Appendix A: Two-Hypothesis Competition: Simple and Variability Hypotheses

Section 3.2 of Text

The posterior probability evaluation metric – that the hypothesis *h* is the correct one, given the data, *d* – is calculated using Bayes' Theorem:

$$p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)} \tag{A1}$$

Under the word-independence assumption, the probability of the set *d* given *h* and *y* (where *h* = *GUJARATI**, *PENULT*, or *GUJARATI*; and *d* is the set of stressed words, with *y* being the underlying unstressed forms) can be expanded as the product of the probability of each member of *d* given *h* and each member of *y*.

$$p(h \mid d) = \frac{p(h)\prod_i p(d_i \mid h, y_i)}{p(d)} \tag{A2}$$

Since one is typically only interested in the relative value of the posterior probability, the ratio of posteriors for any two hypotheses can be taken to determine the winner. Thus *p(d)* can be ignored since it appears on both sides of the ratio, giving

$$\frac{p(h_i \mid d)}{p(h_j \mid d)} = \frac{p(h_i)\prod_x p(d_x \mid h_i, y_x)}{p(h_j)\prod_x p(d_x \mid h_j, y_x)} \tag{A3}$$

For a given three-syllable word, y_x, there are three stress possibilities: 1-initial stress, 2-penultimate stress, and 3-final stress. The set of possible outputs is given by $C = \{1,2,3\}$, and the stress class assigned by H_i is written as a function of the input word: $H_i(y_x) \in C$. For the original Simple Hypothesis space, each hypothesis predicts exactly one stress position per word – that is, assigns all probability to one position. Thus, the probability of stress being in any given position *c* is either 0 or 1.

$$p(c \mid H_i, y_x) = \begin{cases} 1 & c = H_i(y_x) \\ 0 & otherwise \end{cases} \quad \text{(A4)}$$

The Variability versions of the Simple Hypotheses assign some small probability to other stress positions. From a production standpoint, the process can be conceptualized as follows. Stress placement is either decided via rule or at random. The probability that the rule will be used is high. However, the random process will be chosen instead from time to time. This random process (*A*, for Arbitrary) will result in exceptional stress placement 2 out of every 3 times, for 3-syllable words, and will randomly select the same location as *H* 1 out of very 3 times.

$$p(c \mid \mathrm{A}, y_x) = \frac{1}{3}, \forall c \quad \text{(A5)}$$

For the Variability Hypotheses, the probability of stress in any of the three possible locations *c* is given as the weighted sum of the contributions from the two processes:

$$p(c|H_i^{\alpha}, y_x) = w_i p(c|H_i, y_x) + w_a p(c|A, y_x) \quad \text{(A6)}$$

Take 3α ($= w_a$) to be the probability that stress will be assigned randomly (thus, each position has probability α of being stressed under *A*). This leaves $1\text{-}3\alpha$ as the probability with which the normal stress rule is followed ($= w_i$). The probability of stress at each possible location is given in (A7). In the first instance, the two processes agree in the location of stress, at $c_i = H_i(y_x)$. Otherwise, the two processes disagree, and H_i assigns zero probability to each of these locations, $c_{a1}, c_{a2} \neq H_i(y_x)$:

$$\begin{aligned} p(c_i|H_i^{\alpha}, y_x) &= (1-3\alpha)p(c_i|H_i, y_x) + (3\alpha)p(c_i|A, y_x) = 1-2\alpha \quad \text{(A7)} \\ p(c_{a1}|H_i^{\alpha}, y_x) &= (1-3\alpha)p(c_{a1}|H_i, y_x) + (3\alpha)p(c_{a1}|A, y_x) = \alpha \\ p(c_{a2}|H_i^{\alpha}, y_x) &= (1-3\alpha)p(c_{a2}|H_i, y_x) + (3\alpha)p(c_{a2}|A, y_x) = \alpha \end{aligned}$$

The three scenarios can be compactly expressed by the following formula:

$$\underline{H_i^{\alpha}: \text{Variability Version of } H_i} \qquad p(c \mid H_i^{\alpha}, y_x) = \begin{cases} 1-2\alpha & c = H_i(y_x) \\ \alpha & c \neq H_i(y_x) \end{cases} \tag{A8}$$

According to the definition of the Variability Hypotheses in (A8), the probability assigned to any particular surface form is given as *1-2*α if the form is consistent with the categorical version of the given hypothesis, and α if the form is inconsistent. Thus, it is convenient to divide the dataset *d* into two subsets: 1) the set of stressed words that are consistent with *H*, (e.g., $d_i = G^*(y_i)$: the stress that actually appears on word y_i is the same as the stress assigned by hypothesis *GUJARATI** to word y_i) and 2) the set of stressed words that are inconsistent with *H*. Equation (A3) can then be rewritten as

$$\frac{p(d|GUJARATI^{*\alpha})}{p(d|GUJARATI^{\alpha}|d)} = \frac{\prod_{[d_x \neq G^*(y_x)]} \alpha \prod_{[d_x = G^*(y_x)]} (1-2\alpha)}{\prod_{[d_x \neq G(y_x)]} \alpha \prod_{[d_x \neq G(y_x)]} \alpha (1-2\alpha)} \tag{A9}$$

If the prior probability terms are the same ($p(GUJARATI^*) = p(GUJARATI)$), then the ratio of likelihoods in (A9) is equivalent to the ratio of posteriors in (A3).

Derivation of Equation (6):

For any two hypotheses, H_i^{α}, H_j^{α}, the following variables parameters can be defined; *i* = the number of data points consistent with H_i *and* inconsistent with H_j; *j* = the number of data points consistent with H_j *and* inconsistent with H_i; *n* = the number of data points consistent with both hypotheses; and *a* = the number of data points consistent with neither hypothesis. Assuming uniform priors, and rewriting Equation (A9) in terms of these parameters gives

$$\frac{p(H_i^{\alpha} \mid d)}{p(H_j^{\alpha} \mid d)} = \frac{\alpha^{a+j}(1-2\alpha)^{i+n}}{\alpha^{a+i}(1-2\alpha)^{j+n}} \tag{A10}$$

Collecting terms,

$$= \frac{\alpha^{j}\alpha^{a}(1-2\alpha)^{i}(1-2\alpha)^{n}}{\alpha^{i}\alpha^{a}(1-2\alpha)^{j}(1-2\alpha)^{n}} \tag{A11}$$

$$= \frac{\alpha^{j}(1-2\alpha)^{i}}{\alpha^{i}(1-2\alpha)^{j}} \tag{A12}$$

$$= \frac{(1-2\alpha)^{i-j}}{\alpha^{i-j}} \tag{A13}$$

In the special case where there is only one data point difference between *i* and *j* the competition reduces to

$$\frac{p(H_i^{\alpha} \mid d)}{p(H_j^{\alpha} \mid d)} = \frac{(1-2\alpha)}{\alpha} \tag{A14}$$

Appendix B: Mixture Hypotheses

B.1 Derivation of *NO-DIFF* Hypothesis:

NO-DIFF(i/j)$^{\alpha}$ is defined as the hypothesis that a given stressed surface form is as likely to have been generated by H_i^{α} as by H_j^{α}. This hypothesis assigns stress by randomly selecting either H_i^{α} or H_j^{α} in production. Thus, for any particular surface form there are two possible ways it might have been generated. Any actual utterance corresponds to a surface form and a generator pair. The joint probability of a particular surface form and a particular generating sub-grammar is, by definition, the probability of the generator times the probability of the surface form under the generator. Thus, the total probability of all events resulting in a given surface form is determined by the sum of the probability of events in which H_i^{α} was the generating grammar, and the probability of events in which H_j^{α} was the generating grammar

$$p(c|NO\_DIFF(i/j)^{\alpha}, y_x) = w_i p(c|H_i^{\alpha}, y_x) + w_j p(c|H_j^{\alpha}, y_x) \quad \text{(B1)}$$

In the case of *NO-DIFF(i/j)*$^{\alpha}$ each sub-grammar is equally likely to be the generator, and the weights are both set at *.5*.

The actual probability will vary by word type, and by location in word. Using three-syllable words, there are three possible stress locations, and three possible scenarios for each word: i) H_i^{α} and H_j^{α} both assign high probability to that location ii) one of the two assigns high probability, and the other assigns low probability iii) both hypotheses assign low probability.

Scenario i:

$$p(c_I|NO\_DIFF(i/j)^{\alpha}) = \frac{1}{2}(1-2\alpha) + \frac{1}{2}(1-2\alpha) = (1-2\alpha)$$

Scenario ii:

$$p(c_{II}|NO\_DIFF(i/j)^{\alpha}) = \frac{1}{2}(\alpha) + \frac{1}{2}(1-2\alpha) = \frac{1-\alpha}{2}$$

and

$$p(c_{II}|NO\_DIFF(i/j)^{\alpha}) = \frac{1}{2}(1-2\alpha) + \frac{1}{2}(\alpha) = \frac{1-\alpha}{2}$$

Scenario iii:

$$p(c_{II}|NO\_DIFF(i/j)^{\alpha}) = \frac{1}{2}(\alpha) + \frac{1}{2}(\alpha) = \alpha$$

Assessing the descriptive power of *NO-DIFF(i/j)*$^{\alpha}$ over a particular lexicon requires determining what probability *NO-DIFF(i/j)*$^{\alpha}$ assigns to the observed surface forms. For words that are consistent with both H_i and H_j, *NO-DIFF(i/j)*$^{\alpha}$ predicts the correct stress location with the highest probability (corresponding to Scenario i); for words that are consistent with only one of the Simple hypotheses, *NO-DIFF(i/j)*$^{\alpha}$ assigns the intermediate probability under Scenario ii; and for words that are consistent with neither Simple hypothesis, *NO-DIFF(i/j)*$^{\alpha}$ assigns the lowest probability, calculated under Scenario iii. The formula for descriptive power as a function of word type is given in (B2).

NO-DIFF(i/j)$^{\alpha}$: 'No Difference Hypothesis' (B2)

$$p(c|NO\_DIFF(i/j)^{\alpha}, y_x) = \begin{cases} 1-2\alpha & c = H_i(y_x) = H_j(y_x) \\ \frac{1-\alpha}{2} & c = H_i(y_x)\ XOR\ c = H_j(y_x) \\ \alpha & c \neq H_i(y_x)\ \&\ c \neq H_j(y_x) \end{cases}$$

The competition between *NO-DIFF(i/j)*$^{\alpha}$ and H_i^{α}

In Appendix A a set of parameters for a given lexicon was defined: i = the number of data points consistent with H_i and inconsistent with H_j; j = the number of data points consistent with H_j and inconsistent with H_i; n = the number of data points consistent with both hypotheses; and a = the number of data points consistent with neither hypothesis. Thus, following the format in (A13) and (A6) the ratio of descriptive power between *NO-DIFF(i/j)*$^{\alpha}$ and H_i^{α} can be written in the following way

$$\frac{p(d|H_i^{\alpha})}{p(d|NO\_DIFF(i/j)^{\alpha})} = \frac{\alpha^{j+a}(1-2\alpha)^{i+n}}{\left(\frac{1-\alpha}{2}\right)^{i+j}(1-2\alpha)^{n}\alpha^{a}} \tag{B3}$$

Simplifying and collecting terms,

$$= \frac{\alpha^{j}(1-2\alpha)^{i+n}}{(\frac{1}{2})^{i+j}(1-\alpha)^{i+j}(1-2\alpha)^{n}} \tag{B4}$$

$$= \frac{\alpha^{j}(1-2\alpha)^{i}}{(\frac{1}{2})^{i+j}(1-\alpha)^{j}(1-\alpha)^{i}} \tag{B5}$$

$$= 2^{i+j}\left[\frac{\alpha}{1-\alpha}\right]^{j}\left[\frac{1-2\alpha}{1-\alpha}\right]^{i} \tag{B6}$$

For a simple winner-take-all decision metric, and under the assumption of a uniform prior, H_i^{α} wins when the ratio in (B6) is greater than 1. If i is expressed as a function of j ($i = mj$ where $m \geq 1$), it can be determined how much *more* unambiguous data H_i must account for than H_j, as a function of α. Setting (B6) greater to 1 and taking the log of both sides yields:

$$(1+m)\,j\log(2) + j(\log\alpha - \log(1-\alpha)) + mj(\log(1-2\alpha) - \log(1-\alpha)) > 0 \tag{B7}$$

$$m[\log(2) + \log(1-2\alpha) - \log(1-\alpha)] > \log(1-\alpha) - \log\alpha - \log 2 \tag{B8}$$

$$m > \frac{\log\frac{1-\alpha}{2\alpha}}{\log\frac{2(1-2\alpha)}{(1-\alpha)}} \tag{B9}$$

For a given α, *NO-DIFF(i/j)*$^{\alpha}$ is rejected for values of i greater than or equal to $m(\alpha)j$. See Figure B1 (also Figure 1 in text). In order for stress assignment probabilities to remain well-defined α must be less than .5. When $\alpha = 1/3$ all three word positions have an equal probability of being stressed. This also corresponds to an m value of 1: the two hypotheses are exactly equivalent in their descriptiveness of the data, that is, equally bad. Each predicts stress location at chance

levels. As α falls below 1/3, m rises rapidly. For a mid-range α value of 1/6, i must be almost twice j in order for H_j^{α} to beat the No-Difference Hypothesis.

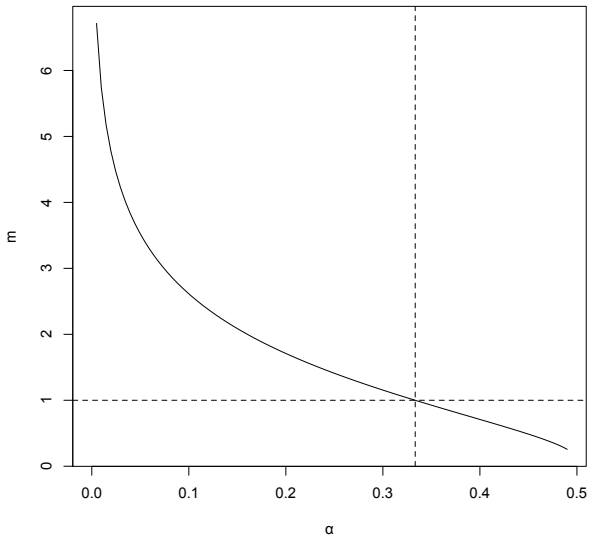

Fig B1
Ratio of unambiguous data ($m=i/j$) as a function of α for a two-hypothesis space.
For a given α, the No-Diff Hypothesis is rejected for points falling above the curve.

B.2 Derivation of *MAX(i/j)*$^{\alpha}$

The Maximum Likelihood Mixture Hypothesis, *MAX(i/j)*$^{\alpha}$, is formulated in the same way as the previously derived No-Diff Hypothesis except the weights are fit from the data, and may take on different values, with the stipulation that $w_i + w_j = 1$.

$$p(c|MAX(i/j)^{\alpha}, y_x) = w_i p(c|H_i^{\alpha}, y_x) + w_j p(c|H_j^{\alpha}, y_x) \quad \text{(B10)}$$

$P(c|H_x^{\alpha}, y_x)$ remains as defined in Equation (A8).

Since the weights for each of the sub-hypotheses are not necessarily the same, there are four unique scenarios to consider with respect to stress placement. The first scenario is the same as above: H_i^{α} and H_j^{α} both assign high probability to the same location; the second is the location to which H_i^{α} assigns high probability, but H_j^{α} assigns low; the third is the reverse scenario: H_j^{α} assigns high probability, but H_i^{α} assigns low. The fourth scenario is also as above: stress locations to which both hypotheses assign low probability.

Scenario i:
$p(c_I|MAX(i/j)^{\alpha}) = w_i(1-2\alpha) + w_j(1-2\alpha) = (w_i + w_j)(1-2\alpha) = 1-2\alpha$

Scenario ii:
$p(c_{II}|MAX(i/j)^{\alpha}) = w_i(1-2\alpha) + w_j(\alpha) = w_i + (w_j - 2w_i)\alpha$

Scenario iii:
$p(c_{III}|MAX(i/j)^{\alpha}) = w_i(\alpha) + w_j(1-2\alpha) = w_j + (w_i - 2w_j)\alpha$

Scenario iv:
$p(c_{IV}|MAX(i/j)^{\alpha}) = w_i(\alpha) + w_j(\alpha) = (w_i + w_j)\alpha = \alpha$

The descriptive power of *MAX(i/j)*$^{\alpha}$ over a particular lexicon is defined as the probability *MAX(i/j)*$^{\alpha}$ assigns to the observed surface forms. For words that are consistent with both H_i and H_j, *MAX(i/j)*$^{\alpha}$ predicts the correct stress location with the highest probability (corresponding to scenario i); for words that are consistent with only one of the Simple Hypotheses, *NO-DIFF(i/j)*$^{\alpha}$ assigns the two different intermediate probabilities under scenario ii or iii; and for words that are

consistent with neither Simple Hypothesis, $MAX(i/j)^{\alpha}$ assigns the lowest probability, calculated under scenario iv. The formula for descriptive power as a function of word type is given in (B11).

MAX(i/j)$^{\alpha}$: 'Maximum Likelihood' Hypotheses (B11)

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

The three parameters w_i, w_j, and α are fit from the observed data so as to maximize the likelihood of the data given $MAX(i/j)^{\alpha}$. As before, i is defined as the number of data points consistent with H_i and inconsistent with H_j; j, as the number of data points consistent with H_j and inconsistent with H_i; n, as the number of data points consistent with both hypotheses, and a, as the number of data points consistent with neither hypothesis.

Using Bayes' Theorem,

$$p(d \mid MAX(i/j)^{\alpha}) = (1-2\alpha)^n \alpha^a (w_i + (w_j - 2w_i)\alpha)^i (w_j + (w_i - 2w_j)\alpha)^j \quad \text{(B12)}$$

This probability is maximized over the data when the derivatives with respect to each free parameter are at zero. Define: $L \equiv (1-2\alpha)^n \alpha^a \left(w_i + \left(w_j - 2w_i\right)\alpha\right)^i (w_j + (w_i - 2w_j)\alpha)^j$

$$\frac{\partial}{\partial w_i} p(d|MAX(i/j)^{\alpha}) = \left[\frac{i(1-2\alpha)}{w_i + (w_j - 2w_i)\alpha}\right] L + \left[\frac{j\alpha}{w_j + (w_i - 2w_j)\alpha}\right] L = 0 \quad \text{(B13)}$$

$$\frac{\partial}{\partial w_j} p(d|MAX(i/j)^{\alpha}) = \left[\frac{i\alpha}{w_i + (w_j - 2w_i)\alpha}\right] L + \left[\frac{j(1-2\alpha)}{w_j + (w_i - 2w_j)\alpha}\right] L = 0 \quad \text{(B14)}$$

Combining (B13) and (B14),

$$\left[\frac{i(1-2\alpha)}{(w_i + (w_j - 2w_i)\alpha)}\right] + \left[\frac{j\alpha}{(w_j + (w_i - 2w_j)\alpha)}\right] = \left[\frac{i\alpha}{(w_i + (w_j - 2w_i)\alpha)}\right] + \left[\frac{j(1-2\alpha)}{(w_j + (w_i - 2w_j)\alpha)}\right] \quad \text{(B15)}$$

$$\left[\frac{i(1-3\alpha)}{(w_i+(w_j-2w_i)\alpha)}\right] = \left[\frac{j(1-3\alpha)}{(w_j+(w_i-2w_j)\alpha)}\right] \tag{B16}$$

$$i\big(w_j + \big(w_i - 2w_j\big)\alpha\big) = j(w_i + (w_j - 2w_i)\alpha) \tag{B17}$$

$$iw_i\alpha + (1-2\alpha)iw_j = jw_j\alpha + (1-2\alpha)jw_i \tag{B18}$$

With the condition $w_i + w_j = 1$ the weight values can be written as

$$w_j = \left[\frac{j-2\alpha j-i\alpha}{i-2\alpha i-j\alpha}\right] w_i = 1 - w_i \tag{B19}$$

$$w_i = \frac{1}{1+\dfrac{j-2\alpha j-i\alpha}{i-2\alpha i-j\alpha}} = \frac{i-2\alpha i-j\alpha}{i-3\alpha i-3\alpha j+j} \tag{B20}$$

Using only the descriptive power metric, $p(d|h)$, $GUJARATI^{*\alpha}$ cannot do better than $MAX(G^*/G)^{\alpha}$. The Mixture Grammar always sets its parameters so as to maximize the likelihood of the training data, and it has more parameters than $GUJARATI^{*\alpha}$. Therefore, the best $GUJARATI^{*\alpha}$ can do is tie, when $w_i = 1$ and $w_j = 0$.

Appendix C: Optimal Bayes Classification

In winner-take-all evaluation, a single hypothesis wins the competition; stress assignment is then determined solely by that hypothesis. By contrast, the Optimal Bayes Learner determines stress assignment by taking a weighted sum of the predictions of all hypotheses in the original space (see, e.g., Mitchell 1997). The weight for a given hypothesis is set to the posterior probability of the hypothesis, given the previously encountered data; thus, the hypotheses are essentially ranked by how plausible they are as generators of the data. For a novel three-syllable word, y_x, each possible stress position is assigned a probability via this weighted sum, as in (C1), where $l \in \{1,2,3\}$.

$$p(c_l \mid d, y_x) = \sum_{H_s} p(c_l \mid H_s, y_x) p(H_s \mid d) \tag{C1}$$

To see how this decision metric changes the previous results, the same exercise can be performed here as was done in the derivation of Equation (6), but with a three-, rather than a two-, hypothesis space: $\{H_i, H_j, H_k\}$.

$$p(c_l \mid d, y_x) = p(c_l \mid H_i^\alpha, y_x) p(H_i^\alpha \mid d) + p(c_l \mid H_j^\alpha, y_x) p(H_j^\alpha \mid d) \tag{C2}$$
$$+ p(c_l \mid H_k^\alpha, y_x) p(H_k^\alpha \mid d)$$

If H_i has only a single data point advantage over both H_j and H_k (i-j = i-k = 1), then Equation (A14) can be used to write

$$\frac{p(H_i^\alpha \mid d)}{p(H_j^\alpha \mid d)} = \frac{(1-2\alpha)}{\alpha} \quad \text{and} \quad \frac{p(H_i^\alpha \mid d)}{p(H_k^\alpha \mid d)} = \frac{(1-2\alpha)}{\alpha} \tag{C3}$$

Substituting $p(H_i^\alpha \mid d)$ into (C2) gives

$$p(c_l \mid d, y) = p(c_l \mid H_i^\alpha, y_x) p(H_i^\alpha \mid d) + p(c_l \mid H_j^\alpha, y_x) \left(\frac{\alpha}{1-2\alpha} \right) p(H_i^\alpha \mid d) \tag{C4}$$
$$+ p(c_l \mid H_k^\alpha, y_x) \left(\frac{\alpha}{1-2\alpha} \right) p(H_i^\alpha \mid d)$$

The area where using Optimal Bayes will make a difference to the calculation is for the following kinds of words: ones where H_j and H_k agree on stress assignment, but disagree with the dominant hypothesis H_i. An example of this type of word comes from row 3 of Table 1 (e.g., /kəʈoro/, where $H_j^{\alpha} = \textit{GUJARATI}^{\alpha}$, and $H_k^{\alpha} = \textit{PENULT}^{\alpha}$. Both assign the highest probability to the second position – the penultimate syllable, which also contains the highest sonority vowel). For words of this type the probability of stress in initial position is given as:

$$p(c_1 \mid d, y_x) = (1-2\alpha)P(H_i^{\alpha} \mid d) + \alpha\frac{\alpha}{1-2\alpha}P(H_i^{\alpha} \mid d) + \alpha\frac{\alpha}{1-2\alpha}P(H_i^{\alpha} \mid d) \tag{C5}$$

And the probability of stress in second position is given as:

$$p(c_2 \mid d, y_x) = (\alpha)P(H_i^{\alpha} \mid d) + (1-2\alpha)\frac{\alpha}{1-2\alpha}P(H_i^{\alpha} \mid d) + (1-2\alpha)\frac{\alpha}{1-2\alpha}P(H_i^{\alpha} \mid d) \tag{C6}$$

Comparing the probability of stress in first versus second position,

$$\frac{p(c_1 \mid d, y_x)}{p(c_2 \mid d, y_x)} = \frac{(1-2\alpha)p(H_i^{\alpha} \mid d) + \alpha\frac{\alpha}{1-2\alpha}p(H_i^{\alpha} \mid d) + \alpha\frac{\alpha}{1-2\alpha}p(H_i^{\alpha} \mid d)}{(\alpha)p(H_i^{\alpha} \mid d) + (1-2\alpha)\frac{\alpha}{1-2\alpha}p(H_i^{\alpha} \mid d) + (1-2\alpha)\frac{\alpha}{1-2\alpha}p(H_i^{\alpha} \mid d)} \tag{C7}$$

Factoring out the $p(H_i^{\alpha} \mid d)$ term and simplifying gives

$$\frac{p(c_1 \mid d, y_x)}{p(c_2 \mid d, y_x)} = \frac{\frac{(1-2\alpha)^2}{(1-2\alpha)} + \alpha\frac{\alpha}{1-2\alpha} + \alpha\frac{\alpha}{1-2\alpha}}{3\alpha} \tag{C8}$$

Collecting terms,

$$\frac{p(c_1 \mid d, y_x)}{p(c_2 \mid d, y_x)} = \frac{(1-2\alpha)^2 + 2\alpha^2}{3\alpha(1-2\alpha)} \tag{C9}$$

$$\frac{p(c_1 \mid d, y_x)}{p(c_2 \mid d, y_x)} = \frac{6\alpha^2 - 4\alpha + 1}{3\alpha(1-2\alpha)} \tag{C10}$$

The behavior of this ratio in the region where $\alpha \leq .33$ is plotted in Figure C1.

In Optimal Bayes, the less dominant hypotheses can, in a sense, collude to move stress to their mutually preferred location. This effect will be strongest when the dominant hypothesis is only slightly better than its competitors (1 data point), and the competitors agree on their stress prediction (e.g., penultimate position). Thus, the formula above illustrates the largest effect size that can be expected.

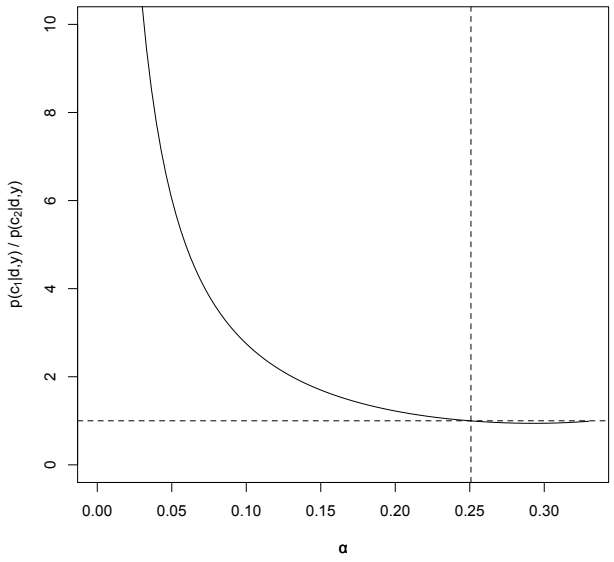

Figure C1
Classification probability ratio: $\frac{p(c_1 \mid d,y)}{p(c_2 \mid d,y)}$ as a function of α, for the three-hypothesis case, with i-j = i-k = 1. Stress in c_2 position is slightly preferred over c_1 for values of $\alpha \geq .25$ (indicated by dashed line).

The gang-up phenomenon, where H_j and H_k agree with each other in opposition to H_i, can be seen to have an appreciable effect in the region $.25 < \alpha < .33$. In this region the two stress positions have roughly equal probabilities. However, recall that α is the probability assigned to *each* of two exceptional stress positions. An α of .25 means that there is only a 50% chance of stress being assigned by rule. For a still high exception rate of 25% (an α of .125), c_1 is more than

twice as likely as c_2, and this discrepancy only increases as α decreases. Thus, it can be seen that using the Optimal Bayes Classifier has relatively little effect on the outcome.

Appendix D: Derivation of Information Theoretic Prior

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 (Rissanen 1989).

$$L(d,H) = L(d \mid H) + L(H) \tag{D1}$$

The relation of (D1) to Bayes' Theorem becomes clear when using the transformation from probability to optimal code length given by

$$L(x) = -\log P(x) \tag{D2}$$

Intuitively, Equation (D2) calls for assigning shorter length codes to higher probability symbols *x*. On average, this will minimize the code length for a string, *d*, of symbols drawn from distribution $P(x)$. For a binary alphabet, the logarithm is taken to be base 2. Re-writing Bayes' Theorem in the following way, $p(H,d) = p(d \mid H)p(H)$, taking the negative logarithm, and applying Equation (D1), returns Equation (D2). The close relationship between the two formalisms lends itself to the mapping of prior probability to hypothesis complexity, or coding length: the more bits it takes to spell out a given hypothesis, the lower its prior probability (and the lower its explanatory power). Under this transformation, $L(H)$ corresponds to $-\log_2 p(H)$ which means that $p(H)$ corresponds to $2^{-L(H)}$.

In the absence of any hypothesis, or stress generating rule, a certain number of bits per word will have to be used to indicate stress location. The coding length of the data will go up. With a hypothesis this cost per word is avoided, because there is a function that can be applied to the underlying form to determine stress placement. The tradeoff is that the hypothesis itself must be described so that the stress location can be computed. A cost is incurred dependent on how many bits it takes to completely specify the hypothesis. In what follows the coding costs for the hypotheses $GUJARATI*^{\alpha}$ and $MAX(G*/G)^{\alpha}$ will be determined. Translated to prior probabilities

(the explanatory power term), these values will be combined with the previously calculated likelihood ratio to determine the conditions under which the anti-markedness grammar defeats the Mixture Grammar using the Bayesian evaluation metric.

To begin, consider the way in which the categorical hypothesis *GUJARATI** assigns stress. The grammar can be conceptualized as a decision tree over underlying forms something like that depicted in Figure D1[1]. In order to be able to specify the correct stress for any three-syllable word the *GUJARATI** hypothesis must allow for at least five consecutive determinations: 1) if the word contains /ə/ in penultimate position, then it will assign stress to that position; if not, 2) if the word has a /ə/ in initial position, then it will assign stress to that position; if not, 3) if the word has /ə/ in final position, then it will assign stress to that position; if not, 4) if the word has a mid-sonority vowel in penultimate position, then it will assign stress to that position; if not, 5) if the word has a mid-sonority vowel in initial position then it will assign stress to that position; if not, the word will be assigned penultimate stress (final stress is only allowed for the lowest sonority vowels in complementarity with the *GUJARATI* grammar in (1)).

[1] In keeping with Kiparsky's conjecture, I have been assuming that *GUJARATI** represents a true reversed-sonority hierarchy language. This entails that the grammar will treat the highest sonority vowels (/a/) as dispreferred stress carriers, even though the inventory of Gujarati′ actually contains no /a/'s, and thus no evidence to the learner regarding their behavior.

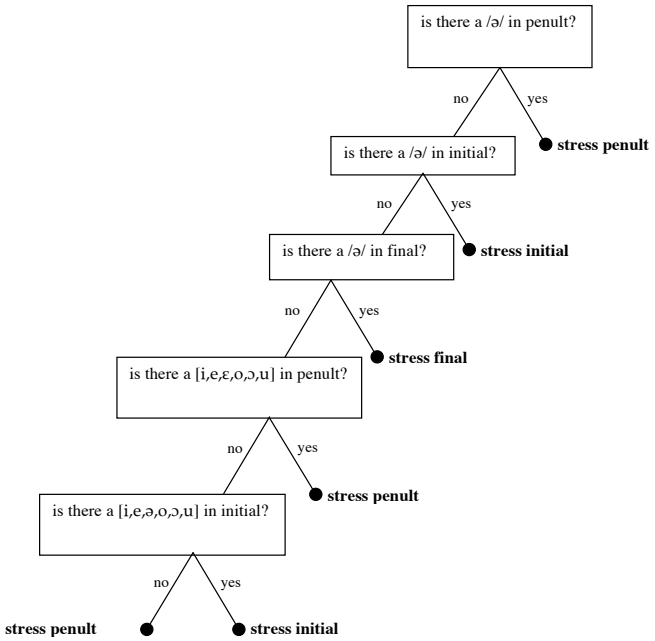

Figure D1
*GUJARATI** Hypothesis represented as a decision tree based on vowel type and position

The decision tree (*T*) in Fig. D1 requires a minimum number of bits to describe, which can be estimated using the binary coding scheme given in Rissanen (1989: section 7.2).

$$L(T) = \log\binom{k_T + m_T - 2}{k_T} \tag{D3}$$

Equation (D3) provides a measure of how much the grammar expressed by *T* compresses its input – or how many classes it must keep track of to produce the correct output. This is a function of k_T, the number of internal nodes of the tree, and m_T, the number of leaf nodes. For the *GUJARATI** grammar, $k_T = 5$ (corresponding to the relevant questions about vowel identity depicted in Figure D1), and $m_T = 6$ (corresponding to the possible stress decisions resulting from the answers to each of those questions).

The Variability version of *GUJARATI** additionally requires the estimation of one parameter: α. In general, for a hypothesis consisting of a set of q free parameters (θ), *L(H)* must

include the cost of estimating those parameters, as well as the length needed to encode the precision of each parameter. Asymptotically, for long strings of training data (large N; $d=\{y_i,\ldots y_N\}$) where precision can be ignored, the optimal code length for the maximum likelihood estimated parameters ($\hat{\theta}$) approaches Equation (D4) (Rissanen 1989: section 3.1).

$$L(\hat{\theta}) = \frac{q}{2}\log N \tag{D4}$$

Combining the length terms for the tree structure (T) and the estimated parameters (θ) gives:

$$L(GUJARATI^{*\alpha}) = \log\binom{k_T + m_T - 2}{k_T} + \frac{1}{2}\log N = \log\binom{9}{5} + \frac{1}{2}\log N \tag{D5}$$

$MAX(G*/G)^{\alpha}$ requires estimation of two parameters: w_{G*} and α (since $w_G = 1 - w_{G*}$, it does not have to be separately estimated from the data). The formulation in (B10) requires grammars for both *GUJARATI** and *GUJARATI*, and a decision node connecting the two trees. The total coding length of $MAX(G*/G)^{\alpha}$ is thus given by

$$L(MAX(G^*/G)^{\alpha}) = \log\binom{21}{10} + \log N \tag{D6}$$

Converting Equations (D5) and (D6) via Equation (D2) determines the ratio of prior probabilities for $GUJARATI^{*\alpha}$ and $MAX(G*/G)^{\alpha}$:

$$\frac{p(GUJARATI^{*\alpha})}{p(MAX(G^*/G)^{\alpha})} = \kappa\sqrt{N} \tag{D7}$$

where $\kappa = 2800$. The ratio of the prior probabilities depends on the length of the string, or the total amount of data to be transmitted. In order to see how factoring in prior probabilities influences the outcome of learning a particular lexicon that is to be learned will have to be specified.

Appendix E: 2-syllable word types

Section 4 of Text

Table E1. Full set of all possible two-syllable word types for stress. Final column gives number of types and hypotheses with which the data are consistent. G* (*GUJARATI**), G (*GUJARATI*), P (*PENULT*). Forms consistent with none of the three hypotheses are denoted *A* (Arbitrary) (M is shorthand for any of the mid-sonority vowel class {i,e,ɛ,o,ɔ,u}). For two-syllable words, there are 8^2, or 64 types.

	Case Gujarati Vowel- Template	Example L > L′	# types H
1	(ə,a)	[pəgár] > [pəgə́r]	1 A
2	(M,a)	[ʃikár] > [ʃikə́r]	6 G*
3	(M,ə)	[díwəs] > [díwəs]	6 G, P
4	(a,a)	[rájja] > [rə́jjə]	51 G, G*,P
	(a,ə)	[gádʒər] > [gə́dʒər]	
	(a,M)	[p^{h}ájdo] > [p^{j}ájdo]	
	(ə,ə)	[bə́kbək] > [bə́kbək]	
	(ə,M)	[məso] > [məso]	
	(M,M)	[lék^{h}e] > [lék^{h}e]	

Appendix F: Competition between *GUJARATI$^{\alpha}$ and *MAX*$(G^*/G)^{\alpha}$ for L'_U**

Section 4 of Text

Calculating the posterior probability ratio between *GUJARATI**$^{\alpha}$ and *MAX*$(G^*/G)^{\alpha}$ will first require setting the following parameters: α for *GUJARATI**$^{\alpha}$, and α, w_{G^*}, and w_G for *MAX*$(G^*/G)^{\alpha}$. By definition, the second set of parameters will be set by maximum likelihood estimation. So as to give *GUJARATI**$^{\alpha}$ the best chance of winning the same method will be used to estimate its free parameter α.

Under L'_U for three-syllable words the proportions from Table 3 are used. To better approximate an adult-sized lexicon, the numbers are scaled up by a factor of 13, resulting in a total lexicon size of 6,656 words. To simplify the calculations a two-hypothesis competition will be used – excluding *PENULT* from consideration for the time being. This yields the following parameter values: $N = 6{,}656$, $a = 273$, $i = 1716$, $j = 1560$, $n = 3107$, $G^* = i+n = 4823$.

Derivation of Maximum Likelihood α under *GUJARATI**$^{\alpha}$:

From Bayes' Theorem (A1), and the definition of Variability Hypotheses in (A8), the likelihood of the data under *GUJARATI**$^{\alpha}$ is given by

$$p(d \mid GUJARATI^{*\alpha}) = \frac{\alpha^{N-G^*}(1-2\alpha)^{G^*}}{p(d)} \tag{F1}$$

Defining $L = \alpha^{N-G^*}(1-2\alpha)^{G^*}\,p(d)$, and taking the derivative with respect to α gives

$$\frac{\partial}{\partial\alpha}p(d \mid GUJARATI^{*\alpha}) = \frac{N-G^*}{\alpha}L - \frac{2G^*}{1-2\alpha}L \tag{F2}$$

The value of α that maximizes the probability of the hypothesis given the observed data occurs when the formula in (F2) is equal to zero,

$$\frac{N-G^*}{\alpha}=\frac{2G^*}{1-2\alpha} \tag{F3}$$

$$N-G^*-2N\alpha\ =0 \tag{F4}$$

$$\alpha\ =\frac{N-G^*}{2N} \tag{F5}$$

Derivation of Maximum Likelihood α under $MAX(G^*/G)^{\alpha}$:

The likelihood of d under $MAX(G^*/G)^{\alpha}$ is given above in (B12), and reproduced in (F6)

$$p(d|MAX(G^*/G)^{\alpha}=(1-2\alpha)^n\alpha^a(w_{G*}+(w_G-2w_{G*})\alpha)^i(w_G+(w_{G*}-2w_G)\alpha)^j \tag{F6}$$

As before, in order to find the maximum likelihood estimate for α, take the partial derivative of (F6) with respect to α and set it to zero. (F6) is maximized, under L'_U, and using the maximum likelihood formulae for w_{G^*} and w_G in B19 and B20, for an α of approximately .026. This is found by plotting the log likelihood in Figure (F1). The curve reaches its maximum at the intersection of the dotted line. This point also corresponds to the maximum likelihood weights $w_{G^*}=52.5\%$, and $w_G=47.5\%$.

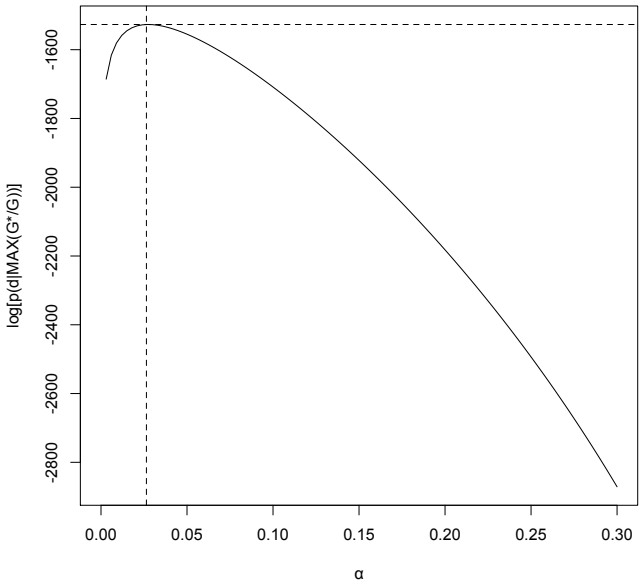

Fig. F1
Log(p(dlMAX(G*/G)) as a function of α. This curve is maximized at the intersection of the dashed lines.

2-Hypothesis Competition

The ratio of posterior probabilities for the competing hypotheses is given by:

$$\frac{p(GUJARATI^{*\alpha} \mid d)}{p(MAX(G^*/G)^{\alpha} \mid d)} = \frac{p(GUJARATI^{*\alpha})}{p(MAX(G^*/G)^{\alpha})} \frac{p(d \mid GUJARATI^{*\alpha})}{p(d \mid MAX(G^*/G)^{\alpha})} \tag{F7}$$

The likelihoods and posteriors will be very small, so logarithms are used to avoid precision loss.

$$\log\left[\frac{p(GUJARATI^{*\alpha} \mid d)}{p(MAX(G^*/G)^{\alpha} \mid d)}\right] = \log\left[\frac{p(GUJARATI^{*\alpha})}{p(MAX(G^*/G)^{\alpha})}\right] + \log\left[\frac{p(d \mid GUJARATI^{*\alpha}}{p(d \mid MAX(G^*/G)^{\alpha})}\right] \tag{F8}$$

Plugging in the Information Theoretic prior from (D7) gives

$$\log\left[\frac{p(GUJARATI^{*\alpha} \mid d)}{p(MAX(G^*/G)^{\alpha} \mid d)}\right] = \log\left[2800\sqrt{N}\right] + \log(p(d \mid GUJARATI^{*\alpha})) - \log(p(d \mid MAX(G^*/G)^{\alpha})) \tag{F9}$$

For $GUJARATI^{*\alpha}$ under L'_U (Eq F5)

$$\alpha = \frac{N - G^*}{2N} \cong .138$$

when $N = 6{,}656$, and $G^* = 4823$.

$$\log[p(d \mid GUJARATI^{*\alpha})] = \log[(\alpha)^{1813}(1-2\alpha)^{4823}] \cong -2253.1$$

The log-likelihood of the data given $MAX(G^*/G)^{\alpha}$ can be read off of Fig (F1), giving:

$$\log\left[\frac{p(GUJARATI^{*\alpha} \mid d)}{p(MAX(G^*/G)^{\alpha} \mid d)}\right] = 5.4 - 2253.1 + 1526.7 \cong -721$$

The log of the posterior ratio is much less than zero, which means that the ratio of the posteriors is much less than 1, and $MAX(G^*/G)^{\alpha}$ is still the winner even with the bias towards $GUJARATI^{*\alpha}$ from the information theoretic prior. While the description-length prior does shift the outcome of the competition by a few orders of magnitude (2.2×10^5), the discrepancy of descriptive power between the two different hypotheses is so large that the overall result is largely unaffected. In order to balance out the lower probability $GUJARATI^{*\alpha}$ assigns to the data, the ratio of the priors would have to be on the order of 10^{726}! The difference in complexity, or description length, between the two hypotheses, as can be seen, doesn't come anywhere close to this value[2].

Another way to think about Equation (F9) is in terms of competition thresholds. For the Simple anti-markedness grammar to defeat the Mixed Grammar hypothesis the posterior probability ratio must be greater than 1 (and thus, the log posterior probability must be greater than 0). Under L'_U this clearly does not occur.

[2] There is at least one caveat related to the calculation of this information-theoretic prior; the value may depend on the particular coding scheme used. 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 (D2) 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)). However, it can be seen that, due to the extremely large numbers involved, small adjustments are unlikely to significantly affect the result.

To determine the lexical conditions which are favorable to the anti-markedness grammar the values of the parameters *j* and *i* can be varied. To simplify things keep *n*, *N* and *a* constant – which will also keep the prior probability ratio constant. For the pure anti-markedness grammar to win under the Bayesian evaluation metric *i* must be significantly greater than *j*. Beginning with the numbers for L'_U and systematically decreasing *j*, while increasing *i* by the same amount (δ), the exact location at which the posterior probability crosses the zero point can be determined. This occurs at δ =1406: i = 3122, and j = 154. This is a data ratio of roughly 20. In order to reject a Mixture Hypothesis where both sonority hierarchies are maintained, *GUJARATI** must account for about twenty times more *unambiguous* data than *GUJARATI*[3].

[3] 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. One is still left, however, with the problem of determining the prior probability distribution over the weights. In order to assess the outcome of learning in the absence of any influence of UG, this needs to be done in a manner which is independent of the linguistic problem at hand.

Appendix G: Lexicons

Section 5.2 of Text

G.1 Sampling

For Section 5.2 a set of lexicons was created by repeatedly sampling (with replacement) from the full set of word types in Tables 3 and E1 at several different rates. Word type here is defined by the unique sequence of vowels within the word. This sampling allows for a kind of tuning of non-uniformity over lexicons. The lexicons are the product of the set of sampled vowel sequences which are randomly assigned consonants to become unique words. However, the more undersampled the space of possible vowel sequences is, the more non-uniform the lexicon is likely to become in sonority space. All else being equal, this should occur symmetrically, such that lexicons are equally likely to skew in any direction. Thus, a set of 1000 such randomly generated lexicons will have a broader distribution, in any given parameter space, the lower the sampling rate.

First, the total number of 3-syllable and 2-syllable words for each lexicon was fixed at 3,072, and 3,840, respectively (these numbers derive from scaling terms applied to the total number of 512 unique 3-syllable word types, and the 64 unique 2-syllable word types such that a roughly equal number of 3- and 2- syllable words result within a reasonably sized vocabulary). The probability over word types was uniformly distributed. For each lexicon, a certain degree of sampling was specified. This Degree indicated how many different word types would be used in the make-up of that lexicon. In the case of under-sampling, some word types were guaranteed to be excluded. This could also happen with full sampling and over-sampling, as the sampling was done with replacement. 4 different degrees of sampling were selected, with each percentage of types a factor of 10 smaller than the Degree below it. Degree 1 sampled three-syllable word

types at 600%, and 2-syllable word types at 6000%. These numbers map directly to the 3,072:3,840 word lexicon. Degree 2 under-sampled 3-syllable word types at 60%, but still over-sampled 2-syllable word types at 600%. Degree 3 under-sampled each at 6%: 60%. The final degree, Degree 4, under-sampled at .06%:.6%. In the case where under-sampling occurred, the number of types were duplicated as necessary to produce the total number of required unique words. For example, the Degree 3 lexicon contained at most 31 out of a possible 512 3-syllable word types, and 38 out of a possible 64 2-syllable word types. For the full-sized lexicon these projected to the fixed 3,072, and 3,840 words, respectively. Thus there was considerable duplication in the represented types, and, correspondingly, duplication in the sonority profiles of the words in Degree 3 lexicons. More fully sampled lexicons can be expected to demonstrate greater variety of types, and thus represent more fully the various sonority profiles illustrated in Tables 3 and E1.

Although the degree of under-sampling gives a measure of how skewed the type distributions can be expected to be for a given lexicon, it doesn't specify the exact nature of that distribution. For example, one lexicon generated with Degree 4 of under-sampling displays the following normalized vowel frequency distribution: a 33%, i 22%, e 16%, ə 16%, ɔ 5%, u 5%, ɛ 0%, o 0%. Identical with respect to degree of sampling, but very different with respect to stress distribution over words, is a lexicon with the following vowel frequencies (where only the vowel identities at each frequency level have changed) u 33%, ɛ 22%, i 16%, o 16%, ə 5%, e 0%, a 0%. Which particular vowels appear with any particular frequency is determined by random selection, and differs for each of the lexicons generated.

The method described above also does not control the *way* in which the lexicons are non-

uniform, only the *degree* to which they are. Although natural language lexical distributions are far from being universally and comprehensively characterized, it has been observed that a number of linguistic units tend to show a specific kind of non-uniform distribution. This distribution is one in which the highest frequency items are observed (in, e.g., a text sample) significantly more often than the next most frequent, and the largest number of different types is found at the lowest rate of occurrence. This kind of distribution has been noted for word and morpheme token frequencies, word lengths, and syllable counts. It has also been suggested for the distribution of phonetic or phonological units, in accordance with principles of articulatory markedness (Zipf 1949). Generally speaking, a distribution in which the absolute frequency of occurrence depends on the relative frequency of occurrence is known as a Zipfian distribution[4]. A particular instantiation of a Zipfian distribution (the standard harmonic) is characterized by the following formula

$$f \propto \frac{1}{r} \tag{G1}$$

which describes a dependency in which the frequency of a type (*f*) is proportional to its rank frequency (*r*). In terms of the types of interest, namely vowels, this means that the second most frequent vowel will occur half as often as the most frequent vowel; the third most frequent vowel will occur one third as often, and so on.

Although the current sample of lexicons contains distributions that are at least as non-uniform as the Zipfian – for a given measure of non-uniformity – there are not necessarily any that are non-uniform in exactly the same way. Accordingly, a fifth set of 1000 lexicons was generated. Each lexicon of this new set was Zipfian in the distribution of its vowels. Sampling occurred over vowels themselves rather than sequences of vowels, but random selection

[4] Thanks to an anonymous reviewer for suggesting consideration of this type of vowel distribution.

determined precisely which vowels corresponded to which frequency rank for each lexicon.

G.2 Homophony avoidant sound change

It should be noted that the competition *GUJARATI** faces from *GUJARATI* is due to the existence of a residue of natural patterns in the post-sound change language: a certain proportion of forms whose surface [ə]'s were historically /ə/'s, rather than deriving from /a/'s. Consider the three-syllable words classified in Table 3, reproduced here as Table G1 for ease of reference. The residual natural pattern is contained in rows 4 and 5, whereas the decisive anti-markedness patterns are evident in rows 2 and 3. The difference between the rates of occurrence of these groups can be roughly characterized as the difference between the rates of occurrence of /a/ and /ə/ in Gujarati. If the frequency of /ə/ is appreciably lower than /a/, then the frequency of words in rows 4 and 5, all else being equal, is analogously less than the frequency of words in rows 2 and 3.

Table G1. Full set of all possible three-syllable word types for stress. Final column gives number of types and hypotheses with which the data are consistent. G* (*GUJARATI**), G (*GUJARATI*), P (*PENULT*). Forms consistent with none of the three hypotheses are denoted A (M is shorthand for any of the mid-sonority vowel class {i,e,ɛ,o,ɔ,u}).

	Case Gujarati Vowel-Template	Example L > L′	# types H
1	(ə,ə,a)	[pəkʃəpát] > [pəkʃəpə́t]	21 A
	(ə,M,a)	[pəɾikʃá] > [pəɾikʃə́]	
	(a,ə,M)	[tábəɖtob] > [tə́bəɖtob]	
	(M,ə,a)	[ucc^{h}əvás] > [ecc^{h}əvə́s]	
	(a,ə,a)	[ɟáɟərman] > [ɟə́ɟərmən]	
	(a,ə,ə)	[pá[nəgər] > [pə́[nəgər]	
2	(M,M,a)	[hoʃijáɾ] > [hoʃijə́ɾ]	84

	(a,M,M)	[ʃáririk] > [ʃəririk]	G*
	(a,M,a)	[háɖohaɖ] > [həɖohəɗ]	
	(a,M,ə)	[pʰấsigər] > [pʰə́sigər]	
3	(M,a,a)	[durácar] > [durəcər]	48 G*,P
	(M,a,ə)	[mubáɾək] > [mubəɾək]	
	(M,a,M)	[betáɭis] > [betəɭis]	
4	(M,M,ə)	[tʃum:ótəɾ] > [tʃum:ótəɾ]	78 G,P
	(ə,M,ə)	[vəríʃtʰə] > [vəríʃtʰə]	
	(ə,M,M)	[kəʈóro] > [kəʈóro]	
5	(M,ə,M)	[kójəldi] > [kójəldi]	42 G
	(M,ə,ə)	[kʃétrəpʰəɭ] > [kʃétrəpʰəɭ]	
6	(a,a,a)	[aw:ánã] > [əw:ə́nə̃]	239 G,G*,P
	(a,a,M)	[amdáni] > [əmdə́ni]	
	(ə,a,a)	[resádar] > [resədər]	
	(ə,a,ə)	[səpʰácət] > [səpʰə́cət]	
	(ə,a,M)	[gʰəʈáɖo] > [gʰəʈə́ɖo]	
	(ə,ə,ə)	[əkbə́ndʰə] > [əkbə́ndʰə]	
	(ə,ə,M)	[cəkcə́kit] > [cəkcə́kit]	
	(M,M,M)	[iʈʰʈʰóter] > [iʈʰʈʰóter]	
	(a,a,ə)	[ɟʰagmágəʈ] > [ɟʰə́gmə́gəʈ]	

The repercussions of a large difference in relative frequency between /ə/ and /a/ can be seen in Table G2. Here, three representative lexicons from each type of sampling are selected: one that results in a Mixture outcome, one that results in a *GUJARATI** outcome, and one that results in a *GUJARATI* outcome. As Table G2 shows, the latter two types of lexicon only occur with Degree 4 and Zipfian lexicons, even for the lowest threshold values. However, for each of the instances that produce the *GUJARATI** outcome /a/ is considerably more frequent than /ə/, whereas in the Mixture outcomes, the two vowels are much closer together in frequency.

Table G2. Normalized Frequencies (Rounded)/Rank Order Frequency

Grammar Type	Lexicon Type	ə	a	e	ɛ	i	ɔ	o	u
MIXTURE	1	.12/2	.12/2	.12/2	.12/2	.12/2	.12/2	.12/2	.13/1
	2	.12/2	.11/3	.13/1	.11/3	.13/1	.11/3	.12/2	.13/1
	3	.11/5	.09/6	.14/2	.12/4	.15/1	.13/3	.11/5	.13/3
	4	.06/3	.11/2	.11/2	.22/1	.11/2	.11/2	.22/1	.06/3
	z	0.07/5	0.09/4	0.05/7	0.36/1	0.12/3	0.04/8	0.18/2	0.06/6
*GUJARATI**	1	--	--	--	--	--	--	--	--
	2	--	--	--	--	--	--	--	--
	3	--	--	--	--	--	--	--	--
	4	.16/3	.33/1	.16/3	0/6	.22/2	.05/4	0/6	.05/4
	z	0.06/6	0.36/1	0.04/8	0.07/5	0.18/2	0.12/3	0.09/4	0.05/7
GUJARATI	1	--	--	--	--	--	--	--	--
	2	--	--	--	--	--	--	--	--
	3	--	--	--	--	--	--	--	--
	4	.28/1	.11/3	.17/2	.17/2	0/5	.17/2	.06/4	.06/4
	z	0.36/1	0.06/6	0.04/8	0.12/3	0.05/7	0.07/5	0.09/4	0.18/2

If lexicons in which /a/ was appreciably more frequent than /ə/ were themselves more likely to occur, then the random sampling assumption of the previous section would not hold, and the estimate for the expected numbers of *GUJARATI** grammars would go up. I can think of no reason why this should be true, however. On the other hand, if the sound change a > ə were more likely to apply to lexicons in which /a/ was appreciably more frequent than /ə/ then the expected *GUJARATI** numbers could also go up. It is that scenario that is now examined.

There is a long-standing intuition in the field that sound changes are more likely to occur if they do not neutralize contrasts (Martinet 1955)[5]. Contrast is achieved by mapping different sounds to different meanings. And while this is not always a one to one mapping, the hypothesis is that the communicative need to reduce ambiguity limits the amount of homophony in any given language. One way this limit can be maintained is by disallowing sound changes that would increase homophony. When an inventory that contains both /ə/ and /a/ is reduced to

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

one that contains only /ə/, a contrast has been removed. Words that differed depending on whether they contained /ə/ or /a/ are now phonologically identical. However, if the original inventory contained few /ə/'s (low token frequency), the amount of neutralization this sound change would introduce is minimal.

Like many of the linguistic ideas already examined, the intuition about homophony avoidance is hard to implement. The exact role that homophony avoidance plays in historic change is not known, or how the allowable level of ambiguity should be measured, or how to use such a measure (see Surendran and Niyogi (2006) for a discussion of these questions). However, it seems safe to assume that a language with *no* prior contrast between /ə/ and /a/ would be unaffected by functional pressures against neutralization. The No-Contrast language will therefore provide a benchmark as potentially the most likely language to undergo the sound change, as well as the language with the least residual data compatible *only* with *GUJARATI* (i.e. none).

For No-Contrast Lexicons, L_{NC}, all data in Gujarati′ are consistent with the *GUJARATI** hypothesis; *GUJARATI** is the clear winner in a simple categorical framework. But, as argued previously, such a learner is incapable of coping robustly with conflicting data. Allowing for exceptions, with Variability grammars allowed into consideration, $PENULT^{\alpha}$ remains a competitor. In this scenario (7 historic vowels, rather than 8), all 343 types of 3-syllable words are stressed consistently with the *GUJARATI** hypothesis, while 265 are also consistent with *PENULT*. 2-syllable words provide somewhat less of an advantage to the anti-markedness grammar with only 3 word types that are consistent with *GUJARATI** alone, at a ratio of 49 to 46.

To approximate an upper bound on the probability of a pure *GUJARATI** outcome further simulations were run. This time only the Zipfian distribution was used, and 1000 lexicons were generated using the 7-vowel inventory, for both 2- and 3- syllable words. Doing so resulted in a 29.5% rate for the reversed sonority-to-stress grammar at the lowest threshold level (1.25). In the other 70.5% of cases *GUJARATI** fails to exceed the minimum threshold level of descriptive advantage over *PENULT*, leading to a Mixture Grammar outcome. Note that this does not take into account the fact that *PENULT* is a simpler hypothesis than *GUJARATI** in information theoretic terms. The upshot being, that even with a non-neutralizing sound change, the anti-markedness outcome is not a clear and compelling winner.

This result cuts in the other direction as well. That is, *GUJARATI* faces the same competition from *PENULT* under both pre-sound change conditions, as well as natural sound change conditions (row 6 of Table 5). Simulations run under L_{NC} (prior to sound change) result in a default *GUJARATI* grammar 28.9% of the time under the 1.25 proportion threshold; the Mixture *GUJARATI*/*PENULT* results in the remaining 71.1% of cases. The outcome will be similar for the full 8-vowel inventory. Because there is such a large proportion of word types with penultimate stress – consistent with both hypotheses – there is typically not enough evidence to reject *PENULT* outright. As before, a *GUJARATI* default only emerges when *GUJARATI* captures significantly more data than *PENULT* alone.