

Likely Input and the Reasonable Learner

Rebecca L. Morley

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. The validity of this hypothesis is assessed by examining the implications of one sound change ($a > \text{ə}$) for a specific phonological process (stress assignment), making explicit assumptions about the structure of words affected by that sound change, and the learner who takes those words as input. Based on the current state of knowledge in the areas of typology, language learning, sound change, and linguistic competence, the conclusion is that restrictions on possible end-point languages are unnecessary – or at least currently unjustified. For a large class of possible lexicons, the outcome of the above sound change is data that are not just ambiguous, but inconsistent, with respect to a single generating rule, or grammar. The larger implications for conventional characterizations of grammatical competence, and for the application of standard probabilistic approaches to phonological learning are discussed.

Keywords: Markedness; Universal Grammar; Evolutionary Phonology; Sonority-Based Stress;

Statistical Learning

1 Introduction

The basic tenet of Evolutionary Phonology is that the observed commonalities in phonological systems of the world arise from the universal commonality of the way listeners and speakers produce, perceive, misproduce and misperceive sound structures (Blevins 2004; Ohala 1981, 1990, 1993). For example, the tendency to run sounds together in natural speech is posited to lead to listener error in parsing features that become spread over multiple segments. This kind of error can produce patterns of place assimilation where formerly /np/ clusters come to be analyzed as /mp/ clusters phonologically. Perhaps counter-intuitively, such a framework, based in phonetic (articulatory and perceptual) naturalness, predicts the possibility of unnatural rules or grammars. These could result, as we will shortly see, from a series of natural changes, each targeting different parts of the sound system of the language and acting independently and autonomously.

Standard constraint-based generative phonological theory is in opposition to this view in that it bans some of these possible grammars. A number of cross-linguistic generalizations about possible and impossible phonological patterns have been projected from many of the same phonetic laws that are taken to drive sound change. Such generalizations make up the universals from which markedness constraints are drawn within Optimality Theory (Prince and Smolensky 1993/2004). For example, the finding that syllables with onsets seem to be preferred while coda consonants are often lacking across languages is the foundation for the set of constraints {ONSET, NOCODA}. Sequences of nasals and voiceless obstruents are often avoided cross-linguistically, but if they are allowed, so will be sequences of nasals and voiced obstruents (*NT) (e.g., Pater 1996). Voiced segments are often neutralized to their voiceless counterparts in final position, supporting positional constraints against voicing (or specification of the laryngeal feature: *LAR) (e.g., Lombardi 1995). And stress is preferentially assigned to high 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). The instantiation of these generalizations either within a constraint set or a constraint hierarchy is a strong prediction about

certain languages that cannot exist, namely, those that would reverse the above implications and others like them such that, e.g., codas were preferred and onsets avoided¹.

It is, in fact, fairly easy to find examples of historical sound changes that disrupted natural patterns by rendering them opaque (that is, non-surface true) (see, e.g., Hooper (1974), Pan'kevich (1938) (cited in Kenstowicz and Kisseberth (1977)), Anderson (1981), Buckley (2000), Bach and Harms (1972)). It is less easy to find examples of systems that are not just messy or idiosyncratic (that is, 'unnatural'), but that actually instantiate patterns that defy phonetic laws as we know them (and are, thus, 'anti-natural').

While sound change may not always optimize or simplify systems as a whole, a limit has been proposed for exactly how much such changes can disrupt well-behaved phonological patterns. In particular, it has been claimed that even if language structures become less natural over time they cannot become anti-natural (de Lacy 2006, Kingston & de Lacy 2006, Kiparsky 2004, 2006, Moreton 2009). That is, a grammar that would violate the typological universals just described, an 'anti-markedness' grammar, is an impossible synchronic system, and thus an impossible outcome of sound change. This claim is based in the Chomskyan idea of Universal Grammar, and, as far as I understand it, hypothesizes innate (although not always synchronically active) forces which would serve to prevent certain sound changes just in case they would result in such a 'disallowed' system.

In fact, without a filter of this kind, Kiparsky argues, common and natural sound changes would frequently produce unnatural, and in fact unobserved, 'anti-markedness' languages (such as a system in which lower sonority vowels were stressed in preference to higher sonority vowels). I will call this the Universal-Grammar-Delimited Hypothesis Space (UG-Delimited \mathcal{H}) Principle. This principle is based on the assumption that the minimal version of Evolutionary Phonology outlined above over-generates in a particular way, predicting higher rates for typologically unattested or rare languages than are actually observed. That is the claim investigated in this paper.

¹ But see Breen and Pensalfini for an argument for preferred VC structure in Arrernte (1999), Hyman (2001) for an argument for *ND in Tswana, and Yu (2004) for an argument for final voicing in Lezgian.

2 Some Notes on Methodology and Argument Structure

One might say that a given linguistic theory is defined largely by which grammars it predicts to be possible and which, impossible (or, which likely and which, unlikely). Perhaps surprisingly then, it is not always straightforward to determine what those predictions are. In the first place, a thorough understanding of the operative principles of sound change is clearly critical to the debate about how unnatural grammars arise. But this is just one piece of the puzzle. Whatever one's views about the principles of sound change, the evidence for any such event will be found in the words which are composed of those sounds. And whatever one's views about the endpoint of learning, the input to the learner (the physical data) will consist of those words (and their combinations). This may seem like an obvious point, but it is worth stressing as it forms the foundation for the arguments in this paper.

Consider the learner who must decide whether the absence of words like [blik] from their lexicon constitutes a systematic avoidance of particular sound combinations or an accidental gap in their inventory. This learner, under one theory, is trying to determine whether or not their phonological grammar should rank a constraint like *bl over faithfulness to such sequences. The only way to decide this question (for the learner as well as the linguist) is to look at the distribution of known words. For example, does the sequence [bl] occur at the beginning of other words? And does it occur as often as other complex clusters? These are the types of considerations that must be taken into account, and they involve assessing the data in order to determine the presence of a regular pattern. Thus, the question about the emergence of an anti-markedness grammar is actually a question about two events: a sound change that gives rise to a lexicon that appears to reflect an underlying grammar of the dispreferred class, and the induction from that data, on the part of the learner, of that specific grammar.

Neither of the above two conditions, in fact, are necessarily obviously true or likely. That is to say, the transition from individual phones, or words, to entire grammars is not a trivial theoretical step. It may be intuitively plausible that a large set of known sound changes (especially acting in unpredictable combination) could easily produce a large variety of idiosyncratic and 'crazy' patterns (Bach and Harms 1972), and ultimately a number of 'anti-natural' ones as well. However, this intuition, and its complement -- that all phonological systems are capable of being generated via natural sound shifts alone -- are rarely made concrete

enough for an assessment of their validity. What must be done is to characterize, explicitly, the properties of the language learner, as well as the data to which that learner is exposed. For this reason, the methodology of this paper is computational modeling and simulation.

There is a large body of work on computational learning that provides a number of useful starting points (e.g., Pearl & Lidz (2009), Jarosz (2006), Tenenbaum & Griffiths (2001), Yang (1999), Gibson & Wexler (1994)). However, the modeled learner is rarely given realistic, lexicon-sized amounts of data over which to compute (although, see Hayes and Wilson (2008), and Albright & Hayes (2003)). Furthermore, although there exist a number of contributions to modeling language change as the result of language acquisition (see, e.g., Yang (2000), Niyogi (1997), Pearl (1997)), directly contradictory (and not just ambiguous) data that could quite easily result from sound change is typically left out of such computations (although see Yang (2005) for the treatment of exceptions). Finally, the far-reaching repercussions for theoretical linguistics as a whole are often absent from the conclusions of the computational work, and the level of explicitness necessary for computational implementations are even more frequently missing from the theoretical linguistics side. This paper seeks to fill in these gaps, and bring together what are often isolated strands of research.

The results of this work should be viewed in the following way: as general findings, addressing the general questions of what cognitive and linguistic structures would be sufficient to produce the phonological patterns that are attested in the world's languages. In order to conduct such an analysis, however, it is necessary to consider a specific scenario: a particular hypothetical phonological system, coupled with a particular hypothetical sound change, resulting in a particular hypothetical lexicon, input to a particular hypothetical learner.

In the following section I will walk through a case study of sonority sensitive stress (for the Gujarati language: a markedness-abiding sonority-to-stress pattern), paying special attention to the lexicon that might be produced after a hypothetical sound change of the type Kiparsky proposes ($a > \text{ə}$: Gujarati \rightarrow Gujarati'). In Section 4 the output of a learning model will be analyzed with regards to its treatment of this hypothetical Gujarati' data; in particular, to determine under what circumstances the anti-markedness grammar, what I will call *GUJARATI**, arises (beating out and replacing the well-behaved *GUJARATI* grammar). While calculations are primarily performed for a single possible lexicon (the uniform distribution case), and a single possible learner (a Bayesian model), these results are subsequently extended. In Section 5 other

members of the class of learners are considered. And in Section 6 a range of lexicons are modeled by varying the probability distributions over word types and the vowels that comprise them. In Section 7 additional relevant modeling concerns are briefly addressed. The paper concludes, in Section 8, with a summary and some provisional conclusions relating to the question of allowed and disallowed languages.

3 Gujarati Phonology

With regards to the theoretical claims under investigation there is nothing particularly special about Gujarati (an Indo-Aryan language of India). Kiparsky introduces it to provide a concrete illustration of the relevant phonological paradigm of a sonority sensitive stress system that respects the posited universal implicational hierarchy. And, in order to follow Kiparsky's argument as closely as possible, it is adopted as the case study in this paper. 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). He describes the stress system as conforming to the following position and sonority dependent rules².

- (1) GUJARATI: Sonority & Position –to-Stress:
- (1) stress penultimate [a] (the most sonorous vowel)
 - (2) otherwise stress ante-penultimate [a]
 - (3) otherwise stress final [a]
 - (4) otherwise stress penultimate mid-sonority vowel (any of [i,e,ɛ,o,ɔ,u])
 - (5) otherwise stress ante-penultimate mid-sonority vowel
 - (6) 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 OT framework (Prince and Smolensky 1993/2004) that utilizes a universal sonority scale aligned with a foot prominence scale (*P-foot/ə >> *P-foot/e >> *P-foot/a: where *P-foot/a is interpreted

² 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.

as the dispreference for having the vowel /a/ as the peak of the foot, that is, the stressed position) (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 alternate approach). These types of stress systems are not uncommon among the world's languages (e.g., Kenstowicz (1996) lists at least six languages whose stress patterns he analyzes 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).

3.1 Gujarati'

One possible sound change that could disrupt the currently natural stress pattern in a language like Gujarati is a change in vowel quality. Since vowel quality is correlated with sonority, a change in the former (from an independent source) could have repercussions for any phonological pattern that is conditioned by the latter. Crucially, this is only true if such a sound change could occur without disrupting the current stress pattern³. If instead, stress shifted in response to a shift in vowel height or backness then the natural pattern would be maintained. Although Kiparsky is not explicit about how such a change would affect stress, I can find no other way to advance the argument, and so attribute the following assumption to the UG-Delimited \mathcal{H} Principle position:

- (2) A common and natural type of sound change is one in which all /a/'s of a language change to /ə/'s, independent of stress.

An explicit characterization of the learning scenario is also missing from the argument. My interpretation of the learner envisioned here corresponds to (3).

- (3) Learners entertain (undecomposed) complex hypotheses (like that in (1)).

³ This could be the case if, for example, stress position had previously become lexicalized

These assumptions will be left largely uncontested (although see Section 7 for a brief discussion). This is not because they are particularly plausible⁴ but because it is clear that assumptions of some kind must be made in order to proceed. It also turns out to be instructive to consider the consequences of such assumptions. It should be kept in mind throughout that this particular sound change is being considered only as a stand-in for a class of possible changes that could produce markedness reversals, and this particular learner for any of a number of possible ones. Future changes to the initial assumptions will affect the modeling results, but they will not change the essential argument, nor invalidate the model itself.

The change in vowel quality given in (2), if it leaves stress placement unaffected, will alter the make-up of the Gujarati lexicon, and introduce the possibility of a system in which stress preferentially falls on the lowest-sonority vowel, [ə] (formerly [a], the most sonorous vowel)⁵. For simplicity's sake it will be further assumed that

(4) Sound change is instantaneous and complete.

This allows the same fully transformed lexicon to act as the input to all learners of Gujarati' at some later time.

What learning means, in generative linguistics, and in this paper, is that the learner has converged on the grammar that generates Gujarati'. This process is often conceptualized as hypothesis selection from a space of candidate hypotheses. Therefore, for the modeler interested in determining the outcome of learning over the transformed lexicon, it is necessary first to characterize the learner's hypothesis space in some way.

A particular set of hypotheses is considered in this paper, just one of many such possible sets. The motivation behind this choice is the following: to capture what is linguistically most relevant about stress placement, and at the same time write down hypotheses in a form that can be easily handled by the Bayesian apparatus. This actually results, at first pass, in a departure

⁴ Assumption (3) affords one of the simplest formulations of the learning process itself (where, in one stage, learners pick the correct hypothesis out of a pre-existing candidate set). I also view this as a minimal departure from the standard formalisms both of linguistic theory and Bayesian learning.

⁵ An alternative traditional generativist account of the post-change language attributes differences in stress assignment to the underlying representations rather than the stress rules. Under this formulation, /a/'s remain /a/'s underlyingly, and an opaque ordering of the stress assignment rule before the vowel shift produces the surface lexicon. If it is assumed that only underlying representations are affected by sound change then rules which start out as natural can never become unnatural (because they never change). This analysis effectively encodes the diachronic change (a > ə) within the synchronic grammar (a → ə), precluding the possibility of an anti-markedness grammar.

from well-established analyses which make use of complicated grammars to capture detailed variation in natural language stress patterns (see, for example, Liberman & Prince (1977), Kager (1989), Halle & Kenstowicz (1991), Pater (2000)). The simpler characterization of hypotheses in this paper is also meant to allow the distinctions between different classes of grammar (markedness abiding and anti-markedness) to be made unambiguously. Following Assumption (3), the hypotheses will be treated as holistic entities. This is a departure from work which treats learning as parameter setting or constraint ranking, where a given hypothesis corresponds to a full set of parameters or ranked constraints (e.g., Gibson & Wexler (1994), Tesar & Smolensky (1998)). Although those approaches change the algorithm and the space of outcomes, as well as potentially increase the complexity of the learning problem (see, e.g., Pearl (submitted)), the issues with respect to the distribution of data encountered by the learner remain the same. Since the primary concern of this paper is the effect of contradictory data on the outcome of learning, Assumption (3) is maintained.

The list in (5) represents the full hypothesis space assumed to be available to the learner. Consideration will be restricted to hypotheses 5.1-5.3 for now, leaving aside the discussion and full definition of hypotheses 5.4 and 5.5 until later sections.

(5) \mathcal{H} : Hypothesis Space

- (1) PENULT: Stress Penultimate vowel
- (2) GUJARATI: Sonority & Position –to-Stress
- (3) GUJARATI^{*}: Reversed-Sonority & Position –to-Stress (as in (1), but with the sonority classes reversed (i.e., /ə/ and /a/ exchanged).
- (4) NULL(G^{*}/G): GUJARATI^{*} and GUJARATI equally likely generators of data
- (5) MAX(G^{*}/G): mixed-grammar of GUJARATI^{*} and GUJARATI with variable weights

3.2 Evidence to the Learner

The hypothetical scenario is the following: a diachronic change affecting the identity (and thus sonority) of a subset of vowels in the inventory (but leaving lexical stress location unchanged) has occurred in the sonority sensitive stress system of Gujarati (or a Gujarati-like language). The question is this: how likely is *GUJARATI*^{*} – the anti-markedness, reversed sonority-to-stress

hypothesis – likely to emerge as the grammar for a subsequent generation of Gujarati speakers (learners of Gujarati')⁶.

The hypothetical lexicon, L' , of Gujarati' depends on the lexicon, L , of the old Gujarati. And, as will become clear shortly, the exact shape of L' will strongly affect the outcome of learning. For modeling purposes, however, what is important is not the actual lexicon of the Gujarati spoken in present-day India. Remember, Gujarati itself is a stand-in for *any* well-behaved phonological system. Instead of a single L , the simulations should run over a sample from the set of all *possible* lexicons for a Gujarati-like language. In Section 6 such a sample of lexicons is constructed. Until then consideration will be restricted to a single member of this set.

For a given possible Gujarati, a possible L is mapped to a possible L' via the sound change $a > ə$. L' is characterized by some distribution over all possible words. However, for the purposes of stress assignment, the only relevant attributes of those words are the vowel identities and their temporal order. Therefore a given word type is defined by the particular sequence of vowels it contains, disregarding the consonants. This significantly reduces the number of possible words that must be listed. Furthermore, only two- and three- syllable words will be included, under the assumption that they comprise the majority of the reliable indicators of stress assignment in Gujarati (de Lacy 2006).

Any particular lexicon, L , can be created by drawing from the set of possible word types, \mathcal{L} , according to some specified sampling process. The particular L employed in the first half of this paper – L_{MU} – is the lexicon that selects exactly one token of each type from \mathcal{L} . L_{MU} is the Minimal Uniform Lexicon : all types occur at the same frequency, and that frequency is once. This also means that all vowels are equally likely, and equally likely in all word positions. For three-syllable words and an eight vowel inventory, there are 8^3 , or 512 distinct types. For two-syllable words, there are 8^2 , or 64 types. Tables 1 and 2 list the word types for three- and two-syllable words respectively. 'Case' refers to the type (vowel sequence) of the word before the hypothetical sound change (where M is shorthand for any of the mid-sonority vowel class {i,e,ɛ,o,ɔ,u}). (Gujarati word examples taken from de Lacy (2006) and Suthar (2003)).

⁶ It should be noted that a very real competitor that is excluded from this space is the "stress ə" grammar: a simpler hypothesis than *GUJARATI'* that makes no reference to relative sonority. This hypothesis is discussed in Section 7.

Table 1. Full set of all possible three-syllable word types for stress (\mathcal{L}). All possible lexicons, L , are created by sampling from this space. These numbers also represent a particular lexicon, L_{MU} , the Uniform Gujarati Lexicon. Final column gives number of types and hypotheses with which the data are consistent. G^* (*GUJARATI**), G (*GUJARATI*), P (*PENULT*). (M is shorthand for any of the mid-sonority vowel class $\{i, e, \varepsilon, o, \varepsilon, u\}$).

	Case Gujarati Vowel- Template	Example $L > L'$	# types H
1	(ə, ə, a)	[pəkʃəpát] > [pəkʃəpát]	21
	(ə, M, a)	[pərikʃá] > [pərikʃá]	
	(a, ə, M)	[tábəɖtob] > [tábəɖtob]	
	(M, ə, a)	[ucc ^h əvás] > [ecc ^h əvəs]	
	(a, ə, a)	[jájərmən] > [jájərmən]	
	(a, ə, ə)	[pá[nəgər] > [pə[nəgər]	
2	(M, M, a)	[hoʃijár] > [hoʃijər]	84 G^*
	(a, M, M)	[járirik] > [járirik]	
	(a, M, a)	[háɖohəɖ] > [háɖohəɖ]	
	(a, M, ə)	[p ^h ásigər] > [p ^h əsigər]	
3	(M, a, a)	[durácar] > [durácər]	48 G^*, P
	(M, a, ə)	[mubárək] > [mubárək]	
	(M, a, M)	[betá[is] > [betá[is]	
4	(M, M, ə)	[tʃum:ótər] > [tʃum:ótər]	78 G, P
	(ə, M, ə)	[vəriʃ ^h ə] > [vəriʃ ^h ə]	
	(ə, M, M)	[kəʃóro] > [kəʃóro]	
5	(M, ə, M)	[kójəldi] > [kójəldi]	42 G
	(M, ə, ə)	[kʃétrəp ^h ə] > [kʃétrəp ^h ə]	
6	(a, a, a)	[aw:ánā] > [əw:ənā]	239 G, G^*, P
	(a, a, M)	[amdáni] > [əmdáni]	
	(ə, a, a)	[resádar] > [resádər]	
	(ə, a, ə)	[səp ^h ácət] > [səp ^h ácət]	
	(ə, a, M)	[g ^h əʃáɖo] > [g ^h əʃáɖo]	
	(ə, ə, ə)	[əkbánd ^h ə] > [əkbánd ^h ə]	
	(ə, ə, M)	[cəkcákit] > [cəkcákit]	
	(M, M, M)	[i ^h ʃ ^h óter] > [i ^h ʃ ^h óter]	
	(a, a, ə)	[j ^h agmágət] > [j ^h əgmágət]	

Table 2. Full set of all possible two-syllable word types for stress (\mathcal{L}). All possible lexicons, L , are created by sampling from this space. These numbers also represent a particular lexicon, L_{MU} , the Uniform Gujarati Lexicon. Final column gives number of types and hypotheses with which the data are consistent. G^* (*GUJARATI**), G (*GUJARATI*), P (*PENULT*). (M is shorthand for any of the mid-sonority vowel class $\{i, e, \varepsilon, o, \omega, u\}$).

	Case Gujarati Vowel-Template	Example $L > L'$	# types H
1	(ə,a)	[pəgár] > [pəgér]	1
2	(M,a)	[ʃikár] > [ʃikór]	6 G^*
3	(M,ə)	[díwəs] > [díwəs]	6 G, P
4	(a,a)	[rája] > [rəjə]	51 G, G^*, P
	(a,ə)	[gádʒər] > [gádʒər]	
	(a,M)	[p ^h ájdo] > [p ^j ájdo]	
	(ə,ə)	[bákbək] > [bákbək]	
	(ə,M)	[máso] > [máso]	
	(M,M)	[lék ^h e] > [lék ^h e]	

Each stress case in L represents positive evidence in L' for some subset of the hypotheses under consideration as grammars for Gujarati'; the hypotheses consistent with a given case are specified in the column headed 'Hypothesis'. At the moment only three hypotheses: (5.3) *GUJARATI** (Reversed-Sonority & Position –to–Stress), (5.2) *GUJARATI* (Sonority & Position –to–Stress) and (5.1) *PENULT* (Stress Penultimate vowel) are considered. As can be seen from Table 1, for example, in Row 3, the word [mubárək] in Gujarati, with stress determined by the sonority-to-stress grammar described in (1) has become [mubérək] in Gujarati', after the hypothetical sound change $a > \varepsilon$ – and 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 pattern is consistent with the reversed sonority-to-stress grammar *GUJARATI**. However, the stress placement in this word is also consistent with the simple penultimate stress grammar *PENULT*. Indicating the number of types that support none of the hypotheses as A (=arbitrary), and the number that support all hypotheses as N (= neutral), then the total type counts in support of each hypothesis are given in Table 3. These values are

calculated by summing across all rows the number of types in each row that support a given hypothesis. For example, from Table 1, for three-syllable words, the number of types from Rows 2, 3, and 6 are summed to give G^* , Rows 4, 5 and 6 are summed to give G ; Row 1 gives A ; Row 6 gives N ; and Rows 3, 4 and 6 are summed to give P .

Table 3. Total vowel-template type numbers in support of each hypothesis for minimal uniform lexicon; N: consistent with all hypotheses; A: consistent with no hypotheses.

	A	G^*	G	N	P	Total
3-syllable words	21	371	359	239	365	512
2-syllable words	1	57	57	51	57	64

4 The Learner

Taking the minimal uniform lexicon, L_{MU} , as the lexicon of Gujarati, Tables 1 and 2 represent the input to the learner of Gujarati' after the hypothetical sound change. These data present a potential challenge to a learner looking for a single consistent pattern. As Table 3 shows, the hypothesis that maximizes the probability of the data — the one consistent with the largest subset of words — is *GUJARATI**, the anti-markedness hypothesis. Consistent with 371 of 512 lexical items in the three-syllable case, this hypothesis however gives an improvement over its nearest competitor (*PENULT*, with 365 consistent items) of just 1.2% of the total data (6 items); in the two-syllable case, the three hypotheses are exactly tied⁷. Tables 4 and 5 show, for three- and two- syllable words, respectively, the performance of the three hypotheses, *GUJARATI**, *GUJARATI* and *PENULT*, on, column 2: all data, on, column 3: 'critical data' (data that can provide evidence to discriminate between two hypotheses, that is, excluding all N and A types), and, column 4: coverage advantage relative to the next-best hypothesis.

⁷ It is worth pointing out, as an anonymous reviewer notes, that this 1.2% coverage difference is the same no matter how large the lexicon gets. That is, there is an inherent ambiguity level in the system that allows for a fixed percentage of discriminating data points. What this means for learning when the lexicon is treated as a sample from an underlying distribution will be discussed in Section 5.

Table 4. Performance of hypotheses on three syllable words over uniform lexicon by percentage, and percentage advantage.

Hypothesis	all consistent	unambiguous consistent	advantage
G*	(371/512) = 72 %	(132/252) = 52 %	1.2 %
P	(365/512) = 71 %	(126/252) = 50 %	1.2 %
G	(359/512) = 70 %	(120/252) = 48 %	--

Table 5. Performance of hypotheses on two syllable words over uniform lexicon by percentage, and percentage advantage.

Hypothesis	all consistent	unambiguous consistent	advantage
G*	(57/64) = 89 %	(6/12) = 50 %	0
P	(57/64) = 89 %	(6/12) = 50 %	0
G	(57/64) = 89 %	(6/12) = 50 %	0

These numbers represent one possible input data scenario; later, in Section 6, the set of possible lexicons will be sampled more fully. For now, the question is, for what kind of learner does this data lead to an outcome where one hypothesis can or will be selected over its competitors? In the rest of this section a particular learner (a Bayesian model) will be examined. This learner is characterized as belonging to a class defined as being ‘conservative’: outcome sensitive to very small differences in data coverage. As a result, for unrestricted hypothesis spaces, output grammars will closely match the distribution of input lexicons. What this means for the central question of this paper, as well as whether such highly sensitive learning models are appropriate to the learning acquisition problem will be discussed in later sections.

4.1 A Bayesian Learner

A reasonable place to start in formalizing the properties of a hypothetical learner is with the Bayesian framework. This model involves a fairly minimal and intuitive apparatus and has been extensively applied to learning scenarios in a number of cognitive domains (e.g., Kemp *et al.* 2007, Tenenbaum *et al.* 2007, Xu & Tenenbaum 2007, Chater *et al.* 2006, Kording & Wolpert 2006, Gopnik *et al.* 2004, Kersten & Yuille 2003, Tenenbaum & Griffiths (2001)). The learner entertains a space of hypotheses, from which single or multiple winners may emerge. Each hypothesis has assigned to it a prior probability, which represents a learner’s bias towards

selecting that hypothesis before actually encountering any data. The prior probability of each hypothesis is combined with the likelihood of the data under that hypothesis: the conditional probability of the input data, given that it was generated by the hypothesis. Hypotheses that assign higher probability to the data have an increased chance of prevailing; this is moderated by their prior probability. The posterior probability, that the hypothesis is the correct one, given the data, can be calculated using Bayes' Theorem:

$$p(h | d) = \frac{p(d | h)p(h)}{p(d)} \quad (6)$$

For the problem at hand the members of d are stress assignments corresponding to each of the T words of the lexicon. 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 Tables 1 and 2) 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 (where $h = GUJARATI^*$, $PENULT$, or $GUJARATI$) can then be expanded as the product of the probability of each member of d given h , and each member of y .

The first and immediate result of modeling this type of learner is that none of the three hypotheses ever has a posterior probability greater than zero. Applying Bayes' Theorem to the minimal uniform lexicon case, with the hypotheses as defined in 5.1-5.3, produces no winner. This is because there is some set of data that is inconsistent with each one of $GUJARATI^*$, $GUJARATI$ and $PENULT$. The data that support hypothesis $GUJARATI$ in rows 4 and 5 of Table 1, for example, directly contradict hypothesis $GUJARATI^*$, and the reverse situation holds for rows 2 and 3. For any word where stress does not occur on the penultimate syllable, hypothesis $PENULT$ assigns zero probability (Rows 1, 2, and 5). There is at least one data point in the set of training data for which each hypothesis assigns zero probability. The Deterministic Bayes approach is unable to deal with contradictory data. Since the total probability is given as the product of the posterior probabilities for each word, all hypotheses zero out.

Presumably, a more informative result can be obtained if some kind of error, or irregularity, term is introduced. To avoid collapse, let us assign a small probability to

contradictory data under each hypothesis and, for simplicity, consider only the case of three-syllable words, ignoring the contribution of two-syllable words until Section 6. A given Non-Deterministic hypothesis H_i is then defined as follows. The correct or ‘consistent’ production of a word (under a given hypothesis) is given as $1-2\alpha$. This leaves the (much smaller) probability that a given word will not be produced in accord with the original (deterministic) hypothesis, but with random stress assignment: 2α . That probability is assumed to be equally distributed among the two inconsistent positions, each with probability α . See (8).

For a given three-syllable word, y_x , there are three stress possibilities: 1-initial stress, 2-penultimate stress, and 3-final stress. The set of possible outputs is given by $C = \{1,2,3\}$, and the stress class assigned by H_i is written as a function of the input word: $H_i(y_x) \in C$. A

deterministic hypothesis, by definition, assigns all its probability mass to the stress position ($c \in C$) that agrees with the hypothesis.

$$p(c | H_i, y_x) = \begin{cases} 1 & c = H_i(y_x) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

The non-deterministic version of H_i divides the probability mass between three possibilities.

$$(8) \quad \underline{H_i^\alpha : \text{Non-Deterministic Version of } H_i}$$

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

The hypothesis space now contains $GUJARATI^{*\alpha}$, $GUJARATI^\alpha$ and $PENULT^\alpha$, the Non-Deterministic counterparts of the previously introduced $GUJARATI^*$, $GUJARATI$ and $PENULT$. Recall that at the moment only three-syllable words act as input to the learner. Additionally, there are currently no biases of any kind, leaving the prior probability uniform over the hypothesis space. Since the measure of interest is the relative probability of any pair of hypotheses, the competition diagnostic will be the ratio of their posteriors. The probability of the data set itself, $p(d)$, (where d is the set of stressed surface forms $\{d_1, \dots, d_t\}$) depends on the unstressed underlying forms $\{y_1, \dots, y_t\}$ that are the inputs to the grammars, and is constant with respect to the calculation in (9), therefore canceling out.

$$\frac{p(GUJARATI^{*\alpha} | d)}{p(PENULT^\alpha | d)} = \frac{\alpha^{T-G^*} (1-2\alpha)^{G^*}}{\alpha^{T-P} (1-2\alpha)^P} \quad (9)$$

In Equation (9) (and throughout) G^* and P in the exponent indicate the total amount of data (stressed surface forms) that is consistent with the $GUJARATI^*$, and $PENULT$ hypothesis, respectively. T is the total number of word types. Below, G will denote the number of items consistent with $GUJARATI$. These numbers are extracted from Table 1, under the minimal lexicon uniformity assumption (and summarized in Table 3). From the definition in (8), a factor of α appears for each form that is inconsistent with a given (deterministic) hypothesis, and a factor of $(1-2\alpha)$ appears for each form consistent with the (deterministic) hypothesis. The same ratio can be constructed for the reversed sonority versus the sonority-to-stress hypothesis. See the Appendix for the full derivation.

As Equation (9) shows, the relative probability advantage is highly dependent on the magnitude of α . To determine the outcome of the competition between $GUJARATI^*$ and $PENULT$, the value, or range of values, of this parameter must be determined. α can be thought of as a type of error term, or an irregularity parameter. As such, it should be much smaller than 1. A tighter bound for α can be obtained by estimating over the training data. In the present scenario, a learner who could fit this parameter based on maximizing hypothesis likelihood could compute an α based on the proportion of forms they have observed which are consistent with $GUJARATI^*$. For the three-syllable uniform lexicon, $G^*=370$, $T=512$, and $\alpha_{ML} = \frac{T-G^*}{2T}$, giving a value for α_{ML} of approximately .14 (see Appendix). For the values of G , G^* , P , and T given for three-syllable words above, and an α of .14, Equation (9) results in the finding that $GUJARATI^{*\alpha}$ is more probable given the data than either $GUJARATI^\alpha$ or $PENULT^\alpha$ by several orders of magnitude: 1.85×10^4 , and 3.4×10^8 times more probable, respectively.

This result seems to provide strong support in favor of the UG-Delimited \mathcal{H} Principle. If it is available to the learner, the universal-violating grammar $GUJARATI^{*\alpha}$ will be strongly favored given the post-sound change learning data of Gujarati'. It will be so strongly favored in fact (by a factor of at least 10,000!), that it seems to be the inevitable outcome for a learner selecting a grammar from the 3-hypothesis space $\{GUJARATI^{*\alpha}, GUJARATI^\alpha, PENULT^\alpha\}$. This is certainly true for a selection procedure based on maximizing posterior probability (a MAP learner). If the

sound change proposed in (2) ($a > \text{ə}$) is plausible, the hypothesis space is correct, the Bayesian learning model is appropriate, and the typological observation of non-occurrence of *GUJARATI** patterns is upheld, then something is needed to prevent that grammar from arising.

Before accepting this result, however, it is instructive to consider the inherent sensitivity of the Bayesian learner in these circumstances. The parameter α introduced to avoid the zero probability problem turns out to give an extremely large weight to each piece of data that distinguishes among the hypotheses. Although the effect is not as pathological as in the deterministic case (where the weighting is effectively infinite), the relatively small value of α is capable of producing extremely disproportionate outcomes in the presence of contradictory data, to which each hypothesis must allot a small portion of the probability mass, as above.

The exact content of the input to this learner is critical, meaning that results derived for simplified artificial data will not necessarily generalize to a more realistic distribution of language data. Therefore the specification of this input is a non-trivial aspect of the overall learning model. The importance of the learner's input has been raised in, for example, Pierrehumbert (2003), and Pearl (submitted), and Pearl and Weinberg (2007), who explicitly propose 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). The issue for the current study, however, centers not on ambiguous data, but on directly contradictory data. For a learner with a proportion-based threshold, in the best case there is no effect on the outcome, and in the worst case, eliminating ambiguous data can actually exacerbate sensitivity to very small differences.

This problem is well illustrated by considering the closest competition possible. It is straightforward to apply Equation (9) to the case where there is only a single data point advantage between any two hypotheses, H_i and H_j . As Equation (10) shows, all the probability mass allocated to data points to which the two hypotheses assign the same probability (i.e., ambiguous data) cancel out when one takes the ratio of their posteriors, leaving only the contribution of the single, critical data point (see Appendix for full derivation):

$$\frac{p(H_i^\alpha | d)}{p(H_j^\alpha | d)} = \frac{(1 - 2\alpha)}{\alpha} \quad (10)$$

This ratio can be quite large, depending on the size of α . For example, for $\alpha=.01$, H_i^α is *a posteriori* 98 times more likely than H_j^α . Even for a relatively (and implausibly) large value for α around .2, Hypothesis H_i^α still has a posterior at least twice that of H_j^α . A learner who maximizes likelihood and picks one hypothesis will always decisively pick this winner even though only a single data point distinguishes between H_i and H_j , and regardless of the size of n : the amount of ambiguous data encountered. (See Appendix for a more thorough examination of this function).

These results are the consequence of the extreme probability distribution over only two types of data (consistent and inconsistent – with values close to 1 in the first case, and close to 0 in the second), and over a very small set of possible outcomes (3 or 2 depending on the length of the word). In the case of a single data point difference as described above, computing the ratio highlights the disparity between coding that data point as consistent (H_i^α , probability $1-2\alpha$) and inconsistent (H_j^α , probability α). Since the probability of an independent collection of outcomes (a particular input lexicon) is computed multiplicatively, each additional discrepancy in the data compounds the single point case, such that the ratio grows exponentially.

Inconsistent data is so dispreferred (assigned such a low probability) that the hypothesis with the fewest observed inconsistencies will always emerge as the winner, by an apparently insurmountable margin. However, this result is not necessarily in accord with linguistic intuitions about the behavior of grammatical systems. Although the alternations analyzed by linguists are often idealized to ignore irregularities or inconsistencies, it seems overly optimistic, for example, to describe a system as a sonority-to-stress grammar when, in fact, a full 48% of words are inconsistent with this description. This is a particular concern in light of the fact that different sub-regions of phonological regularity are often observed to co-exist stably in natural languages (thus providing a certain type of evidence against the emergence of one grammar as an overwhelming winner under conflicting input (e.g., Inkelas et al. (1997))). At the very least, the likelihood ratio might be expected to better reflect the near-equivalence of the two hypotheses.

4.2 The Decision Metric

It is worth investigating whether a more linguistically acceptable result might be achieved with slight modifications to the current learner. So far, all calculations have implicitly assumed a

winner-take-all classification strategy whereby the hypothesis with the highest likelihood given the data is the one selected by the learner, and all others discarded. But consider an alternative, namely the Optimal Bayes Learner. This learner determines the proper treatment of new data by taking a weighted sum of the predictions of all hypotheses in the original space. What this means is that the classifier assigns a category c_m (stress location) to a new data point (three-syllable word) based on the probability assigned to that outcome by *all* hypotheses. As expressed in Equation (11), the probability that a newly encountered word y_x is assigned to category c_m (stress syllable m), given the body of training data d — $p(c_m|d, y_x)$ — is the weighted sum of the probability each hypothesis gives of c_m classification — $p(c_m|H_s, y_x)$ — and the weights are given by the *a posteriori* probability of the particular hypothesis given the training data, $p(H_s|d)$.

$$p(c_m | d, y_x) = \sum_{H_s} p(c_m | H_s, y_x) p(H_s | d) \quad (11)$$

To see how this decision metric changes the previous results, the same exercise can be performed here as was done to derive Equation (10). Instead of two hypotheses, consider three, again with only the slimmest possible difference between the highest likelihood hypothesis and the two competitors. The derivation is performed in the Appendix and results in the following formula:

$$\frac{p(c_1 | d, y_x)}{p(c_2 | d, y_x)} = \frac{6\alpha^2 - 4\alpha + 1}{3\alpha(1 - 2\alpha)} \quad (12)$$

Although the exact dependency is different, this result is, as in Equation (10), highly dependent on the value of the parameter α . Even in a “ganging-up” scenario, where H_j and H_k collude to move stress away from the position preferred by H_i , H_i still carries the day overwhelmingly in all cases except where α is quite large ($.25 < \alpha < .33$). See Fig. 1. The full treatment of this case is given in the Appendix, where it is shown that the Optimal Classifier does not appreciably alter the previous results captured by Equation (9).

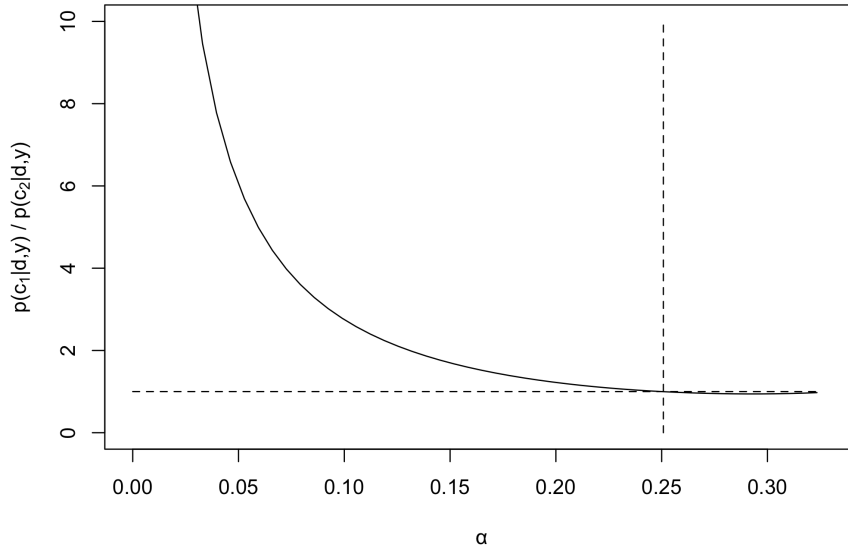


Figure 1

Classification probability ratio: $\frac{p(c_1|d,y)}{p(c_2|d,y)}$ as a function of α , for the three hypothesis case, with a single data point difference in coverage between each pair of hypotheses. Stress in c_2 position preferred over c_1 for values of $\alpha \geq .25$ (indicated by dashed line).

4.3 Complex Hypotheses

The results so far seem to indicate that the weight carried by even a single discriminating data point can cause the anti-markedness reversed sonority-to-stress grammar to arise quite regularly. And, for the minimal uniform lexicon of hypothetical Gujarati', the 6 data point advantage to *GUJARATI** $^{\alpha}$ gives it a substantially greater likelihood than its closest competitor, *PENULT* $^{\alpha}$ (see Equation (8)). Without qualification, these results substantiate the UG-Delimited \mathcal{H} Principle. But the documented sensitivity of the Bayesian learner might be expected to cause a constant see-sawing between various hypotheses as individual lexical items are lost or acquired. In fact, this kind of outcome mirrors some of the undesirable properties of earlier models of trigger learning (cf. Gibson & Wexler 1994). The assumption of such a learner is that all data are consistent, and that once an unambiguous data point has been encountered, the grammar can be set for all time. If a data point is encountered that contradicts that grammar, the parameters must be changed, and so not only is a single data point sufficient to cause this learner to categorically switch hypotheses, the order in which the data is encountered will completely determine the

outcome of learning. One way to avoid outcomes like this is to expand the Bayesian Learner's hypothesis space. This will have immense consequences for the outcome of learning because there exist a number of hypotheses that are better predictors of the data than any considered so far.

In particular, it is instructive to introduce something like a class of hypotheses that explicitly encode an equality relation between any pair of competing alternatives. Firstly, this allows us to select a winning compound, or mixed, hypothesis in the case where the two alternative hypotheses are numerically identical in their coverage of the data. But within the Bayesian framework, the mixed-grammar competitor also allows for a way to avoid selecting an alternative that has only a relatively small advantage over any other. The introduction of this class of hypotheses changes the minimum posterior value that a single-grammar hypothesis must achieve to emerge as the MAP winner. This can be conceptualized as the significance level necessary for the rejection of a default (or null) hypothesis (the no-difference hypothesis). Such a hypothesis is denoted $NULL(i/j)$, and is defined as a mixture, with equal weights, of a given pair of hypotheses, H_i and H_j . See (15).

In fact, an outcome similar to this equally weighted mixture might be exactly what our intuitions led us to expect the optimal Bayes classifier would provide, given that it considers a weighted sum of all hypotheses in its categorization of new data. That this was not the result (see above) is due to the fact 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. Alternatively, one can specify the weight of H in a *mixture* that optimally explains the data (similarly to the way the weights of constraints in a probabilistic OT framework would be assigned (see, e.g., Goldwater & Johnson (2003), Boersma & Hayes (2001))). A version of this approach will be presented momentarily. What immediately follows is the calculation of the competition between the newly introduced set of null hypotheses and the individual alternatives of which they are comprised.

As mentioned above, the posterior probability $NULL(i/j)^a$ assigns to a surface form is calculated by allotting equal probability to selecting the H_i^a or the H_j^a rule to produce an output of that class. For newly encountered words, the speaker is conceptualized as utilizing either

Grammar i or *Grammar j* (at random) in order to determine stress assignment. The outcome of that decision, as a probability distribution, is given in Equation (13), where $w_1 = w_2 = .5$.

$$p(c \mid NULL(i/j)^\alpha, y_x) = w_1 p(c \mid H_i^\alpha, y_x) + w_2 p(c \mid H_j^\alpha, y_x) \quad (13)$$

$$p(c \mid H_x^\alpha, y_z) = \begin{cases} 1 - 2\alpha & c = H_x(y_z) \\ \alpha & \text{otherwise} \end{cases} \quad (14)$$

Adding what has already been defined for individual hypotheses in (8) – re-written as Equation (14) – allows for the determination of the probability of stress assignment c to a particular word, y_x , as a function of α . For a given three-syllable word there are three possible positions for stress. As can be seen from (15), there are also three relevant classes of lexeme: the first where the two hypotheses agree in assigning stress to position c (and, therefore, the highest probability outcome), the second where only one hypothesis selects position c as the location of stress, and the third, where neither assigns stress to position c (the lowest probability outcome) (see Appendix for details).

(15) NULL(i/j)^α: ‘Null Hypothesis’

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

As before, the task is to compare the posterior probabilities over a given set of data, d , for the two grammars under test. In the present case, those grammars are $NULL(i/j)^\alpha$ (which itself contains H_i^α) and H_i^α itself. H_i^α is the winner of this competition for values of the ratio in Equation (16) that are greater than 1. Equation (16) is derived in a manner identical with that of Equation (9), and the variables are defined as follows: n represents the number of data points consistent with both H_i and H_j ; i , the number of data points consistent with H_i but not H_j ; j , the data points consistent with H_j but not H_i , and a , the data points which are inconsistent with both hypotheses.

$$\frac{p(H_i^\alpha | d)}{p(NULL(i | j)^\alpha | d)} = \frac{\alpha^{j+\alpha}(1-2\alpha)^{i+n}}{\left(\frac{1-\alpha}{2}\right)^{i+j} (1-2\alpha)^n \alpha^\alpha} \quad (16)$$

Where $H_i^\alpha = GUJARATI^{*\alpha}$ and $H_j^\alpha = GUJARATI^\alpha$, and the question is whether the anti-markedness grammar arises for a given lexicon (a given dataset), $G^* = i+n$; $G = j+n$; $N = n$, and the maximum likelihood estimate for α is .14⁸. For the numbers as given by L_{MU} , the posterior probability ratio is 1.89×10^{-30} (See Appendix). This represents a very large de facto significance threshold for rejecting $NULL(G^*/G)^\alpha$ in favor of $GUJARATI^{*\alpha}$. Before, for the anti-markedness grammar (or any grammar) to win it only had to do better by *any* amount (even one data point) than each competitor. Now, for $GUJARATI^{*\alpha}$ 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 explaining the data.

To get some idea of the behavior of this new posterior ratio, and under what conditions the null hypothesis might be rejected, Equation (16) can be plotted as a function of $(i-j)$ – as a function of the input data. Doing so shows that the probability of hypothesis $GUJARATI^{*\alpha}$ exceeds that of the null hypothesis for the first time at $(i-j) = 92$ ($j=40$). In other words, for coverage ratios of 132:40, where $GUJARATI^*$ covers over three times as much unambiguous data as $GUJARATI$, the null hypothesis is rejected. Fig. 2. (see Appendix for full derivation).

⁸ This α is calculated with respect to the $GUJARATI^{*\alpha}$ hypothesis, maximizing the likelihood of the forms consistent with $GUJARATI^*$, and giving that hypothesis a larger advantage. Calculating α with respect to the null hypothesis, gives $\alpha_{ML} \approx .029$, giving 1.39×10^{-109} for the ratio in (16) (see Appendix).

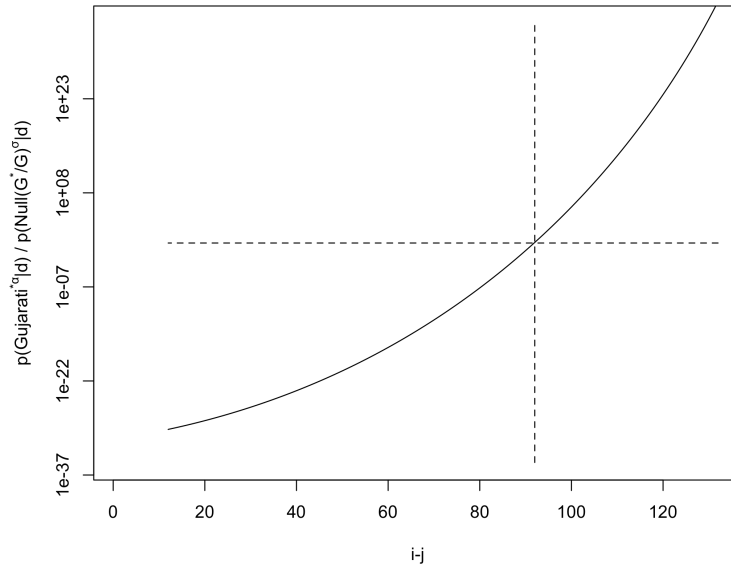


Figure 2

A log plot of the posterior probability ratio: $\frac{p(GUJARATI^{*\alpha} | d)}{p(NULL(G^*/G)^\sigma | d)} = 2^{i+j} \left[\frac{\alpha}{1-\alpha} \right]^j \left[\frac{1-2\alpha}{1-\alpha} \right]^i$ as a function of $i-j$, where $\alpha(i-j)$ is recomputed as the maximum likelihood estimate with respect to $GUJARATI^{*\alpha}$ at each point. The probability of the $GUJARATI^{*\alpha}$ hypothesis first exceeds the probability of the null hypothesis at: $i-j=92$ (indicated by the dashed line).

What the foregoing analysis reveals is that, with a broader consideration of the hypothesis space, the anti-markedness grammar is no longer very likely to emerge as the winner. For the minimal uniform lexicon of hypothetical Gujarati', $GUJARATI^{*\alpha}$ is, in fact, at least 30 orders of magnitude less likely than the null hypothesis that treats $GUJARATI^*$ and $GUJARATI$ as equal sources for the observed data. What this result means in terms of the UG-Delimited \mathcal{H} Principle will be discussed shortly. First, there will be one more class of possible grammar to consider.

Taking the current line of analysis to its logical conclusion prompts the inclusion of a final class of grammars: 'maximum likelihood' grammars. By definition, these provide an even better fit to the observed data, and are straightforwardly calculated by extension of the mixed-grammar formalism in Equation (13). The Null grammar allows fitting to the data by estimation of the irregularity parameter α . It has one degree of freedom. The Maximum likelihood grammar allows fitting to the data by estimation of both the irregularity parameter (re-labeled σ), as well as the weighting factors on the two hypotheses. It has two degrees of freedom. $MAX(i/j)^\sigma$ is defined explicitly below in (17) for any given combination of H_i^σ and H_j^σ . It is computed

identically with (15), the most obvious difference being that, since the weights are not necessarily equal, there are now four relevant classes of lexeme, rather than three, to be considered. However, it can be readily seen that substituting .5 for both weight values reduces $MAX(i/j)^\sigma$ to $NULL(i/j)^\alpha$, and (17) to (15) (derivation provided in Appendix).

(17) $MAX(i/j)^\sigma$: ‘Maximum Likelihood’ Hypotheses

$$p(c \mid MAX(i/j)^\sigma, y_x) = \begin{cases} (1 - 2\sigma) & c = H_i(y_x) = H_j(y_x) \\ w_1 + (w_2 - 2w_1)\sigma & c = H_i(y_x) \ \& \ c \neq H_j(y_x) \\ w_2 + (w_1 - 2w_2)\sigma & c = H_j(y_x) \ \& \ c \neq H_i(y_x) \\ \sigma & c \neq H_i(y_x) \ \& \ c \neq H_j(y_x) \end{cases}$$

This class of hypotheses (with appropriate values for w_1 , w_2 , and σ) will have a higher posterior probability over the data than any other hypothesis considered so far⁹. In fact, under the uniform prior assumption, $GUJARATI^{*\alpha}$ cannot do better than $MAX(i/j)^\sigma$, which always sets its weights so as to maximize the likelihood of the training data, and has more degrees of freedom, and thus more flexibility in doing so than $GUJARATI^{*\alpha}$. $GUJARATI^{*\alpha}$ is actually a special case of $MAX(G^*/G)^\sigma$ (as is $NULL(G^*/G)^\alpha$), and the two hypotheses are identical when there is no unambiguous data in support of $GUJARATI^\alpha$. Therefore, the best outcome for the anti-markedness grammar is a tie, when $w_1=1$ and $w_2=0$ ($j=0$; $\sigma_{ML} = \alpha_{ML} = a/2T$) (see Appendix).

This is not quite the end of the story, as $GUJARATI^{*\alpha}$ can be given some sort of chance of winning if $MAX(G^*/G)^\sigma$ is handicapped in some way. Informally, $MAX(G^*/G)^\sigma$ and $GUJARATI^{*\alpha}$ can be seen to differ in a basic way related to the number of parameters and rules they must each keep track of. 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 smaller hypothesis, this should be reflected in the prior probability distribution. This bias can be formally specified using the information-theoretic notion of coding cost, or description length. As the full calculation is somewhat involved the interested reader is

⁹ Of course, we can do still better if we tailor the ‘maximum likelihood’ hypothesis to individual word types because, in fact, the stress distribution varies for different vowel combinations. And conditioning on individual words, that is, straightforward memorization, 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.

referred to the Appendix. The result can be conceptualized, paralleling earlier discussion, as a significance level. As was done in the case of the competition between $NULL(G^*/G)^\alpha$ and $GUJARATI^{*\alpha}$, it can be shown that, in order to reject a mixed hypothesis where both sonority hierarchies are maintained, $GUJARATI^*$ must account for about seven times more unambiguous data than $GUJARATI$. This condition is clearly not met for the minimum uniform lexicon.

4.4 *The Hypothesis Space Re-Described*

Having shown what the mathematical results are of considering a particular type of learner, over a particular hypothesis space, the next step is to determine what those results mean for linguistic theory. There are two links that will be spelled out here: the mapping from a Bayesian formulation to the more familiar terms of theoretical linguistics, and the mapping from the grammars considered above to the terms ‘anti-markedness’ and ‘markedness abiding’. It was noted above that a speaker who’s grammar consisted of the $NULL(G^*/G)^\alpha$ hypothesis could be described as assigning stress to novel words by consulting either $GUJARATI^*$ or $GUJARATI$ – at random. The speaker of the language instantiated by $MAX(G^*/G)^\sigma$ does exactly the same thing, except the two subset hypotheses are not chosen equally often, but in accordance with the ratio w_1/w_2 .

These kind of explicitly probabilistic hypotheses may be unfamiliar to the reader, but when considered solely as the set of their input-output mappings can be usefully compared to “mixture” formalisms taken from a more traditional generative approach. Under the formulations in Chomsky & Halle (1968) exceptions could be accounted for by specifications in the lexicon that either prevented a general rule from applying to a particular morpheme, or called for a different, ‘minor’ rule to be applied. Lexical Phonology (Kiparsky 1982) assigns different morphemes to different levels, at which different rules apply, or do not. Different classes of words might also be assigned to different ‘strata’ or ‘domains’ in each of which Optimality Theoretic constraints could be ranked in different ways (Ito & Mester 1993). Also within an OT framework, certain words might be underspecified in the lexicon (allowing markedness constraints to determine their surface forms), whereas others would be specified (and thus remain unchanged at surface, due to high-ranking faithfulness constraints) (Inkelas, Orgun and Zoll 1997).

For all of the above formalisms, the learner must correctly partition their data into a number of classes within the lexicon. Assuming that this learning task is tractable (which may not always be the case), the question still remains of how this speaker will treat completely novel words, words for which they must determine not only the correct surface form, but the correct underlying form as well. If there are no possible generalizations to be made over the classes of morphemes, then analogical association is not a possible determiner of class assignment. One possibility is that the minority class will be treated as static, and only the majority class generalized. If there were a large difference in the size of the two classes this could be a plausible strategy. However, in the case where the classes are not significantly different in size, and there is no UG to arbitrate between ‘natural’ and ‘unnatural’, the solution is not obvious – as has been explicitly argued in the preceding sections. Alternatively, a speaker in this situation may select the class at random, even variably¹⁰. But another approach, one that reflects more accurately the speaker’s knowledge, is one that categorizes a new word based on the relative size of the already existing classes. If one class is larger – that is, more likely by a certain degree – then each new word is more likely – by that same degree – to be assigned to that class.

The class of mixed-grammar hypotheses described above performs exactly this type of operation. The two subset grammars define the two relevant classes. Words are assigned stress either by grammar 1 (*GUJARATI**) or grammar 2 (*GUJARATI*). For novel words, which grammar gets to decide stress is determined probabilistically, based on the weighting of each grammar, which, in turn, is based on the size of the categories within the already acquired lexicon¹¹. Under this equivalence, the mixed-grammar hypotheses, $MAX(i/j)^\sigma$, can be understood to stand in for a large number of grammars, all of which account for non-exceptionless patterns in some way (and not all of which are specified enough to evaluate within the Bayesian framework).

By definition, the mixed-grammar hypotheses are distinct from the pure anti-markedness grammar – and they have been shown to win competitions overwhelmingly. Therefore, nothing is currently needed to actively prevent *GUJARATI** from surfacing synchronically, and the strong form of the UG-Delimited \mathcal{H} Principle is rejected. However, the mixed grammar, $MAX(G^*/G)^\sigma$,

¹⁰ In the limit, for a truly lexicalized system, each word belongs to its own unique class. By definition, there are no hypotheses for assigning stress since the learner has simply memorized the stress for each word. In that case stress assignment to novel words could be expected to be entirely random.

¹¹ For the words that the learner already knows (the training set), the stress assignments might be completely determined (specified in the lexicon) - these are the words that were used to select the winning grammar after all. Another theoretical possibility is that the grammar, once learned, supercedes and replaces any memorized forms. In that case, all words are assigned stress according to $MAX(G^*/G)^\sigma$. This kind of completely productive grammar would entail variable stress assignment.

is also not one that can be considered to reflect natural UG laws. The productions of a speaker operating under the rules of this grammar would not be obviously consistent with a reversed sonority-to-stress outcome (since many words would show a stress pattern that is incompatible with that hypothesis), but neither would those productions be inconsistent with such a grammar (since a (slim) majority of words would provide positive evidence for such a hypothesis).

A hybrid system such as this, where even a sub-pattern of ‘unnatural’ forms persists might be considered by some to fall into the class of dispreferred grammars. In that case, the UG-Delimited \mathcal{H} Principle should actively prevent such grammars as well. However, the fact that the present learner prefers these grammars cannot be taken as support for such a principle *absent* evidence from the typology. So far it has been assumed that evidence for grammars like *GUJARATI** is completely missing from the inventory of the world’s languages, past and present. It is quite a different thing to suppose that the same is true for grammars like $MAX(G^*/G)^\sigma$, an assumption that I consider unwarranted given our present state of knowledge.

5 Probabilistic & Categorical Learners

Among other things, the foregoing analysis shows that the Bayesian learner favors by quite large amounts hypotheses that fit the observed frequencies very closely. This property can be characterized as a sensitivity to input that acts to preserve the distribution of the training data. Furthermore, it turns out that a number of different learning algorithms can be characterized in this way, and grouped together as a class of ‘sensitive’, ‘conservative’, or ‘faithful’, learners. Although the grammars selected by the learners in this class will not necessarily be identical, they will be equivalent with respect to the property of interest: they will be mixture grammars of some kind.

Many learners are explicitly formulated to produce mixture grammars, with weights calculated under the assumption that data are generated by a combination of hypotheses. See, for example, the variational model proposed by Yang (1999) or the probabilistic version of Optimality Theory over rankings of Jarosz (2006). Within a rule-based framework, the minimal generalization learner of Albright and Hayes (2003) can be characterized as belonging to this class as well. This learner 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) discovers weighted phonotactic constraints from its input data, keeping track of a large number of correlations of various strengths over various units. Models like these are becoming more and more popular for their ability to capture more of the complex detail of natural language.

However, in shifting to a probabilistic framework it can be difficult to make contact with the core of linguistic theory. The typological claims that this paper seeks to investigate are typically couched in categorical terms: either a language instantiates a stress rule that follows the sonority hierarchy, or it doesn't; either a language encodes a process of inserting /t/ before onsetless vowels, or it doesn't. The data-driven approach of the current model instead leads us towards a more statistical conceptualization of linguistic competence. This in turn forces a reconsideration of what it means for a language to either conform to, or deviate from, a proposed universal. This is not just an expositional shift; as discussed above, this outcome obliges reconsideration of the typological data. It also raises the question of how much categoricity should be imputed to natural language. The resolution of this issue, however, is far beyond the scope of this paper. The goal for this final discussion of the language learner will be to marry the gradient and the categorical in a reasonable way so that the best test of the UG-Delimited \mathcal{H} Principle can be made.

The class of statistics known as the Neyman-Pearson type comprise a number of tests for correlations between data samples, employing either an explicit or implicit threshold – a significance level – to reject the null (no correlation) hypothesis (Neyman & Pearson 1933). In fact, each of these statistical tests can itself be instantiated as a member of our class of learners. In the case of categorical data of the kind under examination in this paper (stress is either in position c , or not), the applicable measure is the χ^2 test of independence (e.g., Rice (1995)). This statistic will allow the learner to determine whether there is a significant correlation between stress and sonority, pitted against the corresponding null hypothesis: namely, that the two factors are independent.

It turns out that such a correlation (in the UG-prohibited direction) can easily be found for a lexicon of size $n=512$ or larger (see Appendix). This results from the characterization of

the input lexicon as a sample from some true underlying distribution of stressed words. For a constant level of departure from independence, a larger sample will have a smaller acceptance region than a smaller sample (see, e.g., Fisher 1958). This means that for large samples the null hypothesis is more likely to be rejected; any difference observed will be more likely to count as a significant difference. This is a desirable attribute up to a point, but can become pathological for samples that are either too large or too small (Berkson 1938, Nunnally 1960, Royall 1986). In particular, for large enough samples the null hypothesis will always be rejected – all deviations become significant, even deviations of a single data point. This sensitivity will be familiar to readers from the discussion of the Bayesian learner in Section 4.1.

What introducing the Neyman-Pearson learner does here is bring the issue of significance to the forefront of the debate. These kinds of statistics are the universally accepted standard in experimental research. However, there are recognized shortcomings in these tests. For one, there is an essential arbitrariness in the setting of the threshold. As discussed, the significance level is determined by the size of the sample, as well as by a parameter which is typically set to minimize the probability of a “false positive” – the probability that a difference discovered between the two distributions is a product of chance (usually set at .05). This threshold is also set independently of the particular population that is being tested. And although statistical methods may find a difference between two populations, it is not always clear whether that difference is actually relevant to the domain of study (cf. Steidl et al. 1997, Johnson 1995, Quinn & Dunham 1983, Tacha et al. 1982). This is the distinction that has been made between ‘statistical significance’ and ‘practical’ or ‘material’ significance (Bakan 1970, Binder 1963, Hodges and Lehmann 1954). For example, material significance in biology might have to do with whether a measured difference between amounts of a particular enzyme actually leads to differences in growth rates among bacteria populations. Material significance in language acquisition attaches to a difference that causes Grammar 1 to be selected over Grammar 2.

What the material significance level for language learners might be (or even whether one exists at all) is not a question to which we currently have an answer. This is another open area of inquiry within linguistics that serves to further complicate the apparently simple question that was posed at the beginning of this paper, namely, in what numbers anti-markedness grammars are predicted to occur under certain conditions of sound change.

Given knowledge of the complete lexicon, the learner may well adopt a frequency matching strategy that preserves the relative proportions of the learning data. The highly populated rule space of Albright and Hayes (2003) is motivated by certain experimental results, namely, the fact that English-speaking listeners are sensitive to details of a nonce word's phonological make-up in judging the appropriateness of possible past tense inflections. This characterization of the learner is supported by other experimental work showing sensitivities to phonotactic regularities and lexical neighborhood effects (e.g., Kelly (1988), Vitevich et al. (1999), Dell et al. (2000), Bailey & Hahn (2001), Bell et al. (2003), Hay et al. (2003)). And a number of theoretical frameworks that offer a probabilistic treatment of phonological grammars can be found (see, for example, Boersma (1998), Zuraw (2000), Pierrehumbert (2003), Antilla (in press)). Gradient representations may be the most likely outcome under most learning scenarios, that is, most kinds of input data distributions. If so, there is some reason to prefer a mixed, or probabilistic formulation of the hypothesis space, and mixed grammars with an anti-markedness subpattern would be the most likely outcome of uncontrolled sound change of the kind described in this paper.

However, in the reverse direction of high regularity, there may also exist cases in which learners generalize over their input. This is predicted to occur given a low enough level of deviation (a small enough number of exceptions), or a low enough level of predictability for those exceptions (this is one view on the evolution of creoles from earlier pidgins (Singleton (1989), Ross and Newport (1996), Senghas and Coppola (2001), Hudson Kam and Newport (2005)). Under the assumption that a (quasi-) categorical grammar should be a possible outcome in certain circumstances, it only remains to specify those circumstances (the relevant significance threshold).

Although the Bayesian learner and the Neyman-Pearson learner can both furnish such a threshold, those thresholds may not accord very well with 'material significance' levels for languages learners (recall that a roughly 7:1 coverage ratio is required for the Bayes Learner to produce a single-grammar outcome). Since there is nothing to exclude the consideration of other possible thresholds, I will do so in the following section. In fact, many language learning algorithms are designed to capture both gradient and categorical phenomena, often with an implicit threshold for categorical behavior (see, e.g., Pearl (submitted), Hayes and Wilson (2006)). Yang (2005) proposes an explicit tolerance level for allowable exceptions to productive

rules. In that spirit, I will choose four competition threshold levels for determining a winning single-grammar hypothesis. These thresholds are essentially arbitrary, but they are chosen as somewhat more plausible thresholds of *material* significance – thresholds that will allow the anti-markedness outcome a better chance of winning.

6 The Lexicon

The current (and final) learner will be characterized as capable both of faithfully reproducing input data (that is, frequency matching) and of converging on a single hypothesis corresponding to a single self-consistent grammar. The latter outcome will obtain under conditions in which a threshold ratio of data coverage is reached. This threshold will separate, on the one side, mixed grammars (no single self-consistent grammar achieving sufficient coverage), and, on the other, single-hypothesis grammars (where a dominant pattern has been identified, and generalized). For this basic Proportion-Based Generalization learner, a set of possible threshold values is proposed, based on the ratio of the total amount of data consistent with each hypothesis¹².

Consider the following “statistical significance” threshold values. In order to defeat the ‘non-generalization’, or mixed-grammar 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 (so *GUJARATI** accounts for considerably more data than *GUJARATI*). 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 lexicon uniformity hypothesis above, none of the thresholds in the range (2.5, 5) is met. The numbers from Section 3.2 for the Gujarati’ lexicon of three-syllable words give values for G^*/P of $370/364 = 1.02$, and G^*/G of $370/358 = 1.03$. The UG-Delimited \mathcal{H} Principle is unsupported.

However, looking back at Table 1, it is easy to imagine a case in which these ratios could fall within the reasonably allowable range. Suppose Gujarati had a lexicon, for whatever fortuitous reasons, in which there were no words that consisted of a (M, M, ə) vowel sequence. Taking this even further, if all lexical types in Gujarati’ that exhibit the *GUJARATI* pattern in Table 1 had never existed (Rows 4-6), then the data would skew towards the *GUJARATI** hypothesis. Of

¹² Ignoring ambiguous data in this case (Pearl submitted) will actually increase sensitivity to very small differences.

course, the reverse scenario is just as imaginable. The question 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?

Up to this point only a single possible lexicon has been considered (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. In order to consider a fuller range of possible lexicons the vowel distribution is allowed to depart from uniformity. However, other simplifying assumptions are maintained, namely those that allow us to ignore any potential effects of phonotactics, and keep a fixed number of 2- and 3- syllable words.

This distribution of lexicons was created by repeatedly under-sampling (with replacement) from the full set of word-types (\mathcal{L}) at several different rates. This procedure allows for a kind of tuning of non-uniformity over lexicons. These lexicons are the product of vowel sequences which are randomly assigned. However, the more undersampled the space is, the higher likelihood there is for highly non-uniform distributions to arise within a set of 1000 such randomly generated lexicons. Thus, the Monte Carlo distribution of lexicons will spread out, in any given parameter space, for lower sampling rates.

Each hypothetical lexicon is defined as consisting of a total of 6,912 words of which roughly half are 3-syllable words (3072), and half, 2-syllable words (3840)¹³. Each rate of under-sampling is defined 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 produced 1000 lexicons at each of these four sampling rates.

Although the degree of under-sampling gives a measure of how skewed the vowel distributions can be expected to be for a given lexicon, it doesn’t specify the exact nature of that distribution. For example, one lexicon generated with Degree 4 of under-sampling displays the following normalized vowel frequency distribution: a 33%, i 22%, e 16%, ə 16%, ɔ 5%, u 5%, ε 0%, o 0% (See Table 7). Identical with respect to degree of sampling, but very different with respect to stress distribution over words, is a lexicon with the following vowel frequencies

¹³ This 1:1.1 ratio corresponds to the distribution found in the online British English corpus CELEX (1993).

(where only the vowel identities at each frequency level have changed) u 33%, ε 22%, i 16%, o 16%, ə 5%, e 0%, a 0%. Which particular vowels appear with any particular frequency is determined by random selection, and differs for each of the lexicons generated.

At this point it is worth considering whether the degrees of non-uniformity, or the ways in which the lexicons are non-uniform, are a good representation of natural language. Although natural language lexical distributions are far from being universally and comprehensively characterized, it has been observed that a number of linguistic units tend to show a very non-uniform distribution. This distribution is one in which the highest frequency items are observed (in, e.g., a text sample) significantly more often than the next most frequent, and the largest number of different types is found at the lowest rate of occurrence. This kind of distribution has been noted for word and morpheme token frequencies, word lengths, and syllable counts. It has also been suggested for the distribution of phonetic or phonological units, in accordance with principles of articulatory markedness (Zipf 1949). Generally speaking, a distribution in which the absolute frequency of occurrence depends on the relative frequency of occurrence is known as a Zipfian distribution¹⁴. A particular instantiation of a Zipfian distribution (the standard harmonic) is characterized by the following formula

$$f \propto \frac{1}{r} \quad (18)$$

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

Although the current sample of lexicons contains distributions that are at least as non-uniform as the Zipfian – for a given measure of non-uniformity – there are not necessarily any that are non-uniform in the same exact way. Accordingly, a fifth set of 1000 Monte Carlo lexicons were generated. Each lexicon of this new set was Zipfian in the distribution of its vowels. Again, the process of random selection determined *exactly* which vowels corresponded to which frequency rank.

¹⁴ Thanks to an anonymous reviewer for suggesting consideration of this type of vowel distribution.

In order to see graphically where these lexicons lie in relationship to possible thresholds for the Proportion-Based Generalization, the results of these simulations were plotted such that the y-axis represents one of the threshold parameters, the ratio G^*/P , and the x-axis represents the other threshold parameter, the ratio G^*/G . Each point corresponds to a possible lexicon of Gujarati (where L was transformed to its L' counterpart via the change $a \rightarrow \text{ə}$). See Fig. 3. Four possible significance levels are plotted (solid black boxes). These correspond to regularization thresholds where $GUJARATI^*$ must explain 2.5, 1.7, and 1.25 times the amount of data that each competitor hypothesis explains. Points interior to each of the boxes represent lexicons for which the $GUJARATI^*$ hypothesis wins under that particular threshold value. All other points represent either mixed-grammar outcomes (area centered around (1,1) in Fig. 3), or outcomes in which $GUJARATI$ or $PENULT$ is the winner.

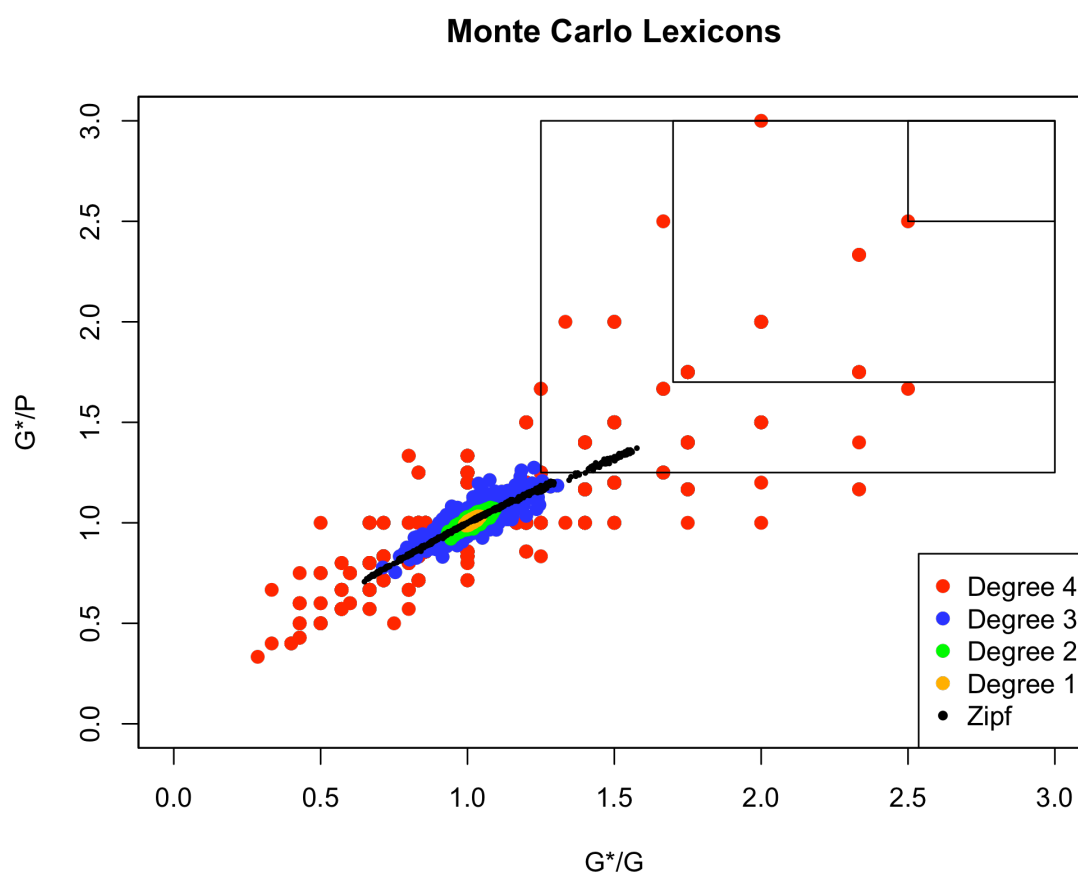


Figure 3

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; higher numbers equate with higher likelihoods. Also displayed are 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. For the learner with a threshold for generalization, the area within each solid-lined black box represents the lexicon space in which the *GUJARATI** hypothesis is chosen. The three difference 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** hypothesis may turn out to be quite uncommon. But, 2) they are not impossible. Table 6 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 at least 25% more data than either *GUJARATI* or *PENULT*), and the highest Degree of skewing from uniformity, over 11% of the lexicons are estimated to lead a learner to adopt *GUJARATI**. At any lesser Degree or more stringent evidence ratio threshold, this estimate drops below 4%. For the Zipfian- distributed lexicons the rate of anti-markedness outcomes at Threshold level 1.25 is lower than the Degree 4 lexicons, at 9.4%. However, there might be reason to weight this conditional probability higher if we believe the prior probability of the Zipfian lexicon to be higher. Clearly, the estimated likelihood of a given type of lexicon will affect the model's predicted numeric values. However, since any type of lexicon can have at most a probability of 1, an upper limit is provided by the numbers in Table 6.

Table 6. Estimated percentage of anti-markedness outcome for Proportion-Based Generalization learner, under 5 different sampling rates (given as [number of 3-syllable,2-syllable word types]), 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

It should be noted that the competition *GUJARATI** faces from *GUJARATI* is due to the existence of a residue of natural patterns in the post-sound change language: a certain proportion

of forms whose surface [ə]’s were historically /ə/’s, rather than deriving from /a/’s. Consider the three-syllable words classified in Table 1. The residual natural pattern is contained in rows 4 and 5, whereas the decisive anti-markedness patterns are evident in rows 2 and 3. The difference between the rates of occurrence of these groups can be roughly characterized as the difference between the rates of occurrence of /a/ and /ə/ in Old Gujarati. If the frequency of /ə/ were appreciably lower than /a/, then the frequency of words in rows 4 and 5, all else being equal, would be analogously less than the frequency of words in rows 2 and 3. The repercussions of a large difference in relative frequency between these two vowels can be seen in Table 7. Here, three representative lexicons from each type of sampling are selected: one that results in a mixed-grammar outcome, one that results in a *GUJARATI** outcome, and one that results in a *GUJARATI* outcome. Of course, as Table 6 shows, the latter two types of lexicon only occur with Degree 4 and Zipfian lexicons, even for the lowest threshold values. However, for each of these instances /a/ is much more frequent than /ə/, whereas in the mixed-grammar outcomes, the two vowels are much closer together in frequency.

Table 7. Normalized Frequencies (Rounded)/Rank Order Frequency

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

7 Several Additional Considerations

It is instructive to model one more type of vowel distribution, derived from a reconsideration of the properties of sound change. So far in this paper, sound change mechanisms have been largely unspecified. Simplifying assumptions have been made for the same reason that all simplifying assumptions have been made – because the actual principles that govern natural sound change are unknown. There are some hypotheses in this domain that can be explored computationally, however, and this section will introduce a few possible modifications to the present model. There is a long-standing intuition in the literature, for example, that the most likely sound changes are those that avoid neutralizing contrasts (Martinet 1955)¹⁵. Contrast is achieved by mapping different sounds to different meanings. And while this is not always a one to one mapping, the hypothesis is that excessive levels of homophony are generally avoided by language users due to communicative pressures. One way that this avoidance ought to manifest itself, then, is in limits placed on sound changes that would introduce more ambiguity.

However, the two works of which I am aware that actually test this intuition produce conflicting findings. Gurevich (2004) argues for widespread avoidance of neutralization across a sample of languages which have undergone historical lenition – neutralizing cases much lower than expected. Surendran & Niyogi (2006), on the other hand, argue that well-defined measures of functional load (degree of contrast) predict that the n/l merger should never have occurred in Mandarin if neutralization was truly avoided in historical change.

Avoidance of neutralization in the Gujarati case means avoidance of vowel quality changes that would create a large number of ambiguous words – homophones. Consider the simple lexical model in which no restrictions apply other than a Zipfian distribution over both vowels and consonants, and a strict CV syllable requirement. Simulations result in from 100-250 words (out of ~7000) which are homophonized when $a \gg \text{ə}$. This number, however, is fairly meaningless. In the first place, there are a number of factors which, in practice, limit the amount of homophony in natural language. These include phonotactic and prosodic constraints that reduce the set of potentially neutralized words. Even complete phonological neutralization does

¹⁵ Thanks to Adam Albright for bringing this to my attention.

not necessarily produce semantic ambiguity when syntactic and pragmatic factors are considered. There is an even more fundamental problem, however, with the numbers we can obtain via simulations. Following Surendran & Niyogi, these results could be converted to a measure of entropy reduction in the lexicon with the relevant (*a/ə*) contrast removed. But, as defined, such a measure would need to be compared to a benchmark. That is to say, whether the result indicated a high prior functional load (and thus a situation that would resist merger), or not, could only be determined by an independently established threshold level for neutralization.

Because it is not entirely clear that avoidance of homophony is a significant factor in sound change, or how it should be measured, or what such a measure would mean, there is no obvious way in which the results of Section 6 should be modified. However, we can be clear that no neutralization will occur in the case in which there was no contrast to neutralize in the first place – in which no */ə/*'s were present in Old Gujarati and thus no homonyms were formed from a process that transforms */ə/*'s to */a/*'s. This new Gujarati' contains no words with */a/*'s, and all words containing */ə/*'s contain stressed */ə/*'s since all */ə/*'s were formerly */a/*'s. All data is consistent with the *GUJARATI** hypothesis. In these circumstances *GUJARATI** might be expected to emerge as the clear winner. But in order to assess this, a so-far unspecified property of the learner must be decided: whether or not the lack of conflicting data ought to be an overwhelming factor in hypothesis selection. If the fact that no words are in disagreement with *GUJARATI** (whereas some data contradict *PENULT*) is more important than the relative data coverage between any two hypotheses, then *GUJARATI** will win. This is the standard theoretical assumption upon which the Trigger Learner is based (Gibson & Wexler 1994).

But such a learner is incapable of coping robustly with contradictory data of the kind we expect to see in natural language, especially under the types of sound change described in this paper. Of necessity, we fall back once again on our coverage ratio measure. Under the non-neutralizing scenario, Gujarati has 7 vowels (rather than 8); for 3-syllable words, all 343 types support the *GUJARATI**^a hypothesis, while 265 are also consistent with *PENULT*^a. And $G^*/P = 1.3$. 2-syllable words will provide somewhat less of an advantage to the anti-markedness grammar (49:46~1.13); taken together, the ratio over the entire lexicon is roughly 1.2. And finally, although this fact has been hitherto ignored, the *PENULT* hypothesis is simpler than the

*GUJARATI** hypothesis, and as such ought to be preferred in some way. Due to these considerations, the case for a clear *GUJARATI** outcome is weakened.

For comparison, this discussion can be projected backward to the case of the original Gujarati instantiated by L_{MU} . For that 8-vowel system prior to sound change, all 512 3-syllable word types are consistent with *GUJARATI*^a, while 365 are also consistent with *PENULT*^a. $G/P = 1.4$ for 3-syllable words, and 1.12 (64:57) for 2-syllable words, giving a weighted average of 1.25. The description in de Lacy (2006) of the actual Gujarati spoken in the Indian state of Gujarat is of a sonority-respecting stress system. And that is what this paper has assumed throughout. However, the data, as given in Tables 1 and 2, are actually ambiguous with respect to whether the speaker's grammar consists of something like the *GUJARATI* hypothesis, or something more like the *PENULT* hypothesis, with a set of memorized exceptions. The position of this paper (as throughout) is that this alternative cannot be trivially dismissed, and that deciding the question requires external evidence in addition to the conventional distributional evidence.

For the sake of the current argument, however, let us provisionally accept the premise that the 'relatively simple' grammar which achieves full coverage will win. Let us allow *GUJARATI** to be the clear winner under conditions of maximally non-neutralizing sound change. Whether this outcome is considered a dispreferred anti-markedness outcome of the kind Kiparsky had in mind, however, is once again unclear. The reason for this ambiguity hinges on the aptness of describing the *GUJARATI** hypothesis as a reversed sonority-to-stress scale. In either instantiation of Gujarati' (deriving either from the 7- or 8-vowel system) there are only two expressed sonority categories: $\{M, \emptyset\}$. Stressing /ə/ preferentially over a higher-sonority mid vowel is already dispreferred behavior from a universalist perspective, but it is qualitatively different from a hypothesis that targets sonority as the deciding factor (rather than vowel identity). The latter hypothesis – the true reversed sonority scale – would avoid stressing a newly encountered /a/, for example, precisely because of the high sonority of that vowel. The likelihood of learning this 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. By the same token, the reasonable learner (lacking UG) should require not only more data consistent with *GUJARATI**, but more data of a particular kind. What drives the

selection of a more complex hypothesis over a simpler one is the presence of complexity in the data – complexity that will be missing when all /a/’s are missing from the lexicon of Gujarati’. A grammar that prescribes “stress ə, and the penultimate syllable when no ə is present” will serve the speaker just as well as a grammar that prescribes “stress the lowest sonority vowel in the word, and otherwise the penultimate syllable when all vowels are of equal sonority”. Under the provision of the beginning of this paragraph, we must select the “stress ə” grammar. The conclusion about the correct specification of *GUJARATI** will be the same for the original case of non-neutralizing sound change. If the “stress ə” grammar is unattested and theoretically dispreferred, then we have discovered the best case for the UG-Delimited \mathcal{H} Principle. We cannot quite stop here, however, but must introduce yet another caveat to the argument.

Now that the plausibility of different types of sound change has been introduced to the discussion we cannot avoid another set of even more problematic issues. 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 (phonetically identical) surface /ə/’s, *and which leaves stress placement unchanged*. 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 decomposability influencing when and whether certain segments will shift (Phillips 1984, Bybee 2001, 2003, Pierrehumbert 2001).

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 2000). 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, would be highly correlated with whether they were 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. And 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 having shifted to a different, higher energy location in the word even before such a vowel quality shift had been completed. Finally, in a synchronic grammar that lacks a contrast between /a/ and /ə/, 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, quite possibly recapitulating the sonority-stress relation at the phonetic level (Colarusso 1988, Choi 1992, Kondo 1994, Van Bergen 1994).

8 Discussion & Conclusion

At this point, the reader may well feel that nothing can ever really be concluded under the research methodology described and implemented in this paper. This feeling has no doubt been inculcated by the sheer number of factors and competing possibilities that have been presented in the preceding 40 odd pages. But it is important to be clear that this modeling work, rather than introducing untold complexity, is actually an exercise in drawing the curtain back from the complexity that already exists, and underlies every theoretical claim about possible typologies. 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 woefully inadequate – underspecified to a truly staggering degree.

The reader may also harbor the suspicion that arguments presented *against* the very assumptions underlying the models of this paper serve to invalidate the arguments of this paper. In actuality, such counter-arguments serve as an illustration of areas of the model that must be specified in a plausible manner. If the specific type of sound change (or even the particular class of sound change) is rejected, then the model changes, and the predictions of the model change. This paper has made particular theoretical commitments, not all of which will stand up to future

scrutiny. The particular specifications are up to the particular linguist. What has been definitively argued for in this paper is *what* must be specified. The work done in describing the parts of the model, and how these parts must fit together, will be applicable to any fully specified model that comes after. What should be kept in mind is that the arguable nature of practically every hypothesis in this paper is not a shortcoming of the present work, or of modeling work in general – it simply reflects our very imperfect knowledge in almost every sub-domain of linguistics. 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 a number of gaps in our knowledge are, and where certain traditional ways of viewing the problem are not completely compatible with one another.

Since so much remains unknown there are no iron-clad conclusions to be drawn. But this does not mean that no conclusions whatsoever can be drawn. Under the assumptions of this model we can, for example, see that there is in no way an overwhelming likelihood of an anti-markedness outcome. While particular parameters can be manipulated endlessly, it has been argued that this result speaks to a reasonably large class of models. That is, the learners for which this outcome holds is a set comprised of probabilistic models which match their data closely, but allow for some sort of significance threshold. The lexicons considered have spanned a number of different distributions, sampled fairly extensively. And the hypothesis space of the learners, although superficially small, can be shown to encompass a range of grammars, from fully lexicalized patterns, through ‘simple’ grammars (fully general descriptions), ‘complex grammars’ (more complicated and specific), and mixtures of grammars which can involve descriptions of complex grammars that apply to different subsets of the data.

This set of models, although not particularly simple, under the standard definition of that term, is simple in a very specific way. These models lack any mechanism which could serve to instantiate an innate and universal substantive bias. Taking the modeling results at face value, diachronic changes, acting without regard to their synchronic repercussions, and coupled with the right kind of lexical statistics, do create systems that seem to violate universal grammatical principles. The ‘reasonable’ learner provided with ‘likely’ input produces a mixed-grammar output the overwhelming majority of the time. However, under certain circumstances, it will converge on the *GUJARATI* grammar. And under complementary circumstances, it will converge on the anti-markedness *GUJARATI** grammar.

If we had absolute confidence that *GUJARATI** is never, has never been, and can never be, attested, then this result could be taken as evidence for a force that limits possible synchronic grammars: a UG-Delimited \mathcal{H} Principle. Similarly, if we believed the same to be true even of grammars that only exhibited such a dispreferred grammar over a subset of their words, then we could also conclude that such a force was necessary. However, it remains inconclusive that the model actually predicts these systems to arise more frequently than they do, for the very simple reason that the typological facts are not so clearly established. This is true even for the pure anti-markedness grammars¹⁶, and it is even more true for potential mixed-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 perceived as patternless behavior might not belong to this set? In fact, the situation is more uncertain even than this. Evans & Levinson (2009) estimate that current typological knowledge comprises at most 2% of possible linguistic diversity. If this is even close to correct, strong claims about attested phonological patterns become extremely problematic.

Therefore, the final conclusion is that the burden of proof must rest on those who wish to complicate the current ‘simple’ model, by specifying particular UG constraints, either on the hypothesis space, or on the learner, or on sound change itself. Both sides of the debate, however, are responsible for specifying the model which they are assuming to the point where it can actually be tested – a non-trivial task as argued by this paper.

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. We have also had to devise some way to deal with issues which have been largely unexplored, at least from a formal perspective, such as the proper treatment of contradictory data. Once we have a better understanding of these outstanding issues we will be in a position to fully examine the larger question of the interaction between synchrony and diachrony. The type of modeling and simulation work exemplified in this paper strikes me as a very promising methodology for providing better insight into the question of what human grammars really consist of, and how

¹⁶ 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.

much of linguistic knowledge can be induced *a posteriori* over the data available to the learner, versus how much must come *a priori* from the mind of the learner.

Acknowledgments

This work was supported by an NSF IGERT grant and a Department of Education Javits Fellowship. I would like to thank Paul Smolensky and Colin Wilson, without whom this paper would never have come about. Thanks to Simon Fischer-Baum and the members of Math City for their invaluable input. Thank you to Mary Beckman for her encouragement, and to Matt Goldrick for his generous assistance and advice. Thanks also go to Adam Albright for his extensive and extremely helpful comments.

(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.

Anderson, S. R. (1981). Why phonology isn't "natural". *Linguistic Inquiry* 12(4): 493-539.

Antilla, A. (in press). Gradient phonotactics and the complexity hypothesis. *Natural Language and Linguistic Theory*.

Bach, E. and R. T. Harms (1972). How do languages get crazy rules? In R. P. Stockwell and R. K. S. Macaulay (eds.) *Linguistic Change and Generative Theory*: Indiana University Press. 1-21.

Bailey, T. M. and U. Hahn (2001). Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language* 44: 568–591.

Bakan, D. (1970). The test of significance in psychological research. In D. E. Morrison and R. E. Henkel (eds.) *The Significance Test Controversy*: Aldine Publishing Co. 231-251.

Bell, A., D. Jurafsky, et al. (2003). Effects of disfluencies, predictability and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113(2): 1001-1024.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the Chi-Square test. *Journal of the American Statistical Association* 33(203): 526-536.

- Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review* 70(1): 107-115.
- Blevins, J. (2004). *Evolutionary Phonology: the emergence of sound patterns*. New York, Cambridge University Press.
- 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.
- Buckley, E. (2000). On the naturalness of unnatural rules. Proceedings from the second workshop on American Indigenous Languages, UCSB Working Papers in Linguistics.
- Bybee, J. (2001). *Phonology and Language Use*, Cambridge University Press.
- Bybee, J. (2003). Mechanisms of change as universals of language. ms.
- Chater, N., J. B. Tenenbaum, et al. (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. 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.
- Dell, G. S., K. D. Reed, et al. (2000). Speech errors, phonotactic constraints, and implicit learning: a study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory and Cognition* 26(6): 1355-1367.
- 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.
- Fisher, R. A. (1958). *Statistical Methods for Research Workers*. New York, Hafner Publishing Co Inc.
- Gibson, E. and K. Wexler (1994). Triggers. *Linguistic Inquiry* 25(3): 407-454.
- 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.

Gopnik, A. and e. al. (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.

Gurevich, N. (2004). *Lenition and Contrast: The Functional Consequences of Certain Phonetically Conditioned Sound Changes*, 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, et al. (2003). *Speech Perception, Well-formedness and the Statistics of the Lexicon*. In *Papers in Laboratory Phonology VI*: Cambridge University Press. 58-74.

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

Hooper, J. B. (1974). Rule morphologization in Natural Generative Phonology. In A. Bruck, R. A. Fox and M. W. La Galy (eds.) *Papers from the Parasession on Natural Phonology*.

Hudson Kam, C. L. and E. L. Newport (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development* 1(2): 151-195.

Hyman, L. M. (1975). *Phonology: theory and analysis*, Holt, Rinehart and Winston.

Hyman, L. M. (2001). The limits of phonetic determinism in phonology: *NC Revisited. In E. Hume and K. Johnson (eds.) *The Role of speech perception in phonology*: Academic Press. 141-185.

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

Inkelas, S., C. O. Orgun, et al. (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*, New York City.

Johnson, D. H. (1995). Statistical sirens: the allure of nonparametrics. *Ecology* 76(6): 1998-2000.

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

Kelly, M. H. (1988). Phonological Biases in Grammatical Category Shifts. *Journal of Memory and Language* 27: 343-358.

Kemp, C., A. Perfors, et al. (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.

Kenstowicz, M. and C. Kisseberth (1977). *Topics in Phonological Theory*, Academic Press.

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

Kingston, J. and P. de Lacy (2006). *Synchronic Explanation*. ms.

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

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

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

Kondo, Y. (1994). Targetless schwa: is that how we get the impression of stress timing in English? *Edinburgh Linguistics Department Conference*.

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

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

Lieberman, 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.

Lombardi, L. (1995). Laryngeal neutralization and syllable wellformedness. *Natural Language and Linguistic Theory* 13(1): 39-74.

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

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

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

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.

Niyogi, P. and R. C. Berwick (1997). A dynamical systems model for language change. *Complex Systems* 11: 161–204.

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement* 20(4): 641-650.

Ohala, J. (1981). *The Listener as a source of sound change*. Parasession on Language and Behavior: Chicago Linguistics Society, Chicago.

Ohala, J. (1990). The phonetics and phonology of aspects of assimilation. In J. Kingston and M. Beckman (eds.) *Papers in Laboratory Phonology I: Between the grammar and physics of speech*: Cambridge University Press. 258-275.

Ohala, J. (1993). The phonetics of sound change. In C. Jones (ed.) *Historical Linguistics*. 237-278.

Pan'kevich, I. (1938). *Ukrain'ski hovory pidkarpac'koji rusi i sumexnyx oblastej*. Prague.

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

Pearl, L. (2010). When unbiased probabilistic learning is not enough: Acquiring a parametric system of metrical phonology. Submitted.

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. S. (1997). *Necessary bias in natural language learning*. Linguistics, University of Maryland.

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*: John Benjamins. 137-158.

Pierrehumbert, J. B. (2003). *Probabilistic Phonology: Discrimination and Robustness*. In R. Bod, J. Hay and S. Jannedy (eds.) *Probabilistic Linguistics*. Cambridge: The MIT Press. 177-228.

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

Quinn, J. F. and A. E. Dunham (1983). On hypothesis testing in ecology and evolution. *American Naturalist* 122: 602-617.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*. Belmont, CA, Duxbury Press.

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.

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician* 40(4): 313-315.

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.

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.

Tacha, T. C., W. D. Warde, et al. (1982). Use and interpretation of statistics in wildlife journals. *Wildlife Society Bulletin* 10: 355-362.

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

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

Tesar, B. and P. Smolenksy (1998). Learnability in Optimality Theory. *Linguistic Inquiry* 29(2): 229-268.

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

Vitevitch, M. S., P. A. Luce, et al. (1999). Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language* 68(1-2): 306-311.

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

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

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.

Yu, A. C. L. (2004). Explaining final obstruent voicing in Lezgian: phonetics and history. *Language* 80(1): 73-97.

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

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