, what causes that distribution to change from one generation to the next? Why don't children mimic the frequencies of their elders? We have seen that in some cases—generally cases of morphosyntactic change—the shift in frequency appear to reflect a reanalysis (e.g., Frisch's *not* case, Tabor's *sort of* case): younger speakers use a construction at a different rate because they assign a different interpretation to it. In other cases—generally cases of phonological change—younger speakers may use the older or the newer variant according to the requirements of the social setting (i.e., formal vs. informal), indicating that they control a grammar that is qualitatively similar to their elders', but that assigns different probabilities to different variants. Biber and Finegan (1989) argue that stylistic shifts in written English over the last four centuries similarly reflect sociological, rather than structural, motivations.

Another possible source of frequency shift that has been proposed is the conflict between frequency and learnability: some variable situations could be inherently unstable, depending on learners' bias in dealing with ambiguous utterances. Yang (2000) and Briscoe (1999) both explore this idea within a principles and parameters framework (Chomsky 1981), where acquisition is the process of parameter setting.

For Yang, no parameters are preset—all settings must be learned. The learner has a finite number of grammars to choose from, each having an associated weight that the learner maintains.[19] In each learning trial, the learner receives an input sentence and probabilistically selects one grammar, with higher-weighted grammars more likely to be chosen. If the grammar selected can parse the sentence, then the learner augments its weight and decrements the weight of the other grammars. If the grammar selected cannot parse the sentence, then the learner decrement its weight and augment the weights of all the other grammars.

The final weight that each grammar will attain depends in part on the distribution of grammars among the adults providing the learning data, but also on how many ambiguous sentences occur and what the learner does with them. For example, adults using a V2 (verb-second) grammar will produce a high proportion of sentences that are compatible with an SVO grammar.

Yang shows how this system can cause a drift in grammar probabilities. Suppose that the learning environment contains two grammars G_i and G_j, and that

a proportion α of G_i's sentences are incompatible with G_j (this is G_i's *advantage*—the proportion of G_i-generated sentences that unambiguously lead the learner to strengthen the weight of G_i), and a proportion β of G_j's sentences are incompatible with G_i (G_j's *advantage*). These proportions vary according to the specifics of the grammars and according to the likelihood of various utterances—for example, the likelihood that an unambiguously V2 sentence is uttered given a V2 grammar may be quite different from the likelihood that an unambiguously SVO sentence is uttered given an SVO grammar. At generation n, the linguistic environment contains some proportion p of adult utterances from G_i and some proportion q of adult utterances from G_j ($p + q = 1$).

The probability that grammar G_i will have its weight incremented in any learning trial is αp, and the probability that G_j will have its weight incremented is βq. The learners will therefore tend to converge on new weights $p' = \alpha p / (\alpha p + \beta q)$ for G_i and $q' = \beta q / (\alpha p + \beta q)$ for G_j. This means that the weights have been learned unfaithfully ($p' \neq p$ and $q' \neq q$), except in the special case of $\alpha = \beta$.

For G_j to overtake G_i, q needs to grow at p's expense. This means that $p'/q' < p / q$ (the ratio of p to q decreases), or, coalescing the equations for p' and q' obtained above, $\alpha p / \beta q < p / q$, or $\alpha < \beta$: G_i's advantage must be smaller than G_j's. Yang examines corpora in two case studies to see if $\alpha < \beta$ does indeed cause G_j to overtake G_i.

The first case is the change from Old French's V2 grammar to Modern French's SVO grammar. The SVO grammar must have generated more sentences not analyzable as V2 (i.e., SXVO and XSVO sentences) than the V2 grammar generated sentences not analyzable as SVO (i.e., XVSO and OVS sentences). Certain sentences would have been ambiguous: SVO, SVOX, SVXO. To get an idea of how many unambiguous sentences an SVO grammar would generate, Yang looks at modern SVO English and finds that 10% of sentences are SXVO or XSVO (SVO's advantage). Looking at modern V2 languages, Yang finds that the combined proportion of XVSO and OVS sentences (V2's advantage) is 30%. If these advantages also held for competing SVO and V2 grammars in transitional French, then SVO should not have been able to overtake V2. Yang proposes that the solution lies in Old French's null-subjecthood: null-subject XVS sentences would be produced XV, which is also compatible with an SVO analysis (the XV sentence would be interpreted as XSV). Taking null subjects into account, V2's advantage is only 5-18%. If it fell below about 10%, then SVO would begin to take over.

The second case study is the change from V2 in Middle English to SVO in Modern English. The problem is similar to that in the French case: why would SVO take over? Yang proposes that the answer here is Middle English's pronominal proclitics, which resulted in some XSVO and OSV sentences ("V3"). When cliticization waned and these pronouns had to be reanalyzed as real DPs,

the V3 sentences would have been compatible only with an SVO grammar, adding to its advantage.

In Briscoe's (1999) model, the instability of a variable grammar comes not from the frequency of ambiguous sentences, but from the (overturnable) pre-setting of certain parameter values.[20] On the one hand, a more frequent variant has more opportunities to shape the learner's grammar; but on the other hand, the more learnable variant—the one that uses more default parameter settings—has an advantage from the start. Briscoe simulates changes in word order, using a generalized categorial grammar framework, in which the syntactic rules are weighted so that different well-formed sentences have different probabilities of being uttered under a particular grammar. The parameters of the grammar include the default head-complement order, the order of subject and verb, the order of verb and object, and several others.

Certain parameter settings are associated with prior probabilities, intended to reflect innate markedness. Parameter settings are changed only when the learner's current grammar fails to parse an incoming sentence. The learner tries changing some settings, and if this makes the input sentence parsable, those potential new settings are strengthened, though not adopted right away. If the strength of a setting exceeds a threshold value, the new setting is adopted, though it can be changed back if contrary evidence is encountered later. Even after a setting is adopted, its strength continues to be updated; this determines how easy

it will be to reverse the setting later. Thus, during learning each parameter has an innate prior probability, a posterior probability derived from learning, and a current setting.

Briscoe's approach differs from Yang's in that, although the learner keeps several grammars under consideration, only one grammar is used at a time, and thus individual adults' outputs will not be variable. The learner chooses the most probable grammar according to Bayes' Law. Letting g be a grammar, G the space of possible grammars, and t_n a triggering input (= the set of sentences seen so far), the probability of a grammar g given a triggering input t_n is given in (17):

(17) $$P(g \in G \mid t_n) = \frac{P(g) \cdot P(t_n \mid g)}{P(t_n)}$$

The prior probability of $P(g)$ is equal to the product of the probabilities of all its parameter settings. $P(t_n \mid g)$, the probability that a given grammar produces the set of sentences seen, is derived from the rule weights of each grammar. The denominator $P(t_n)$ of (17) is unimportant, because it is the same in all grammars being compared. The grammar that the learner uses to try to parse incoming sentences, and that the learner will use when speaking, is simply the most probable $g \in G$.

Briscoe's model, like Yang's, is sensitive to the amount of overlap in triggers (e.g., surface *SVO* as evidence for either an SVO grammar or a V2 grammar). Briscoe found that in a population of mainly SOV+V2-speaking adults ("German") and some SVO-speaking adults, learners reliably converged to SOV+V2 as long as the percentage of unambiguously SVO triggers did not exceed 15% (the number depends on the strength of the default settings, if any).[21] If the percentage exceeded 15%, a drift towards SVO could begin.

The learner's response to a variable environment is crucial to language change. But in order to ensure that learners do not merely replicate the frequencies around them, there must be some persistent bias at work, whether it comes from social motivations, from learnability, or from ambiguity.

4.3. Language change under competing forces

We have seen several forces that may be at work in probabilistic language change: innate parameter settings, frequencies of ambiguous utterances, frequencies of individual lexical items or constructions, variation due to language or dialect contact. In real cases of language change, however, it can be difficult to tease apart these factors to determine which are necessary or sufficient triggers for various changes. As Dras et al. (2001) note, simulation studies provide a way to perform diachronic experiments on language, altering the strength of each force and observing the effects.

Few such simulations have been undertaken that attempt to model real changes, but this line of inquiry seems promising. This section concludes by reviewing preliminary results from one simulation project (Dras et al. 2001) that investigates the effects of various forces on changes in vowel harmony. Dras et al. collected corpus data on Old Turkic—in which 100% of words were harmonic for palatality and backness—and several of its descendants, from the 9th century to the present. Some of the contemporary languages have maintained very high rates of harmony, while others' rates of harmony have fallen drastically. Dras et al. identify internal and external factors that could affect rates of vowel harmony: vowel co-articulation, inherent markedness of certain vowels, consonantal effects, merger of vowels (collapsing harmony pairs), the introduction of disharmonic loanwords, and language contact, and model several of them.[22] Agents in the simulation exchange words with randomly selected neighbors, updating their lexical entries to reflect what they have heard. When speaking, an agent may mispronounce a word or co-articulation; when listening, an agent may mishear, ignore co-articulate, or adjust an interlocutor's pronunciation before adding it to the lexicon. Agents may also mutate their lexical entries at an agent-specific rate; if a vowel is to be mutated, there is an agent-specific probability that it will be made harmonic.

If factors favoring harmony are strong enough, harmony can increase, following roughly an S-shaped curve. Dras et al. find that vowel merger alone is

not sufficient to eliminate harmony, nor is the addition of disharmonic loanwords. Though Dras et al. emphasize that their results are preliminary, and that the model needs to be enriched with several more factors, this study shows a promising direction for future work: using probabilistic simulation tools and real historical data to model the effects of a variety of internal and external factors on language change. Such simulations should help us determine which factors, alone or in conjunction, are strong enough to cause and continue language change.

5. Conclusion

Many linguists are interested in language change because of what it can tell us about synchronic language. For example, the types of reanalysis that are common may tell us about the learning mechanism, and the way a change spreads through the speech community may tell us about the social function of language.

Because the study of language change draws on all areas of linguistics, and particularly on probabilistic approaches to all areas of linguistics, the study of language change also has something to contribute to the study of probabilistic linguistics: models of synchronic acquisition, representation, and use of language must be consistent with observed diachronic facts.

We have seen that the probabilistic behavior of learners, speakers, and listeners can shape language change, and that simulation studies can help us explore how this happens. Simulation studies can also support or undermine

models of how agents represent and use linguistic knowledge: if a model yields a good match to the known facts concerning language change, it is to be preferred over one that does not. In Tabor (1994), for example, a connectionist model of syntactic knowledge is supported by its ability to model frequency linkage and the relationship between frequency changes and reanalysis. In Zuraw (2000), a probabilistic model of knowledge of lexical regularities is supported by its ability to model the incorporation of new words into the lexicon.

Language change can be a testing ground, then, for probabilistic models of learning, speaking, and listening. It is to be hoped that current advances in our understanding of the probabilistic nature of the language faculty will have much to contribute to the study of language change over the coming years, and that the study of language change can return the favor.

FIGURES

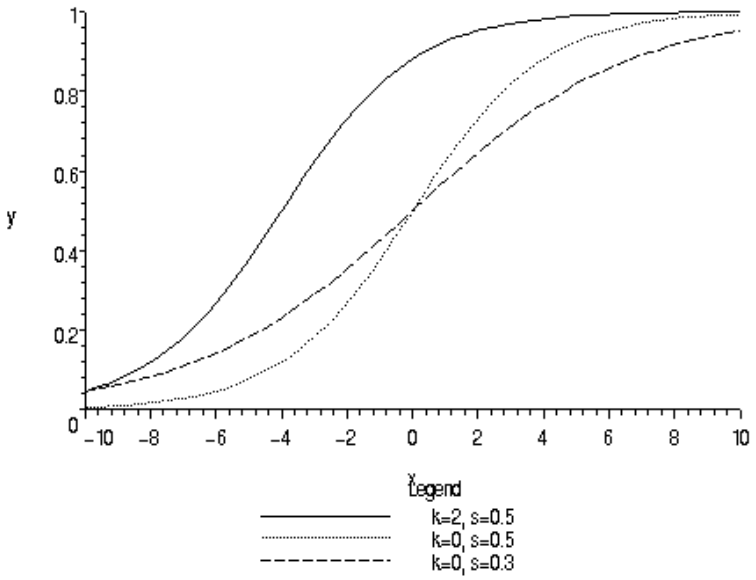
(1)

		German																	
		f	Ø	h	b	v	ʃ	k	z	r	l	n	g	m	t	ts	d	pf	etc.
English	s	0	0	1	0	0	5	1	6	1	0	0	0	0	0	0	0	0	14
	b	1	0	0	5	0	1	1	0	1	0	0	1	0	0	0	0	0	10
	h	0	0	6	0	1	0	1	0	0	0	0	0	1	0	0	0	0	9
	Ø	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
	n	0	0	1	0	1	0	1	0	0	0	5	0	0	0	0	0	0	8
	f	8	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	12
	w	1	1	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	7
	l	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	5
	m	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	5
	t	0	0	0	1	0	1	0	0	0	0	0	0	0	0	3	0	0	5
	k	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0	4
	r	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	4
	d	0	0	1	0	0	1	0	0	0	0	0	0	0	2	0	0	0	4
	g	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	3
	j	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2
	ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
	p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	total	11	9	9	8	8	8	7	7	5	9	5	5	4	2	3	2	1	103

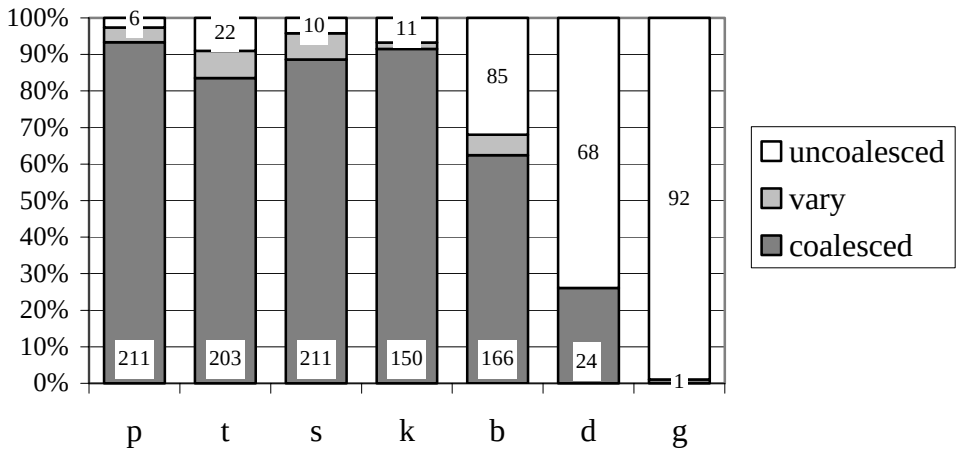
(2)

		German																	
		f	Ø	h	b	v	ʃ	k	z	r	l	n	g	m	t	ts	d	pf	sum
English	s	1.5	1.2	1.2	1.1	1.1	1.1	1.0	1.0	0.7	1.2	0.7	0.7	0.5	0.3	0.4	0.3	0.1	14.1
	b	1.1	0.9	0.9	0.8	0.8	0.8	0.7	0.7	0.5	0.9	0.5	0.5	0.4	0.2	0.3	0.2	0.1	10.3
	h	1.0	0.8	0.8	0.7	0.7	0.7	0.6	0.6	0.4	0.8	0.4	0.4	0.4	0.2	0.3	0.2	0.1	9.1
	Ø	0.9	0.7	0.7	0.6	0.6	0.6	0.5	0.5	0.4	0.7	0.4	0.4	0.3	0.2	0.2	0.2	0.1	8
	n	0.9	0.7	0.7	0.6	0.6	0.6	0.5	0.5	0.4	0.7	0.4	0.4	0.3	0.2	0.2	0.2	0.1	8
	f	1.3	1.0	1.0	0.9	0.9	0.9	0.8	0.8	0.6	1.0	0.6	0.6	0.5	0.2	0.4	0.2	0.1	11.8
	w	0.7	0.6	0.6	0.5	0.5	0.5	0.5	0.5	0.3	0.6	0.3	0.3	0.3	0.1	0.2	0.1	0.1	6.7
	l	0.5	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.6
	m	0.5	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.6
	t	0.5	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.6
	k	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.2
	r	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.2
	d	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.2	0.4	0.2	0.2	0.2	0.1	0.1	0.1	0.0	4.2
	g	0.3	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.0	2.9
	j	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.1	0.0	0.0	2.1
	ð	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0	0.1	0.0	0.0	2.1
	p	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9
	sum	10.9	9.1	9.1	8	8	8	6.9	6.9	4.8	9.1	4.8	4.8	4.2	2.1	2.9	2.1	0.7	102.4

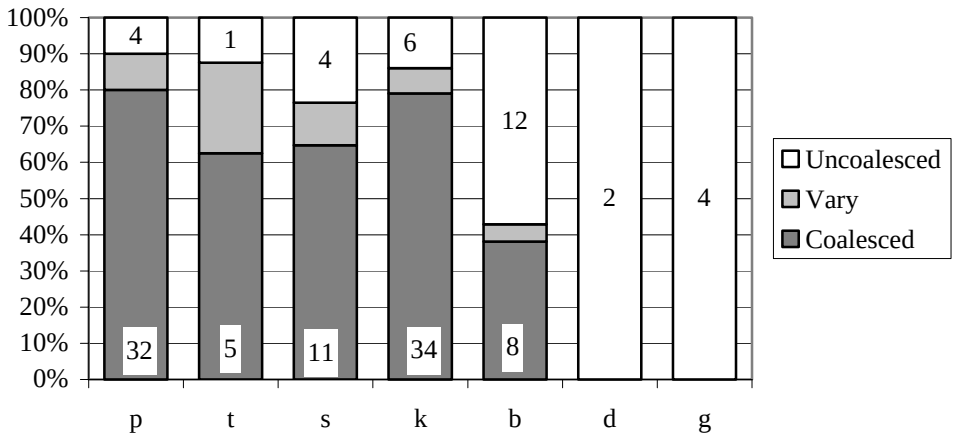
(6)



(13)


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
6
22
10
11
85
68
92
211
203
211
150
166
24
1
p
t
s
k
b
d
g
uncoalesced
vary
coalesced

(14)


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
4
1
4
6
12
2
4
32
5
11
34
8
p
t
s
k
b
d
g
Uncoalesced
Vary
Coalesced

(16)

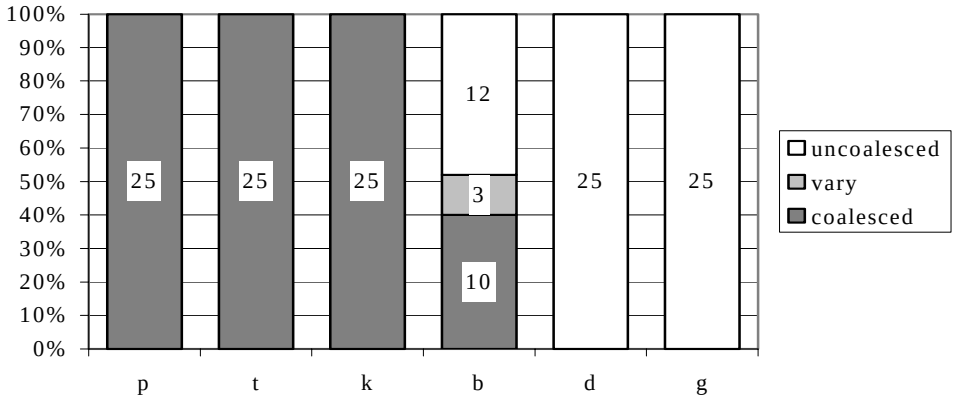

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
25
25
25
12
3
10
25
25
p
t
k
b
d
g
uncoalesced
vary
coalesced

FIGURE CAPTIONS

(1)

Observed values for initial-consonant correspondences in English and German (Ringe 1992, pp. 22-23)

(2)

Expected values for initial-consonant correspondences in English and German

(6)

Three logistic functions: the solid and dotted lines have the same logit slope (0.5), but different logit intercepts (2 and 0, respectively); the dashed line has the same logit intercept as the dotted line (0), but a different logit slope (0.3).

(13)

Rates of nasal coalescence in the native Tagalog lexicon, broken down by stem-initial consonant.

(14)

Rates of nasal coalescence in Spanish loans.

(16)

Rates of coalescence on new words in simulated speech community.

ENDNOTES

---

[1] Many thanks to the editors, Rens Bod, Jennifer Hay, and Stefanie Jannedy, and to Bryan Zuraw, for their substantial feedback. The greatest thanks goes to Norma Mendoza-Denton and Chris Manning, whose comments significantly shaped the form and content of this chapter.

[2] When a group of languages is known to be related, we may wonder how closely. Although answering this question requires quantitative techniques, it generally does not involve probabilistic tools. Determining degree of relatedness involves establishing a similarity metric and a clustering technique to group the most similar languages most closely. See Embleton (1986) for a review of quantitative techniques in tree reconstruction, and Guy (1980a, 1980b) for some interesting computer experiments. Current work in comparative dialectology (e.g., Kessler 1995, Nerbonne and Heeringa to appear) similarly explores the questions of similarity metrics and clustering techniques, although this literature generally does not seek to establish facts about genetic relatedness but rather seeks to quantify the degree of current similarity between dialects in a way that might, for example, be useful to language planners and educators.

[3] The tools of probability can do little to help us identify loans; obvious loans could be excluded from word-lists, but, as Kessler (2001) observes, borrowings that took place in the distant past may be impossible to detect.

[4] An analogous non-linguistic example is the "lottery fallacy": even though you will almost certainly not win the lottery this week (even if you buy a ticket), it is quite likely that someone will win. It is perhaps the high likelihood of the general event that makes the specific event seem within reach.

[5] $\binom{p}{q}$, "*p* choose *q*", is the number of ways that a subset set with *q* elements can be chosen from a set of *p* elements. $\binom{p}{q} = \frac{p!}{(p-q)!q!}$, where *n*!, "*n* factorial", is equal to $n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot ... \cdot 3 \cdot 2 \cdot 1$.

[6] χ^2 is the sum of $\frac{(O-E)^2}{E}$ for each cell, where *O* is the observed value, and *E* is the expected value (the value that the cell would have if the proportion of entries per row were the same for each column and vice versa).

[7] In the shift test, for a list of length *n*, only *n* permutations are considered: shifting list *B* down by 0 lines, by 1 line, by 2 lines, etc.

[8] Or, equivalently, $P = \frac{e^{k+st}}{1+e^{k+st}}$.

[9] "ln" stands for "natural logarithm". $\ln(x) = y$ means that $x^y = e$, where $e \approx 2.72$ is the so-called "natural number".

[10] This is the source of Tabor's "Q-divergence": changes in a word's category are accompanied or even preceded by changes in the frequency at which it appears in an ambiguous context.

[11] A potential counterargument is that if *ne* was lost for phonological reasons, it might have been preserved in the written record for a deceptively long time.

[12] Niyogi and Berwick make the point that only the 5-parameter system, not a similar 3-parameter system, tends towards loss of V2. The 3-parameter system actually produces a tendency towards increasing V2. Therefore, Niyogi and Berwick argue, diachronic simulations such as theirs can be a way of investigating the plausibility of substantive proposals in linguistic theory.

---

[13] See also Bybee (2001).

[14] Phillips finds that the stress shift illustrated by *convíct* (noun or verb) becoming *cónvict* (noun) / *convíct* (verb) affects infrequent words first. She also finds, however, that final stress on verbs ending in *–ate* in British English developed first on frequent words. Phillips suggests that the *–ate* shift is not really a morphological rule in Bybee's sense. Applying the stress shift does not require analyzing the morphological category or morphemic structure of a word; rather, it involves ignoring *–ate*'s status as a suffix.

[15] Bybee (2000) proposes that just as frequent words can retain archaic phonology or morphology, frequent words and constructions can retain archaic syntax. Bybee proposes this as the explanation for why the English modal auxiliaries (*can, might, should*, etc.) retain their archaic behavior with respect to negation and question-formation: *He should not, Should he?* vs. *He does not drive, Does he drive?* The frequency effect could be thought of as complementary to, or a driving force for, the traditional explanation that only modals occupy Infl.

---

[16] A growing literature simulates "language evolution", that is, change whose starting point is a speech community that lacks any shared linguistic knowledge. See Steels (2000) for a very brief overview; for further reading see Kirby (1999) and many of the articles in Hurford et al. (1998) and Knight et al. (2000).

[17] Like most agent-based simulations of language change, the model lacks social or spatial structure: each agent has an equal probability of interacting with any other agent. Because leadership in a language change seems to be correlated with the characteristics of the speaker's social network (see Milroy 1980, Labov 2001), a simulation that attempts to model the spread of a sociolinguistically correlated change through the speech community would require a more realistic social structure. Social structure is probably irrelevant, however, to merely demonstrating the transmissibility of a lexical pattern.

[18] In the simulation reported in Zuraw (2000), pronunciations do compete directly—when one is augmented, the other is diminished. The results here are from a later version of the simulation.

[19] In many situations, it is unrealistic for the learner to maintain a full list of possible grammars. In a principles-and-parameters model, the number of possible grammars is v^p, where v is the number of values that each parameter can take, and p is the number of parameters. Even with binary parameters ($v = 2$), this number

grows very quickly as the number of parameters grows. In Optimality Theory, the situation is even worse: the number of grammars is *c*!, where *c* is the number of constraints. A model in which individual parameter settings or constraint rankings is probabilistic is more tractable (see Boersma 1998, Boersma and Hayes 2001 for probabilistically ranked constraints within a single grammar).

[20] This paper also includes some interesting simulations of creole genesis.

[21] The source of the variation could be dialect contact, social prestige, or random mislearning by the older generation.

[22] Co-articulation, inventory structure, lexical patterns, vowel merger, fixed non-harmonic suffixes, and disharmonic loanwords.