Can Phonological Universals be Emergent?
Modeling the space of sound change, lexical distribution, and hypothesis selection.

Rebecca L. Morley
The Ohio State University
Department of Linguistics
200 Oxley Hall
1712 Neil Ave.
Columbus, OH 43210

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. Such a claim is based on the following assumptions: that phonological typology shows a highly asymmetric distribution, and that such a distribution cannot possibly arise 'naturally' – that is, without a UG-based restriction of the learner's hypothesis space. Attempting to test this claim reveals a number of open issues in linguistic theory. In the first place, there exist critical aspects of synchronic theory that are not specified explicitly enough to implement computationally. Secondly, there remain many aspects of linguistic competence, language acquisition, sound change, and even typology that are still unknown. It is not currently possible, therefore, to reach a definitive conclusion about the necessity, or lack thereof, of an innate substantive grammar module. This paper thus serves two main functions; acting as a pointer to the areas of phonological theory that require further development – especially at the overlap between traditionally separate sub-domains; and as a template for the type of argumentation required to defend or attack claims about phonological universals.

Keywords: Evaluation Metric; Universal Grammar; Sonority-Based Stress; Statistical Learning; Sound Change; Computational Models

1 Introduction

One of the central goals of linguistic theory is the characterization of the set of universal properties of human language. Because of wide-spread surface dissimilarities it is generally assumed that such universals are to be found in deep/abstract grammatical properties. These posited universals may reside in domain-general aspects of human cognition: perception, memory, or learning, or they may reside in domain-specific linguistic structures. Language-specific universals may be algorithmic (e.g., a violation-minimizing function for assigning grammaticality judgments), or substantive (e.g., a violable constraint against sequential consonants). Proposed universals may be absolute, or they may be better characterized as tendencies or biases.

One way to discover universal properties of language is through logical inference: given what is observed to be true of natural language as a whole, what *must* be true of the universal language endowment? A unique answer proves the existence of one or more universals. The jumping off point for this paper comes from the following intuition: without a mechanism that directly limits possible human languages, significantly more diversity in linguistic typology would be observed. Another way to frame this argument is that a curb on the results of 'blind' sound change is required to explain the fact that 'unnatural' systems do not proliferate (e.g., de Lacy 2006; de Lacy & Kingston 2013; Kiparsky 2006, 2008; Moreton 2009). The necessity of Universal Grammar follows from this if one assumes that there is no mechanism, or set of mechanisms, that could indirectly shape the typology in the necessary way. In other words, the above argument relies on the assumption that universals cannot be emergent from the cycle of language change and language acquisition. I will call this the UG-Delimited $\mathcal{H}$ Principle. In fact, a plausible test of this principle has yet to be undertaken. The present study will attempt to do exactly that.

In the next section a specific case study will be introduced in order to illustrate what is involved in testing the UG-Delimited $\mathcal{H}$ Principle. The focus will be on the conditions under which supposedly unnatural, unobserved, and, specifically, 'anti-markedness' grammars might arise. An anti-markedness grammar is one that directly contradicts a purportedly absolute implicational hierarchy. It will be important to keep in mind that the question about the emergence of such a grammar – or indeed any grammar -- is really a question about two events: a sound change that gives rise to a lexicon that *could have been* generated by the grammar in question, and the induction from that lexical data, on the part of the learner, of that specific grammar.

In Section 3 the behavior of a set of learners is examined; these are variations on a basic statistical learner operating over a series of (more inclusive) hypothesis spaces. In implementing these various learners it quickly becomes clear that there is a basic incongruity between the conventions of theoretical linguistics and the products of a linguistically naive learner. That is, it is not possible to map the conditional probabilities assigned to various hypotheses directly to an unambiguous answer regarding the likelihood of an anti-markedness grammar. One of the reasons for this is the lack of explicitness in the definition of critical linguistic terms, most notably the uncertain status of the 'lexical exception'; another is the incompatibility of the tools of statistical learning with standard linguistic assumptions (even under probabilistic formulations).

In order to coerce an interpretable mapping, a set of post-hoc significance thresholds is set in Section 5. Via Monte Carlo simulations over lexical vowel frequencies the conditions under which the learner will arrive at a grammar that is markedness-abiding, markedness-violating, or a mixture of the two is determined for this set of thresholds. For a final assessment of the likelihood of each type of grammar, the evidence for the sound change and the typological claim are re-evaluated in Section 6. In Section 7 the results are summarized and discussed.

What this paper demonstrates is that in order to even approach the goal of testing the UG-Delimited $\mathcal{H}$ Principle many and larger questions must be answered. These include questions about the exact make-up of the input to the learner, the learner's hypothesis space, and the precise nature of any sound changes that could lead to a grammar that is predicted to be impossible. In fact, our collective knowledge in each of these domains is far too incomplete for any definitive conclusions to be drawn. The results vary considerably, depending on which of a large set of possible assumptions are chosen. Thus, via computational modeling, the actual predictions of the theories under test are shown to be significantly underspecified. Ultimately, the major contribution of this work lies in illustrating the necessity of explicitness (explicitness to the point of computational implementation) in linguistic theory. The paper concludes in Section 8, with a discussion of the ramifications more generally for formally testing linguistic theories.

2 The Case Study

Although the results of this work are intended as general ones, the first step will be to select a particular case study: a particular hypothetical phonological system, coupled with a particular hypothetical sound change, resulting in a particular hypothetical lexicon, input to a particular hypothetical learner. For this purpose a scenario outlined in Kiparsky (2008) will be adopted. This scenario is argued to provide compelling evidence for the logical necessity of restricting the space of learnable grammars.

An Indo-Aryan language spoken in India, Gujarati is chosen as a concrete illustration of a sonority sensitive stress system that respects the posited universal implicational hierarchy; stress is preferentially assigned to higher sonority over lower sonority vowels, such that if a given vowel is a possible stress carrier, then any higher sonority vowels will also be possible stress carriers (e.g., Kenstowicz 1996).

According to de Lacy (2006) 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].

(1) GUJARATI: Sonority & Position –to-Stress:

(a) stress penultimate [a] (the most sonorous vowel)

[1] The situation is possibly more complicated than this. Doctor (2004) lists 10 basic vowels for Gujarati, including low-high front and back vowels. He also reports that nasalization, breathiness and length contrast on a subset of these vowels. Furthermore, Doctor cites syllable weight and morphology as factors in stress assignment. For the purposes of the general argument in this paper these complications will be ignored.

(b) otherwise stress ante-penultimate [a]

(c) otherwise stress final [a]

(d) otherwise stress penultimate mid-sonority vowel (any of [i,e,ɛ,o,ɔ,u])

(e) otherwise stress ante-penultimate mid-sonority vowel

(f) otherwise stress the penultimate position (which must be [ə], the lowest sonority vowel)

Sonority-sensitive stress systems in general are easily describable within a standard Optimality Theoretic framework that utilizes a universal sonority scale aligned with a foot prominence scale. Sonority is correlated with both height and peripherality. The expanded scale is given as: *P-foot/ə >> *P-foot/i,u >>*P-foot/e,o >> *P-foot/ɛ,ɔ >> *P-foot/a: where *P-foot/V is interpreted as the dispreference for having the vowel V as the peak of the foot, that is, the stressed position; since the scale is expressed negatively, the lowest ranked constraint expresses the strongest preference for stress (see, for example, Kenstowicz (1996), as well as Crosswhite (2000) and Smith (2000) for related phenomena, and Prince and Smolensky (1993/2004) for a more general discussion of prominence scales, but de Lacy (2006) for an alternative approach). These types of stress systems are not uncommon among the world's languages; between them Kenstowicz (1996) and de Lacy (2007) list at least 13 languages whose stress patterns they analyze as being sensitive to vowel sonority in accordance with this scale. Crucially, however, the reverse type of system, in which lower sonority vowels are the ones that attract stress, is so far unattested, and considered to be impossible within some theoretical frameworks (e.g., the foregoing OT account which disallows any re-ranking that would change the relative order of the above constraints).

The description of the stress rules of Gujarati can be characterized as the *GUJARATI* grammar. This grammar is a combination of sonority sensitivity, penultimate position bias, and avoidance of final stress. The grammar that assigns stress strictly to penultimate word position – *PENULT* – is thus a subset of the

GUJARATI grammar. This grammar will be considered as a competitor during the learning task described in Section 3.1. Of primary interest is the hypothetical reversed sonority-to-stress grammar, *GUJARATI**, and whether it is able, or likely, to emerge under unrestricted sound change. These three grammars will comprise the preliminary Simple Hypothesis space, defined in (2).

(2) $\mathcal{H}_i$: Simple Hypothesis Space

(a) PENULT: Stress Penultimate vowel

(b) GUJARATI: Sonority & Position –to-Stress [see (1)]

(c) GUJARATI*: Reversed-Sonority & Position –to-Stress [as in (1), but with the sonority classes reversed (i.e., /ə/ and /a/ exchanged].

The learning problem is conceptualized as the choice of a winning grammar from this space of competitors. Each competitor represents a hypothesis about the generative process underlying the set of surface forms to which the learner is exposed. By definition, a given learner can only learn a grammar if it has been included in the hypothesis space. The UG-Delimited $\mathcal{H}$ Principle requires *GUJARATI** to be excluded from this space. This assumes that such 'incorrect' hypotheses – if included – would be the clear winners under certain conditions of sound change, thus producing unattested and anti-markedness languages.

2.1 Sound Change

The specific sound change proposed in Kiparsky (2008) is one that would render the highest sonority vowels into the lowest sonority vowels: a > ə. If an unconditioned sound change such as this went to completion, the argument goes, the stress pattern of Gujarati would reflect a sonority *dispreference*. I interpret this argument to rely on the following chain of logic:

(3) UG-Delimited $\mathcal{H}$ Principle as applied to Gujarati case study

> Since the proposed sound change is widely attested, and since sonority-preference languages are similarly common, there is a high probability that the two will co-occur. Thus, a certain number of synchronic languages that are the result of this historic trajectory should be observable. In the absence of UG the grammars of these languages would reflect the changed pattern, becoming sonority-dispreferring

> grammars (*GUJARATI**). Since such grammars are, in fact, unobserved the existence of UG is corroborated.

In what follows each aspect of the argument in (3) will be scrutinized in detail. The unsupported assumption that UG is the only mechanism capable of performing the necessary filtering function has already been alluded to. Additionally, arguments based on absence from typology are problematic for a number of reasons. In the first case, the fraction of documented languages is quite small, making it likely that a particular instance could be missing from the sample by accident. Nor can rarity of occurrence be distinguished from impossibility of occurrence for the same reason.

The expectation of occurrence of the given historical trajectory (sonority-dependent stress system, affected by a > ə sound change at some point in its history) relies on a statistical argument. As such, it must take into account the size of the sample; an actual value for the estimated probability of each of the events must also be provided. Since the probability of the combined event is the multiplicand of the two probabilities (under assumptions of independence), this number might prove to be small enough to render it unlikely to occur in the given sample. In fact, the likelihood of the proposed sound change at all must also be called into question.

The argument in (3) also relies on the assumption that, despite the fact that vowel qualities have changed, and thus sonority, stress location does not shift. Low vowels are typically produced with greater duration than high vowels, and the ə symbol is often assigned to the most temporally reduced vowel in the inventory, as such shortening tends to result in a centralized realization (Lindblom 1963; Lehiste 1970; Kondo 1994). Tokens of /a/ that are less fully realized (e.g., shorter) may well merge with productions of the somewhat more centralized, shorter /ə/'s. That is, such tokens will be more likely to undergo the change than more fully /a/-like tokens. The phonetic realization of particular /a/'s, in turn, should be highly correlated with whether they are stress carriers or not (as stressed vowels tend to be more fully realized and longer than their unstressed counterparts (Lindblom 1963). Furthermore, energy is affected by vowel length, and sonority is at least dependent on energy, if not equivalent to it: thus the (universal) implicational sonority hierarchy. By the same token, Gordon (2006) finds total acoustic energy to be a fairly robust predictor of stress.

The point is that height, length, and sonority are not independent dimensions of variation. Any /a/'s which are likely to become /ə/'s (higher, shorter, less sonorous), are also less likely to be stress-carriers in the first place. Stress, therefore, has a very high probability of shifting to a different, higher energy location in the word during such a vowel quality shift, or even as a necessary precursor to such a shift.

However, despite the list of questionable assumptions upon which the argument in (3) depends, I argue that it is worth pursuing this case study. Although the specific sound change proposed may be implausible there could exist a scenario of historic change that would produce a similar result. Furthermore, in any scenario of sound change, the issues that arise with respect to learning, and selecting a winning hypothesis from a candidate space, are the same. They are the same, in fact, for synchronic linguistic theory as a whole, even disregarding the contribution of diachronic forces. One of the issues that looms large is the fact that grammars cannot actually be observed. The 'observation' of a particular type of grammar relies on the assumptions of the linguist, and two linguists may not agree on the correct analysis of a set of data. Thus, the scenario in (3) will be pursued, allowing an in-depth exploration of both the learning and the analysis problem. Additional problematic assumptions about the typology and natural sound change will be left aside until the end of the paper.

2.2 Word Change

The sound change a > ə, applying equally to all words of Gujarati, transforms the original lexicon, *L*, to the hypothetical future lexicon *L′*, belonging to the hypothetical language Gujarati′. Take the example of the word mubáɾək. In Gujarati′ it becomes [mubə́ɾək] (critically, in the absence of any repair involving a shift in the location of stress). This form now exhibits stress on the lowest (rather than the highest) sonority vowel in the word. This form could not be produced by the grammar given in (1). It could, however, have been generated by a rule that preferentially stresses lower sonority vowels (the *GUJARATI** grammar*)*.

The situation, however, is more complicated than this. The stress placement in [mubə́ɾək] is also consistent with a penultimate stress rule (the *PENULT* grammar). Furthermore, not all words of Gujarati will exhibit the {Mid, High, Low} vowel sonority profile that led to this outcome. In Table 1 the full set of stress types is given for three-syllable words (Gujarati word examples taken from de Lacy (2006) and Suthar (2003)). *L′* will consist of words with stress placement consistent with all of *GUJARATI*, *GUJARATI** and *PENULT* (row 1); words consistent with both *GUJARATI** and *PENULT* (row 2); words consistent with both *GUJARATI* and *PENULT* (row 3); words consistent with *GUJARATI* only (row 4); words consistent with *GUJARATI** only (row 5); and words whose stress is not predicted by any of the three grammars (row 6).

Table 1: All stress types for three-syllable words.
H: hypotheses from (2) consistent with stress type in that row under *L′*

	Example a > ə : L > L′	H
1	[g^{h}əʈáɖo] > [g^{h}əʈə́ɖo]	G,G*,P
2	[mubáɾək] > [mubə́ɾək]	G*,P
3	[kəʈóro] > [kəʈóro]	G,P
4	[kʃétrəp^{h}əɭ] > [kʃétrəp^{h}əɭ]	G
5	[háɖohaɖ] > [hə́ɖohəɗ]	G*
6	[tábəɖtob] > [tə́bəɖtob]	--

Table 1 illustrates that *L′* will contain ambiguous and contradictory evidence for a learner attempting to infer a generating stress grammar. These data could be analyzed as the product of a *GUJARATI* grammar, accompanied by a set of lexical exceptions (rows 2, 5 and 6). Such an analysis, in principle, poses no problem for current markedness theories, and argues against the UG-Delimited $\mathcal{H}$ Principle. But equally could these data be analyzed as the product of a *GUJARATI** grammar, accompanied by a set of lexical exceptions (rows 3, 4 and 6 this time). Thus, in order to proceed, it must be possible to determine which analysis would be chosen by a learner (equally, which analysis is 'better' in terms of linguistic theory).

In fact, synchronic stress patterns in a number of languages (if not all) exhibit exceptionality, sub-regularities, and inconsistencies. Liberman & Prince (1977), Kager (1989), Halle & Kenstowicz (1991), and Pater (2000), among others, are well-established analyses which make use of complex representational devices in order to achieve good descriptive adequacy for natural language stress patterns in English and other languages. However, a description of English stress as quantity sensitive (closed syllables heavy), rightmost extra-metricality, right-edge binary-foot parsing, and foot-left stress (Halle & Vergnaud (1987) cited in Pearl (2011)) misses a considerable amount of the data. Pearl (2011) found that a learning algorithm parameterized for the above dimensions failed to learn the "correct" stress rules of English when trained on a large corpus of transcribed speech. Even the best grammar was only correct on roughly 67% of the data. This finding suggests that the idealized generative analysis of English stress is, in fact, an inaccurate or incomplete description of speakers' competence. At the very least, the cited analysis requires at least 33% of all words to be labeled as exceptions.

Criteria for designating forms as exceptional are critical to the analysis problem. They are "essential for making lexical phonology work" Goldsmith (2002). In his terms, the question is "how much redundancy (i.e. patterning) must there be in the lexicon to make it "worthwhile" for the lexical phonology to set up a rule…?" This is the same question that must be answered in order to determine the outcome of learning over the forms of Gujarati′ which, in turn, will allow a test of the UG-Delimited $\mathcal{H}$ Principle. However, there currently exists no agreed upon formal definition of a grammatically exceptional item. A large part of the remainder of this paper will be devoted, therefore, to formulating a working definition, and a way to implement it. In the following section a number of current phonological theories will be evaluated with respect to their treatment of inconsistent data. The strict requirements of robust learning and a substantively unbiased learner will lead to the adoption of stochastic grammars. Competing grammars of this type will then be assessed by an evaluation metric based on the probability with which each grammar predicts the observed data.

3 The Learner

Consider first a parameterized learning space and a learner that sets each parameter to either 'on/off', 'yes/no', based on observed forms (such as the Trigger Learner of Gibson and Wexler (1994)). A single data point which is inconsistent with a given parameter setting will cause that parameter to be set to the opposite value. For example, a word like [am.dá.ni] would cause the hypothetical parameter Weight-To-Stress to be set to NO (given that the heavy initial syllable is unstressed). The assumption upon which such a learner is based is that all data are consistent; once an unambiguous data point has been encountered, the grammar can be set for all time. The Trigger Learner does not deal robustly with exceptions – a single data point is sufficient to cause this learner to categorically switch hypotheses. Faced with a finite set of randomly ordered inconsistent data, this learner will switch back and forth between hypotheses several times, finally settling on the hypothesis that is compatible with the last data point observed. What this means is that all learners of a language like Gujarati′ are not guaranteed to converge on the same stress grammar unless they all encounter the data in the same order.

The current learning problem is one in which all Simple Hypotheses (including those listed in (2)) are subject to numerous exceptions. Furthermore, those exceptions are not random; the stress position of most words, while irregular with respect to one hypothesis, is regular with respect to another. Even more problematic for most learning models, most of the irregular items for a given hypothesis can be characterized as being in direct contradiction to that hypothesis. That is, this is not a case of a rule applying or failing to apply, or of a set of predictable affixes which apply to disjoint sets of stems. Whatever learning algorithm is adopted for this problem must be able to strike a balance between faithfully accounting for each lexical form (descriptive adequacy), and generalizing enough to be able to predict stress placement for novel forms (explanatory adequacy).

The treatment of exceptions has a long history in linguistic theory. In Chomsky & Halle (1968) exceptions are dealt with by diacritics in the lexicon that either indicate the non-application of a general rule for a particular morpheme, or call for a different, 'minor' rule to be applied. Lexical Phonology (Kiparsky 1982) assigns different morphemes to different levels based in part on the order of their affixation; rules are marked as applying at a given level or not. Different classes of words can also be assigned to different 'strata' or 'domains' in Optimality Theory. Domain-specific constraint ranking produces the observed differences in the behavior of different sets of lexical items (Ito & Mester 2001). Also within an OT framework, certain words can be underspecified in the lexicon (allowing markedness constraints to determine their surface forms), whereas others are fully specified (and thus remain unchanged at surface, due to high-ranking faithfulness constraints) (Inkelas, Orgun, & Zoll 1997). However, in all the above frameworks, what belongs to which class is determined *post hoc*. That is to say, given the data, it is possible to formulate a description that achieves adequacy in each given framework. It cannot be known, in principle, however, whether descriptive adequacy could also be achieved with a different sorting of items into classes. There is, similarly, no way for the grammar to determine which class an item belongs to when encountering it for the first time.

It is not unreasonable to suppose that language learners form provisional hypotheses during the course of learning. The learner must be able to alter or revise these hypotheses as necessary, as new data are encountered. Thus, each new input constitutes a decision point. The learner can decide to maintain the current hypothesis, alter it somewhat, or discard it altogether. The first outcome would occur if the phonological form was consistent with the current hypothesis; the second or third outcome would result if an inconsistent form was encountered. The problem is in determining which of outcomes 2 or 3 will prevail. The learner has to decide if the new inconsistent form can be marked as an exception to the current hypothesis, or if it requires discarding the current hypothesis in favor of another that achieves better descriptive adequacy. This revision could involve switching the marks for certain previously learned forms or, in the limit, deciding to store all forms in the lexicon (in the absence of a sufficiently predictable rule).

If full lexical specification were the inevitable outcome of learning over the type of contradictory data of Guajrati′ then the UG-Delimited $\mathcal{H}$ Principle for this case study could be rejected. The same would be true if learning could be shown to result in an opaque interaction between the sonority sensitive stress assignment rule and a subsequently ordered synchronic rule of vowel shift: /a/→[ə]. That is, certain surface [ə]'s would derive from underlying /a/'s, while others would be underlyingly /ə/'s. This grammar requires the learner to posit a vowel in underlying representation that never surfaces. Since I know of no obvious reason to prefer either of these two outcomes, the remainder of this paper will focus on the possible outcomes which *could* motivate Kiparsky's argument, allowing the most favorable test of the UG-Delimited $\mathcal{H}$ Principle.

A stochastic grammar that can update its weights allows for a stress rule that can change as it receives new input. Such a grammar can decide how to classify new words based on the forms it already knows. The highest probability stress location will be the one that accords with the majority pattern, but stress will be assigned according to minority patterns with non-zero probability; and even stress locations that are incompatible with all the patterns in the data will have some probability of being stressed in order to account for the possibility of true exceptions.

There exist frameworks in linguistics designed to express variability or gradience in acceptability which allow for fully productive grammars in this way. In Stochastic OT constraints are given a continuously-valued weight that can be perturbed during evaluation. This effectively allows constraints to switch places in the ranking order (Boersma 1998; see Zuraw (2010) for an application to what she calls "patterned exceptionality"). Another class of phonological learning models selects grammars that are mixtures of multiple rules, constraints, or entire grammars that act on the same forms. The Minimal Generalization learner of Albright and Hayes (2003), for example, constructs rules iteratively over its training set, compiling a collection of context-dependent transformation rules ranging from the very specific (applicable to a single word), to the completely general (applicable to all words). Critically, rules of all levels of specificity and consistency are retained in the learning space, with reliability as a weighting factor. Within a constraint-based framework, the Maximum Entropy learner of Hayes and Wilson (2008) does something similar, discovering weighted phonotactic constraints from its input data and keeping track of a large number of correlations of various strengths over various units. Both models produce probability distributions over a set of possible outputs. The variational model proposed by Yang (1999, 2000) is one in which probabilities are assigned to grammars, and competence is modeled by a weighted distribution of multiple grammars, in the same way that the Expectation Maximizing learner of Jarosz (2006) stores probabilities, each associated with a complete constraint ranking and lexicon set.

For the purposes of this paper, it is desirable to adopt the most general learner, leaning as little as possible towards any particular theoretical position, and thus assuming no pre-existing set of properties or biases. Four types of general learner will be considered. These are differentiated on one dimension by representational complexity, and on a second dimension, by productivity. 'Exceptions' and 'Variability' hypotheses are Simple Hypotheses that allow for inconsistent data. In the former case, certain words are marked underlyingly as exceptions. However, this is a static class, and does not affect stress assignment to novel forms. In contrast, the ‘Variability’ hypotheses allow for the Simple Grammar to fail to apply with a certain probability to each new word.

‘Partitions’ and ‘Mixtures’ are the complex counterparts of ‘Exception’ and ‘Variability’ hypotheses, respectively. Complex is used here in a specific sense to denote the fact that Complex Grammars are made up of multiple Simple Grammars. Under a ‘Partition’ hypothesis each of these sub-grammars applies to a set of pre-specified forms. Whereas in ‘Mixtures’ the generating sub-grammar is selected stochastically for a given token.

These general learners encompass more specific theories. This can be seen in Table 2, where the foregoing linguistic frameworks have been sorted into equivalence classes. As stated above, ‘static’ hypotheses that limit exceptions to those known beforehand are not sufficient to model the learning problem, and thus neither Exception nor Partition Grammars will be considered further. In other words, despite the other ways in which these theories may differ from one another, all those listed in the corresponding cells are eliminated for the same reason of non-generality. In section 3.2 the fully general stochastic Variability and Mixture Hypotheses will be implemented in a computational framework that allows for the grammars to be tested against each other as the ‘best’ generator of the observed data.

Table 2. 4-way hypothesis space distinction by complexity, then productivity. The ‘Equivalence Class’ column lists the linguistic theories with the corresponding degree of productivity and complexity.

	Hypothesis	Lexical Items	Novel Items	Equivalence Class
Unpatterned Exceptions				
‘Exceptions’	Simple Grammar	Marked exceptions	Simple Grammar	Chomsky & Halle (1968) Inkelas, Orgun & Zoll (1997)
‘Variability’	Simple Grammar + Random Term	Random application parameter	H_i applies with probability $1- \alpha$; random output with probability α	
Patterned Exceptions				
‘Partition’	Collection of Simple Grammars	Marked as taking Grammar n	Majority grammar (?)	Kiparsky(1982) Ito & Mester (2001)

‘Mixture’	Collection of weighted Simple Grammars/rules/ constraints	Weights of each grammar/rule/ constraint	H_i^α applies with probability w_i	Boersma (1998)
				Zuraw(2010)
				Yang(1999)
				Jarosz(2006)

			Set of possible outputs with associated probabilities	Albright & Hayes (2003) Hayes & Wilson (2008)

3.1 Bayesian Inference

The learner’s objective is to determine the grammar of stress assignment in Gujarati′ based on observation of a set of stressed lexical items, here labeled as L'. The learner is assumed to have knowledge of the phonological form of each word, to perceive the location of stress unambiguously, and to be aware of the relative sonority of the vowels in their inventory. Crucially, however, this learner has no preset parameters, no knowledge of relative markedness, and no UG-based prior preference for certain grammars. The learning task is conceptualized as a competition between various stress-assigning hypotheses. Each hypothesis is assigned a score based on how well it individually accounts for L'. The winner is determined based on this score. Of particular interest is the grammar that represents an anti-markedness outcome. If this grammar can be shown to win then part of the argument in (3) for the UG-Delimited $\mathcal{H}$ Principle is supported. The next step, therefore, is to decide on the score that will be used to determine the winning grammar.

Numerous evaluation metrics are possible for calculating this score. Under the most stringent requirements for descriptive adequacy, the winning hypothesis must correctly predict *all* words. Explanatory adequacy is only considered for the set of hypotheses that pass this first criterion (see Chomsky 1965; Chomsky & Halle 1968). In practice, however, this is too stringent a requirement, especially in the present case where each Simple Hypothesis fails to correctly predict stress in a non-trivial number of words. Instead, each hypothesis can be given one score for its ‘degree of descriptive power’, and another for its ‘degree of explanatory power’.

It turns out that a general-purpose algorithm for combining these two scores is readily available. The two linguistically-based measures map relatively transparently to a widely used mathematical framework for learning: Bayesian inference (e.g., Kemp, Perfors & Tenenbaum 2007; Tenenbaum *et al*. 2006; Xu & Tenenbaum 2007; Chater *et al*. 2006; Kording & Wolpert 2006; Gopnik *et al*. 2004; Kersten & Yuille 2003; Tenenbaum & Griffiths 2001). Under this mapping, descriptive power is defined as the probability with which a given hypothesis predicts the observed set of data [$p(h|d)$]; explanatory power as the probability of the hypothesis itself [$p(h)$]. The evaluation metric multiplies the two quantities in the following expression that derives from basic principles of probability theory. This relationship is known as Bayes' Theorem, and is given in (4).

$$p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)} \qquad (4)$$

The resulting score is defined as the probability with which the *data* predict each *hypothesis*. One way to determine the winning hypothesis is to simply calculate the ratio of $p(h|d)$ for each pair of hypotheses. This method also eliminates the necessity of determining $p(d)$; as it is constant across the hypothesis space, it cancels out in the ratio[2].

For the problem at hand the members of *d* are stress assignments corresponding to each of the words of the lexicon *L′*. The conditional probability of a particular stress assignment for a given word, 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 (or underlying, unstressed form). As is usual, it will be assumed that the conditional probability of each surface stressed form is independent of any other. The probability of the set *d* given *h* and *y* can then be expanded as the product of the probability of each stressed surface form, given a particular grammar.

[2] This ratio score is also known as the Likelihood Ratio test, a statistical tool for determining which model provides a better fit of the data (e.g., Neyman & Pearson 1933).

3.2 The Hypothesis Space

The proper characterization of the hypothesis space is arguably the most complex problem that will be encountered in this paper. Therefore this section is rather long and mathematically intensive. In order to make the argumentation as clear as possible most of the derivations are confined to Appendices A and B. Additionally, implementation will be broken up into several stages in order to gain a sense of the behavior of the learning function. At each stage the current hypothesis space will be assessed as to whether or not it exhibits appropriate behavior with respect to the language learning task.

What will be shown in this section is that grammars that treat exceptional stress placement as a random function (Variability Grammars) offer little improvement in descriptive power over Simple categorical grammars (such as those in (2)). Grammars that determine stress placement via a random (weighted) selection of generating hypothesis (Mixture Grammars), on the other hand, can achieve much better descriptive power. In fact, as long as the weights can be fit from the data, Mixture Grammars are guaranteed to equal or exceed the descriptive power of any Variability or Simple Grammar. If the winner depends only on descriptive power, then Mixture Grammars will win.

Mixture Grammars are capable of fitting the data more closely than the other two types of grammar, and, concomitantly, they are more complex. To complete the analysis, the effect of this increased complexity will be calculated by assigning an explanatory power score to these hypotheses. Explanatory power will be calculated based on complexity, or hypothesis length. The higher the complexity – longer the description length – , the less explanatory power. As this calculation depends on the amount of data being learned, this will lead directly to consideration of the lexicon itself in Section 5.2.

There is no principled reason to assume that no new forms encountered by a speaker/learner will be exceptional. Therefore all hypotheses must be able to handle such forms. Effectively, what this means is that each hypothesis must predict the occurrence of such forms with non-zero probability. In terms of production, such hypotheses will assign different stresses with differing probabilities. This will allow them to closer approximate the actual distribution. These requirements are exactly what motivated the introduction of 'Variability' and 'Mixture' hypotheses in Table 2. Instantiations specific to the current case study are given in (5). These Variability Hypotheses are the counterparts of the Simple Hypotheses in (2). Variability Hypotheses are predicated on the assumption that there is a single majority rule of stress assignment, but that there exists a non-zero probability that stress might be assigned by some other rule. In this case, the 'exceptional rule' is taken to be random assignment of stress.

(5) $\mathcal{H}^{\alpha}$: Variability Hypothesis Space

(a) PENULT$^{\alpha}$: Stress Penultimate vowel, but with probability α of stress in each of other two possible locations

(b) GUJARATI$^{\alpha}$: Sonority & Position –to-Stress, but with probability α of stress in each of other two possible locations

(c) GUJARATI*$^{\alpha}$: Reversed-Sonority & Position –to-Stress, but with probability α of stress in each of other two possible locations

The calculation of descriptive power for the hypotheses in (5) will depend strongly on the value chosen for α. However, it can be shown that the choice of winning hypothesis does *not* depend on the total amount of data, or the number of exceptions to the dominant hypothesis[3]. The only other number that affects the ratio score is the *difference* in number of exceptions between the competitor hypotheses. To illustrate this dependence, the problem is simplified somewhat to consider only two competitor grammars: *GUJARATI*$^{*\alpha}$ and *GUJARATI*$^{\alpha}$. For convenience a single α value will be used for both hypotheses. In the special case where the difference in number of exceptions between the two hypotheses is a single word (the difference in number of words belonging to (Rows 3 & 4) versus (Rows 2 & 5) of Table 1), the ratio score reduces to:

[3] To improve learning of difficult-to-set parameters Pearl (2011), and Pearl and Weinberg (2007), have proposed consideration of unambiguous data only. The repercussions of allowing different types of ambiguous data into the learner's input is also examined in Pearl and Lidz (2009). This approach is, unfortunately, unproductive in the current case. Typically, Bayesian learning research focuses on the subset/superset problem. Does the learner choose the most specific hypothesis that fits their training data, or the most general, or something in between? Such a learner is only given data that are equally consistent with more than one grammar. With data that are consistent with mutually exclusive grammars, excluding ambiguous data makes no difference.

$$\frac{p(Gujarati^{*\alpha} \mid d)}{p(Gujarati^{\alpha} \mid d)} = \frac{(1-2\alpha)}{\alpha} \qquad (6)$$

See Appendix A for the full derivation of Equation (6). As α gets smaller, this ratio gets larger – meaning that the less likely exceptions are, the more a single one counts against a given hypothesis. For an α in the middle of the range (=1/6), a GUJARATI$^{*\alpha}$ with one fewer exception than GUJARATI$^{\alpha}$ is the winner by a factor of 4. The ratio also grows exponentially by the difference in exceptions between the two hypotheses. Although inconsistent forms are now allowed, these exceptional data are so dispreferred by the Variability Hypotheses (assigned such a low probability) that the hypothesis with the fewest observed inconsistencies will always emerge as the winner – by an apparently insurmountable margin[4].

In fact, analogously with Simple Hypotheses, failure of descriptive adequacy still comes down to a single word – one treated as an exception by one hypothesis, but as non-exceptional by the other. This could be an acceptable result under the following two conditions: exceptions are truly random, and all possible competitors have been included in the hypothesis space. However, neither of these conditions is met. What the Variability Hypotheses ignore is the fact that exceptional stress placement is not completely arbitrary. Mixture Hypotheses allow the grammar to capture the patterning of the exceptional-stress words.

(7) $\mathcal{H}_{ijk}^{\alpha}$: Mixture Hypothesis Space

Stress is assigned by H_i^{α}, with probability w_i ; by H_j^{α}, with probability w_j; or by H_k^{α} with probability w_k

Using Variability Hypotheses as the component hypotheses allows Mixture Grammars to remain robust in the face of forms that are exceptions to *all* sub-hypotheses, while capturing the fact that certain forms are entirely consistent with one (and only one) of the component hypotheses. The Mixture Hypotheses defined in (7) assign stress based on the stochastic selection of one of the simple

[4] This is true even for a different classification technique known as Optimal Bayes, which retains all hypotheses, assigning them weights determined by how well they predict a set of training data. See Appendix C for details.

sub-hypotheses. This means that identical inputs will not always be stressed identically. On average, the likelihood of a given stress location is given by the probability of that location receiving stress under all component sub-hypotheses. Effectively, a set of candidate outputs with associated weights are generated for each input. Such weights can be interpreted as degree of well-formedness, and/or taken to correspond to variable stress placement. Thus, any candidate output with non-zero probability can be taken as a possible production of the speaker possessing a Mixture Grammar. Whether such variability applies at the lexical level or the word-type level depends on whether one assumes that learners learn a single ‘correct’ stress location for a subset of items or not.

In what follows certain types of Mixture Grammars will be selected for comparison in order to illustrate how their inclusion drastically changes the learning results. What will be found is that every time the complexity of a hypothesis is increased in order to increase its descriptive power it becomes the new winner. This is because there is currently nothing that penalizes complexity in hypotheses. Excluding Mixture grammars altogether from the candidate space is one way to accomplish this. Functionally, this amounts to assigning an explanatory power score of zero – such hypotheses are not considered to be valid linguistic hypotheses. This is a possible position to hold, but it must be justified on independent grounds, and it must be formally expressible. In general, linguistic theory must decide the appropriate criterion level for explanatory adequacy. This will be discussed in Section 5.1.

If Mixture Grammars are, in fact, excluded from the hypothesis space then the learner must pick a single Simple generating grammar. Under this scenario, the grammar with the fewest exceptions will always be selected. It doesn’t matter if over half the data are exceptions for that grammar, as long as there are more exceptions for every other Simple Grammar. In the case where two grammars have *exactly* the same number of exceptions there will be a tie. But once one grammar gains even a single word advantage the second grammar is effectively discarded; it is unproductive, and plays no role in determining stress for novel words. There is no way to express the fact that two grammars have *approximately* the same number of exceptions and should both be retained as generators of stressed forms. Mixture Grammars are capable of expressing this scenario, as well as any other distribution of component grammars.

The Mixture Grammar that expresses a tie between the two contradictory grammars is defined as an equal mix of *GUJARATI* and *GUJARATI**. This grammar will be referred to as *NO-DIFF*(*G*/G)*$^{\alpha}$, the ‘no difference’ hypothesis. A tie means that it is as likely for any given lexical form to have stress consistent with *GUJARATI* as with *GUJARATI**. In terms of production, it is equally likely that stress will be assigned by the *GUJARATI* grammar or by the *GUJARATI** grammar. It is straightforward to write down the mathematical expression for the probability that *NO-DIFF*(*G*/G)*$^{\alpha}$ assigns to any stress location for a given word type. All possible stress locations within a given 3-syllable word type fall under one of three cases: i) *GUJARATI**$^{\alpha}$ and *GUJARATI* $^{\alpha}$ both assign high probability to that location ii) one of the two assigns high probability, and the other assigns low probability, to that location iii) both hypotheses assign low probability to that location.

For example, take the 3-syllable word /ʃəririk/. Stress on the first syllable corresponds to Scenario ii: *GUJARATI**$^{\alpha}$ assigns high probability to this stress location, but *GUJARATI*$^{\alpha}$ assigns low probability; stress on the second syllable also corresponds to Scenario ii: this time *GUJARATI**$^{\alpha}$ assigns low probability to the stress location, and *GUJARATI*$^{\alpha}$ assigns high probability. Stress on the third syllable corresponds to Scenario iii: neither hypothesis assigns high probability to this stress location. For the word /əwːənə̃/, by comparison, stress on the first and third syllable both fall under Scenario iii, while stress on the second syllable corresponds to Scenario i: both hypotheses assign high probability to this stress location. See Appendix B.1 for the mathematical formalism.

In a hypothesis space that contains only *GUJARATI**$^{\alpha}$ and *GUJARATI*$^{\alpha}$, a difference of even a single ‘exception’ between the two grammars produces an overwhelming winner. For a three-member hypothesis space that includes *NO-DIFF*(*G*/G)*$^{\alpha}$ this outcome changes. Now, for the Simple Variability hypothesis *GUJARATI**$^{\alpha}$ to win it must do “significantly” better than *GUJARATI*$^{\alpha}$; it must do better than the hypothesis that assigns output probabilities as though *GUJARATI** and *GUJARATI* are equally good at describing the data. When the ratio of descriptive power for the two hypotheses is equal to one, *GUJARATI*$*^{\alpha}$ and *NO-DIFF*$(G*/G)^{\alpha}$ tie (continuing to assume, for the moment, that all hypotheses have identical explanatory power). The set of parameter values for which this ratio is greater than 1 are those under which *GUJARATI*$*^{\alpha}$ wins.

Let the variable *m* represent the ratio of the parameters *i* and *j*, where *i* is defined as the number of forms that are exceptions for *GUJARATI*$^{\alpha}$, but not for *GUJARATI*$*^{\alpha}$, and *j* is defined as the number of forms that are exceptions for *GUJARATI*$*^{\alpha}$, but not for *GUJARATI*$^{\alpha}$ (in Table 1, *i* is given by the total number of words belonging to Rows 3 & 4; *j* by the number belonging to Rows 2 & 5). For values of *i/j* falling above the curve in Figure 1, *GUJARATI*$*^{\alpha}$ is the winner in the hypothesis space containing *GUJARATI*$*^{\alpha}$, *GUJARATI*$^{\alpha}$, and *NO-DIFF*$(G*/G)^{\alpha}$. See Appendix B.1 for the derivation of this function.

An α value of 1/3 (dashed line in Figure 1) means that word stress is assigned completely by chance. A more plausible α value is perhaps half of that. For values of α less than 1/6, Figure 1 shows that the anti-markedness grammar (*GUJARATI*$*^{\alpha}$) can only win if *i* is more than twice as big as *j*. The same criterion holds, of course, for *GUJARATI*$^{\alpha}$: *j* must be more than twice as big as *i* in order for it to defeat *NO-DIFF*$(G*/G)$.

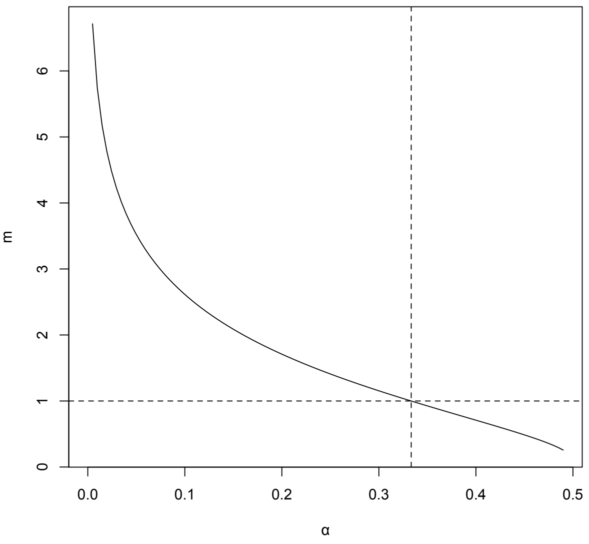

Fig 1

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.

Thus, with a broader consideration of the hypothesis space, the pure anti-markedness grammar (or any other Simple Grammar) is no longer very likely to emerge as the winner. In fact, in can be shown that there is a particular type of mixture grammar that will always win any competition with *GUJARATI*$*^{\alpha}$. Rather than assigning equal weights to its component grammars, this hypothesis assigns weights based on the proportions of each stress type observed in the data[5]. This is similar to the way responsibility for a given output would be apportioned to a set of weighted constraints in a probabilistic OT framework (see, e.g., Goldwater & Johnson 2003; Boersma & Hayes 2001).

Setting the weights such that the Mixture Grammar fits the known data as closely as possible results in the class of what will be called ‘Maximum

[5] This formalization differs from the Optimal Bayes classifier in that the weights, in that framework, are specified by the posteriors, which are calculated individually for each hypothesis, with no reference to the rest of the space, or the sense of a partition of responsibility based on how often hypotheses agree. That is to say, in the Bayes’ Optimal mixture, the weight of *H* reflects how well *H* all by itself can explain the data. The Mixture hypothesis effectively counts the successful predictions of all hypotheses without penalizing a given component hypothesis for getting forms wrong.

Likelihood' grammars[6]. *Max*$(G^*/G)^{\alpha}$ is defined analogously with *No-Diff*$(G^*/G)^{\alpha}$, with the difference being that the weights on the sub-hypotheses are free to vary. Since the weights are no longer guaranteed to be equal, there are four, rather than three, relevant stress scenarios to be considered. However, it can be readily seen that substituting .5 for both weight values reduces *Max*$(G^*/G)^{\alpha}$ to *No-Diff*$(G^*/G)^{\alpha}$ (see the derivation provided in Appendix B.2).

The Maximum Likelihood Grammar has a higher descriptive power score than any other hypothesis considered so far[7]. In fact, under the assumption of uniform explanatory power, *Gujarati*$^{*\alpha}$ cannot do better than *Max*$(G^*/G)^{\alpha}$. This is because *Max*$(G^*/G)^{\alpha}$ always sets its weights so as to maximize descriptive power over the training data, and has more degrees of freedom, and thus more flexibility in doing so. *Gujarati*$^{*\alpha}$ is actually a special case of *Max*$(G^*/G)^{\alpha}$, and the two hypotheses are identical when there are no unambiguous data in support of *Gujarati*$^{\alpha}$. Therefore, the best outcome for the anti-markedness grammar is a tie (see Appendix B.2).

However, *Gujarati*$^{*\alpha}$ can be given a chance of winning if *Max*$(G^*/G)^{\alpha}$ is handicapped in some way. Informally, *Max*$(G^*/G)^{\alpha}$ 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. Ignoring this difference implies that the two hypotheses are *a priori* equivalently acceptable to the learner as candidate grammars. If, instead, the learner has some bias towards the less complex, or shorter hypothesis, this can

[6] In machine learning terms, such grammars over-fit the data. That is, they do not allow for the possibility of noise in the signal, and thus have the potential to do poorly with novel data when those are different in any way from the training data. Throughout the paper it has been assumed that all data are part of the signal, and thus should be learned veridically. However, one can think of words that do not conform to a majority grammar as a type of 'noise' that obscures the general pattern. The problem in doing so is the exact same problem encountered when trying to explicitly formulate linguistic descriptions, or to define a learning mechanism that will produce a consistent and plausible result. There is no way to decide *a priori* – in the absence of UG – what constitutes 'noise', and what 'signal'. The MAX class of hypotheses can be handicapped to prevent over-fitting, but within the Bayesian framework, and with the extreme probability distributions, that handicap would have to be very large. See the remainder of the discussion on information theoretic priors in Appendix D.

[7] Of course, one can do still better if the 'maximum likelihood' hypotheses are tailored to individual word types. This is because the stress distribution varies for different vowel combinations. The best one can do is conditioning on individual words, that is, straighforward memorization, which will result in zero error over the observed data. However, this strategy leaves the learner without any explicit mechanism for deciding stress placement for novel words – without a true grammar. Therefore, this possibility will be excluded from further consideration.

be reflected in the explanatory power term. This bias can be formally defined using the information-theoretic notion of coding cost. Within this framework, the ratio of explanatory power for the two hypotheses is estimated by the following equation (see Appendix D for the derivation).

$$\frac{p(GUJARATI^{*\alpha})}{p(MAX(G^{*}/G)^{\alpha})} = \kappa\sqrt{N} \qquad (8)$$

In Equation (8), κ is a constant on the order of 10^3, reflecting the greater complexity of the Mixture Hypothesis. This ratio depends on the total amount of data, N, as does the ratio of the descriptive power. In order to determine under exactly what conditions $GUJARATI^{*\alpha}$ will beat $MAX(G^{*}/G)^{\alpha}$ it will be necessary to specify a particular lexicon that is to be learned.

4 The Lexicon

Table 1 provides the set of word types that are relevant to the analysis; what it does not provide are the relative proportions of those types. The outcome of learning, however, depends very strongly on the actual proportions in the original lexicon. To achieve a general result, several possible lexicons will be considered in section 5.2. For now, however, a single member of that set will be chosen in order to clearly illustrate the behavior of the learning algorithm. That lexicon is the one in which all word types occur in equal proportions: the uniform lexicon: (L_U>) L'_U.

Maintaining the assumption that consonants are irrelevant to stress placement, each word of L_U can be represented as a sequence of vowels. Furthermore, since there are only three sonority classes, those representations can be further reduced to a sequence from the set {ə, a, and M}, where M is any of the vowels {i,e,ɛ,o,ɔ,u}. Restricting the analysis to three-syllable words for simplicity generates 8^3, or 512 distinct types (from an eight vowel inventory). Table 3 is an expansion of Table 1, providing the total number of word types for each stress scenario. The ‘Case’ column of Table 3 contains the exhaustive list of sonority profiles. The '# Types' column contains the count of words with a given sonority profile, but a unique sequence of vowels. The set of sonority profiles is divided up into six classes of words, determined by which Simple Hypotheses those forms are consistent with with respect to stress placement (see Appendix E for the analogous table for two-syllable words).

Table 3. Full set of all possible three-syllable word types with respect to 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}).

	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)	[páʈnəgər] > [pə́ʈnəgər]	
2	(M,M,a)	[hoʃijáɾ] > [hoʃijə́ɾ]	84 G*
	(a,M,M)	[ʃáririk] > [ʃə́ririk]	
	(a,M,a)	[háɖohaɖ] > [hə́ɖohəɗ]	
	(a,M,ə)	[p^{h}ắsigər] > [p^{h}ə̃́sigər]	
3	(M,a,a)	[durácar] > [durə́cər]	48 G*,P
	(M,a,ə)	[mubáɾək] > [mubə́ɾək]	
	(M,a,M)	[betáɭis] > [betə́ɭis]	
4	(M,M,ə)	[tʃumːótəɾ] > [tʃumːótəɾ]	78 G,P
	(ə,M,ə)	[vəríʃʈ^{h}ə] > [vəríʃʈ^{h}ə]	
	(ə,M,M)	[kəʈóro] > [kəʈóro]	
5	(M,ə,M)	[kójəldi] > [kójəldi]	42 G
	(M,ə,ə)	[kʃétrəp^{h}əɭ] > [kʃétrəp^{h}əɭ]	
6	(a,a,a)	[awːánã] > [əwːə́nə̃]	239 G,G*,P
	(a,a,M)	[amdáni] > [əmdə́ni]	
	(ə,a,a)	[resádar] > [resə́dər]	

	(ə,a,ə)	[səp^{h}ácət] > [səp^{h}ə́cət]	
	(ə,a,M)	[g^{h}əʈáɖo] > [g^{h}əʈə́ɖo]	
	(ə,ə,ə)	[əkbə́nd^{h}ə] > [əkbə́nd^{h}ə]	
	(ə,ə,M)	[cəkcə́kit] > [cəkcə́kit]	
	(M,M,M)	[iʈ^{h}ʈ^{h}óter] > [iʈ^{h}ʈ^{h}óter]	
	(a,a,ə)	[ɟ^{h}agmágəʈ] > [ɟ^{h}ə́gmə́gəʈ]	

As illustrated previously, the word [mubáɾək] in Gujarati becomes [mubə́ɾək] in Gujarati′, providing equal evidence for *GUJARATI** and *PENULT*. Its three vowels are (u,a,ə), which corresponds to the stress template (M,a,ə). The hypothetical Gujarati word [mebáɾək] has a different set of vowels, but is also a member of that stress template, meaning stress will occur in the same location – the penultimate /a/, becoming a penultimate /ə/ in Gujarati′. The hypothetical words [bumáɾət] and [sebáɾəm] also belong to this set – changes to the consonants in a word will not affect stress placement. The same is true for words sharing the vowel template (M,a,M), such as the word [betáɭis], also in Row 3. In total there are 48 3-syllable word types (allowing for all values of M) that are ambiguous with respect to their generating grammar in exactly this way.

The Uniform Lexicon is defined as maintaining the ratio of types from Table 2 within an adult-size vocabulary of unique words. Scaling the 512 word types by a factor of 13 yields a reasonably sized lexicon of 6,656 words. This lexicon is used to calculate the winner of the competition between *GUJARATI*$^{\alpha}$, *GUJARATI**$^{\alpha}$, and *MAX*(*G*/G*)$^{\alpha}$ (momentarily excluding *PENULT* for the sake of simplicity). This lexicon and hypothesis space is defined by the following parameter values: the number of words consistent with both hypotheses, $n = 3107$; the number of words consistent with neither hypothesis, $a = 273$; the number of words consistent with only *GUJARATI**, $i = 1716$; the number of words consistent with only *GUJARATI*, $j = 1560$. α and w are set to the maximum likelihood estimates over the Uniform Lexicon. The information-theoretic explanatory power ratio given in (8) – with an *N* of 6,656 –, boosts the combined ratio score for *GUJARATI*$*^{\alpha}$ by about 5 orders of magnitude. However, the descriptive power ratio is so skewed to *MAX*$(G*/G)^{\alpha}$ under L'_U (on the order of 10^{726}!) that this makes effectively no difference (compare the maximum descriptive power of *MAX*$(G*/G)^{\alpha}$ in Fig. F1 to the descriptive power of *GUJARATI**: $\alpha^{N-G*}(1-2\alpha)^{G*}$, for $\alpha \approx .138$, in Appendix F). The result can be conceptualized, paralleling earlier discussion, as a significance level. It can be shown that, 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*.

5 Likely Input and a Reasonable Learner

This paper began with categorical hypotheses that tolerated no exceptions (see (2)). These were quickly seen to be inadequate for fully capturing naturalistic language data. Once variable hypotheses were allowed into consideration, however, there was no obvious reason for limiting their capacity to reflect the underlying distribution. The Mixture Grammar with weights fit to maximize descriptive power over L' was seen to win under a large range of possible values for *i*/*j* in L'. This grammar is exceptionless in the sense that all forms contribute to the final weighting of its component hypotheses[8].

The overwhelming advantage to the Maximum Likelihood Mixture hypothesis is a result of the simple statistical learning architecture represented by the Bayesian learner. However, with an evaluation metric that also factors in explanatory power, there still exists a narrow range of values for which the Simple Variability Grammar can win the competition. It is straightforward to construct such a scenario. Start with a 6,656 word lexicon in which 3122 words (*i*) are consistent with only *GUJARATI**, 3107 (*n*) are ambiguous (consistent with both *GUJARATI** and *GUJARATI*), and 273 (*a*) are consistent with neither hypothesis. For

[8] In fact, there are no exceptions unless they are exceptions for *both* grammars. These are the datapoints designated ‘(A)rbitrary’ stress in Table 3 (row 1). One can determine the maximum allowable number of these exceptions that can be tolerated and still allow this particular Mixture Grammar to win; the calculation, however, would involve consideration of additional hypotheses and is not as directly relevant to the discussion as are datapoints which are exceptional under only one of the two component grammars of the Mixture.

this set of parameter values, the absolute maximum allowable value for *j* is 154; Thus, the simple Variability hypothesis *GUJARATI** results when there are fewer than 155 forms (*j*) that are consistent only with *GUJARATI* (exceptions to *GUJARATI**) (and for the given values of *N*, *n*, *i*, and *a*). Once *j* reaches 155, however, words that were formerly ‘exceptions’ become regular outputs of the *GUJARATI*$^{\alpha}$ sub-grammar contained within the winning Mixture Grammar. For the case of the uniform lexicon L'_U , *j* is much larger than 155, and the Mixture Grammar is the clear winner, specifically, the Mixture Grammar that assigns stress according to *GUJARATI*$^{\alpha}$ approximately 47.5% of the time; and, according to *GUJARATI*$^{*\alpha}$ approximately 52.5% of the time (see Appendix F).

It was necessary to explore the hypothesis space in this manner because it was not clear at the start of this simulation what the outcome of learning would be. In the absence of principled reasons (independent from the theory under test) for restricting the learner’s hypothesis space, to do so is to prejudice the outcome. Equally dangerous, such a move makes the learning problem appear less complex than it actually is. Whereas the foregoing discussion has hopefully brought ought the true complexity of the problem.

It was necessary to tackle this difficult problem in order to answer our original question: given a well-behaved stress system and a sound change of a certain type, what kind of stress system results? If it can be shown that an anti-markedness grammar is the inevitable result (stress on lower sonority vowels in preference to higher) then there must be some other mechanism to prevent it (and all other anti-markedness grammars that could result from similar historical trajectories). This extra mechanism is required under the assumption that such grammars are entirely non-existent in languages of the word. If both things are true, the UG-Delimited $\mathcal{H}$ Principle is supported. Alternatively, if anti-markedness grammars do not inevitably result from learning, then it cannot be argued that some other mechanism is required, and the UG-Delimited $\mathcal{H}$ Principle is not supported.

Despite the fact that the chain of reasoning above is relatively easy to state, and logically consistent, there is more than one problem with testing it explicitly, several of which have already been discussed. An additional problem that surfaces at this point in the analysis is that ‘anti-markedness grammar’ is not strictly enough defined. It seems to be assumed that the grammars in question are deterministic, categorical, and exceptionless. But the actual data over which the grammar must be learned do not support hypotheses of that kind. The preceding section demonstrates that Simple Variability Hypotheses should rarely prevail (requiring a narrow range of lexical conditions). However, the theoretically dispreferred *GUJARATI** comprises a subset of the winning Mixture Grammar. Quite likely, this Mixture Grammar would also be banned by proponents of the UG-Delimited $\mathcal{H}$ Principle. The fact that it is the clear winner under the foregoing analysis, however, does not necessarily provide support for that principle. Our question has not yet been answered – once again, there are additional nuances of the problem to take into account.

Before the typological status of Mixture Grammars is taken up in sections 6 and 7, a new set of results will be derived from a more sophisticated model. Alternatives to the threshold level derived in section 4 that may have more psychological validity will be considered. The effect of lexicon on outcome will be modeled in section 5.2 over a wide range of possible lexicons. In section 6 results will be derived from treating the probability of the sound change as a function of lexical contrast. Finally, the phonetic component of the problem will be re-visited.

In the first place, it is not clear that the Bayesian significance threshold is the correct one for language learners. Despite the fact that much about the language acquisition problem is unknown, there may be some consensus about the range of possible outcomes, and the general conditions under which they come about. For example, a regular pattern with a large number of instances, and a small number of exceptions relative to those instances, might be expected to induce a Simple Hypothesis outcome. This implies that there are a certain number of forms that are tolerated as true exceptions. It might also be reasonable to expect that the type of ‘exception’ would influence the outcome. ‘Patterned exceptions’ (see Zuraw 2010) might inhibit the adoption of a single ‘default’ rule, whereas truly random exceptions might promote it (as has been proposed in the evolution of creoles from pidgin precursors (Singleton,1989; Ross & Newport 1996; Senghas & Coppola 2001; Hudson Kam & Newport 2005).

Yang (2005) has a proposal for calculating the cross-over point from detailed sub-patterns to default rule – what he calls the Tolerance Level. His metric for complexity is expressed in terms of processing time, rather than description length. In his formulation the speaker must check each word to see if it falls into any known class of exceptions before applying the productive rule as default. The point at which there is no longer any time savings associated with having a general rule is the point at which generalization fails. He estimates the largest allowable number of exceptions as $\frac{N}{\ln N}$, where N is the size of the dataset[9]. For any lexicon of size 6,656 this metric tolerates 756 exceptions. Subtracting the contribution of *a* type words gives a maximum *j* value of 483.

Yang's metric provides an alternative to Bayes for calculating a significance threshold. Furthermore, both metrics can be implemented in more than one way. One can count the number of exceptions as the total number of words whose stress is not predicted by the dominant grammar, or as a ratio of words that are consistent with the dominant grammar versus the secondary grammar (G^*/G), or as a ratio of only the unambiguous data (i/j). For total numbers, Yang's metric tolerates about 3 times as many exceptions as the Bayesian learner. For the ratio of consistent data, the Bayesian learner requires *GUJARATI** to account for about 2 times as much data as *GUJARATI*, whereas Yang's metric requires 1.6 times as much. For the unambiguous data ratio, the Bayesian learner's threshold ratio is a factor of 20, while Yang's morphological learner requires about 6 times as much data for *GUJARATI** to win.

[9] Yang (2005) is explicitly concerned with determining whether morphological patterns can be said to have default rules. He considers rules like "add –d to make the English past tense". These rules are assumed to have a constant processing time. Rules are not evaluated against each other with respect to explanatory or descriptive power. The only competition is for default status; this is achieved when one rule class is significantly larger than all other classes *combined*. Yang distinguishes between productive and default rules. There may be a number of productive irregular rules which are not defaults. If there are enough regular forms as compared to irregulars, then a default rule can be defined. If not, then essentially all words are irregulars; either the past tense form of each word must be memorized, or the irregular rule class to which it belongs.

5.1 A Reasonable Learner

Clearly the final results of the learning competition will depend on exactly how the test statistic is calculated, as well as where the significance level is set. At the moment, I am aware of no empirical findings that can be used to validate one choice over another. However, a favorable test of the UG-Delimited $\mathcal{H}$ Principle can still be conducted, given our present state of knowledge. To do so, this section will adopt a range of threshold values much lower than that given by the Bayesian Learner or even Yang's Tolerance Level. Doing so allows the outcome of learning to result in a simple Variability grammar for a broader range of data[10]. This gives both the anti-markedness, as well as the markedness-abiding, grammar the best chance of winning. This is also more in accord with the intuitions that led to the formulation of the UG-Delimited $\mathcal{H}$ Principle in the first place, as well as the description of present-day Gujarati as the grammar *GUJARATI*, ignoring possible exceptions or sub-patterns.

It should be made clear both that the chosen levels are completely arbitrary, and that they are no more arbitrary than any other level that could be chosen. Probability theory provides a principled way to set such a threshold. But that is not a guarantee that human learners set their thresholds in the same way. Nor is it known exactly how ambiguous data figures into the grammar selection process. In general, language learning algorithms that deal with gradient phenomena often contain implicit thresholds for categorical behavior (see, e.g., Pearl (2011), Hayes & Wilson (2008)). Here the levels will be explicitly chosen as a set of relatively permissive criteria for adoption of a Simple Hypothesis.

The measure for which these levels are set is taken to be a simple ratio, dubbed the Proportion-Based Evaluation Metric to distinguish it from the Bayesian evaluation metric employed throughout sections 3 and 4. The hypothetical learner may select a grammar that closely matches the input data via the combination of multiple grammars, but they may also sacrifice a certain amount of descriptive power in order to select a single hypothesis corresponding

[10] Analogously, one could fix a given prior probability distribution – select a given 'over-hypothesis' based on the expected linguistic outcome (see, e.g., Kemp, Perfors & Tenenbaum 2007).

to a Simple Grammar (a default). The latter outcome will obtain under conditions in which a threshold ratio of data coverage is reached. Including ambiguous data in the calculation will act to decrease sensitivity to very small differences, therefore the ratio of the total amount of data consistent with each hypothesis will be used.

Consider the following "statistical significance" threshold values. In order to defeat the Mixture Hypothesis, the total amount of data that is consistent with *GUJARATI** must be, say, from 1.25 to 5 times as large as the amount of *GUJARATI*-consistent data. By the same token, *GUJARATI**-consistent data should also be from 1.25 to 5 times as large as the amount of *PENULT*-consistent data. Under the Uniform Lexicon (L'_U), the relaxed thresholds do not affect the results. The numbers from section 4 for three-syllable words give values for G^*/P of 4823/4745 = 1.02, and G^*/G of 4823/4667 = 1.03. None of the thresholds in the range (1.25, 5) is met, and the Mixture Grammar is still the winner.

However, looking back at Table 3, it is easy to imagine a case in which the relevant ratios could reach criterion. Suppose Gujarati had a lexicon, for whatever fortuitous reasons, in which there were no words that consisted of the sonority profile (M, M, ə). Taking this even further, if all lexical types in Gujarati′ that exhibit the *GUJARATI* pattern in Table 3 had never existed (Rows 4-6), then the data would skew towards the *GUJARATI** hypothesis. Of course, the reverse scenario is just as imaginable. The question now becomes: under what lexical conditions will the 'reasonable' thresholds be met, such that the anti-markedness *GUJARATI** hypothesis dominates, and how likely are those lexical conditions to arise?

5.2 Likely Input

Up to this point only a single possible lexicon has been considered in any detail (and, in fact, a single lexicon comprised of only three-syllable words). This was done partially for ease of exposition, but also to avoid the additional modeling assumptions that would have to be made about lexical distributions. It is now clear, however, that the exact make-up of the lexicon will strongly affect the outcome of learning. Therefore, the space of lexicons as a function of word-type distribution will now be systematically explored. To isolate the effect of vowel frequency on the outcome of learning, other simplifying assumptions will be maintained, namely the assumptions regarding segmental independence within words, and across words of varying length (i.e., possible phonotactic or prosodic constraints on words will be ignored).

Counting by possible word types, the data in Gujarati′ are almost equally split in support of *GUJARATI** versus *GUJARATI*. However, the more non-uniform a lexicon is with respect to the actual number of words that fall into each row of Table 3, the more likely that a default grammar (of one type of the other) will result. Accordingly, the space of word-types was sampled to varying extents to create a distribution of lexicons that spanned a range of distances from the Uniform Lexicon.

Each such lexicon consists of a total of 6,912 words of which roughly half are 3-syllable words (3072), and half, 2-syllable words (3840)[11]. How much each lexicon departed from uniformity with respect to the rows of Table 3 (along any of the 6 dimensions) is given as a 'Degree' of biasing away from uniformity. There are four such degrees: Degree 1 corresponding to the most uniform vowel distributions; Degree 4 to the least. Monte Carlo simulations were run for 1000 lexicons at each of these four degrees (the details of the sampling method are given in Appendix G.1). One important characteristic of the method is that the resulting vowel distributions are not directly controlled.

For comparison, an additional 1000 lexicons were generated in a different way, setting the vowel distributions to a specific function. This final set of lexicons was given Zipfian distributions over the inventory of vowels {a,i,e,ɛ,o,ɔ,u,ə}. Each of the 1000 lexicons reflected such a distribution, the only differences being which vowels occupied which rank order in frequency of occurrence. See Appendix G.1 for details.

[11] This is a fairly arbitrary ratio, set to reflect the distribution found in the online British English corpus CELEX (1993), as it was readily available at the time of writing.

In Figure 2, each lexicon is represented by two numbers; the y-axis represents the threshold parameter *G*/P*, and the x-axis represents the threshold parameter *G*/G*. Each set of lexicons is represented by a different color. From Figure 2 it can be determined how many simple anti-markedness grammars are predicted to win, for any given lexicon distribution. The four proposed significance thresholds are plotted as solid black boxes. Points interior to each of the boxes correspond to lexicons in which *GUJARATI** accounts for 2.5, 1.7, and 1.25 times the amount of data as each competitor hypothesis (from smallest to largest box). All other points represent either Mixture outcomes (area centered around (1,1) in Fig. 2), or outcomes in which *GUJARATI*$^{\alpha}$ or *PENULT*$^{\alpha}$ is the winner. The exact weighting values for each predicted Mixture Grammar winner depend on the particular lexicon.

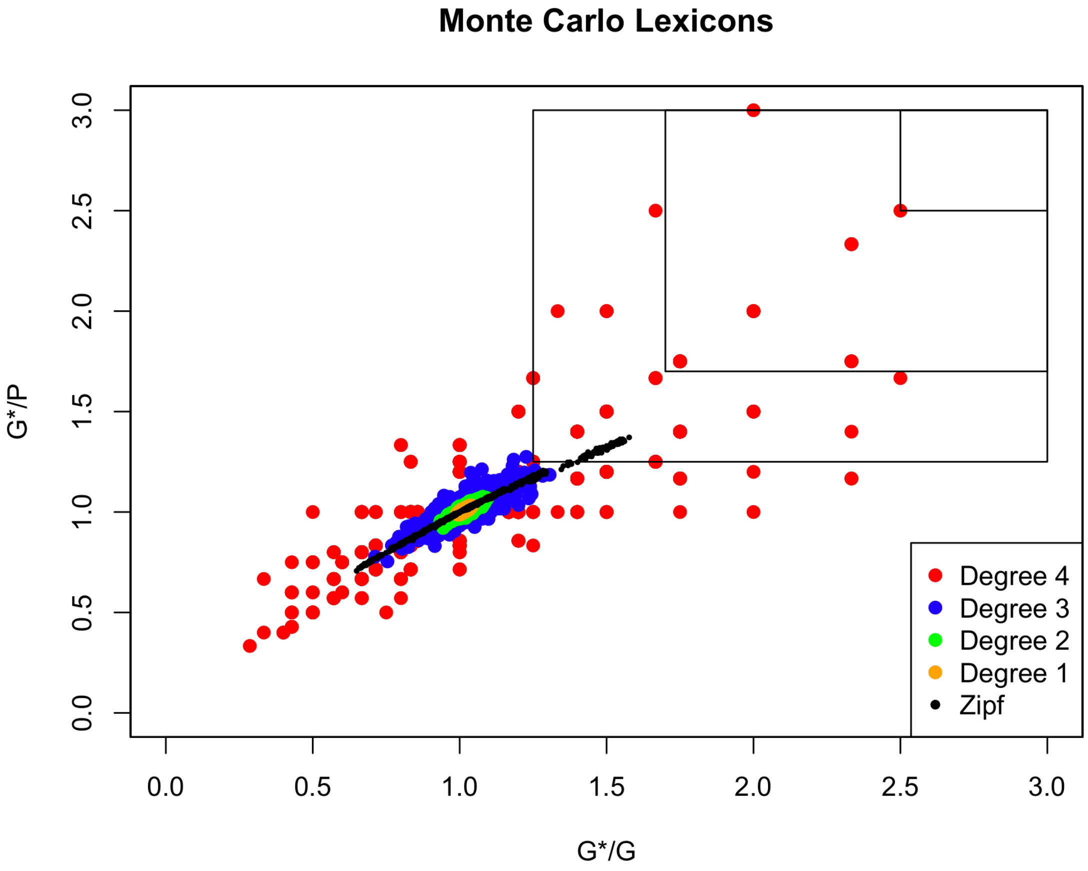

Figure 2

Monte Carlo Simulations of 1000 Lexicons generated at each of 4 degrees of random sampling. Degree of sampling corresponds to likelihood of departure from a uniform vowel distribution; higher numbers equate with higher skewing. Also displayed is a set of lexicons generated with a randomly assigned Zipfian distribution over the vowels (see text). Each lexicon consisted of 6072 words of which roughly half were 3-syllable words (3072), and half, 2-syllable words (3840). The axes are the ratios $G*/G$ and $G*/P$, respectively, which are defined as threshold variables. The area within each solid-lined black box represents the lexicon space in which the GUJARATI*$^{\alpha}$ hypothesis is chosen: the threshold significance value is reached. The three different boxes correspond to three different possible threshold values: levels where GUJARATI* covers 2.5, 1.7, and 1.25 times as much data as each of its competitor hypotheses (*GUJARATI* and *PENULT*).

These simulations reveal two things: 1) under assumptions of random sampling and reasonable decision thresholds, lexicons that support a GUJARATI*$^{\alpha}$ hypothesis are quite uncommon. But, 2) they are not impossible. Table 4 gives probability estimates based on the number of lexicons out of 1000 at each Degree that fall above each of the given thresholds. With the least stringent threshold (GUJARATI* consistent with only 1.25 times as much data as either *GUJARATI* or *PENULT*), and the highest Degree of non-uniformity, over 11% of the lexicons are estimated to lead a learner to adopt GUJARATI*$^{\alpha}$. At any lesser Degree, or more stringent evidence ratio threshold, this estimate drops below 4%.

Table 4. Estimated percentage of anti-markedness outcome for Proportion-Based Evaluation learner, under 5 different sampling rates, for four different threshold values ($G*/G$, $G*/P$). Calculated from 1000 Monte Carlo simulations.

Vowel Distribution	Percentage of lexicons with $G*/G$ & $G*/P$ >			
	5	2.5	1.7	1.25
Degree 4	0	0	3.5%	11.5%
Zipfian	0	0	0	9.4%
Degree 3	0	0	0	0
Degree 2	0	0	0	0
Degree 1	0	0	0	0

If vowels have Zipfian distributions, and human learners have a significance threshold greater than 1.25, the probability of an anti-markedness grammar arising from the proposed sound change is very small: less than 1 in 1000. For the 1.25 threshold, such default grammars are predicted at a rate of 9.4%. All other Zipfian Lexicons (black dots) result in Mixture Grammars. Lexicons that fall within the area bounded by (1,1.25) on the x-axis, and (1, 1.25) on the y-axis are Mixture Grammars with a slightly higher weighting for the GUJARATI*$^{\alpha}$ sub-grammar with respect to the GUJARATI$^{\alpha}$ and PENULT$^{\alpha}$ sub-grammars. Lexicons that fall within the area bounded by (.8,1) on the x-axis are Mixture Grammars with a slightly higher weighting for *GUJARATI*$^{\alpha}$ with respect to *GUJARATI**$^{\alpha}$; those that also fall within the range (.8,1) on the y-axis have slightly higher weightings for *PENULT*$^{\alpha}$ than *GUJARATI**$^{\alpha}$. Lexicons that fall below .8 on both axes are Mixture Grammars of *PENULT*$^{\alpha}$ and *GUJARATI*$^{\alpha}$ only, some with higher weighting for *PENULT*$^{\alpha}$ than *GUJARATI*$^{\alpha}$, and some with the opposite relative weighting. There are no lexicons for which either *GUJARATI*$^{\alpha}$ or *PENULT*$^{\alpha}$ is the Simple default; thus such default grammars are predicted to occur less than 1 time out of a 1000, for any significance threshold.

6 Interpretation of Results

The above results rely on a number of assumptions, one of which is that all lexicons are equally likely to undergo the change a > ə. However, there are reasons to think that the a > ə sound change might be more likely under certain conditions than others. Neutralizing sound change can be considered undesirable to the extent that it creates communicative ambiguity. The degree of potential ambiguity can be approximated by the number of homophones created by the loss of the contrast. This, in turn, can be estimated by the number of words containing the pre-merged segments. The expectation is thus that lexicons containing fewer words, with fewer /ə/'s, would be more likely to undergo the neutralizing sound change than lexicons with high occurrence of /ə/. A subset of this class of lexicons are ones in which Gujarati has no /ə/'s at all: the No-Contrast Lexicon: L_{NC}. L'_{NC} is missing the types of words that provide unambiguous support for *GUJARATI* : Row 5 of Table 2, reproduced below in (9) for reference. Thus, this scenario also provides the most favorable conditions for *GUJARATI**$^{\alpha}$, and an upper bound on the likelihood of a *GUJARATI**$^{\alpha}$ outcome.

(9)

5	(M,ə,M)	[kójəldi] > [kójəldi]	42 G
	(M,ə,ə)	[kʃétrəpʰəɭ] > [kʃétrəpʰəɭ]	

Simulations of the No-Contrast condition do result in a relatively high rate of *GUJARATI*$^{*\alpha}$ outcome, at 29.5% – but only at the lowest threshold level (see Appendix G.2). However, the Mixture grammar *GUJARATI*/ PENULT* is more than twice as likely (70.5%). Despite the fact that there are exceptions to the penultimate stress rule in L'_{NC}, there is still not enough of a difference between the descriptive power of the two hypotheses to rule out the Mixture grammar. This result reveals a heretofore unexamined property of the Variability Hypothesis Space. Allotting a certain amount of probability mass to possible exceptions has the consequence of downgrading the advantage of an *exceptionless* grammar. In traditional phonological analysis a single data point that can decide between two grammars also satisfies the criterion for doing so. The burden of proof is now higher, since such critical forms can now be treated as allowable 'exceptions'.

This result holds in the other direction as well. It has been the assumption from the beginning of the paper that *GUJARATI*, as defined in (2), is the correct grammar for the Gujarati language. However, the Proportion-Based Evaluation Metric will only come out in favor of *GUJARATI*$^{\alpha}$ under certain lexical conditions. Under others, the *GUJARATI/PENULT* Mixture grammar is predicted to be the winner. Under L_{NC} (prior to sound change) and the 1.25 threshold, *GUJARATI*$^{\alpha}$ wins 28.9% of the simulations, while *GUJARATI/PENULT* is the result in the remaining 71.1% of cases.

The relatively low rate of the *GUJARATI* grammar may be a theoretically dispreferred result, but it cannot be assumed to be incorrect. This result is the direct outgrowth of modifying the model of linguistic competence in order to deal with exceptions at all. The data themselves are ambiguous with respect to the generating grammar because the status of any given form is ambiguous as to whether it is an allowable exception to H_i, or the deciding piece of evidence in favor of H_j.

In fact, in the No-Contrast condition, a given form may be ambiguous with respect to whether it is an exception to *PENULT*$^{\alpha}$, evidence for *GUJARATI*$^{*\alpha}$, or evidence for a simpler hypothesis, namely, 'Always Stress ə'. The original hypothesis space was chosen with *GUJARATI*$^{*\alpha}$ as the competitor of interest because the question as first posed had to do with the emergence of a fully generative reversed sonority-to-stress grammar. However, data that are consistent with this abstract hypothesis are also consistent with a more concrete hypothesis that refers to vowel identity. It is important to note that there are only two expressed sonority categories in L'_{NC}: {M,ə}. What this means for the analysis is that a grammar that instantiates the rule 'Stress ə' has the same descriptive adequacy as the *GUJARATI*$^{*\alpha}$ grammar. The likelihood of selecting the true anti-markedness grammar is lower than that of learning the 'Stress ə' rule from an information theoretic perspective, as it involves a more complex representation: multiple categories corresponding to sonority-tier membership and the relations between them. Since the coverage ratio is exactly the same for these two hypotheses the simpler one must win.

This type of competitor is not specific to the sonority-to-stress case study but must be considered for any hypothesis describing a harmonic scale. That is, if prior (possibly innate) knowledge of the scale is not assumed, then it is not guaranteed that learners will be able to infer the entire range from exposure to only a small subset of its members. Thus, for example, a pattern in which /k/ palatalized before /e/ but not before /i/ could be the result of a 'Reversed Palatalization' grammar, but also of the simpler 'Palatalize Before e' grammar[12].

[12] Wilson (2006) has, in fact, experimentally demonstrated an implicational bias for palatalizing before /i/ after being exposed to a palatalization alternation before /e/. This suggests that learners may well infer an entire harmonic scale from limited instances. However, it has not been demonstrated that such a bias is *necessary* for learning, which is the argument examined in this paper.

While the ‘Stress ə’/‘Palatalize Before e’ grammar can be considered a partial anti-markedness grammar, it does not make theoretically undesirable predictions to the extent that the full anti-markedness grammar does. The true reversed scale calls for avoiding the placement of stress on the /a/ in a newly encountered word, for example, precisely because of the high sonority of that vowel (or failing to palatalize before a newly encountered /ɪ/ vowel, precisely because of the more front/high qualities of that vowel).

In either case, however, there is a more basic problem with determining results under L'_{NC}. In a synchronic grammar that lacks a contrast between /a/ and /ə/, it is unlikely that any ‘unnatural’ grammar would be observable. Without the possibility of ambiguity, the surface realization of the /ə/ vowel will be freer to vary between phonetically longer tokens (in stressed position) which are more /a/-like, and phonetically shorter tokens. These phonetic context effects are likely to restore the markedness-abiding sonority-stress relation (Colarusso 1988; Choi 1992; Kondo 1994; Van Bergen 1994)[13].

The upshot of the preceding discussion is that the No-Contrast condition may not, in fact, provide the most favorable conditions for a *GUJARATI*$^{*\alpha}$ outcome. Even though the conflicting forms that favor *GUJARATI*$^{\alpha}$ are removed, so is an entire level of the sonority hierarchy. The latter is arguably necessary for learning a sonority-based grammar, as well as for maintaining a distinction between stressed /ə/ and unstressed /a/. *GUJARATI*$^{*\alpha}$ may well never occur under those

[13] This observation returns full circle to the originally proposed sound change. The scenario was described at the outset of this paper in the following way: a completely exceptionless, completely context-free sound change in which surface [a]’s become realized as surface [ə]’s, *under which stress placement does not change*. In fact, it is not clear how likely are internally motivated language changes of a completely general nature. Arguably more likely is that such changes would depend heavily on context. A large body of work within the frequentist and exemplar-based frameworks presents a strong case for non-uniform sound change, with factors such as word frequency, phonetic conditioning, and morphological decomposability influencing when and whether certain segments will shift (Phillips 1984; Bybee 2001, 2006; Pierrehumbert 2001; Hay *et al.* 2003). As alluded to at the very beginning of this paper, any /a/’s which are likely to become /ə/’s (higher, shorter, less sonorous), are also less likely to be stress-carriers in the first place. Stress, therefore, has a very high probability of shifting to a different, higher energy location in the word during such a vowel quality shift, or even as a necessary precursor to such a shift.

conditions. If L'_{NC} is the lexicon class under which the a > ə sound change is most likely to occur, then the likelihood of *GUJARATI*$^{*\alpha}$ is actually reduced overall.

In section 5 *GUJARATI*$^{*\alpha}$ was estimated to result in from 0 to 11.5% of Gujarati-type languages that had undergone the sound change (the vast majority of which exhibited a contrast between the two vowels prior to the merger). The actual predicted rate of occurrence of *GUJARATI*$^{*\alpha}$, however, depends additionally on the probability of the lexicons themselves, as well as the probability of each lexicon to undergo the sound change. That is, there is an 11.5% chance that a particular set of lexicons will reflect a generative grammar of the type *GUJARATI*$^{*\alpha}$. But there is also some probability associated with the actual attestation of those lexicons (as opposed to other possible ones), and an additional probability associated with those lexicons undergoing the sound change (as opposed to failing to undergo, or undergoing some other sound changes instead, or in addition). For example, if all generated lexicons are equally likely and exhaust the space of possible types, and if a's are uniformly predicted to shift to ə's in 10% of all languages with any contrast between the two, then 1.2% of historic Gujarti-like languages should be expected to convert to synchronic *GUJARATI*$^{*\alpha}$ languages. This rate is low enough that it may be plausible that such languages exist but are missing from the limited typological sample currently available to us.

What remains to be accounted for is the overwhelmingly predicted Mixture Grammar, *GUJARATI*/GUJARATI*. However, the interpretation of this result still rests on whether or not the Mixture Grammar that *includes* the anti-markedness grammar has the same status as the Simple anti-markedness grammar. Whether or not the Mixture is conspicuously absent from the typology should be an empirical question, but it requires consensus on what constitutes evidence for the existence of such a grammar. The lexicon would have to contain a minimum number of words whose stress was consistent with *GUJARATI** only, and a minimum number of words whose stress was consistent with *GUJARATI* only. That is fairly straightforward. But the full details of the lexicon are not always available from descriptive grammars. Forms that are not consistent with a Simple Hypothesis may be pre-judged as exceptions; in such cases they are unlikely to be individually listed, and are often not mentioned at all. Given a possible bias on the part of the analyst in favor of the markedness-abiding grammar, it may be the case that potential Mixture Grammars have been systematically mis-classified or ignored.

7 Summary & Discussion

A number of issues have been raised in the preceding analysis. Not all have clear solutions. This is simply a reflection of our very imperfect knowledge in almost every sub-domain of linguistics. It is important to realize that this work has not introduced unnecessary complexity. Rather, it has drawn the curtain back from the complexity that already exists – and underlies every theoretical claim about linguistic universals. Once this fact has been absorbed, it becomes clear that 'intuitive' predictions, on both sides of the debate, as plausible as they may seem at first glance, are massively inadequate.

This paper is not an attempt to answer long-standing and difficult questions about learning, linguistic competence, or sound change. The goal is the more modest, although still foundational, one of pointing out where gaps in our current theories exist, especially at the intersections of historically separate subfields. Computational modeling forces a degree of explicitness in formulating theoretical questions that is difficult, if not impossible, to adequately anticipate. In fact, it must be concluded that a definitive test of the UG-Delimited $\mathcal{H}$ Principle can only be made once sufficiently worked out theories of language acquisition and sound change are available. In the absence of such theories there are no iron-clad conclusions to be drawn. However, this does not mean that no conclusions whatsoever can be drawn.

Computational simulations allow for sampling of the space of possible learners and lexicons. The results provide a sense of which constellation of parameters will lead to which outcomes. Both necessary and sufficient conditions can also be determined *conditionally*. That is, *if* (x and y) *then* (w or z). For example, if anti-markedness grammars are allowed into the learner's hypothesis space, *and* vowels are uniformly distributed in the input data, *then* the anti-markedness grammar is the hands-down winner. But if Mixture Grammars are allowed into the hypothesis space, with a winner-take-all Bayesian learning algorithm and an Information Theoretic prior, *then* the anti-markedness grammar can only rarely win.

Table 5 is a summary of the different learning scenarios that have been considered in this paper. For each row, 'Mixture Grammar' is a stand-in for the class of Mixture Grammars containing as sub-grammars the Simple Variability grammars listed in the corresponding 'Hypothesis Space' column. Any particular Mixture Grammar winner will have a unique set of weights associated with those sub-grammars. The first two rows of Table 5 illustrate the Bayesian learning results over the Uniform Lexicon; row 1: without including Mixture Grammars; row 2: allowing Mixture Grammars into the hypothesis space. The remaining rows are results derived from the Proportion-Based Evaluation Metric. Rows 3 and 4 each test the results of different lexical conditions. In row 3 the Uniform Lexicon is replaced with a set of lexicons in which the frequencies of the 8 vowels follow a Zipfian distribution. In row 4, /ə/ has been eliminated from the inventory, resulting in the set of No-Contrast Lexicons. In row 5 the hypothesis space is expanded once more to include the 'Stress ə' hypothesis. And in row 6, the sound change is modified to better reflect phonetic naturalness.

Table 5: Summary of learning results under varying hypothesis spaces, learners, lexicons, and types of sound change. For the Proportion-Based Evaluation Metric, the numbers correspond to the simulation results plotted in Fig. 2, specifically, the black dots corresponding to the Zipfian vowel distributions.

	Sound Change	Lexicon	Hypothesis Space	Evaluation Metric	Winner
1	a > ə	L_U	*GUJARATI**$^{\alpha}$ *GUJARATI*$^{\alpha}$ *PENULT*$^{\alpha}$	Bayesian	*GUJARATI**$^{\alpha}$ 100%
2	a > ə	L_U	*GUJARATI**$^{\alpha}$ *GUJARATI*$^{\alpha}$ *PENULT*$^{\alpha}$ *MAX*(*G**/*G*)$^{\alpha}$	Bayesian	*MAX*(*G**/*G*)$^{\alpha}$ 100%
3	a > ə	Zipfian vowel	*GUJARATI**$^{\alpha}$ *GUJARATI* $^{\alpha}$	Proportion-Based	*GUJARATI**$^{\alpha}$ *9.4%*

		distribution	PENULT$^{\alpha}$ MIXTURE	[1.25 level]	MIXTURE 90.6%
4	a > ə	L_{NC} (Zipfian)	GUJARATI*$^{\alpha}$ GUJARATI$^{\alpha}$ PENULT$^{\alpha}$ MIXTURE	Proportion-Based [1.25 level]	GUJARATI*$^{\alpha}$ 29.5% MIXTURE 70.5%
5	a > ə	L_{NC} (Zipfian)	GUJARATI*$^{\alpha}$ GUJARATI$^{\alpha}$ PENULT$^{\alpha}$ MIXTURE STRESS ə	Proportion-Based [1.25 level]	STRESS-ə 29.5% MIXTURE 70.5%
6	a > ə ˈa > ˈa	L_{NC} (Zipfian)	GUJARATI*$^{\alpha}$ GUJARATI$^{\alpha}$ PENULT$^{\alpha}$ MIXTURE	Proportion-Based [1.25 level]	GUJARATI$^{\alpha}$ 28.9% MIXTURE 71.1%

The results of rows 1 and 4 only obtain due to the exclusion of certain hypotheses from the space. In row 1, GUJARATI*$^{\alpha}$ only wins because it does not have to compete against a Mixture Hypothesis. And in row 4, GUJARATI*$^{\alpha}$ only wins because it does not have to compete with the 'Stress ə' hypothesis. Rows 3, 5, and 6, (bolded rows) represent the most informative solutions under differing models of sound change. If the sound change is independent of word and vowel inventories, then the results in row 3 hold (Fig. 2). However, an arguably better model is one in which the sound change applies with the highest likelihood to lexicons with few or no prior /ə/'s, providing the results in row 5. An even better model, however, is one in which the sound change is significantly revised.

In row 6, it is only the unstressed /a/'s that shift to /ə/'s – or, equivalently, all /a/'s shift, with a subsequent shift of stressed /ə/'s back to /a/'s. This scenario results in almost a complete return to the pre-change lexicon – the only difference being that there are no longer any unstressed surface /a/'s. Unstressed /a/'s only appeared in words with multiple /a/'s prior to sound change, so this class of words is relatively small in any case. The outcome in row 6 is reversion to a markedness-abiding grammar, offering no support for the UG-Delimited $\mathcal{H}$ Principle. The model predicts, however, a preponderance of GUJARATI/PENULT

Mixture Grammars, an outcome which is difficult to disprove, requiring detailed typological work.

8 Conclusion

The problem of rules that apply within restricted domains has been a topic of much study (see, e.g., Selkirk 1980; Nespor & Vogel 1986; McCarthy & Prince 1995; Inkelas, Orgun & Zoll 1997; Inkelas 1998; Lubowicz 2002). The enterprise has typically focused on finding an analysis that eliminates the need for exceptions, or at least minimizes them to the degree possible. But what if more than a mere handful of words diverge from a dominant pattern? What if there are subsets of data that belong to directly contradictory grammars, rather than opaque patterns in which a rule over- or under- applies? If there is no clear, or optimal, way for a learner to pick a single grammar, what does that mean for questions about linguistic universals?

As can be seen from Table 5, Mixture Grammars comprise the vast majority of simulation outcomes. As formally defined, Mixture Grammars contain at least two independent grammars. In many of the cases that have been investigated those grammars consist of both the universally preferred type: *GUJARATI*$^{\alpha}$, and the theoretically banned type: *GUJARATI*$*^{\alpha}$. Thus, the Mixture Grammar outcome should be excluded from the hypothesis space, according to the UG-Delimited $\mathcal{H}$ Principle. This step is only justified, however, if the presence of those grammars leads to inaccurate typological predictions. As far as I know, it is not currently possible to verify this. The typological facts are simply not clearly enough established. This is true even for the pure anti-markedness grammars[14], and it is even more true for potential Mixture-Grammar languages. For example, who is to say that systems that have been analyzed as exhibiting a high degree of lexical exceptionality, or gone largely unanalyzed due to what is

[14] There is some evidence for a collection of languages that do seem to violate universal markedness implications: Arrernte (codas preferred over onsets) (Breen & Pensalfini 1999); Sea Dayak (nasalized vowels allowed in nasal but not oral contexts) (Court 1970); Eastern Pomo (neutralization to aspirated (rather than plain voiceless) stop in coda position (McLendon 1975); Buryat (epenthesis of /g/, rather than /t/) (Poppe 1960), etc.

perceived as patternless behavior, might not belong to this set?[15]

Some may have the intuition that exceptionless grammars should never be defeated by Mixture Grammars. In which case, the current model will need to be modified in some way. However, the point must be made once more that we simply don't know. Typological evidence in its current state does not determine the form the hypothesis space should take; statistical learning tools are not guaranteed to provide psychologically appropriate significance levels; and it is not always possible to satisfy linguistic intuitions in a way that is completely self-consistent, and not ad hoc.

In attempting to determine the distribution of theoretically predicted grammars, we have come up against several general questions in linguistics which remain unresolved, such as whether probabilistic rule (or constraint) selection is the right model of linguistic knowledge; what statistical significance level (if any) is used by language learners; how strong homophony avoidance is as a force in sound change; etc. It has also been necessary to devise some way of dealing with issues which have been largely unexplored, at least from a formal perspective, such as the proper treatment of contradictory data. Once phonetics are incorporated into the model things change considerably; a reversed sonority-to-stress grammar becomes rather implausible – at least under the type of sound change proposed. Perhaps phonetic forces will cause anti-markedness languages to instantaneously revert to a natural (markedness-abiding) state. Or perhaps there is a brief moment in time when such pathological grammars can be observed.

This paper has focused on one particular type of system: a Gujarati-like language, under a context-free vowel shift, and the effect on a sonority-sensitive stress system. It is certainly possible that one could find a better scenario to argue for the necessity of the UG-Delimited $\mathcal{H}$ Principle. However, this case study

[15] In fact, the situation is more uncertain even than this. It has been estimated that only about 10% of currently spoken languages have been adequately documented; furthermore the 5,000-8,000 currently spoken languages must be considered a sample of possible human languages, both past and future. Given what is known about language families and historical changes, the number of languages that have ever existed can be estimated at about half a million, leaving us with typological knowledge of .02% of possible linguistic diversity (Evans & Levinson 2009). If this is even close to correct, strong claims about attested phonological patterns become extremely problematic.

illustrates a number of properties that are general to the study of linguistic systems. Among the most notable of these is the ability of sound change to leave a system in an intermediate state without a majority rule. In order for any proposed scenario to be compelling it must provide a model of the language learner that is capable of dealing with such data. Without explicitly specifying this learner one runs the risk of assuming properties which are mutually incompatible, inconsistent with empirical data, or even inconsistent with one's theoretical position.

It is far from a foregone conclusion that there is a single, unique mechanism that is necessary to account for the observed typology. Principles of phonetics may be enough to shape the distribution of phonological grammars to the degree that we observe. Principles of learning, or principles of word formation may also function in a similar way. Thus, the burden of proof rests with those who wish to argue for the *necessity* of UG in limiting the learner's hypothesis space in particular ways. It is my hope that this paper may serve as a general guide as to what that burden entails, and to what foundational issues must be resolved within linguistic theory in order to reach a final conclusion.

References

(1993). CELEX English database (Release E25) [On-line], Nijmegen: Center for Lexical Information [Producer and Distributor].

Albright, A. and B. Hayes (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 90: 119-161.

Boersma, P. (1998). *Functional Phonology*. University of Amsterdam.

Boersma, P. and B. Hayes (2001). Empirical Tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32(1): 45-86.

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

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning* 34: 71-105.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press.

Bybee, J. (2006). Language change and universals. *Linguistic universals*, 179-194. Cambridge University Press.

Chater, N., J. B. Tenenbaum, and A. Yuille (2006). Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Science* 10(7): 287-291.

Choi, J. D. (1992). *Phonetic Underspecification and Target Interpolation: An Acoustic Study of Marshallese Vowel Allophony*. UCLA.

Chomsky, N. (1965). *Aspects of the Theory of Syntax. The MIT Press.*

Chomsky, N. and M. Halle (1968). *The Sound Pattern Of English*. Harper & Row.

Colarusso, J. (1988). *The Northwest Caucasian Languages: A Phonological Survey*. Garland Publishing.

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

Crosswhite, K. M. (2000). Sonority Driven Reduction. Proceedings of the 26th Berkeley Linguistics Society Meeting.

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

de Lacy, P. (2007). The interaction of tone, sonority, and prosodic structure. In
P. de Lacy (ed.) The Cambridge handbook of phonology, pp. 281-307. Cambridge University Press.

de Lacy, P., and J. Kingston (2013). Synchronic explanation. *Natural Language & Linguistic Theory* 31(2): 287-355.

Doctor, R. (2004). *A grammar of Gujarati*. Lincom Europa.

Evans, N. and S. C. Levinson (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32: 429-492.

Gibson, E. and K. Wexler (1994). Triggers. *Linguistic Inquiry* 25(3): 407-454.

Goldsmith, J. (2002). Probabilistic models of grammar: Phonology as information minimization. *Phonological Studies* 5: 21-46.

Goldwater, S. and M. Johnson (2003). Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, Stockholm University, pp. 111-120.

Gopnik, A. , C. Glymour, D.M. Sobel, L.E. Schulz, T. Kushner and D. Danks (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review* 111: 3-32.

Gordon, M. (2006). *Syllable Weight: Phonetics, Phonology, Typology*. Routledge.

Halle, M. and M. Kenstowicz (1991). The Free Element Condition and cyclic versus non-cyclic stress. *Linguistic Inquiry* 22: 457-501.

Hay, J., J. B. Pierrehumbert, and M. Beckman (2003). Speech Perception, Well-formedness and the Statistics of the Lexicon. In *Papers in Laboratory Phonology VI* (pp. 58-74). Cambridge University Press.

Hayes, B. and C. Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3): 379-440.

Hudson Kam, C.L. and Newport, E.L. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and

change
Language Learning and Development 1, 151-195.

Inkelas, S. (1998). The theoretical status of morphologically conditioned phonology: a case study of dominance effects. In *Yearbook of Morphology 1997* (pp. 121-155). Springer Netherlands.

Inkelas, S., O. Orgun and C. Zoll (1997). The implications of lexical exceptions for the nature of grammar. In *Derivations and Constraints in Phonology*. 393-418.

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*, pp 50-59. New York City.

Kager, R. (1989). *A metrical theory of stress and destressing in English and Dutch*. Dordrecht: Foris.

Kemp, C., A. Perfors, and J.B. Tenenbaum (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science* 10(3): 307-321.

Kenstowicz, M. (1996). Quality-sensitive Stress. *Rivista di Linguistica* 9: 157-187.

Kersten, D. and A. Yuille (2003). Bayesian models of object perception. *Current Opinions in Neurobiology* 13: 150-158.

Kiparsky, P. (1982). From cyclic phonology to lexical phonology. In H. V. D. Hulst and N. Smith (eds.) *The structure of phonological representations* (pp.131-172). Foris.

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

Kiparsky, P. (2008). Universals constrain change; change results in typological generalizations. *Linguistic universals and language change*, 23-53.

Kondo, Y. (1994). Targetless schwa: is that how we get the impression of stress
timing in English? In *Proceedings of the Edinburgh linguistics department conference,* pp. 63-76.

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

Lehiste, I. (1970). *Suprasegmentals*. MIT Press.

Liberman, M. and A. Prince (1977). On stress and linguistic rhythm. *Linguistic Inquiry* 8: 249-336.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the*
Acoustical Society of America 35: 1773-1781.

Lubowicz, A. (2002). Derived environment effects in Optimality Theory. *Lingua* 112: 243-280.

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

McCarthy, J. and A. Prince (1995). Faithfulness and reduplicative identity. *ScholarWorks@UMass Amherst [http://scholarworks.umass.edu/cgi/oai2.cgi] (United States), ScholarWorks@UMass Amherst*

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

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

Moreton, E. (2009). Underphonologization and modularity bias. In S. Parker (ed.) *Phonological Argumentation: Essays on Evidence and Motivation*. London: Equinox.

Nespor, M. and I. Vogel (1986). *Prosodic Phonology. Foris Publications.*

Neyman, J. and E.S. Pearson (1933). On the problem of the most efficient test of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A* 231: 289-337.

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

Pearl, L. (2011). When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. *Language Acquisition* 18(2): 87-120

Pearl, L. and J. Lidz (2009). When domain-general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development* 5(4): 235-265.

Pearl, L. & A. Weinberg (2007). Input Filtering in Syntactic Acquisition: Answers from Language Change Modeling. *Language Learning and Development* 3(1): 43-72.

Phillips, B. (1984). Word frequency and the actuation of sound change. *Language* 60(2): 320-342.

Pierrehumbert, J. B. (2001). Exemplar dynamics: word frequency, lenition and contrast. In J. Bybee and P. J. Hopper (eds.) *Frequency effects and the emergence of linguistic structure* (pp. 137-158). John Benjamins.

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

Prince, A. and P. Smolensky (2004). Optimality Theory. *Blackwell Publishing.*

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

Ross, D. S. and E. L. Newport (1996). The development of language from non-native linguistic input. *Proceedings of the 20th annual Boston University Conference on Language Development*, Boston: Cascadilla.

Selkirk, E.O. (1980). Prosodic domains in phonology: Sanskrit revisited. In M. Aronoff, M. & M.-L. Kean (eds.) *Juncture, Anma Libri,* 107-129.

Senghas, A. and M. Coppola (2001). The creation of Nicaraguan Sign Language by children: Language genesis as language acquisition. *Psychological Science* 12: 323–328.

Singleton, J. L. (1989). *Restructuring of language from impoverished input: Evidence for linguistic compensation*. University of Illinois at Urbana-Champaign.

Smith, J. L. (2000). Prominence, augmentation and neutralization in phonology. *Proceedings of the 26th Berkeley Linguistics Society Meeting* (pp. 247-257).

Surendran, D. and P. Niyogi (2006). Quantifying the functional load of phonemic oppositions, distinctive features, and suprasegmentals. In O. Nedergaard Thomsen (ed.) *Current trends in the theory of linguistic change. In commemoration of Eugenio Coseriu (1921-2002)*, Benjamins.

Suthar, B. (2003). *Gujarati-English Learner's Dictionary*. Nirman Foundation.

Tenenbaum, J. B. and T. L. Griffiths (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences* 24: 629-641.

Tenenbaum, J. B., T.L. Griffiths and C. Kemp (2006). Theory-based Bayesian models of inductive reasoning. *Trends in Cognitive Sciences* 10(7): 309-318.

Van Bergen, D. (1994). A model of coarticulatory effects on the schwa. *Speech Communication* 14: 143-162.

Wilson, C. (2006) Learning Phonology with Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cognitive Science* 30: 945-982.

Xu, F. and J. B. Tenenbaum (2007). Word learning as Bayesian inference. *Psychological Review* 114(2): 245-272.

Yang, C. (1999). A Selectionist Theory of Language Acquisition. *27th Annual Meeting of the Association for Computational Linguistics,* 429-435.

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

Yang, C. (2005). On Productivity. *Linguistic Variation Yearbook* 5: 265-302.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

Zuraw, K. R. (2000). *Patterned Exceptions in Phonology*. UCLA.