Bayesian Learning Over Conflicting Data: Predictions for language change

Rebecca Morley
Cognitive Science Department
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218
morley@cogsci.jhu.edu

Abstract

This paper is an analysis of the claim that a universal ban on certain (‘anti-markedness’) grammars is necessary in order to explain their non-occurrence in the languages of the world. To assess the validity of this hypothesis I examine the implications of one sound change (a > ə) for learning in a specific phonological domain (stress assignment), making explicit assumptions about the type of data that results, and the learning function that computes over that data. The preliminary conclusion is that restrictions on possible end-point languages are unneeded, and that the most likely outcome of change is a lexicon that is inconsistent with respect to a single generating rule.

1 Introduction

The basic tenet of Evolutionary Phonology is that the observed universal commonalities in phonological systems of the world arise from the universal commonality of the way listeners and speakers produce and perceive sound structures (Blevins, 2004). Diachronic processes operating via the transmission of the speech signal act without regard for the subsequent system they create. Alternate theories in the tradition of Chomsky argue for universal prohibitions which would serve to ban or repair certain changes just in case they would result in a ‘disallowed’ system (Kiparsky 2004, 2006). In Optimality Theoretic terms, this would be a grammar that violates the canonical set of universal markedness constraints. I will call this claim the Universal-Grammar-Delimited Hypothesis Space (UG-Delimited $\mathcal{H}$) Principle.

Without this check, Kiparsky argues, common and natural sound changes (‘blind’ Evolutionary Phonology) would frequently produce unnatural and in fact unobserved ‘anti-markedness’ languages (such as a system in which lower sonority vowels were stressed in preference to higher sonority vowels).

An analysis of the properties of possible grammars is an analysis that involves explicitly characterizing the properties of the learner, as well as of the data to which the learner is exposed. The work in this paper is, to my knowledge, the first attempt to do exactly this kind of analysis, for exactly the type of scenario in which a dispreferred, but hypothetically learnable, grammar might arise.

Diachronic changes that are caused by factors outside of the grammar have the capability of disrupting a categorical rule system, introducing irregularities into a previously regular pattern. These irregularities may have an ‘unnatural’ or anti-markedness character, but typically, they will coexist alongside remnants of the prior natural pattern. That is the first observation. The second is that if learners are allowed to adopt mixed-grammar hypotheses (‘co-phonologies’ (Inkelas 1997), ‘stratal faithfulness’ (Ito and Mester 2001), ‘lexical indexation’ (Pater 2000)), then under a posterior-maximizing learning model, these hybrid systems are the most likely outcome (rather than a categorical ‘anti-markedness’ grammar).

I will work through a case study of sonority sensitive stress, paying special attention to the lexicon that would be produced after a hypothetical sound change of the type Kiparsky proposes. By examining the output of Bayesian hypothesis testing in this domain I will conclude that for the pure anti-markedness grammar to arise, not only is a certain type of diachronic change necessary, but also a certain type of non-uniform lexical distribution. To first approximation, this confluence of circumstances appears rather rare, leading me to tentatively reject the hypothesis that categorical bans on allowed grammars are necessary to explain the distribution of the world's languages.

2 Gujarati Phonology

Kiparsky uses Gujarati to provide a concrete illustration of the relevant phonological paradigm: a sonority-sensitive stress system that respects the posited universal implicational hierarchy. There are eight vowels in Gujarati, corresponding to three sonority tiers: low: (ə), mid: (i,e,ɛ,o,ɔ,u), and high: (a). The stress system is described as conforming to the following position- and sonority- dependent rules.

[1] *GUJARATI*: Sonority & Position -to-Stress

- stress penultimate [a] (the most sonorous vowel)
- otherwise stress ante-penultimate [a]
- otherwise stress final [a]
- otherwise stress penultimate mid-sonority vowel (any of [i,e,ɛ,o,ɔ,u])
- otherwise stress ante-penultimate mid-sonority vowel
- otherwise stress the penultimate position (which must be [ə] (the lowest sonority vowel))

This type of system is easily describable within a standard OT framework (Prince and Smolensky 1993/2004) that utilizes a universally ordered sonority scale with respect to the markedness of (or dispreference for) stressing a particular vowel. Crucially, however, the reverse type of system, in which lower sonority vowels are the ones that attract stress, is so far unattested, and predicted, within the same framework, to be impossible.

2.1 Gujarati′

In stating his claim about the necessity of intrinsic bans on possible grammars, Kiparsky makes the following assumption: A common and natural type of sound change is one in which all a's of a language change to ə's [1]. I will adopt this assumption as well for the sake of argument, leaving aside a discussion of the evidence for how plausible it may be. It should be kept in mind that this particular change is being considered only as a stand-in for a class of possible sound changes that could produce similar outcomes with respect to markedness implications.

A change in vowel quality (with unchanged stress placement) will alter the make-up of the Gujarati lexicon, and raise the possibility of a system in which stress preferentially falls on the lowest-sonority vowel, [ə] (formerly [a], the most sonorous vowel)[2]. This new lexicon will, in turn, act as the input to the learner of Gujarati′. To determine the outcome of learning over this data set, some sort of characterization of the learner's hypothesis space is necessary. The list in [2] represents the full hypothesis set considered in this paper[3]. To begin, I will consider only hypotheses 1)-3), leaving aside the discussion of hypotheses 4) and 5) until Section 3.3.

[2] $\mathcal{H}$:Hypothesis Space

1) *PENULT*: Stress Penultimate Vowel
2) *GUJARATI*: Sonority & Position -to-Stress
3) *GUJARATI**: Reversed-Sonority & Position -to-Stress[4]
4) *NULL*(*G*/G*): *GUJARATI** and *G UJARATI* equally likely generators of data
5) *MAX*(*G*/G*): mixed-grammar of *GUJARATI** and *GUJARATI* with variable weights

[1] In fact, it is not clear how likely an internally motivated la n-guage change of a completely general nature is. What might be more plausible is that such changes would depend very heavily on context, with tokens that were less fully realized (e.g., shorter) being more likely to undergo the change than more fully /a/-like tokens. This, of course, would be correlated with their stress status.

[2] An alternative traditional generativist account, rather than admitting an anti-markedness hypothesis, might propose a difference between stress-attracting ə's and non-stress-attracting ə's based on differences in their underlying representations, effectively encoding the diachronic change within the synchronic grammar. This type of analytic bias will impede or prevent changes from affecting the rule system (grammar) of a language, and thus it is not pursued in the present work.

[3] This is clearly far from the only way in which the learning problem can be formulated. Given that this is, to my knowledge, the first study of its kind, a number of somewhat arbitrary representational decisions had to be made. For the purposes of this work the given $\mathcal{H}$-space is the result of what I view as a minimal departure from the standard formalisms both of linguistic theory and Bayesian learning.

[4] As in [1], but with the sonority classes reversed.

2.2 Evidence to the Learner: Gujarati Lexicon

The hypothetical lexicon of Gujarati′ (L') depends on the inventory of the old Gujarati (L). For a given possible Gujarati, L is mapped to L' via the sound change a > ə. To construct the space of L I start by making a list of all possible word types, where the type depends on features that are relevant to the hypotheses under consideration, namely the vowel identities. This listing also corresponds to a particular lexicon $L_{MU} \in L$; this is the word inventory under what I will call the Minimal Lexicon Uniformity assumption: that all types are represented in equal numbers, and each type occurs exactly once. For 3-syllable words and an 8 vowel inventory, there are 8^3, or 512 distinct types. For 2-syllable words, there are 8^2, or 64 types.

Table 1 lists the word types for 3-syllable words. ‘Case’ refers to the type (vowel make-up) of the word before the hypothetical sound change (where M indicates any of the mid-sonority vowel class {i,e,ɛ,o,ɔ,u}). We will restrict ourselves for the moment to considering only the first three hypotheses in the space: *PENULT*(P), *GUJARATI*(G), and *GUJARATI**(G*)).

	Case	Gujarati Example L > L′	# types H
1	(ə,(ə,M),a) (M,ə,a) (a,ə,M) (a,ə,(ə,a))	[pəɾikʃá]>[pəɾikʃə́]	21
2	(M,M,a) (a,M,(M,a,ə))	[hoʃijáɾ]>[hoʃijə́r]	84 G*
3	(M,a,(ə,M,a))	[mubáɾək]>[mubə́ɾək]	48 G*,P
4	((ə,M), M,ə) (ə,M,M)	[tʃumːótəɾ]>[tʃumːótəɾ]	78 G,P
5	(M,ə,M) (M,ə,ə)	[kójəldi]>[kójəldi]	42 G
6	(a,a,(a,M)) (ə,(a,ə),(a,ə,M)) (M,M,M)	[awːánã]>[əwːə́nə̃]	239 G,G*,P

Table 1. Uniform Gujarati Lexicon: three-syllable words (words taken from de Lacy (2006))

Each row represents positive evidence for some subset of the three hypotheses under consideration; the hypotheses consistent with a given case are specified in the last column below the type counts. For example, in Row 3, the word [mubáɾək] in Gujarati, with stress determined by the markedness-abiding grammar described in [1] has become [mubə́ɾək] in Gujarati′. This form now exhibits stress on the lowest (rather than the highest) sonority vowel in the word. This pattern is consistent with the anti-markedness grammar *GUJARATI**. However, the stress placement in this word is also consistent with the simple positional grammar *PENULT*. If we indicate the number of types that support none of the hypotheses as A (=arbitrary), and the number that support all hypotheses as N (= neutral), then we can calculate the total type counts in support of each hypothesis ($A=21;\ G^*=371;\ G=359;\ N=239;\ P=365;\ T=512$). Note that G* exceeds P by six word types.

3 The Bayesian Learner

The numbers in Table 1 represent the make-up of a possible lexicon of Gujarati′, namely, L_{MU}'. This will act as the initial input to our Bayesian learner (for simplicity, all calculations in this section will be performed only for 3-syllable words).

The Bayesian model has been extensively applied to learning scenarios in a number of cognitive domains (e.g., Chater et al., 2006; Kemp et al., 2007; Kording and Wolpert, 2006; Tenenbaum et al., 2007), and involves a fairly minimal and intuitive apparatus. Bayes theorem, which provides a formula for computing the posterior probability of a hypothesis given the data, and thus a method for evaluating competing grammars, is given in (1).

$$p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)} \qquad (1)$$

For the problem at hand, the members of d are stress assignments corresponding to each of the n words of the lexicon. The conditional probability of a stress assignment d_i under hypothesis h is more properly written as $p(d_i|h,y_i)$, where stress assignment (as can be seen from Table 1) depends on the particular word type y_i. I will assume that the conditional probability of each surface stressed form is independent of any other. The probability of the set d given h and y (where h = *GUJARATI**, *PENULT,* or *GUJARATI*) can then be expanded as the product of the probability of each member of d

given h and each member of y (see Equation (2)).

3.1 'Non-Deterministic' Hypothesis Space

Applying Bayes Theorem to the first three hypotheses of [2] returns a value of $p(h|d)=0$ for each grammar. To avoid this collapse (due to the existence of contradictory data), let us assign a small probability of error (2α) under each hypothesis. For a given 3-syllable word type, y, there are three stress possibilities: $C = \{1,2,3\}$, and the stress class assigned by a given hypothesis H_i is written as a function of the input word type: $H_i(y) \in C$. For the Non-Deterministic version of the same hypothesis, written as H_i^{α}, stress will be assigned to the consistent position ($c=H_i(y)$) with probability $1-2\alpha$, and to either of the two inconsistent positions with probability α. See [3].

[3] $\underline{H_i^{\alpha}}$: Non-Deterministic Version of H_i

$$p(c \mid H_i^{\alpha}, y) = \begin{cases} 1-2\alpha & c = H_i(y) \\ \alpha & c \neq H_i(y) \end{cases}$$

We are assessing the consequences of learning with no markedness biases, so we will let the prior probability in Equation (1) be uniform over the hypothesis space. Since we are concerned with the winner in any two-hypothesis competition, we will work with the ratio of their posteriors. Here the hypotheses *GUJARATI**$^{\alpha}$, *GUJARATI*$^{\alpha}$ and *PENULT*$^{\alpha}$ are the Non-Deterministic counterparts of the previously introduced hypotheses of the same names, and the numerical values of $G*$, P and T are extracted from Table 1, under L_{MU}' (and given at the end of Section 2.1).

$$\frac{p(GUJARATI^{*\alpha} \mid d)}{p(PENULT^{\alpha} \mid d)} = \frac{\prod_i p(d_i \mid GUJARATI^{*\alpha}, y_i)}{\prod_i p(d_i \mid PENULT^{\alpha}, y_i)}$$

$$\frac{\prod_{[d_i \neq G^*(y_i)]} \alpha \prod_{[d_i = G^*(y_i)]} (1-2\alpha)}{\prod_{[d_i \neq P(y_i)]} \alpha \prod_{[d_i = P(y_i)]} (1-2\alpha)} = \frac{\alpha^{T-G^*}(1-2\alpha)^{G^*}}{\alpha^{T-P}(1-2\alpha)^{P}} \qquad (2)$$

As we can see from Equation (2), the relative probability advantage is highly dependent on the magnitude of α. Since α is an error term, it should remain relatively small. Within this constraint, we could allow the learner to fit this parameter based on maximizing hypothesis likelihood. For the 3-syllable uniform lexicon, α_{ML} computed with respect to *GUJARATI** is approximately .14. Using this value in Equation (2) we find that *GUJARATI**$^{\alpha}$ wins out over both *GUJARATI*$^{\alpha}$ and *PENULT*$^{\alpha}$ by several orders of magnitude: $\frac{p(G^* \mid d)}{p(P \mid d)} \approx 1.85 \times 10^4$; $\frac{p(G^* \mid d)}{p(G \mid d)} \approx 3.4 \times 10^8$.

This initial result seems to provide strong support for The UG-Delimited $\mathcal{H}$ Principle: the *GUJARATI** grammar seems overwhelmingly likely to arise, and yet is unattested. However, it is instructive to consider the inherent sensitivity of the Bayesian learner to quite small differences between the linguistic hypotheses in question. A discrepancy between data coverage of a mere 6 words, as seen in the above case, can lead to a hypothesis advantage of four orders of magnitude. And, in fact, a discrepancy of even 1 word can give a posterior advantage on the order of a factor of 5 or greater (depending on the value of α). This result is the consequence of the extreme probability distribution over only two types of data (consistent and inconsistent -- with values close to 1 in the first case, and close to 0 in the second). Since the probability of an independent collection of outcomes (a particular input lexicon) is computed via multiplication, each additional difference in data coverage compounds the single point case, such that the ratio grows exponentially.

If this behavior is indeed a problem for our linguistic domain (where different sub-regions of phonological regularity are often observed to coexist stably in natural language (Inkelas 1997)) then there are various means at our disposal to modify the learning model. In the following section I will consider an alternative weighted decision metric; in Section 3.3 I will expand the hypothesis space to include mixed-grammar competitors; and in Section 4 I will alter the parameters of the learning rule to provide a more stringent threshold for success in hypothesis competition.

3.2 Optimal Bayes Classifier

So far, we have been implicitly assuming a winner-take-all classification strategy whereby the hypothesis with the highest likelihood given the data is the one selected by the learner, and all others discarded. Let us now consider, instead, the Optimal Bayes Classifier which categorizes new instances of data by taking a weighted sum of the predictions of all hypotheses in the space.

As expressed in Equation (3), the probability that a new word *y* will be assigned to category c_m (stress syllable *m*), given the body of training data *d* — $p(c_m|d,y)$ — is the weighted sum of the probability each hypothesis gives of c_m classification — $p(c_m|H_s,y)$ — where each of these terms is weighted by the a posteriori probability of the particular hypothesis given the training data, $p(H_s \mid d)$.

$$p(c_m \mid d,y) = \sum_{H_s} p(c_m \mid H_s, y)\, p(H_s \mid d) \qquad (3)$$

Consider now the situation where there are three hypotheses in the space: H_i^{α}, H_j^{α}, and H_k^{α}. The formulation of the selector function in Equation (3) allows for the possibility of a ganging-up effect whereby H_j^{α} and H_k^{α}, even if they individually have lower posterior probability over *d* than does H_i^{α}, can act together to influence the classification of a new data point *y*. We can choose the lexicon in this example so as to showcase the largest possible effect these two subordinate rules could have by making the difference in consistent data between the (deterministic) hypotheses as small as possible, such that H_i has a coverage advantage of only one data point over both H_j and H_k. We will also consider those words for which H_j and H_k differ from the classification predicted by H_i ($H_i(y)=c_1$), but agree with one another in selecting c_2 with the highest probability ($H_j(y)= H_k(y)= c_2$).

From Equation (2), with $G^*\text{-}P=1$,

$$p(H_{j/k}^{\alpha} \mid d) = \frac{\alpha}{1-2\alpha}\, p(H_i^{\alpha} \mid d) \qquad (4a)$$

Substituting (4a) into Equation (3) gives the probability that classification will occur in line with the dominant hypothesis H_i:

$$p(c_1 \mid d,y) = (1-2\alpha)P(H_i^{\alpha} \mid d) + \alpha\frac{\alpha}{1-2\alpha}P(H_i^{\alpha} \mid d)$$
$$+\ \alpha\frac{\alpha}{1-2\alpha}P(H_i^{\alpha} \mid d) \qquad (4b)$$

And the probability that classification will occur in line with the subordinate, but mutually reinforcing, H_j and H_k can be calculated similarly.

The ratio of the probability of categorizing the new item consistently with H_i to that of categorizing consistently with H_j and H_k can then be shown to be

$$\frac{p(c_1 \mid d,y)}{p(c_2 \mid d,y)} = \frac{6\alpha^2-4\alpha+1}{3\alpha(1-2\alpha)} \qquad (5)$$

Now take H_i = GUJARATI*, H_j= GUJARATI, and H_k = PENULT; *y* is a new word of the type in Row 4 of Table 1. The gang-up phenomenon, where GUJARATI and PENULT collude to move stress away from the position preferred by GUJARATI*, may be seen to have any kind of appreciable effect (where $\frac{p(c_1 \mid d,y)}{p(c_2 \mid d,y)} \leq 1.5$) only in the region $.17 < \alpha < .4$ (relatively large values for α). Outside of this region GUJARATI* dominates. And keep in mind, the advantage to GUJARATI* only gets higher for larger differences in coverage (in Equation (5) only a single data point separates the three hypotheses), and for instances of lexical items where GUJARATI and PENULT disagree (Row 5 of Table 1).

So far we have seen that the Bayesian framework exhibits a potential over-sensitivity when applied to problems of the type formulated in this paper: learning over a space of quasi-categorical, contradictory hypotheses. This is true whether we consider learning to result in a single winner-take-all hypothesis, or instead opt for the weighted decision metric of the Optimal Bayes Classifier. We will return to this issue in Section 4. First, however, I will expand the hypothesis space under consideration, in Section 3.3, and introduce, in Section 3.4, a non-uniform prior, adding principled biases on the selection of those different hypotheses.

3.3 Mixed-Grammar Hypotheses

Before we can assess the performance of the Bayesian learner with respect to the UG-Delimited $\mathcal{H}$ Principle we must make sure we consider all potential competitor hypotheses that might be better predictors of the data than those examined so far. In particular, it is instructive to introduce something like a class of null hypotheses: hybrid grammars which explicitly encode equality between any pair of competing alternatives' ability to explain the data[5].

[5] The effect of mixed-grammar hypotheses can also be realized by allowing a selection procedure over a set of simple grammars, as described in Section 3.2, but, crucially, with the weights calculated under the assumption that data are generated by a combination of grammars (see, for example, the variational model proposed by Yang (1999), or the

I define this class as follows: the posterior probability that the hypothesis *NULL*$(i/j)^{\alpha}$ assigns to a stress class c is calculated by allotting equal probability to selecting the H_i^{α} or the H_j^{α} rule to produce an output of that class:

$$p(c \mid NULL(i/j)^{\alpha}, y) = w_i p(c \mid H_i^{\alpha}, y) + w_j p(c \mid H_j^{\alpha}, y) \quad (6)$$

where w_i = w_j = .5. From Equation (6) and the definition in [3], we can compute the probability distribution of stress assignment c given the application of *NULL*$(i/j)^{\alpha}$ to a particular word, y

[4] <u>*NULL*$(i/j)^{\alpha}$: 'Null Hypothesis'</u>

$$p(c \mid NULL(i/j)^{\alpha}, y) = \begin{cases} 1-2\alpha & c = H_i(y) = H_j(y) \\ \dfrac{1-\alpha}{2} & c = H_i(y) \;\; XOR \; c = H_j(y) \\ \alpha & c \neq H_i(y) \; \& \; c \neq H_j(y) \end{cases}$$

It can be shown that, for L_{MU}' (the Gujarati′ lexicon generated from the Gujarati minimum uniform lexicon), the null hypothesis, *NULL*$(G^*/G)^{\alpha}$, is the decisive winner over *GUJARATI*$*^{\alpha}$ (by approximately 30 orders of magnitude). With this broader consideration of the hypothesis space, the anti-markedness grammar is no longer the outcome of learning. And it turns out that we can specify another hypothesis that gives an even higher likelihood over the data.

The 'maximum likelihood' hypotheses are specified by allowing all three parameters (w_i, w_j, and α (now σ)) in Equation (6) to be estimated from the data. *MAX*$(i/j)^{\sigma}$ is defined explicitly below in [5] for any given weighted combination of H_i^{σ} and H_j^{σ}.

[5] <u>*MAX*$(i/j)^{\sigma}$: 'Maximum Likelihood'</u>

$$p(c \mid MAX(i/j)^{\sigma}, y) =$$
$$\begin{cases} (w_i + w_j)(1-2\sigma) & c = H_i(y) = H_j(y) \\ (1-2\sigma)w_i + \sigma w_j & c = H_i(y) \; \& \; c \neq H_j(y) \\ (1-2\sigma)w_j + \sigma w_i & c = H_j(y) \; \& \; c \neq H_i(y) \\ (w_i + w_j)\sigma & c \neq H_i(y) \; \& \; c \neq H_j(y) \end{cases}$$

When H_i = *GUJARATI** and H_j = *GUJARATI*, *MAX*$(G^*/G)^{\sigma}$ assigns the highest posterior of any we have seen so far (approximately 56 orders of magnitude larger than G^*). This is because, within the space of candidates, it gives the highest likelihood to the observed data, and the prior probability (assumed so far to be uniform) plays no role in this calculation. As the hypotheses we are considering become more complicated, however, we are led to consider an alternative to this assumption, one in which hypotheses with longer description lengths, or greater complexity, are penalized (Rissanen 1989).

3.4 Non-Uniform Prior: Hypothesis Description Length

Under the uniform prior assumption, only with a lexicon in which *GUJARATI** accounts for at least 44 times as much data as does *GUJARATI* will *MAX*$(G^*/G)^{\sigma}$ be defeated. In this section I will show how that result would be altered by considering a better approximation to the prior probability distribution over those hypotheses. *MAX*$(G^*/G)^{\sigma}$ and *GUJARATI*$*^{\alpha}$ can be seen to differ in a basic way related to the number of parameters and rules they must each keep track of. A domain-independent means of determining a prior probability based on this difference in size, or complexity, can be found in the information theoretic notion of coding cost, or description length.

Each hypothesis uses a particular labeling strategy to encode the input data (which can be quantified by the number of binary pieces of information, or bits needed to transmit that information to a waiting decoder). In addition, a certain number of bits is needed to encode the hypothesis itself. The total description length for a string (or set of data) d and a particular hypothesis H is given by the following general formula for two-part coding.

$$L(d,H) = L(d \mid H) + L(H) \quad (7)$$

The relation of (7) to Bayes Theorem becomes clear when we introduce the fundamental transformation from probability to optimal code length given by

$$L(x) = -\log P(x) \quad (8)$$

Intuitively, Equation (8) calls for assigning shorter length codes to higher probability symbols x which, on average, will minimize the code length for a string, d, of symbols drawn from distribution $P(x)$. The ability to transform between length and probability allows for the conceptualization of the prior probabilities over the hypothesis space as biases against complexity.

probabilistic version of Optimality Theory over rankings utilized by Jarosz (2006)).

We can think of the hypotheses in $\mathcal{H}$ as decision trees which produce stressed outputs from input words. In order to encode such decision trees we need something like the binary coding scheme given in Rissanen (1989, section 7.2).

$$L(T) = \log\binom{k_T + m_T - 2}{k_T} \tag{9}$$

Here k_T is the number of internal (non-terminal) nodes in the tree and m_T is the number of leaf (terminal) nodes. Equation (9) provides a measure of how much the grammar compresses its input – or how many classes it must keep track of to produce the correct output. For a series of decisions, based on querying for a series of features at a series of internal nodes, there will be a particular outcome at a particular leaf node. For the *GUJARATI** grammar, k_T=5 (corresponding to the relevant questions about vowel identity listed in definition [1] above), and m_T=6 (corresponding to the possible stress decisions resulting from the answers to each of those questions).

Additionally, all Non-Deterministic hypotheses require the estimation of at least one error term. I will approximate the coding length for a set of k free parameters ($\hat{\theta}$), estimated over a string of length n, by Equation (10) (Rissanen 1989, section 3.1).

$$L(\hat{\theta}) = \frac{k}{2}\log n \tag{10}$$

Since I am only interested here in computing the length associated with the hypotheses themselves (the negative log of their prior probability), we will focus on the second term of Equation (7), which can be written as the sum of (9) and (10).

$MAX(G^*/G)^{\sigma}$ consists of a decision tree that is twice as large as that of $GUJARATI^{*\alpha}$ (since it keeps track of both $GUJARATI^{*\alpha}$ and $GUJARATI^{\alpha}$). Additionally, the combination hypothesis makes use of one more estimated parameter (w_{G^*}).

Under L_{MU}', where n=512 words, the prior probability ratio[6] of $MAX(G^*/G)^{\sigma}$ to $GUJARATI^{*\alpha}$ is 1.7x10^4. From this result we can calculate that the type of lexicon in which the mixed-grammar hypothesis would be rejected is one in which the *GUJARATI** hypothesis accounts for at least eight times more data than does *GUJARATI* (G^*/G = 8).

This value must be regarded as an approximation due to its dependence on the particular coding scheme used[7]. It is, however, likely the best and most principled estimate of the linguistic-bias-free prior we can achieve[8].

Under the information theoretic treatment, its lower probability prior is still not enough to prevent $MAX(G^*/G)^{\sigma}$ from winning under L_{MU}' (by 52 orders of magnitude over $GUJARATI^{*\alpha}$). The productions of a learner who has converged on this grammar would not be obviously consistent with a reversed sonority-to-stress output (since many words would show a stress pattern that is incompatible with that hypothesis), but neither would those productions be inconsistent with such a grammar (since a (slim) majority of words provide positive evidence for such a hypothesis). The typological status of such languages will be discussed in the following section.

4 Discussion & Conclusion

The foregoing analysis has served to address the question of whether the observed frequency of occurrence (approximately never) of anti-markedness systems (such as a grammar with a preference for stressing low sonority vowels over high) requires an active constraint that removes those grammars from the learner's hypothesis space. The central claim within this paper has been that attempts to answer this question must involve a careful examination and specification of the learning process, as well as the inputs to the learner.

Given that systems, at any particular time, tend

[6] the contribution of the hypotheses lengths, converted back to probability via Equation (8)

[7] In practice, a code length exactly equal to the negative log of the probability of a particular symbol may be unattainable, and the relationship in Equation (8) becomes an approximation which may be better in some cases than others. Due to this limitation, it is not clear how much the exact magnitude of a result obtained with this method can be relied upon (for a brief discussion of this issue see, for example, Brent (1999).)

[8] An alternative to this approach is to imagine all grammars as potential mixtures, and to stipulate a prior probability distribution over the possible weight values. Each grammar in this view is equally complex, but certain weight combinations may be more likely than others (such as the 'simple' 0/100% distribution over weights). Conceptually this seems at least as reasonable as the current approach. We are still left, however, with the problem of determining the prior probability distribution over the weights, in a manner which, ideally, would be independent of the problem at hand.

to be in a state in which higher sonority vowels attract stress (due to assumed perceptual factors), the hypothetical sound change that disrupts the natural order must act over forms that are originally markedness-abiding. Thus, there will be a residue of those forms in the language even after the change has occurred (those in which /ə/'s *not* derived from /a/'s fail to attract stress in the presence of mid-sonority vowels). If this residue is small enough then the anti-markedness hypothesis might emerge as the winner. In turn, for this residue to be small, the lexicon before the change must exhibit a certain make-up, such that some word types either fail to appear or occur with much lower frequency than others.

In order to approximate these conditions I created 1000 (x5) simulated lexicons by sampling (without replacement) from the uniform word inventory (L_{MU}) at five different rates; for 3-syllable words: 1% (=5 types), 3% (=15 types), 5% (=26 types), 7% (=36 types), and 10% (=51 types). Higher sampling rates meant a greater likelihood of reproducing the underlying uniform type distribution over the 1000 trials, while lower sampling rates (under-sampling) allowed for a higher likelihood of departure from uniformity, and a greater chance for skewed, or outlier, lexicons to emerge.

These simulations were done for the full set of both 3-syllable and 2-syllable words (a more realistic distribution of input to the learner). To combine the two word lengths, with differing numbers of types, I scaled selection from the two classes. A cursory examination of the online English database CELEX (1993) gives a count of 45,652 for 3-syllable words, and 61,738 for 2-syllable words, a 1:1.4 relationship. Using this as a rough guide, and since the ratio of total types between 3-syllable and 2-syllable words is 512:64, a 1:10 scale was used (giving a proportion of 512:640=1:1.25). Each of the five sampling rates maintained this 1:10 scaling factor, such that the lexicon containing 3-syllable word types sampled at 7%, also contained 2-syllable word types sampled at 70%; this is the lexicon of 36 3-syllable word types (out of a possible total of 512) and 45 2-syllable word types (out of a possible total of 64) (Row 5: [36,45] in Table 2).

Each lexicon, L, at a particular sampling rate, was transformed to its L' counterpart (via the change a>ə), and the coverage ratio between hypotheses *GUJARATI** and *GUJARATI* over L' was computed. As given at the end of Section 3.4 for the description-length prior, a value greater than $G^*/G = 8$ is needed for a *GUJARATI** outcome. Here, due to concerns about the sensitivity of the Bayesian learner, and the degree of uncertainty in the calculation of the prior, I relax this criterion. The last four columns of Table 2 correspond to four (largely arbitrary) values for the G^*/G ratio which were stipulated as thresholds (or possible prior probability ratios) that would allow *GUJARATI** to beat $MAX(G^*/G)^{\sigma}$. Each cell contains the percentage of anti-markedness outcomes (calculated from 1000 runs) for a given threshold, at a given sampling rate.

Sampling	[3,2]-syllable	G*/G			
Rate	word types	5	2.5	1.7	1.25
1%,10%	[5,6]	0	0	.4%	6.4%
3%,30%	[15,19]	0	0	0	.9%
5%,50%	[26,32]	0	0	0	.1%
7%,70%	[36,45]	0	0	0	0
10%,100%	[51,64]	0	0	0	0

Table 2: Estimated probabilities of learned anti-markedness grammar: under 5 different sampling rates (given as [number of 3-syllable,2-syllable word types]), for four different threshold coverage ratios.

The very low occurrence rates of Table 2 show that changing our assumptions about the make-up of the lexicon (departing from uniformity) do not qualitatively alter the results of the previous sections. A pure anti-markedness grammar (*GUJARATI*)* seems to be a relatively rare outcome as compared to a mixed-grammar competitor ($MAX(G^*/G)^{\sigma}$), even under relaxed acceptance criteria.

The above work relies heavily on the existence of a residue of natural patterns in a post-sound change language. Under circumstances in which sound change is non-neutralizing (that is, ə is absent from the inventory of Gujarati before the sound change), there will be no contradictory evidence to the learner of Gujarati′: all data is consistent with the *GUJARATI** hypothesis. Furthermore, there is a long-standing intuition in the literature that the most likely sound changes might actually be of this type (Martinet 1955)[9].

Under these circumstances we might expect GUJARATI* to emerge as the clear winner. This will depend critically on whether or not we consider the lack of conflicting data to be an overwhelming factor in hypothesis selection. If, instead, we maintain our space of non-deterministic hypotheses, then there is still competition from the mixed-grammar alternatives. Under the non-neutralizing scenario, Gujarati has 7 vowels (rather than 8); for 3-syllable words, all 343 types support the GUJARATI*$^{\alpha}$ hypothesis, while 265 are also consistent with PENULT$^{\alpha}$. And $G*/P$ = 1.3. 2-syllable words will provide somewhat less of an advantage to the anti-markedness grammar (49:46~1.13), and with a larger weight (10 times greater frequency to approximate the CELEX ratios), giving an adjusted ratio of roughly 1.15. Whether this is enough of an advantage to cause GUJARATI* to be selected will depend on the parameters of our learner, as well as the prior probability ratio between the two hypotheses: the difference in complexity between the GUJARATI* rule, which computes stress location based on both position and sonority, and the PENULT rule, which only computes over position.

What the above discussion illustrates is that the actual form of common or likely sound changes can significantly alter the outcome of analysis. If non-neutralizing sound changes are the norm, then the dispreferred grammar might have a higher predicted likelihood than that calculated here. Alternatively, if chain shifts predominate, whereby all the vowels in the system undergo related incremental changes in quality, the outcome might be different again. And if realistic sound changes operate on a word by word basis, as predicted by Evolutionary Phonology, such that results are even less consistent in terms of sonority class, an even lower likelihood for a true anti-markedness grammar might be the result[10].

This work has been a preliminary attempt to accurately lay out the methodological requirements for addressing questions of how grammars arise. Further research ought to be concerned with exactly the complications to the question just raised. For present purposes, however, there are two general points to be made. The first is that, in order to determine what any theory predicts in this domain, one has to make assumptions about what constitutes a realistic language learner, as well as establish estimates of the normal state of lexical statistics. The second point is that determining those predictions tells us what the relevant typological facts are. The work here suggests that it is the occurrence, not so much of pure anti-markedness systems, but of partial anti-markedness (mixed-grammar) systems that is the critical issue. It may turn out to be the case that these systems are also very rare, and the over-prediction claim holds in its revised form. However, the true distribution of these types of languages seems far from clear at the present time, and work will have to be done to establish the fact of the matter[11].

Acknowledgments

This work was supported by an NSF IGERT grant and a Department of Education Javits Fellowship. I would like to thank Paul Smolensky, Colin Wilson, and Simon Fischer-Baum for their invaluable assistance. Thanks also go to the three reviewers of this paper, especially Adam Albright for his extensive and extremely helpful comments.

[9] Thanks to Adam Albright for bringing this to my attention.

[10] Another issue so far undiscussed is the aptness of describing the GUJARATI* hypothesis as a reversed sonority-to-stress scale. In either instantiation of Gujarati′ (deriving either from the 7- or 8-vowel system) there are only two operable sonority categories {MID,ə}. Stressing ə preferentially over a higher-sonority mid vowel is already dispreferred behavior from a universalist perspective, but it is qualitatively different than a hypothesis that targets sonority as the deciding factor (rather than vowel identity). This second hypothesis, for example, would avoid stressing newly encountered a's, precisely because of the high sonority of the vowel. The likelihood of achieving a true sonority scale reversal seems even lower than that of learning the 'stress-ə' rule. This is because the strongest evidence for a sonority-sensitive scale involves multiple tiers or classes of sonority (probably at least three). However, the more different classes of vowels (the more complications to the calculation of stress) the less likely it seems that an indirect sound change (one that does not target sonority itself) will produce a clean reversal of the pattern. Again, disorder, or proliferating 'co-phonologies' seem more likely to carry the day.

[11] In the first place, it is not a given that pure anti-markedness systems are completely non-occurring (see, for example, Poppe (1960); McLendon (1975); Breen and Pensalfini (1999)). As for potential mixed-grammar languages, these might include systems that have been analyzed as exhibiting high degrees of lexical exceptionality, or gone largely unanalyzed due to what is perceived as patternless behavior.

References

Blevins, J. (2004). Evolutionary Phonology: the emergence of sound patterns. New York, Cambridge University Press.

Breen, G. and R. Pensalfini (1999). "Arrernte: a language with no syllable onsets." Linguistic Inquiry 30(1): 1-25.

Chater, N., J. B. Tenenbaum, et al. (2006). "Probabilistic models of cognition: conceptual foundations." Trends in Cognitive Science 10(7): 287-291.

Court, C. (1970). Nasal harmony and some indonesian sound laws. Pacific Linguistics Series C No.13. S. A. Wurm and C. Laycock.

de Lacy, P. (2006). Markedness: Reduction and Preservation in Phonology, Cambridge University Press.

Inkelas, S. (1997). The theoretical status of morphologically conditioned phonology: a case study of dominance effects. Yearbook of Morphology. G. Booij and J. van Marle, Kluwer Academic Publishers: 121-155.

Ito, J. and A. Mester (2001). "Covert generalizations in Optimality Theory: the role of stratal faithfulness constraints." Studies in Phonetics, Phonology and Morphology 7: 273-299.

Jarosz, G. (2006). Richness of the base and probabilistic unsupervised learning in Optimality Theory. Proceedings of the Eighth Meeting of the ACL Special Interest Group in Computational Phonology and Morphology, New York City.

Kemp, C., A. Perfors, et al. (2007). "Learning overhypotheses with hierarchical Bayesian models." Developmental Science 10(3): 307-321.

Kiparsky, P. (2004). "Universals constrain change; change results in typological generalizations." ms.

Kiparsky, P. (2006). "The Amphichronic Program vs. Evolutionary Phonology." Theoretical Linguistics 32: 217-236.

Kording, K. P. and D. M. Wolper (2006). "Bayesian decision theory in sensorimotor control." Trends in Cognitive Science 10(7): 319-326.

Martinet, A. (1955). Economie des changements phonetiques. Bern, Francke.

McLendon, S. (1975). A Grammar of Eastern Pomo, University of California Press.

Mitchell, T. M. (1997). Machine Learning, McGraw-Hill.

Pater, J. (2000). "Non-uniformity in English seconday stress: the role of ranked and lexically specific constraints." Phonology 17: 237-274.

Poppe, N. N. (1960). Buriat Grammar, Indiana University Publications.

Prince, A. and P. Smolenksy (1993/2004). Optimality Theory, Blackwell Publishing.

Rissanen, J. (1989). Stochastic Complexity in Statistical Enquiry, World Scientific Publishing Co.

Tenenbaum, J. B., C. Kemp, et al. (2007). Theory-based Bayesian models of inductive reasoning. Inductive Reasoning. A. Feeney and E. Heit, Cambridge University Press.

Yang, C. (1999). A Selectionist Theory of Language Acquisition. 27th Annual Meeting of the Association for Computational Linguistics, College Park, MD.