The phonetics-phonology interface

John Kingston
Linguistics Department
University of Massachusetts
226 South College, 150 Hicks Way
Amherst, MA 01003-9274
jkingston@linguist.umass.edu
1-413-545-6833, fax -2792

Submitted to Handbook of Phonology, Paul de Lacy (editor), Cambridge University Press.

Please don’t cite or quote without permission.

4 September 2005
1 Introduction

Phonetics interfaces with phonology in three ways. First, phonetics defines distinctive features. Second, phonetics explains many phonological patterns. These two interfaces constitute what has come to be called the “substantive grounding” of phonology (Archangeli & Pulleyblank, 1994). Finally, phonetics implements phonological representations.

The number and depth of these interfaces is so great that one is naturally moved to ask how autonomous phonetics and phonology are from one another and whether one can be largely reduced to the other. The answers to these questions in the current literature could not differ more. At one extreme, Ohala (1990) argues that there is in fact no interface between phonetics and phonology because the latter can largely if not completely be reduced to the former. At the opposite extreme, Hale & Reiss (2000) argue for excluding phonetics entirely from phonology because the latter is about computation, while the former is about something else. Between these extremes are a large variety of other answers to these questions, including Blevins’s (2004) claim that phonetics motivates sound change but does not otherwise regulate synchronic sound patterns to Browman & Goldstein’s (1995) assertion that phonological representations are merely an assemblage of phonetic units of a grain coarse enough to be reliably categorized (see also Hayes, 1999).

I don’t intend these examples to be a comprehensive list of current points of view, nor even to represent the principal alternatives from which one might choose. For the moment, my purpose instead is merely to show that the field has reached no consensus about what the interface is, nor has it even agreed that one exists at all. The field therefore cannot agree about how distinctive features are defined, phonological patterns explained, or phonological representations implemented by the phonetics. The confident assertions about the three interfaces with which I began this paper are not self-evident truths to everyone, much less any particular phonetic definitions, explanations, and implementations.

These disagreements could be nothing more than the consequences of lacking the evidence needed to choose between competing hypotheses, and indeed there is much feverish effort to find such evidence. But I think the disagreements reflect something more than the commonplace struggle between hypotheses. There is in fact a broader and deeper dispute here about what the phenomena are and how they should be investigated. Given the current state of our knowledge, this dispute cannot be resolved by asserting that one’s own account is to be preferred over another’s because it more economically divides the labor between phonetics and phonology. Occam’s razor is an essential tool when such divisions can be resolved empirically, but it cuts
nothing all when the would-be dividers can't agree about what the phenomena are. Instead we need to ask how if at all can one reliably separate the phonetic from the phonological.

In this chapter, I lay out some of the difficulties one encounters in trying to answer this question for each of the three ways that phonetics interfaces with phonology: definition, explanation, and implementation. I do so by displaying some of the enormous richness of the interchange between phonetics and phonology. This richness is what forestalls any simple solution to the division of labor between these two components of the grammar. Nonetheless, where it's possible to do so, I show solutions, even incomplete ones to the problem of dividing labor between phonetics and phonology.¹

The solution to the problem of defining distinctive features may arise by starting with the phonetics and working up to the phonology, while the problem of explaining phonological patterns may be solved by recognizing that more than one force is at work even in apparently simple cases, that these forces are phonological as well as phonetic, and finally that the various forces may compete with one another. The phonological pattern that results represents the often delicate resolution of this competition. Finally, the solution to the problem of implementing phonological representations is not to be found by reversing the solution to defining their constituents and working down from the phonology to a pronunciation or percept. Instead, the phonetic implementation also determines what kind of phonological representation is possible in the first place.

2 Definition
2.1 Resolving the variability problem

Phoneticians and phonologists have expended considerable effort in defining distinctive features phonetically. Landmarks in this effort are the acoustic-auditory definitions proposed in Preliminaries to Speech Analysis (Jakobson, Fant, & Halle, 1952), the articulatory alternatives laid out in Chapter 7 of the Sound Pattern of English (Chomsky & Halle, 1968), and most recently, the combined acoustic and articulatory definitions in The Sounds of the World’s Languages (Ladefoged & Maddieson, 1996) and Acoustic Phonetics (Stevens, 1998). Distinctive features would be easy to define phonetically if some phonetic property or properties could be observed in articulatory or acoustic records of an utterance every time a distinctive feature takes on a particular value in that utterance’s phonological representation. Unfortunately, this is not the case. Rather than invariant phonetic realizations, distinctive feature values are realized differently in different languages, contexts, speaking styles, and even speakers (Kingston & Diehl, 1994; cf. Stevens & Blumstein, 1978; Sussman, McCaffrey, & Matthews, 1991). How then can distinctive features be defined phonetically?
Some research suggests that distinctive feature values are in fact polymorphous, in that their phonetic realizations at best bear a family resemblance to one another (Kluender, 1994; Kingston, 2003). However, here I outline a pair of alternative approaches, both of which argue that invariance can be found, so long as one steps away appropriately from the detail of a particular utterance’s phonetic realization. The two approaches differ in the direction they recommend one should step away to find this invariance: articulatory phonology recommends one step back from the utterance’s articulatory detail to the speaker’s plan for the utterance (Browman & Goldstein, 1995), while auditorism instead recommends one step forward from the utterance’s acoustic detail to the acoustic properties’ auditory effects (Kingston & Diehl, 1994, 1995; Kingston, Diehl, Kirk, & Castleman, in preparation). Despite these approaches’ recommending that one step away from an utterance’s phonetic detail in exactly opposite directions, they both find invariance by moving to a description of the utterance with many fewer dimensions than are necessary to describe its physical realization, either as a set of articulations or as a set of acoustic properties.

In articulatory phonology, an utterance is represented as a collection of gestures. Gestures specify that the vocal tract be constricted at a particular location to a particular degree for particular interval of time. As such, a gesture specifies the speaker’s goal in the interval of time during which the gesture is active, and this specification evokes the coordinated action of the various articulators whose movements achieve that goal. Because the gesture specifies the goal rather than the movements of individual articulators, an articulator can contribute different amounts to achieving that goal in different contexts. For example, when the goal is to close the lips, as in the pronunciation of [b], [p], or [m], the upper and lower lips and the jaw all move, but each of these articulators moves to a different extent depending on neighboring vowel (Sussman, MacNeilage, & Hanson, 1973; Macchi, 1988). The individual articulators contribute different amounts to achieving the lip closure next to different vowels because the gestures for the neighboring vowel are active at the same time as the lip closing gesture and they compete for control over these articulators. This competition is resolved by the task dynamics, which calculates articulator movements in terms of the demands imposed by all the gestures that are active at any moment in time. The different combinations of upper and low lip and jaw movement that contribute to closing the lips are, however, motor equivalent, because they all succeed in achieving that goal. The gesture that specifies that goal in the first place is then the desired step back from the variable realization of that goal by the different combinations of individual articulator movements to the invariant specification of the goal itself.

Auditorism finds invariance in the listener’s percepts rather than the speaker’s goals. Just as the individual articulators’ movements vary from token to token of a distinctive feature value, so too do their acoustic consequences. Listeners could
therefore perceive each token as a different value, yet they don’t do so. They don’t
because different arrays of acoustic properties are perceptually equivalent to one
another. For example, a stop with a relatively short delay in voice onset following the
release of the oral closure and a relatively high onset frequency for the first formant (F1)
of the following vowel is equally likely to be perceived by an English speaker as [+voice]
as one with a relatively longer voice onset delay and a relatively lower F1 onset
Kingston, et al. (in preparation; see also Kingston & Diehl, 1994, 1995; Kingston &
Macmillan, Kingston, Thorburn, Walsh Dickey, & Bartels, 1999) argue that acoustic
properties can be perceptually equivalent like this when their auditory effects are similar
enough that they integrate perceptually with one another. In the example, a shorter
delay in voice onset and a lower F1 onset frequency both create the percept that low
frequency energy occurs near the stop release.

Perceptual equivalence could arise from another source than the auditory
similarity of acoustic properties. The properties could be perceptually equivalent simply
because listeners have experienced them covarying reliably: stops with shorter voice
onset delays usually have lower F1 onset frequencies, too. Kingston, et al. (in
preparation, also Kingston & Diehl, 1995) show, however, that perceptual equivalence
comes from auditory similarity rather than experience of covariation by obtaining the
same responses from listeners to non-speech analogues in which acoustic properties
are manipulated in the same way as in the speech signals. Because these stimuli aren’t
recognized as speech, they should not evoke the listeners’ experience with the
covariation of acoustic properties in the speech signals they mimic. Listeners should
thus only respond in the same way to the non-speech analogues as the original speech
signals if their acoustic properties are auditorily similar enough to integrate perceptually.
Moreover, if speech sounds are to contrast reliably with one another, speakers may be
enjoined to produce articulations whose acoustic correlates integrate perceptually with
one another.

Perceptual integration thus achieves the same result as the motor equivalence
embodied in gestures: a variable, high-dimensional description is reduced to an
invariant, low-dimensional one whose units corresponds to the contrastive units of
which phonological representations are composed. Distinctive feature values and the
distinctive features themselves may therefore emerge out of human’s speaking or
listening behavior, i.e. either out of the motor equivalence of different combinations of
articulations or out of the perceptual integration of different combinations of acoustic
properties. If this is correct, then distinctive features can be obtained without the
phonological component of the grammar having to impose formal constraints requiring
structural symmetry such as those argued for by Hayes (1999). At a gestural or
auditory level of description, much of the phonetic particularity that phonological constraints typically ignore has been lost.

2.2 Articulatory or auditory targets?

We are burdened here with an embarrassment of riches, two ways of getting distinctive features to emerge out of the phonetics. Is there any reason to choose one over the other, i.e. are speakers’ targets articulatory or auditory? This question can be answered in favor of auditory rather than articulatory targets by examining compensation for artificial perturbations of articulations and natural covariation between articulations.

As long as the perturbations of articulations are not too extreme, speakers immediately and successfully compensate for them. For example, when a bite block is inserted between the molars to prevent the jaw from moving in the pronunciation of vowels, speakers still constrict the vocal tract in the same locations and to the same degree, and the vowel produced differs very little acoustically from that produced without the bite block (Lindblom, Lubker, & Gay, 1979; Fowler & Turvey, 1980; Kelso & Tuller, 1983). Similarly, if a light load is randomly and infrequently applied to the lower lip at the moment when a bilabial closure is initiated, the speaker exerts more force to lower the upper lip more as well as to overcome the load on the lower lip (Abbs, Gracco, & Cole, 1984). Both results demonstrate that different combinations of articulations are motor equivalent, and that speakers’ targets are local constrictions of the vocal tract, as are the gestures in articulatory phonology.

However, when the upper lip is prevented from protruding in a rounded vowel, speakers compensate by lowering their larynges more (Riordan, 1977). Speakers already lower their larynges to some degree in pronouncing rounded vowels (Lindblom & Sundberg, 1971), so this additional lowering simply exaggerates an articulatory movement they already make. Nonetheless, this finding suggests that the speakers’ target is a global vocal tract configuration – a long resonating cavity – rather than a local constriction at the lips.

In examinations of other sounds that are made by moving two independent articulators, Perkell, Matthies, Svirsky, & Jordan (1993) and Perkell, Matthies, & Zandipour (1998) demonstrated that tongue backing traded off with lip rounding in the unperturbed pronunciation of [u] and [j] by American English speakers. This tradeoff produces a sufficiently but not excessively long resonator between the lingual constriction and the lips, i.e. a global vocal tract configuration that lowers the second formant (F2) in [u] and the first resonance in [j] to the desired value. The tradeoff also ensures that this resonator’s length varies relatively little from token to token, despite
possibly large individual variation in lip protrusion or tongue backing.

Speakers’ compensation for bite blocks or lip loading cannot be interpreted as evidence that they’re seeking to produce a global vocal tract configuration rather than a local constriction. The more extreme tongue movements required to achieve the desired local constriction when the jaw is prevented from moving the bite block in fact produce global vocal tract configurations that are quite different from those produced without the bite block. For example, the bite block forces the speaker to raise the tongue body more to constrict the vocal tract closely at the front and back of the palate in the high vowels [i] and [u], and as a result the pharynx is expanded considerably more than when no bite block is present. The additional lowering of the upper lip when the lower lip is loaded does no more than ensure that the local articulatory target, bilabial closure, is reached.

All these results can, however, be interpreted as evidence that the speaker is trying to produce a particular acoustic or auditory effect. The close palatal constriction in the high vowels [i] and [u] lowers the vowel’s F1, the bilabial closure lowers all formant frequencies at the edges of adjacent vowels, as does larynx lowering and lip protrusion in rounded vowels, and as already noted the tradeoff between lip rounding and tongue backing in [u] and [j] lowers the vowel’s F2 and the consonant’s first resonance frequency, respectively.

Even more compelling evidence that the speakers’ target is acoustic or auditory rather than articulatory comes from a study of variation in the lingual articulations of American English [j] (Guenther, Espy-Wilson, Boyce, Matthies, Zandipour, & Perkell, 1999). Earlier work (Delattre & Freeman, 1968; Westbury, Hashi, & Lindstrom, 1995; Alwan, Narayanan, & Haker, 1997) showed that speakers either curl or retroflex the tongue tip back and up toward the palate or bunch the tongue body toward the palate in pronouncing this sound. All seven speakers studied by Guenther, et al. chose a bunched articulation for [j] after the lingual consonants [d, g] but a retroflexed pronunciation after [a, b, v]. Tongue bunching is more efficient than the retroflexing after a lingual consonant because it requires the tongue to move a shorter distance and to change its shape less. When the tongue isn’t used to make the preceding consonantal constriction, speakers are free to choose the retroflex pronunciation instead.

[j] can be pronounced in these two ways because both lower the third formant (F3) extremely. Retroflexing creates a large sublingual cavity from which a low F3 arises, and bunching constricts the vocal tract along a long stretch of the palate where the F3 has a velocity maximum, lowering its frequency. Neither the local constrictions nor the global vocal tract configurations are the same in these two ways of pronouncing
The reason why the articulation lowers the F3’s frequency is also not the same in the two pronunciations. Thus, the speakers’ target in all pronunciations of [i] is the acoustic or auditory effect produced by these articulations and not the articulations themselves, even though speakers choose the more efficient articulation in a particular context.

The most compelling evidence that speakers’ targets are acoustic or auditory rather than articulatory comes from studies in which auditory feedback about the sound is perturbed rather than its articulation (Houde & Jordan, 1998, 2002; Jones & Munhall, 2002, 2003). Houde & Jordan (1998, 2002) altered auditory feedback to listeners gradually such that the vowel [e] in pet came to sound increasingly like the higher vowel [i]. In response, speakers shifted their articulations to the lower vowels [æ] or [a], undoing the alteration. No speaker compensated completely for the altered feedback, and some speakers compensated far less than others. Speakers also compensated by shifting their articulations on trials where the feedback about pet was replaced by noise. They shifted their pronunciations for [e] in words with this vowel other than pet and for other vowels, too, even though feedback wasn’t altered for [e] other than in pet or for other vowels. These shifts in pronunciation show that speakers have acoustic or auditory rather than articulatory targets, and that these targets are determined in relation to the auditory targets of other sounds in the same class.

All these results are compatible with the hypothesis that speakers’ targets are auditory rather than articulatory, while only some of them are compatible with the opposing hypothesis. They thus suggest that the invariants from which distinctive feature values emerge are the auditorily similar effects of covarying acoustic properties and not the motor equivalences of different combinations of articulations.

3 Explanation
3.1 Introduction

Phonetic explanations of phonological patterns are built from physical, physiological, and/or psychological properties of speaking and listening. For example, the absence of the voiced velar stop /g/ in languages such as Dutch or Thai where /b/ and /d/ are phonemes is explained by how much harder it is to keep air flowing up through the glottis when the stop closure is velar rather than bilabial or alveolar (Ohala, 1976; Javkin, 1977). The intrusion of stops between nasals or laterals and following fricatives in many American English speakers’ pronunciations of words such as warm[p]th, prin[t]ce, leng[k]th, and el[t]se is explained as a timing error in which voicing ceases and in the case of the nasal-fricative sequences the soft palate rises before the oral articulators move to the fricative configuration (Ohala, 1971, 1974, 1981). The palatalization of the velar stop [k] to [kʰ] before [i] is explained by the consonant’s
coarticulation with the vowel and its eventual affrication to [tʃ] is explained by the auditory similarity of [k] to [tʃ] (Plauché, Delogu, & Ohala, 1997; Guion, 1998; Chang, Plauché, & Ohala, 2001).

Although all of these phonological patterns are peculiar to particular speech communities or even individuals (many languages have /g/ as well as /b, d/; stops don’t intrude between nasals or laterals and fricatives in South African English, Fourakis & Port, 1986, and [k] remains unpalatalized and unaffricated before [i] in many languages), they recur in unrelated speech communities and in individuals who have no contact with one another, and they are phonetically possible in all speech communities and individuals (/g/ is present only because speakers make the articulatory adjustments needed to keep air flowing up through the glottis, speakers of South African English must coordinate oral articulations precisely with laryngeal and palatal ones to avoid intrusive stops, and [k] is fronted before [i] even if not fully palatalized). They recur and are always phonetically possible because all humans who aren’t suffering from some speech or hearing pathology possess essentially the same apparatus for speaking and listening. Indeed, as Ohala has repeatedly shown, these and many other phonological patterns can be reproduced in the laboratory with speakers and listeners whose languages don’t (yet) exhibit them. Explanations of this kind are highly valued because they are built on generalizations of properties that can be observed any time the affected sound or sounds are uttered or heard, and they are in many instances built on generalizations of properties that can be observed in other domains than speaking and listening.

This section has three parts. I first discuss the iambic-trochaic law, a strong correlation between quantity sensitivity and the choice of iambic rather than trochaic feet, that appeared to have a straightforward explanation in terms of how listeners prefer to pair non-speech sounds alternating in duration vs intensity (Hayes, 1985). Subsequent work by Kusumoto & Moreton (1997) and Patel, Iversen, & Ohguchi (2004) has ruled out this explanation by showing the rhythmic properties of a listener’s native language influence their grouping of non-speech sounds alternating in duration rather than vice versa. The second section explains why particular subsets of languages’ segment inventories contain the segments that they do. These subsets are the pulmonic and glottalic stops and the oral, nasal, and reduced vowels. The purpose of this discussion is to show how the contents of even a compact subset of a segment inventory is multiply determined. The final section discusses how sound changes can be explained phonetically and whether their phonetic motivation persists in the synchronic phonology once they’ve been phonologized.

3.2 The iambic-trochaic law: An explanation that failed
Hayes observes that when syllable weight\textsuperscript{5} determines the choice of the stressed syllable, i.e. when stress is quantity-sensitive, then those syllables are much more likely to grouped into weak-strong or iambic feet than into strong-weak or trochaic feet. Languages with iambic feet also create a durational contrast between the weak and strong syllables in a foot composed of two short syllables by lengthening the strong syllable or shortening or reducing the weak syllable. No adjustments of this kind are observed in languages with trochaic feet. This association between quantity-sensitivity and iambic footing vs quantity-insensitivity and trochaic footing has came to be called the “iambic-trochaic law”.

Hayes argues that this association arises because listeners prefer to group sounds that alternate in duration into short-long pairs, while they prefer to group sounds that alternate in intensity into loud-soft pairs. Listeners show these preferences when asked to group alternating sequences of pure tones, buzzes, and noise intervals. Hayes argues they show the same preferences when grouping pairs of syllables into feet. When syllables’ weight or quantity determines which syllable is stressed, listeners prefer the short-long over the long-short grouping. Because the longer syllable is more prominent than the shorter one, the result is an iambic foot. When the choice of the stressed syllable is indifferent to quantity, some property other than duration distinguishes stressed from unstressed syllables. Intensity is the most likely candidate, so it’s not surprising that listeners prefer to group syllables in such languages into loud-soft rather than soft-loud pairs.\textsuperscript{6}

This is an admirably straightforward example of a phonetic explanation for a phonological pattern of the kind described above: a quite general property of humans’ behavior – preferences for pairing non-speech sounds in alternating sequences – appears to correspond neatly to a specific linguistic pattern – preferences for pairing syllables alternating in stress into feet. Unfortunately, this explanation turns out to be wrong, because it incorrectly predicts that the rhythmic properties of a listener’s native language will not influence his or her preferences for grouping sequences of non-speech sounds that alternate in duration or intensity.

Kusumoto & Moreton (1997) presented speakers of English, Tokyo Japanese, and Kansai Japanese with tone sequences alternating in duration and intensity to be grouped into pairs.\textsuperscript{7} Listeners from all three groups significantly preferred to group sequences of alternating intensity into loud-soft pairs, i.e. pairs that correspond to the strong-weak pairing of syllables in trochees. This result matches those obtained in the studies cited by Hayes. The uniformity of the responses from the three different linguistic groups shows that the rhythmic differences between the listeners’ native languages don’t influence their preference for pairing tones of alternating intensity. Because this preference is uninfluenced by linguistic experience, it also plausibly
explains why quantity-insensitive footing produces trochees much more often than iambs. Unfortunately, listeners’ responses to sequences alternating in duration weren’t equally uniform: English listeners significantly preferred to group such sequences into short-long pairs but neither Kansai nor Tokyo Japanese listeners did so, too. The two groups of Japanese listeners instead showed a non-significant preference to group such sequences into long-short pairs. Patel, Iversen, & Ohguchi (2004) replicated Kusumoto & Moreton’s experiment with speakers of English and Kansai Japanese. Their results differ from Kusumoto & Moreton’s only in that their Japanese listeners’ preference for the long-short grouping in sequences alternating in duration was significant for the two shorter duration ratios they used; for the longest duration ratio, these listeners significantly preferred the short-long grouping. These two studies show, contrary to expectation, that a listener’s native language does influence their grouping preferences.

Both Kusumoto & Moreton (1997) and Patel, et al. (2004) point to a number of rhythmic differences between English and Japanese words that may explain why listeners from the two languages differ in their preference for grouping sequences alternating in duration. Patel, et al. report that the durations of the two syllables are about equal in two-syllable English words with first syllable stress like other, but that the second syllable is 3 to 4 times longer than the first in words with second syllable stress like about. In two-syllable Japanese words with the same number of moras in both syllables, i.e. either short-short or long-long, syllable durations are about equal, while in long-short and short-long words, the long syllable is roughly twice as long as the short one. Kusumoto & Moreton observe that English prohibits short full vowels from occurring in final open syllables – the vowels that occur in such syllables are either reduced or long. They also point out that five times as many two-syllable English words have a short vowel in the first syllable and a long one in the second rather than long vowel in the first and short one in the second. Again, Japanese is more balanced. Words ending in short full vowels are common in Japanese. According to Patel, et al., of the 50 most frequent two-syllable Japanese words, 39% are short-short, and 15% are long-short, for a total 54% with a short second syllable. Of the remaining words in the top 50, 20% are short-long and 26% long-long. All these differences between the two languages suggest that English speakers prefer to group sequences alternating in duration into short-long pairs because there is a strong tendency for the second syllable to be longer than the first in two-syllable words in that language, while Japanese listeners don’t show the same preference because no comparable durational asymmetry exists between the syllables of its two-syllable words. They also suggest that the explanation for listeners’ preferences for grouping alternating sequences of non-speech sounds are determined by rhythmic properties of the languages they speak rather than vice versa.

I have presented this example at some length to sound a note of caution about
how phonetic explanations for phonological patterns should be constructed. It’s not enough to show a correspondence between a phonological pattern and a phonetic or non-linguistic behavior as Hayes (1985) did. It’s also necessary to test the explanation’s predictions as Kusumoto & Moreton (1997) and Patel, et al. (2004) did.

3.3 Explaining inventory content
3.3.1 Introduction

There is considerable evidence that the contents of segment inventories can be explained phonetically. An explanation is first developed for the inventories of the pulmonic and glottalic stops commonly found in languages. This explanation depends equally on efficient use of distinctive features and the aerodynamics of laryngeal articulations. I then turn to explaining why languages have the oral, nasal, and reduced vowels they do. These explanations depend on sufficient dispersion of vowels in the vowel space, the perceptual salience of particular vowel qualities, vowel duration, and other as yet undiscovered factors. The lesson in both sets of explanations is that the contents of segment inventories, even apparently compact subsets of inventories such as these, are determined by many more than just one factor. We will also see that these factors may conflict with one another, and that a balance must be struck between them when they do.

3.3.2 Stops

Table I lists the frequency of pulmonic and glottalic stops\(^8\) at the three major places of articulation in the 451 languages in the genetically and areally balanced UCLA Phonological Segment Inventory Database (UPSID, Maddieson & Precoda, 1992).

<table>
<thead>
<tr>
<th>Airstream, Laryngeal</th>
<th>Labial (Bilabial)</th>
<th>Coronal (Dental, Alveolar)(^9)</th>
<th>Dorsal (Velar)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulmonic</td>
<td>Voiced</td>
<td>287</td>
<td>279</td>
</tr>
<tr>
<td></td>
<td>Voiceless</td>
<td>375</td>
<td>413</td>
</tr>
<tr>
<td></td>
<td>Voiced/Voiceless</td>
<td>0.765</td>
<td>0.676</td>
</tr>
<tr>
<td>Glottalic</td>
<td>Implosive</td>
<td>64</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Ejective</td>
<td>44</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Implosive/Ejective</td>
<td>1.455</td>
<td>0.911</td>
</tr>
</tbody>
</table>

Table I. The number of languages with voiced and voiceless (unaspirated) pulmonic stops and implosive and ejective glottalic stops at the three major places of articulation.
in UPSID, and voiced/voiceless and implosive/ejective ratios.

The raw frequencies in the top of this table show that voiceless unaspirated stops are more common at all three major places of articulation than the corresponding voiced stops. They also show that voiced stops become less frequent the farther back in the mouth the stop is articulated, while voiceless stops are less frequent at the very front of the mouth than farther back. The front-to-back differences observed in the pulmonic stops are exaggerated in the corresponding glottalic stops, so much so that only a very small number of languages have a velar implosive and bilabial impulsives actually outnumber bilabial ejectives. The voiced/voiceless and implosive/ejective ratios make these differences between places of articulation even more obvious: these ratios are largest for bilabial stops and smallest for velar stops, with dental and alveolar ratios falling in between.

These facts suggest that a language may lack a velar voiced stop or implosive, while having a voiceless stop or ejective at this place of articulation, and at the other end of the oral cavity, it may lack a bilabial voiceless stop or ejective, while having the corresponding voiced stop or implosive. This expectation can be tested by counting the number of languages in which a voiced stop or implosive is missing at the velar place but not at more anterior places or a voiceless stop or ejective is missing at the bilabial place but not at more posterior places (Table II).

<table>
<thead>
<tr>
<th>Place Pairs</th>
<th>Voiced</th>
<th>Voiceless</th>
<th>Implosive</th>
<th>Ejective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velar, No Bilabial</td>
<td>9</td>
<td>35</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Bilabial, No Velar</td>
<td>43</td>
<td>7</td>
<td>59</td>
<td>3</td>
</tr>
<tr>
<td>Bilabial, No Coronal</td>
<td>15</td>
<td>1</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Coronal, No Bilabial</td>
<td>7</td>
<td>39</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Velar, No Coronal</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Coronal, No Velar</td>
<td>34</td>
<td>11</td>
<td>46</td>
<td>4</td>
</tr>
</tbody>
</table>

Table II. Frequency of gaps in pulmonic or glottalic stop inventories at one place of articulation but not at a more anterior or posterior place of articulation, for all possible pairs of major places.

The top two rows in Table II show that a language is much more likely to have a bilabial voiced stop or implosive and lack their velar counterparts than vice versa, and it is much more likely to have a velar voiceless stop or ejective and lack their bilabial counterparts than vice versa. More anterior and more posterior stops pattern in the same way in the
middle two rows of the table where bilabial and coronal places are compared. The last
two rows, where coronal and velar places are compared, show that a language is also
much more likely to have a coronal voiced stop or implosive and lack their velar
counterparts than vice versa. However, it is also more likely to have a coronal voiceless
stop and lack its velar counterpart than vice versa, and it is not much more likely to have
a velar ejective while lacking its coronal counterpart than vice versa.

These facts lead to the following implications, where “x > y” means that the
presence of the sound x in a language implies the presence of sound y.

<table>
<thead>
<tr>
<th>Voicing</th>
<th>Airstream</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced</td>
<td>Pulmonic</td>
<td>g &gt; d &gt; b</td>
</tr>
<tr>
<td></td>
<td>Glottalic</td>
<td>g' &gt; d' &gt; b</td>
</tr>
<tr>
<td>Voiceless</td>
<td>Pulmonic</td>
<td>p &gt; t, p &gt; k, but k &gt; t</td>
</tr>
<tr>
<td></td>
<td>Glottalic</td>
<td>p' &gt; t' &gt; k'</td>
</tr>
</tbody>
</table>

Table III. Implications of the frequencies in Tables I and II.

Implicational statements like these are commonplace in linguistics, but they are
seldom substantiated by anything more than observations like those just presented. It’s
possible to do better. Patterns of co-occurrence can be arranged in the form of a
contingency table, and the (in)dependence of one sound’s occurrence on another’s can
be assessed statistically. In such analyses, a statistical model of the data is constructed
from which one calculates how often two sounds are expected to co-occur, i.e. how
many languages are expected to have both sounds. These expected frequencies of co-
ocurrence are then compared to the observed frequencies. If the expected and
observed frequencies are close, then the model fits the data and can be used to
interpret it. Otherwise, another model is tried. The starting point for modeling is to
assume that one sound’s occurrence is independent of the other’s. If this is the case,
then the observed frequency of their co-occurrence should be close to the product of
their individual probabilities of occurrence. If the observed frequency differs
considerably from this product, then one must instead consider the possibility that one
sound’s occurrence depends on the other’s, and construct a model including that
interaction. That model is then used to calculate a new expected frequency of co-
ocurrence, which is compared to the observed frequency. Interactions are added to
the model until there’s no longer any large discrepancy between observed and expected
frequencies of co-occurrence.

To see whether the occurrence of a stop with one laryngeal articulation or
airstream mechanism depended on the occurrence of a stop with the same place of articulation but another laryngeal articulation or airstream mechanism, hierarchical loglinear models were constructed for contingency tables of six dimensions that combined (1) voiced and voiceless unaspirated stops, (2) voiced and voiceless aspirated stops, (3) voiceless unaspirated and voiceless aspirated stops, (4) voiced stops and implosives, or (5) voiceless unaspirated and ejective stops. Only four of the fifteen possible two-way interactions between stops contrasting in their laryngeal articulations or airstream mechanisms at the same place of articulation, e.g. between /b/ and /p/, /l/ and /pʰ/, etc., reached significance. The occurrence of a voiceless aspirated stop at a particular place of articulation depended on the occurrence of a voiceless unaspirated stop at that place, but not in the same way for all three places of articulation. Voiceless aspirated stops /tʰ, kʰ/ were significantly more likely to occur when the voiceless unaspirated stops /t, k/ did not occur, but /pʰ/ was significantly more likely to occur when /p/ also did. The only other significant interaction was between the coronal ejective /t'/ and its voiceless unaspirated counterpart /t/, where /t'/ is significantly more likely to occur when /t/ does not. The lack of significance for the other interactions shows that the occurrence of a voiced stop at a particular place of articulation doesn’t depend on the occurrence of a voiceless unaspirated or aspirated stop at that place, nor does the occurrence of an implosive at a particular place depend on the occurrence of the corresponding voiced stop, and the occurrence of bilabial and velar ejectives doesn’t depend on the occurrence of corresponding voiceless unaspirated stops. In other words, stops with one of these laryngeal articulations or airstream mechanisms do or don’t occur regardless of what other laryngeal articulations or airstream mechanisms a language uses to produce a contrasting stop series. Because most of these interactions were not significant and the significant ones weren’t systematic and because three-way tables are far easier to interpret than six-way tables, voiced, voiceless unaspirated, voiceless aspirated, implosive, and ejective co-occurrence patterns are analyzed separately.

A hierarchy of loglinear models\(^{10}\) was built and evaluated, adding interactions one after another until observed and expected frequencies no longer differed substantially from one another – at this point the model “fit” the data (Fienberg, 1980). The columns and rows labeled “yes” and “no” in the top halves of Tables IVA-c show the frequencies with which the individual voice, voiceless unaspirated, or voiceless aspirated pulmonic stops, /b, d, g/, /p, t, k/, and /pʰ, tʰ, kʰ/ do and don’t co-occur with other stops with the same laryngeal articulation. Immediately below the observed frequencies are the expected frequencies obtained from the best-fitting model. In the bottom half of these tables are listed the statistics for the successive models in the hierarchy. \(G^2\) is the measure of the difference between the expected and observed frequencies for a particular model, which has a probability (\(p\)) of occurring by chance for a particular number of degrees of freedom (\(df\)) in the model. Modeling continues through the
hierarchy until $p > 0.05$ and/or the difference in $G^2$ between the current model and the previous one in the hierarchy is no longer significant for the difference in degrees of freedom between the two models. Modeling also stops once the model is saturated. The successive models in the hierarchy are represented as follows: $x, y, z =$ whether one sound occurs is independent of the other sounds; $xy =$ whether $x$ occurs depends on $y$ but whether $x$ or $y$ occur is independent of $z$; $xz =$ whether $x$ occurs depends on both $y$ and $z$ but whether $y$ occurs is independent of $z$, and so on. The saturated model would be represented as $xyz$, which indicates that the occurrence of any one of the sounds depends on the occurrence of the other two. Modeling stops before the saturated model is reached because its expected frequencies exactly match the observed frequencies.

<table>
<thead>
<tr>
<th>(a) Voiced</th>
<th>b</th>
<th></th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>Yes</td>
<td>240</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>241</td>
<td>3</td>
</tr>
<tr>
<td>g</td>
<td>Yes</td>
<td>32</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>31</td>
<td>12</td>
</tr>
</tbody>
</table>

| (b) Voiceless | p | | |
|---|---|---|
| | Yes | No | | Yes | No |
| t | Yes | 367 | 1 | 35 | 0 |
| | No | 368 | 0 | 34 | 1 |

<table>
<thead>
<tr>
<th>Model</th>
<th>$G^2$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>b,d,g</td>
<td>782.4</td>
<td>4</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>bd,bg</td>
<td>52.1</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>bd,dg</td>
<td>11.5</td>
<td>2</td>
<td>= 0.003</td>
</tr>
<tr>
<td>bg,dg</td>
<td>129.2</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>bd,bg,dg</td>
<td>1.8</td>
<td>1</td>
<td>= 0.175</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>$G^2$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p,t,k</td>
<td>352.0</td>
<td>4</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>pt,pk</td>
<td>79.1</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>pt,tk</td>
<td>15.4</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>pk,tk</td>
<td>25.6</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>pt,pk,tk</td>
<td>5.6</td>
<td>1</td>
<td>0.018</td>
</tr>
</tbody>
</table>
The best fitting model for all three kinds of stops is that containing all two-way interactions. No voiced, voiceless unaspirated, nor voiceless unaspirated stop is likely to occur at any place of articulation in a language without the other two stops also occurring in that language. In other words, if a language has a stop with one of these laryngeal articulations at one major place of articulation, it is more likely to have all of them than to have a gap at one or both of the other major places of articulation.

Tables Va, b show the observed and expected frequencies and the model hierarchies for the glottalic stops:

<table>
<thead>
<tr>
<th>Model</th>
<th>$G^2$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^h, th, kh$</td>
<td>789.3</td>
<td>4</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$p^{th}, p^{th}k^h$</td>
<td>44.9</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$p^{th}, thk^h$</td>
<td>16.4</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$p^{kh}, thk^h$</td>
<td>60.6</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$p^{th}, p^{th}k^h, thk^h$</td>
<td>5.4</td>
<td>1</td>
<td>= .020</td>
</tr>
</tbody>
</table>
Modeling building stops sooner than expected for the implosives in Table Va because /g/ doesn't occur in any language that doesn't have both /b/ and /d/. As a result, models including the $bg$ and $dg$ interactions are degenerate. The model with the $bd$ interaction significantly improves the fit over the one in which all the stops are independent, which shows that it's more likely for implosives to occur at both of these places of articulation than at just one of them. Finally, the outcome for the ejectives is like that for the pulmonic stops: if a language has an ejective at any major place of articulation, then it's significantly more likely to have ejectives at all major places than to have a gap at one or both of the other major places.

In short, voiced, voiceless unaspirated, and voiceless aspirated stops as well as ejectives are more likely to occur at all the major places of articulation at once or at none of them than to occur at just one or two major places. However, a velar implosive is only likely to occur in a language which has both bilabial and coronal implosives, too, and an implosive is only likely to occur at one of these places if an implosive also occurs.
at the other.

Although the pre-statistical analysis suggested that voiced stops and implosives were more likely to be missing at the velar place of articulation than more anterior places and that voiceless stops and ejectives were more likely to be missing at the bilabial place of articulation than more anterior places, the statistical analysis shows that only one of these gaps is not in fact significantly less likely to occur than expected, the absence of a velar implosives in languages which have more anterior implosives. Otherwise, for the other four kinds of stops, the occurrence of a stop at any major place of articulation predicts that stops of the same kind will occur at both the other two major places of articulation. So the first piece of the explanation for the contents of the pulmonic and ejective stop inventories is that languages use these laryngeal articulations maximally. See Ohala (1980) for discussion of the maximal use of distinctive features as a general property of consonant inventories.

The analyses of the six-way tables showed that with one unsystematic exception, the occurrence of stops produced with one laryngeal articulation or airstream mechanism is independent of the occurrence of stops produced with another. This result is important because it shows that whether a language uses an ostensibly marked laryngeal articulation or airstream mechanism doesn’t depend on whether it does or doesn’t use the contrasting unmarked laryngeal articulation or airstream mechanism.

Because the glottalic stops are much rarer than their pulmonic counterparts, they are often treated as marked, yet these analyses show, too, that the implosives or ejectives do or don’t occur in a language regardless of whether their unmarked pulmonic counterparts, voiced and voiceless unaspirated stops, respectively, also occur. Even more telling is the statistical independence of the voiced pulmonic stops from either voiceless unaspirated or aspirated stops. Voicing in pulmonic stops is widely claimed to be a marked feature: Table I above shows that voiced pulmonic stops occur noticeably less frequently than voiceless unaspirated stops, and many languages devise voiced stops or prohibit their occurrence in particular contexts (for discussion see Lombardi, 1991; Steriade, 1999; Itô & Mester, 2003 among many others). The one piece of statistical evidence that the presence of a stop with one laryngeal articulation depends on the presence of another is the significant interactions between voiceless aspirated and unaspirated stops, but these interactions aren’t systematic: if a stop occurs with one of these laryngeal articulations at the coronal or velar places of articulation, it’s significantly less likely to occur there with the other laryngeal articulation, while stops with these two laryngeal articulations are significantly more likely to co-occur at the bilabial place of articulation. So even for this pair of laryngeal articulations, it’s impossible to claim that the marked laryngeal articulation – presumvably aspiration – can only occur when it contrasts with the unmarked one.
Since the marked laryngeal articulations and airstream mechanisms contrast with the unmarked ones no more often that would be expected if one’s occurrence is independent of the other’s, it’s necessary to ask whether only the marked member of the contrast is specified phonologically for the relevant distinctive feature, e.g. whether [voice] contrasts with nothing in the representations of voiced stops vs voiceless unaspirated stops, respectively. Such representations are called “privative”. The independence of marked and unmarked series instead recommends more egalitarian phonological representations in which each laryngeal articulation or airstream mechanism is phonologically specified, in “equipollent” rather than privative representations. Equipollent representations also agree better with the phonetic facts in that pronouncing unmarked sounds positively requires the speaker to produce a distinct array of phonetic properties from those in marked sounds and not merely to omit the phonetic properties of marked sounds. Of course, the contrasts found in inventories are not the only evidence relevant to deciding how sounds should be represented phonologically; their behaviors, as reflected in phonological processes and constraints, must also be taken into account. If those behaviors appear to reflect privative representations, then we have to work out why contrast apparently reflects equipollent representations instead. Resolving this possible paradox is a challenge for future research.

Another challenge for future research is explaining why laryngeal articulations and airstream mechanisms behave differently from place of articulation, where the statistics show that a stop with a particular laryngeal articulation or airstream mechanism is likely to occur at one major place only if the same kind of stop also occurs at both other major places. Does this mean that the phonological representation of place contrasts is formally different from that of laryngeal and airstream contrasts? There is some evidence that they behave differently in phonological processes and constraints (Lombardi, 2001), but again there isn’t enough evidence to answer the question.

The second fact to be explained is: why is /g/ so rare? The answer is that it is aerodynamically difficult to distinguish /g/ from /g/ (see Javkin, 1977, for extensive discussion). The speaker’s goal in an implosive is to constrict the glottis and then expand the oral cavity as much as possible while the stop closure is being held. This expansion keeps intraoral air pressure from rising and may in extreme cases even create a negative intraoral air pressure.

Some expansion is necessary in pulmonic voiced stops, too, if they are to remain voiced, because otherwise intraoral air pressure (Po) will quickly build up enough to approach subglottal air pressure (Ps) and air will stop flowing up through the glottis. As it is this air flow through the glottis that causes the vocal folds to vibrate, oral cavity must
be expanded to slow Po's rise. The oral cavity is expanded both actively (Westbury, 1983) and passively (Ohala & Riordan, 1979). Active expansion is achieved primarily by advancing the tongue root and lowering the larynx, although the soft palate may also be raised and the tongue body lowered. The oral cavity expands passively because the rising Po pushes on any surfaces of compliant tissue lining the vocal tract, principally the lips, cheeks, and soft palate.

Producing voicing in a velar stop is a challenge because the articulation that produces the closure at the palate severely constraints both active and passive expansion. Raising the tongue body to close the oral cavity at the palate inhibits both tongue root advancement and larynx raising – the latter by pulling upward on the hyoid bone, from which the larynx is suspended. Passive expansion is constrained because the largest surface of compliant tissue, the lips and the cheeks, is downstream from a velar stop closure, and at least part of the remaining surface, the soft palate, is actually covered by the tongue body. There is an additional difficulty, the cavity between the oral closure and glottis in a velar stop is the smallest of the three major places of articulation, having only about half the volume of that behind a bilabial closure. For a constant flow of air up through the glottis, Po will rise faster in a smaller than larger cavity behind the stop closure. Po thus approaches Ps fastest in a velar stop, which is also the place of articulation that handicaps the speaker most in expanding the oral cavity, either actively or passively, to slow Po’s rise. The speaker may be able to expand the oral cavity enough to produce a pulmonic voiced stop, but there is little capacity for the additional expansion that would produce a glottalic one.

The velar implosive /g/ is so rare, and is absent in a language such as Sre that has more anterior implosives because it is aerodynamically challenging to distinguish it from /g/. This aerodynamic challenge also explains why languages such as Dutch lack the pulmonic voiced velar stop, /g/. And these relationships between the anatomy, physiology, and aerodynamics also explain why K’ekchi and Arabic have ejectives and voiceless unaspirated stops, respectively, at all major places but the bilabial. What’s an advantage when one wants to keep intraoral air pressure from rising too fast is a disadvantage when one instead wants Po to rise quickly to produce the more intense bursts that, among other properties, distinguish glottalic and pulmonic voiceless stops from their voiced counterparts. Recall, however, that all these gaps but that for velar implosives occur significantly less often than expected; therefore, the requirement to use distinctive features maximally usually overcomes these aerodynamic challenges. This requirement is essentially phonological not phonetic, and its greater potency shows that the phonology may regulate the phonetics just as much as the phonetics regulates the phonology (see Ohala, 1983b, for other evidence of mutual regulation).

3.3.3 Oral vowels
Even a casual look at the vowel inventories of just a few randomly chosen languages shows that they resemble one another closely. Liljencrantz & Lindblom (1972) and subsequent work by Lindblom (1986) established that these resemblances are determined by two factors: the vowel space’s limits and mutual repulsion of one vowel from another within that space. Following up proposals in Stevens (1989), Schwartz, Boë, Vallée, & Abry (1997a,b) added a third factor: vowels are preferred that are made salient when two of their formants are close in frequency, an effect called “focalization”. This section begins with a description of the recurrent properties of the oral vowel inventories in the 451 languages collected in UPSID (Maddieson & Precoda, 1992) to show what exactly needs to be explained. Then an attempt is made to explain the recurrence of these properties in terms mutual repulsion and focalization. These forces successfully predict that inventories of certain sizes are preferred over both smaller and larger ones but fall short in predicting which vowels are most likely to occur in an inventory of a given size.

The short and long oral vowels were extracted from each of the 451 languages in the UCLA Phonological Segment Inventory Database (UPSID). All secondary articulations were stripped off, and the distinct short and long vowel qualities were counted. If a language’s long vowels distinguished more vowel qualities, their number was used to represent how many vowels that language had; otherwise, the number of distinct short vowel qualities was used.

The smallest number of distinct vowel qualities was three and the largest number thirteen. The histogram in Figure 1 shows how many languages have each number of vowels within these two extremes.
Figure 1. The number of languages having between three and thirteen oral vowel qualities, based on the short and long oral vowel inventories in UPSID.

There is a very clear mode at five vowels, which are found in 136 languages. Even though large numbers of languages have either six or seven vowels, only about 65% as many have six vowels as have five vowels and only about 56% as many have seven. This is the first fact to be explained.

Fully two-thirds of the languages in the sample (300/451) have the three most common numbers of vowels, five-seven. The strength of this preference is emphasized
by how few languages have four or eight vowels: in each instance just under 29% as many languages as have five or seven vowels, respectively. This precipitous drop also needs to be explained. Given the drop off in frequency between seven and eight vowels, why do noticeably more languages have nine than eight vowels?

Aside from these preferences and dispreferences for a certain number of vowels, the remaining facts to be explained are the preferred arrangements of vowels in the vowel space for inventories of different sizes. The 44 oral vowel qualities distinguished in the UPSID sample were divided into 15 peripheral vowels and 29 central qualities. A vowel quality is defined as peripheral if there is no quality closer to the front, back, or bottom of the vowel space than it. At the top of the vowel space, only front unrounded and back rounded vowels, [i] and [u], are peripheral. For non-low peripheral vowels, the tongue body is either as far forward or backward as possible and the lips are unrounded if the tongue is the front but rounded if the tongue is back. In low peripheral vowels, the tongue is as low as it can get. This cluster of properties shows that peripheral vowels are actually defined acoustically rather than articulatorily: if they are non-low, they have the highest or lowest F2 and F3 frequency values for their tongue height, or if they’re low, they have the highest F1 frequency values. Central vowels include those articulated with the tongue body in a central position as well as vowels in which the tongue body is fully front or back, but the lips are rounded when the tongue is front or unrounded when it’s back. This definition is again acoustic rather than articulatory as the F2 and F3 frequencies are neither lowered nor raised particularly when the tongue body is central or when the lips are rounded in front vowels or unrounded in back ones. The essential observation then is that peripheral vowels are acoustically farther apart from one another in the vowel space than the central vowels.15

The vowel inventories for each of the languages with from three to ten vowels were then classified into patterns by whether they had any central vowels, and if so, how many – as there are only 8 languages with more than 10 vowels, they are ignored in the rest of this discussion. A pattern is identified by a formula of the form “P+C”, where “P” represents the number of peripheral vowels and “C” the number of central vowels. The results of this classification are shown in Figure 2, where each panel corresponds to a number of vowels and the heights of the bars indicate the proportion of all the languages with that many vowels that have a particular pattern.16
Figure 2. Proportions of languages with from three to ten vowels that exhibit particular common patterns of peripheral and central vowels for languages. Some proportions don’t add up to 1 because patterns aren’t shown that appear in only a very few languages or that can’t be classified as one of these patterns.
In the vast majority of five vowel languages, all five vowels are peripheral (Figure 2c). In all but three of the 128 languages with this 5+0 pattern, front unrounded and back rounded high and mid vowels contrast and there is one low vowel. The remaining 8 languages with five vowels have four peripheral vowels and one central vowel, the 4+1 pattern. Figure 2d shows that the most common patterns among six vowel languages is 5+1, which occurs in 64 languages. The second most common pattern among the six vowel languages is 6+0, which is only a third as common as the 5+1 pattern, occurring in 22 languages. The most common pattern among the seven vowel languages is 7+0 (Figure 2e), in 36 languages. Of these 36 languages, 26 distinguish higher mid from lower mid vowels in both the front and back, /i, ɛ, ə, a, o, u/. The second most common pattern among the seven vowel languages is 5+2, which is not all that much less common than the 7+0 pattern, in 28 vs 36 languages. Languages with five peripheral vowels and two central vowels are much less common than those with five peripheral vowels and just one central vowel, 28 vs 64. A dozen languages have the third most common pattern among the seven vowel languages, 6+1.

This summary shows that many languages with six or seven vowels have the same peripheral vowel inventory as the very large number of languages with the 5+0 pattern. All the details aren’t presented here, the heights and other properties of the central vowels in these languages are unpredictable from what peripheral vowels they have, suggesting that central and peripheral inventories are independent of one another.

These patterns and frequencies suggest that following evolutionary paths. The base inventory is the exceptionally common 5+0 pattern. Vowels can be added to this pattern in two ways. Either one or two central vowels can be added, creating the 5+1 and 5+2 patterns, or one or two peripheral mid vowels can be added, creating the 6+0 and 7+0 patterns. Adding one or two central vowels doesn’t change the distribution of peripheral vowels in a language, but adding one peripheral mid vowel, in either the front or the back, usually entails adding the other at the same height, as well as shifting the existing mid vowel’s height to equalize the intervals between vowels of different heights. Finally, the 6+1 pattern is created by adding one peripheral and one central vowel.\textsuperscript{17}

A briefer look at the inventories of languages with fewer than five vowels and more than seven confirms this evolutionary interpretation.

All 26 languages with just three vowels have the 3+0 pattern (Figure 2a).\textsuperscript{18} 22 of them are missing the two mid vowels that occur in the 5+0 pattern, while just 4 are missing one or both high vowels. The 39 languages with four vowels divide unevenly
into 24 with four peripheral qualities and 14 with three peripheral qualities and one central quality – the remaining four-vowel inventory /i, ɪ, a, o/ is unclassifiable (Figure 2b).

Turning now to the languages with more than seven vowels, Figures 2f-h show that the most common patterns have odd numbers of peripheral vowels, five, seven, or even nine. More often than not, the peripheral vowels in larger inventories also symmetrically contrast front unrounded with back rounded vowels at all non-low heights. Among the 22 languages with eight vowels, 8 have the 7+1 pattern, where the peripheral vowels are those found in the common 7+0 inventories. Otherwise, 5 have one of the common 5+0 peripheral inventories plus three central vowels. 4 languages have the 8+0 pattern, and of the remaining 4, 2 each have the 6+2 pattern and 4+4 patterns. There is an even stronger preference for an odd number of peripheral vowels among those languages with nine vowels, where 19 languages add two central vowels to one of the 7+0 peripheral inventories, and 9 languages have the 9+0 pattern, created by adding a third mid height. Notice that 7+2 pattern occurs more than twice as often as the 9+0 pattern. 6 of the nine-vowel languages arrange their vowels into the 6+3 pattern, and the remaining 5 languages have the 8+1 pattern. Finally, among the languages with ten vowels, those with an odd number of peripheral vowels again outnumber those with an even number: 6 languages fall into the 9+1 pattern and 5 into the 7+3 pattern, while only 4 fall into the 8+2 pattern and 2 into the 10+0 pattern. Ten-vowel languages with seven peripheral vowels are nearly as common as those with nine. The remaining 2 ten-vowel languages fall into the 6+4 and 5+5 patterns, respectively.

Summing up, the most common vowel patterns contrast front unrounded with back rounded peripheral vowels at all but the lowest height, where only a single, usually central vowel is found. Larger vowel inventories differ from smaller ones in two ways: they may have central vowels that are absent in the smaller inventories and/or they may instead have more contrasts between the high and low extremes. The most common inventory by far is the 5+0 pattern. The next smallest inventory is more likely to have lost one of the mid vowels, producing the 4+0 pattern in 24 languages, than to lose both and add one central vowel, producing the 3+1 pattern in only 14 languages. It’s far more likely that a language will add one central vowel, producing the 5+1 six-vowel inventory in 64 languages, than to add one peripheral vowel, producing the 6+0 inventory in just 22 languages. However, when two or more vowels are added to 5+0 pattern, it is far more likely that a single front:back pair will be added before any other vowels. Among languages with seven, eight, and nine vowels, 7+0 > 5+2, 7+1 > 5+3, and 7+2 > 9+0, respectively, and for languages with ten vowels, 7+3 occurs in nearly as many languages as 9+1. Once this additional pair of peripheral vowels is added, any more vowels are likely to be central.
Why are certain total numbers of vowels and particular patterns within each total number preferred over others? Following Liljencrantz & Lindblom (1972) and Lindblom (1986), I propose that five to seven vowels are preferred over fewer or more vowels because these numbers of vowels divide the vowel space efficiently. Fewer vowels than five are dispreferred because the space can be divided more finely without crowding the vowels so close together that they’re likely to be confused, while more vowels than seven are dispreferred because above that number the vowels are crowded too closely together. This outcome would be obtained if vowels are required to contrast sufficiently but not maximally within the limits of the vowel space. Up to a point, height contrasts can be multiply among the peripheral vowels without the vowels coming too close together, but central vowels are resorted to when a vowel inventory gets so large enough that a yet finer division of the height continuum among the peripheral vowels pulls adjacent vowels below the threshold for sufficient contrast.

These are all functional explanations. On the one hand, if a vowel inventory is too small, more consonants or longer strings of segments will have to be used to create distinct messages (see also Flemming 2001, 2004). On the other hand, if an inventory is too large or its members are acoustically too close to one another, then distinct messages will be confused with one another. Although these explanations are certainly quite plausible, they remain rather hypothetical, and other kinds of explanations can be imagined. For example, languages may prefer to have a back rounded vowel for every front unrounded non-low vowel it has because languages prefer symmetry. This alternative is not implausible because symmetry is not in this instance an abstract, geometric property of a vowel inventory but instead a requirement that a language efficiently use all the possible combinations of distinctive feature values. If a language has a front unrounded vowel of height $n$ and it also has a back rounded vowel at that height, then it combines height $n$ with both [-back] and [+back] rather than with just one value of this feature.\(^1\)

Very small inventories aren’t dispreferred for phonetic reasons but instead because the consequences of having too few vowels are too many consonants or too long messages. On perceptual grounds alone, a small inventory would in fact be preferred because its members would seldom if ever be confused with another, so long as they’re sufficiently dispersed in the vowel space and their pronunciations aren’t so variable that they often overlap with one another.

The rest of this discussion focuses on how the numbers and patterns of vowel inventories are chosen so as to have a sufficient number of distinguishable vowels. Three quantities for each inventory: the extent to which its members are dispersed in the vowel space and how many focal vowels it has.
Dispersion is a property of an entire vowel inventory. The sum of the auditory distances between also pairs of vowels are calculated, and then the reciprocal is taken of this sum. The resulting value is larger when the vowels are crowded together in an inventory and smaller when they’re more dispersed, so it can be conceived of as measuring the energy with which the vowels mutually repel one another. It is called the “dispersion energy”.

Some vowel qualities, e.g. /i/, occur so often, in inventories of different sizes and compositions, that they appear to be favored intrinsically and not just for their auditory distance from other vowels. What may make these vowels special is that two of their formants are so close in frequency that they merge auditorily into a single, relatively narrow yet intense spectral prominence. Formants come together in this way when a change in articulation changes the cavity the formant comes from (Stevens, 1989). This greater intensity or “focalization” makes these vowels more salient than acoustically similar vowels whose formants are farther apart (Schwartz, et al., 1997b). Focalization values are calculated for each vowel as a function of how close adjacent formants are to one another, these values are then summed, and the reciprocal of this sum is taken. The resulting value is smaller for languages with more focal vowels, and larger for those with fewer. It is called the “focalization energy” and like the dispersion energy it increases for less favored inventories.

The procedures for calculating dispersion and focalization energies are those presented by Schwartz, et al. (1997b), which should be consulted for the details. These two energy values, and their sum, the “total energy”, can be calculated for any inventory of the vowels and used to answer the question: are less energetic inventories more frequent? Figure 3 shows median focalization, dispersion, and total energy (F, D, and T in the figure) for vowel inventories with from three to ten vowels.20
Figure 3. Median focalization energy (F), dispersion energy (D), and total energy (T) for languages with from three to ten vowels.

Unsurprisingly, dispersion energy grows with the number of vowels in the inventory. Inventories with more vowels apparently also include more in which two formants are close together than one with fewer vowels, and focalization energy drops steadily as the number of vowels increases. Up to seven vowels, this steady drop in focalization energy offsets the growth in dispersion energy, and total energy remains relatively unchanged. Indeed, focalization energy drops more than dispersion energy grows between four and five vowels, and as a result total energy is somewhat lower in a typical five than four vowel inventory, which may contribute to five vowels being more
popular than four. Total energy then grows only modestly from five to seven vowels. However, as the number of vowels in an inventory increases beyond seven, dispersion energy grows much faster than focalization energy drops, and total energy climbs with increasing steepness. This jump in crowdedness probably explains the markedly lower frequency of languages with eight or more vowels compared to seven or fewer. Total energy grows less steeply between eight and nine vowels than between seven and eight or nine and ten vowels. This shallower growth may partly explain why nine vowel inventories are surprisingly popular: they contain more vowels which are made salient by the closeness in frequency of adjacent formants but which are not crowded excessively closely together. (Languages with fewer than five vowels are probably less frequent for a very different reason: they underuse the capacity of the vowel space to distinguish messages reliably from one another.)

The minimum and maximum values as well as the median value of focalization (F), dispersion (D), and total energy (T) are shown in Figure 4 for inventories of different sizes and compositions. The minimum and maximum energies show how variable the energy values are for the different inventories that fall into a particular pattern – this interpretation is only meaningful when 10 or more languages exhibit a particular pattern. If the median is close to the minimum or maximum, then more of the inventories with that pattern have lower or higher energies.
Figure 4. Minimum (bottom whisker), median (X), and maximum (top whisker) values of focalization (F, left), dispersion (D, middle), and total energies (T, right) for inventories of three to ten vowels, broken down by pattern. The arrangement matches that in Figure 2. The numbers in each division of a panel are the proportions of languages with that inventory pattern; they correspond to the values that are displayed graphically in Figure 2.
For inventories consisting of four to seven vowels (Figures 4a-e), the results are at first glance disappointing: the more popular patterns don’t have noticeably lower energies. To be sure, the maximum dispersion and total energies are smaller for the more popular 5+1 pattern than the less popular 6+0 pattern (Figure 4d). However, the median also lies very close to the minimum rather than the maximum, which shows that very few languages with the less popular pattern have the higher energy versions of this pattern. Moreover, energy values differ very little between the wildly popular 5+0 pattern and the decidedly unpopular 4+1 pattern (Figure 4c), and they differ equally little between the 4+0 vs 3+1 and 7+0 or 5+2 vs 6+1 patterns, despite their marked differences in popularity (Figures 4b,e). Energies are also not uniformly lower for the more popular patterns in languages with from eight to ten vowels (Figures 4f-h).

To see if some relationship might nonetheless be hidden in the data, the proportions with each pattern occurred were correlated with the median focalization, dispersion, and total energies for vowel inventories containing four to ten vowels – three-vowel inventories are left out because they all fall into just one pattern and the proportion is 1. If the more popular inventories have lower energies, then all these correlations should be negative. The correlations were significantly negative for dispersion and total energies (dispersion $r(21) = -0.472, p = 0.031$; total $r(18) = -0.447, p = 0.042$; two-tailed), but curiously, they were significantly positive for focalization energy ($r(21) = 0.455, p = 0.038$). This correlation turns out positive because focalization energy drops as inventory size increases, and larger inventories are divided into more patterns, each making up a smaller proportion of the total than do the fewer divisions of smaller inventories. The correlations with dispersion or total energy are also influenced by this artifact, but it’s hidden in their case because it works in the same direction as the prediction. Accordingly, the correlations were recalculated using only the proportions of the most popular pattern for each inventory size. The results are quite similar: the most popular patterns' proportions correlate negatively with dispersion and total energies (dispersion $r(7) = -0.741, p = 0.056$; total $r(7) = -0.752, p = 0.051$) and positively with focalization energy ($r(7) = 0.777, p = 0.040$), except that the correlations with dispersion and total energy are now only marginally significant. These correlations show that the more popular patterns are less energetic but also that the energy differences between them and the less popular inventories aren’t enormous.

This exercise has shown how the differences in popularity between inventories of different sizes and compositions might be explained, as preferences for vowels that dispersed enough in the vowel space as not to be easily confused with one another and for vowels whose adjacent formants are close enough together to produce prominent spectral peaks. Both properties reduce the energy of a vowel inventory. That five to seven vowels are the most popular inventory sizes does appear to be well explained by a preference for vowels that reduce an inventory’s energy. Energy increases little from
three to seven vowels and then begins to rise steeply as more vowels are added, because beyond seven vowels dispersion energy – the cost of crowding – grows much faster than focalization energy drops. However, energy differences explain much less of the detail about why certain patterns are preferred within inventories of a given sizes. There is at best a marginally significant trend toward more popular inventories being less energetic.

3.3.4 Nasal vowels

Nasalization is the only property other than length that distinguishes vowels of the same quality in more than a very few languages. Nasal vowel inventories were extracted from UPSID in exactly the same way as the oral inventories: the larger of the long or short nasal inventories in a language represents its nasal inventory. 102 languages in UPSID have at least one nasal vowel. Analyses like those for oral vowels shows that nasal inventories are structured much like oral inventories, except they’re often smaller. Figure 5 shows that, just like oral vowels, many more languages have five nasal vowels than any other number. The figure also shows in a number of ways, however, that languages have fewer nasal than oral vowels: (1) one more language has three or four nasal vowels than has six or seven, six languages have fewer than three nasal vowels, and no languages have more than nine (cf. Figure 1).
Nasal vowels never occur in an inventory without oral vowels, so their presence unequivocally implies the presence of oral vowels. But is the size or composition of a language’s nasal inventory related in any further way to its oral inventory? The answer is “yes” for both size and composition.
Taking up size first, a little over half the languages with nasal vowels, 53 of 102, have fewer nasal than oral vowels in their inventories, and none have more nasal than oral vowels. The languages with more oral than nasal vowels have on average 2-3 more oral than nasal vowels, and some have as many as 6 more.

What vowels are missing in the nasal inventories that are found in the corresponding oral inventories? In two languages, all the oral vowel qualities are missing from the nasal inventories, which each have a single central nasal vowel: Zoque’s oral vowels are /i, e, a, o, u/ and its sole nasal vowel is /\~{y}/, and Cherokee’s oral vowels are /i, e, a, o, u/ and its sole nasal vowel is /\~{a}/. Otherwise, one or more mid nasal vowels are missing in 41 languages (“gutless” inventories), one or more high nasal vowels are missing in 20 languages (“headless”), and one or more low nasal vowels are missing in just 6 languages (“footless”). Senadi exemplifies the gutless type, with oral /i, e, æ, a, Æ, a/, vs nasal /\~{i}, \~{e}, \~{æ}/, Amuzgo is headless, with oral /i, e, æ, a, Æ, o, u/ vs nasal /\~{e}, \~{æ}, \~{a}, \~{o}/, and Chatino shows what the rare footless type is like, with oral /i, e, a, o, u/ vs nasal /\~{i}, \~{e}, \~{o}/. Some languages lack nasal counterparts to their oral vowels at more than one these three height divisions: of the 6 footless languages, 2 are also headless, 1 is also gutless, and 1 is also headless and gutless, while 12 of the 20 headless languages are also gutless. In short, nasalization reduces height contrasts, and it does so most often by eliminating mid vowels.

Why should it do so? The answer lies in the perceptual consequences of acoustically coupling the nasal to the oral cavity. Coupling adds pairs of poles and zeroes to the poles produced in the oral cavity. The lowest nasal pole (N1) and zero (Z1) occur close the lowest oral pole (F1) and change both the center of gravity and the bandwidth of this lowest spectral prominence.

N1 is below F1 when the F1 is high but above it when F1 is low. If N1 is intense, as in a heavily nasalized vowel, it changes the center of gravity of this prominence, lowering the center of gravity when N1 is below F1 but raising it when N1 is above F1. Lowering the center of gravity makes the vowel will sound higher, while raising it makes the vowel sound lower. Z1 is just above N1. When N1 is below F1, Z1 is likely to coincide with F1 and attenuate it. This attenuation lowers the center of gravity of the lowest spectral prominence and makes the vowel sound higher. When both N1 and Z1 are above F1, the center of gravity of the lowest spectral prominence is instead likely to be raised, which should make the vowel sound lower. Headless inventories such as Amuzgo’s may be more common than footless inventories such as Chatino’s because adding N1 and Z1 more often raises than lowers the lowest spectral prominence’s center of gravity.

N1 and Z1 also increase the bandwidth of this lowest spectral prominence, and
this effect may be perceptually more important in a less heavily nasalized vowel than the raising or lowering of this prominence’s center of gravity. A less intense N1 broadens the distribution of energy in this prominence regardless of whether it is above or below F1, and Z1 broadens the prominence’s bandwidth simply be attenuating its energy. A broader bandwidth may make the vowel sound lower, but it also probably obscures a nasal vowel’s height by making it harder to perceive the center of gravity of the lowest spectral prominence. Gutless nasal inventories such as Senadi’s may be most common simply because fine distinctions in height between mid vowels or between mid and high or low vowels may very hard to detect when an bandwidth increase obscures the this prominence’s center of gravity. This prominence’s center of gravity differs enough in the remaining high and low vowels that they’re preserved. Evidence compatible with this interpretation is reported in Wright (1986), where it’s shown that *ceteris paribus* nasal vowels are perceptually closer to one another than their oral counterparts.

Perceptual results reported in Kingston (1991), Kingston & Macmillan (1995), and Macmillan, et al. (1999), add to the explanation of gutless inventories’ greater frequency. Listeners in these studies identified and discriminated vowels in which vowel height and nasalization were manipulated independently. They were more likely to identify a vowel as high when it was more nasalized, and more likely to identify a vowel as oral when it was lower. Listeners were also consistently better at discriminating a higher, more nasalized vowel from a lower, less nasalized one than at discriminating a higher, less nasalized vowel from a lower, more nasalized one.²³ Both results show that height and nasalization integrate perceptually, and their integration disfavors intermediate percepts for both height and nasalization.

These findings predict incorrectly that low nasal vowels should often denasalize because a lower vowel is more likely to be identified as oral. Two factors keep low vowels nasalized. First, the soft palate is actually permitted to lower more in lower than higher nasalized vowels (Clumeck, 1976; Bell-Berti, Baer, Harris, & Niimi, 1979; Al-Bamerni, 1983; Henderson, 1984) and is actively kept high in higher than low vowels by contracting the levator palatini and relaxing the palatoglossus (Moll & Shriner, 1967; Lubker, 1968; Fritzell, 1969; Lubker, Fritzell, & Lindqvist, 1970; Bell-Berti, 1976; Kuehn, Folkins, & Cutting, 1982; Henderson, 1984). Second, low vowels are longer than higher vowels, apparently because the jaw must lower more (Lehiste, 1970; Westbury & Keating, 1980), and even light nasalization is easier to detect when the vowel lasts longer (Whalen & Beddor, 1989; Hajek, 1997). Indeed, nasal vowels of a given height are often longer than the corresponding oral vowels (Whalen & Beddor, 1989). Speakers may deliberately prolong them to ensure that nasalization is detected. Whether a low vowel lasts longer merely because the jaw moves slowly or it’s also deliberately prolonged, its greater duration compensates for their height’s reducing the
perceptibility of nasalization by making whatever nasalization is there easier to detect. Although this compensation appears to be a serendipitous consequence of the jaw’s inertia, it’s apparently robust enough to reliably preserve nasalization in low vowels.

In this case, too, we observe a delicate tradeoff among competing factors, all of them phonetic in this instance: the obscuring of the lowest spectral prominence’s center of gravity by the increase in its bandwidth, the perceptual integration of the vowel height and nasalization, and the compensatory effects of low vowel’s greater duration. We also observe robust differences in priority: the compensatory effects of greater duration ensure that low vowels remain nasalized despite perceptual integration. In the next section, we observe the consequences for the contents of vowel inventories of having to shorten a vowel. Rather than dispersing vowels in terms of height, shortening compresses them upward.

3.3.5 Vowel reduction

In many languages, fewer vowels contrast in unstressed than stressed syllables. The proper characterization of unstressed vowel reduction has raised fundamental questions about how the phonetics influences phonology. At least three proposals can be distinguished in the recent literature, Crosswhite’s (2001, 2004), Barnes’s (2002), and Flemming’s (2001, 2004, submitted). The proposals agree that vowel contrasts are reduced in unstressed syllables because these syllables are shorter than stressed syllables, but they disagree in how unstressed syllables’ shorter duration brings about this reduction.

Crosswhite distinguishes between two kinds of vowel reduction, contrast-enhancing and prominence-reducing. She exemplifies contrast-enhancing reduction with Italian (1a) and immediately pre-tonic syllables in Standard Russian (1b), and prominence-reducing reduction with (Eastern)24 Bulgarian (1c) and all other unstressed syllables in Standard Russian (1d):
(1) Stressed Unstressed
a. Italian i, u i, u
e, o e, o
ɛ, ɔ e, o
a a
b. Russian i, u i, u immediately pre-tonic
e i
o a
a a
c. Bulgarian i, u i, u
e, o i, u
ə ə
a ə
d. Russian i, u i, u all other unstressed syllables
e i
o, a ə

The reduction patterns in Italian and in immediately pre-tonic syllables in Standard Russian are contrast-enhancing because the neutralization of height contrasts involving mid vowels leaves the remaining contrasting vowels farther apart – they might better be called “contrast-dispersing” patterns. The raising of lower mid /ɛ, ɔ/ to higher mid [e, o] in Italian increases the distance between /a/ and the remaining mid vowels /e, o/ and the raising of /ɛ/ to [i] and the lowering of /ɔ/ to [a] in immediately pre-tonic syllables in Standard Russian eliminates the mid vowels altogether, leaving behind only high and low vowels. Crosswhite argues that contrasts are enhanced in this way in unstressed syllables because their short duration makes it hard to maintain small differences in vowel height, particularly for vowels that aren’t at the corners of the vowel space.

The reduction patterns in Bulgarian and in all other unstressed syllables in Standard Russian are prominence-reducing neutralizations because the low vowel is raised, in both cases to a mid central unrounded vowel with the quality of [ə]. Raising lowers the vowel’s F1 and shortens it – higher vowels are shorter than lower ones (Lehiste, 1970). Both these changes reduce the vowel’s overall intensity and presumably its prominence; in doing so, these changes make the resulting vowel more compatible with its prosodically weak position and enhance the phonetic difference between unstressed and stressed syllables. In prominence-reducing reduction, the shorter duration of unstressed syllables is a byproduct of reduction, rather than being its principal motivation as it was in contrasting-enhancing reduction. However, it’s likely that even those vowels whose qualities aren’t affected by reduction are shorter in unstressed than stressed syllables in these languages, so the vowels whose qualities are affected are probably shortened by occurring in unstressed syllables as well as by
reduction itself.

Diagnosing which of the two kinds of reduction has occurred in a particular language depends on how the low vowel behaves, because the high vowels remain unchanged and mid vowels rise in both kinds. When reduction enhances contrasts, the low vowel remains low, but when it instead reduces prominence, the low vowel raises.

Crosswhite’s analyses of both kinds of reduction are explicitly functional: contrast-enhancing reduction maintains only those vowel height contrasts whose members can be reliably distinguished in the short span of an unstressed syllable and prominence-reducing reduction ensures that unstressed syllables aren’t so prominent that they’re mistaken for stressed ones. This functionalism is, moreover, built explicitly into the phonological formalizations of the analyses as a licensing constraint and a stress/vowel quality compatibility scale, respectively.

Neither Barnes (2002) nor Flemming (2001, 2004, submitted) distinguish between two kinds of reduction, and both treat the shortening of unstressed vowels as reduction’s *primum mobile*. Both accounts rely on Lindblom’s (1963) finding that the F1 frequencies of Swedish non-high vowels decrease with vowel duration, i.e. speakers undershoot the non-high vowel height targets when there’s too little time to reach them. Shortening causes undershoot because speakers don’t speed up articulatory movements, particularly of the massive jaw, to reach those targets in the time available. The result is that all the non-high vowels are raised, compressing the vowel space upward from bottom. This outcome resembles Crosswhite’s prominence-reducing reduction but lacks its functional motivation. Indeed, Barnes’s and Flemming’s accounts reverse the *explicanda* and *explicandum*. Instead of speakers’ raising a vowel to lower it F1 frequency, shorten it, and thereby reduce its intensity and prominence, the vowel is raised, more or less automatically, because it’s shortened.

What of contrast-enhancing reduction of the sort Crosswhite ascribes to Italian or immediately pre-tonic syllables in Standard Russian? Both Barnes and Flemming argue that this may not in fact be a distinct kind of reduction. They both cite instrumental studies of Italian (Farnetani & Vayra, 1991; Albano Leoni, Caputo, Cerrato, Cutugno, Maturi, & Savy, 1995) which show that in unstressed syllables the low vowel [a] is realized with a considerably lower F1 frequency, i.e. as a higher vowel, perhaps as [ε]. For Standard Russian, Barnes shows that the low vowel produced by reduction in immediately pre-tonic syllables does not have categorically different quality from that produced in other unstressed syllables. Instead, the reduced vowels in these two kinds of unstressed syllables differ in duration. Those in immediately pre-tonic syllables last 50-120 ms, while those in other unstressed syllables last 20-60 ms. F1’s frequency is
uncorrelated with vowel duration in immediately pre-tonic syllables, while it correlates positively with duration in other unstressed syllables. Most importantly, the long tokens of low vowels in other unstressed syllables whose durations overlap with the duration range of the immediately pre-tonic syllables have F1 frequencies which are just as high as those in immediately pre-tonic syllables. These results suggest that the low vowels in immediately pre-tonic syllables last long enough that speakers have the time needed to lower the jaw and tongue and raise the F1’s frequency to a value that sounds like [a]. They also have enough time to produce a vowel with this quality when the low vowels in other unstressed syllables last at least as long as the shortest immediately pre-tonic syllables. The majority of tokens in other unstressed syllables are shorter than this minimum, however, and speakers don’t have enough time to lower tongue and jaw as much. As a result, the high target frequency for the F1 are undershot to the extent that the vowel’s duration falls short of the minimum time needed. Thus, in both Italian and Standard Russian, the phonetic evidence indicates that the low vowel is raised when reduced, as in so-called prominence-reducing reduction, rather than remaining low as would be expected if reduction were contrast-enhancing. In other words, there may be only one kind of vowel reduction after all, in which vowels are raised when articulations are undershot as result of shortening.

At this point, Barnes and Flemming part ways. Barnes argues that the undershooting of articulations that occurs when a vowel is shortened in an unstressed syllable can be phonologized, in the form of categorical alternations between vowel qualities produced in stressed syllables and the higher vowel qualities heard in unstressed ones. Once phonologization occurs, reduction is no longer governed by the phonetic constraints that originally motivated it, and the vowels that participate in the alternation may freely undergo further sound changes. In making these claims, Barnes agrees with the proposal developed at great length in Blevins (2004) that sound change is phonetically motivated but subsequent patterning of the affected sounds in synchronic phonologies is no longer governed by the phonetic factors that originally brought about the sound change. As I evaluate this proposal in some detail in §3.4.4 below, I won’t say anything more about Barnes’s account here.

Unlike Barnes, Flemming does build the phonetic motivation for vowel reduction into its phonology. Flemming’s account differs fundamentally from Crosswhite’s in using constraints on what contrasts may occur in a language in place of licensing or markedness constraints limiting the circumstances in which individual segments may occur. Two kinds of constraints regulate contrasts: the first requires that contrasting sounds be some minimal distance apart within the phonetic space occupied by the sounds (MinDist=n) and the second requires that the number of contrasts be maximized (MaxCon). These constraints obviously conflict with one another, as requiring that contrasting sounds be far apart limits the number of possible contrasts,
while requiring a large number of contrasts forces contrasting sounds close together. This conflict is resolved as is usual in optimality theory by ranking the contrast maximization constraint relative to the minimal distance constraints.

Imagine the range of vowel heights is divided into seven, equally spaced heights: (1) high [i, u], (2) lowered high [ɪ, ʊ], (3) raised mid [ɛ, ə], (4) mid [e, o], (5) lowered mid [ɛ, ə], (6) raised low [æ, ɛ, ɘ], and (6) low [a, a, ɘ]. A language with the seven vowels /i, u, ɛ, ə, e, o, a/ found in Italian stressed syllables would then obey the minimal distance constraint requiring that contrasting vowels differ by at least two steps (MINDIST=2). This constraint is ranked in Italian immediately above the constraint requiring that the number of contrasts be maximized. MAXCON is a positive requirement rather than a prohibition, and a ✗ is listed in the tableau for each contrasting category:

<table>
<thead>
<tr>
<th></th>
<th>MINDIST=1</th>
<th>MINDIST=2</th>
<th>MAXCON</th>
<th>MINDIST=3</th>
<th>MINDIST=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) i:a</td>
<td>✓</td>
<td>✗</td>
<td><em>!(i:e)</em>(e:a)</td>
<td><em>!(i:e)</em>(e:a)</td>
<td><em>!(i:e)</em>(e:a)</td>
</tr>
<tr>
<td>(b) i:e:a</td>
<td>✓</td>
<td>✓</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
</tr>
<tr>
<td>(c) i:e:e:a</td>
<td>✓</td>
<td>✓</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
</tr>
<tr>
<td>(d) i:e:e:a</td>
<td>✓</td>
<td>✓</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
</tr>
</tbody>
</table>

How does this approach produce the vowel inventory in Italian’s unstressed syllables, /i, u, ɛ, ə, ɘ/? In his 2004 paper, Flemming adds a constraint prohibiting short low vowels (*SHORTLOW) and ranks it above MINDIST=2:

<table>
<thead>
<tr>
<th></th>
<th>*SHORTLOW</th>
<th>MINDIST=2</th>
<th>MAXCON</th>
<th>MINDIST=3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) i:ə:ɘ</td>
<td>✓</td>
<td>✓</td>
<td><em>!(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
</tr>
<tr>
<td>(b) i:ə:e:a</td>
<td>✓</td>
<td>✓</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
</tr>
<tr>
<td>(c) i:ə:ə:e</td>
<td>✓</td>
<td>✓</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
<td><em>(i:e)</em>(e:e)*(e:a)</td>
</tr>
</tbody>
</table>

The constraint prohibiting short low vowels rules out the candidate with [a] (3b), while that requiring that the minimal distance between contrasting categories be at least 2 steps rules out the candidate which retains [ɘ] (3c), because the raised low vowel [ɘ] is only one step away from it. What’s left as the optimal candidate has only three heights and a raised low vowel [ɘ] (aa).

Flemming (submitted) lays out a more explicitly quantitative account of vowel reduction. Rather than requiring that all contrasting vowels be at least some minimum
distance apart, the minimum distance between any pair of vowels now must be maximized. This approach closely resembles the one spelled out in §3.3.3 above, except now an inventory’s dispersion energy is measured by its closest pair of vowels, and the distance between them is the measure of the inventory’s goodness. This requirement also has the effect of spacing vowels evenly in the vowel space because the distance be one pair of vowels can only be increased so long as doing so doesn’t force some other pair of vowels closer together than the pair being adjusted. As in Flemming’s (2004) account, unstressed and stressed vowels are subject to the same distance requirements, so the minimum distance is the smallest distance found in either inventory. The new account retains the requirement that the number of contrasts be maximized.

Following Lindblom (1963), the actual formant frequencies of the vowels in the full and reduced inventories are predicted with functions that undershot their target frequencies as exponential functions of the vowel’s duration. With the appropriate settings of parameters, these functions correctly predict that height contrasts neutralize in unstressed vowels and the raising of the unstressed low vowel relative to its stressed counterpart. The parameter settings also ensures that F2 is undershot less than F1, which coincides with observed patterns of vowel reduction. Reduction doesn’t neutralize front-back and/or unrounded-rounded contrasts except in the most extreme cases, where it neutralizes all vowel contrasts, as, for example, in English where all vowels reduce to [a]. Flemming observes that front-back and unrounded-rounded contrasts are largely preserved in reduced vowels because consonants coarticulate with the tongue backness and lip rounding of adjacent vowels, even if to different degrees depending on the consonant’s place of articulation. Height contrasts, however, aren’t preserved because a consonant’s degree of constriction is rarely if ever affected by the height of adjacent vowels. Instead vowels coarticulate with the close constriction of the consonant and when too short undershoot their targets and reduce to higher vowel qualities.

Flemming’s new account invokes the phonetic motivation for and mechanisms of vowel reduction far more directly that his earlier account did. Though he now implements minimum distance and maximize contrast constraints with continuous rather than discrete mathematics, he does not, however, dispense with ranking them. Their ranking is expressed by varying their relative weights. Even so, the only explicitly phonological aspect of this account is its outcome: fewer vowels contrast in the reduced inventories. But that’s an emergent property, not one that’s in any way preordained by obedience to some phonological constraint.

3.3.6 The phonological consequences of vowel reduction vs nasalization
As noted at the end of §3.3.4, the effect of vowel reduction on height is quite different from that of nasalization: reduction raises vowels by compressing them upward, while nasalization either disperses vowels toward the top and bottom of the vowel space or more rarely lowers them. Their effects differ because the speaker’s goals are different.

The goal in pronouncing an unstressed vowel is to produce one that’s shorter than a stressed vowel. Meeting this goal leads to the vowel’s target being undershot. Undershooting is greater for a vowel whose articulation differs more in degree of constriction from the flanking consonants, i.e. the more open articulations of lower vowels are undershot more and they end up raised.

Shortening is no part of the goal in pronouncing a nasalized vowel, quite the reverse. The speaker’s goal is instead to convey the vowel’s height and nasalization, which demands that the vowel last long enough for the spectral modification caused by nasalization to be detected. Detecting this modification may even require speakers to prolong the vowel (Whalen & Beddor, 1989). Sometimes, the speaker succeeds in conveying both height and nasalization, but other times nasalization changes the center of gravity and bandwidth of the vowel’s lowest spectral prominence enough to alter and/or obscure its height. These effects are exacerbated by the perceptual integration of vowel height and nasalization, which disfavors percepts of intermediate height and nasalization. Mid vowels disappear as a result. Low nasal vowels, which integration would otherwise denasalize, remain nasal because they last long enough for listeners to detect whatever nasalization may be present and because the speaker actually lowers the soft palate more in low than high vowels.

Vowel nasalization’s dependence on the vowel’s duration has phonological consequences, some of which interact with reduction. Hajek (1997) shows that contrastive vowel nasalization only develops on long vowels in Northern Italian dialects; he also cites Hombert’s (1987) observation that a vowel must be long to become contrastively nasalized in Teke. Similarly, in Copala Trique (Hollenbach, 1977) and Guarani (Gregores & Suárez, 1967), vowels may only contrast for nasalization in stressed syllables, presumably because they’re longer. Finally, in Brazilian Portuguese, the contrast between higher mid /e/, o/ and lower mid /e, o/ reduces categorically to [e, o] in unstressed syllables (Major, 1985). The resulting mid vowels then reduce further to [i, u] both pre- and post-tonically in casual or exceptionally rapid speech, both circumstances in which they’re likely to be quite short. Under these circumstances, the mid nasal vowels /ê, ô/ only raise to [i, ü] post-tonically. They resist raising pre-tonically because that’s a position where vowels tend to remain longer, and because they are longer by virtue of being nasalized.
Despite these differences between reduction and nasalization in their effects on vowel height, both reduced and nasal vowels are still dispersed within the spaces available to them. Both spaces have shorter distances between the top and bottom than in stressed or oral vowels, but this distance is shortened in the reduced vowel space by raising the bottom toward the top, while it’s shortened in the nasal vowel space by lowering the top as well as raising the bottom.

3.3.7 Postscript

In the preceding sections, I have developed incomplete explanations of the patterns commonly observed in inventories of pulmonic and glottalic stops, and of oral, nasal, and reduced vowels. For these subsets of an inventory, the first goal was to establish just what the facts were that needed to be explained. Only once that was accomplished was any attempt made to explain them. In each case, the explanations appeal to multiple forces, some phonetic, others phonological, and in each case, the explanations would have been even more incomplete than all of these still are if a contributing force had been excluded by an overly fastidious concern to keep the phonetics out of the phonology or vice versa. The approach to explaining sound change in the next section is no different.

3.4 Explaining sound change
3.4.1 Introduction

Many perhaps most of the sound changes languages undergo are phonetically motivated. What does this mean? It means that something about how the speaker pronounces the sound that changes, how that sound is transmitted to the listener, or how the listener perceives that sound eventually leads to that sound being phonologically different at some later time in its history than it was at some earlier time. It does not mean that the sound’s pronunciation, transmission, or perception caused it to change, if what we mean by “cause” is that pronouncing, transmitting, or perceiving the sound in that way will always result in that sound change. Phonetically motivated sound change is not that deterministic. Instead a sound’s pronunciation, transmission, or perception makes it’s possible or perhaps even likely that it will undergo that sound change, if two other, essentially non-phonetic conditions are also met.

To understand the how these two other conditions make a sound change actually happen, it’s necessary to identify just what it is about a sound’s pronunciation, transmission, or perception that makes sound change possible in the first place. It is phonetic variation. At one time, the sound may coarticulate more with a neighboring sound, it may be transmitted against a noisier background, or it may produce different perceptual effects than at another time. This phonetic variation is the material needed
to initiate and eventually complete the sound change. Without it, there’d be neither the reason nor the means for a language to change.

The first of the non-phonetic conditions is social: the particular variants that point to the sound change must be taken up by and eventually spread through and be exaggerated by the speech community. No matter how phonetically likely the sound change is, it will never be “actuated” if the speech community doesn’t come to prefer these variants (Labov, 2001). The second non-phonetic condition is statistical: these variants must also spread to other instances of a word and from one word to another. They may do so if speakers choose to hypo- or hyper-articulate consistently in particular circumstances (Lindblom, 1990; Moon & Lindblom 1994, 2003), or if they consistently choose distinct allophones in different segmental or prosodic contexts (Kingston & Diehl, 1994; Guenther, et al., 1999; Fougeron & Keating, 1997; Keating, Cho, Fougeron, & Hsu, 1999; Cho & Keating, 2001). This generalization or diffusion through the lexicon is equally necessary if the sound change is to be fully implemented (Cheng & Wang, 1977; Wang & Cheng, 1977; Bybee, 2001).

I describe tonogenesis in Athabaskan as an example of a phonetically motivated sound change in the next section. Discussing this case is also an occasion in the following section to evaluate Steriade’s (1999) proposal that contrasts are maintained in those contexts where their phonetic correlates or cues are easy to detect and neutralized in contexts where their cues hard to detect. This approach to explaining where contrasts are maintained vs neutralized is called “licensing by cue”; it was introduced without discussion in the presentation of Crosswhite’s account of contrast-enhancing reduction in §3.3.5 above. I conclude with a critique of Blevins’s (2004) argument that once a phonetically motivated sound change is phonologized, the affected sounds are no longer regulated by the phonetic properties that originally motivated the sound change.

3.4.2 The phonetics of Athabaskan tonogenesis

In Proto-Athabaskan, glottalic and non-glottalic stops, affricates, nasals, and glides contrasted at the ends as well as the beginnings of stems (Krauss, in press). These contrasts are maintained at the ends of stems in conservative languages spoken in Alaska and along the Pacific Coast in northern California and Oregon, but in stems ending in stops and affricates (henceforth just “stops”), they have been replaced by tone contrasts on the preceding vowel in most of the rest of the family – a few peripheral Alaskan languages have lost to stem-final contrast between glottalic and non-glottalic stops without replacing it with a tone contrast.

The development of tone from an earlier contrast in laryngeal articulations of an
adjacent consonant is an extremely common sound change (Hombert, Ohala, & Ewan, 1979), particularly in the language families of East and Southeast Asia. It can occur because one of the phonetic correlates of a laryngeal contrast in consonants is differences in the fundamental frequency (F0) of adjacent vowel. These F0 differences become tone contrasts in the vowels and replace the original laryngeal contrast between the consonants when the other phonetic correlates of the laryngeal contrast are lost from the consonants.

Explaining tonogenesis in Athabaskan is complicated by three factors. The first was hinted at in the narrow reference to stops in the next to last paragraph. Regardless of whether they develop tone, nearly all Athabaskan languages maintain the contrast between stem-final glottalic and non-glottalic sonorants and between stems ending in glottal stop vs a vowel. Even so, the same tone appears in stems ending in glottalic sonorants and glottal stop as in stems that once ended in glottalic stops, and the other tone appears in stems ending in non-glottalic sonorants or a vowel as in stems that once ended in non-glottalic stops. Because the contrast between glottalic and non-glottalic consonants has only been lost in stem-final stops, tone is only contrastive in such stems, while in stems ending in sonorants, glottal stop, or a vowel, the F0 differences remain predictable from other properties of the end of the stem and tone is redundant rather than contrastive.

The second complication is that the tone which developed in stems that ended in glottalic stops in the protolanguage only did so when the stem vowel was short (“reduced” in Athabaskanist terminology). When the vowel was long (“full”), the tone that otherwise developed in stems that ended in non-glottalic stops appeared instead. However, if the stem-final stop was spirantized, the same tone developed on long vowels as on short vowels before a stem-final stop. Similarly, this tone also develops on long as well as short vowels in stems ending in glottalic sonorants. In other words, vowel length doesn’t matter if the stem-final consonant is not a stop. There is no contrast between a short and long vowel before stem-final glottal stop; the historical developments of the vowels in such stems uniformly indicate that the vowel is long. Nonetheless, the same tone appears in these stems as in those in which a short vowel preceded a glottalic stop in the protolanguage or those ending in a glottalic sonorant.

Kingston (1985, in press) accounts for both complications by invoking a difference in the relative timing of laryngeal and oral articulations in stops vs other manners of articulation. The relative timing of the two articulations differs because a stop closure’s release is acoustically a very different event than its onset. A brief but intense burst of noise occurs when the stop is released as the air that had been trapped behind the stop closure quickly rushes out. No comparable acoustic event, certainly none with such potential for salience, occurs at the onset of the stop closure. Nor does
anything comparable occur at the onsets or releases of fricatives and sonorants; in these manners of articulation, these events are acoustically mere mirror images of one another. The stop burst is salient enough that the stop’s laryngeal articulation is timed to coincide with it. This timing allows the laryngeal articulation to modify the burst itself in ways that are specific to that laryngeal articulation and thus convey its nature to the listener (Kingston, 1985, 1990). Because no similarly salient acoustic event occurs at either the onset or release of the oral constriction in fricatives or sonorants, the timing of laryngeal articulations relative to oral ones is freer in these manners of articulation. In many languages, however, the laryngeal articulation is timed to coincide with the onset of the oral constriction in sonorants and fricatives (Kingston, 1985, 1990).

If the laryngeal articulation coincides with the release of the oral constriction in a stop, then it is farther from a preceding vowel than it would be in a fricative or sonorant, at least when the laryngeal articulation coincides with the onset of the constriction for the latter manners of articulation. Because the laryngeal articulation is farther away, it coarticulates less with the preceding vowel before a stop. The tone developments in Athabaskan suggest that coarticulation of the vowel with the following stop was still extensive enough to alter the pronunciation of a short preceding vowel but not a long one. The laryngeal articulation is closer to the preceding vowel before a fricative or sonorant, and more of the vowel would be altered by coarticulating with the consonant, enough apparently to change a long as well as a short vowel. The timing difference also explains why tone replaced the glottalic:non-glottalic contrast in stem-final stops but merely supplements it in stem-final sonorants. If the stop weren’t released in some contexts or the release were inaudible, the principal cue to the consonant’s identity would become the acoustic effects of its coarticulation with the preceding vowel. If the release became absent or inaudible in the majority of stop pronunciations, the contrast would effectively have shifted from the consonant to the vowel. The absence or inaudibility of the release in a sonorant would be of little consequence for conveying its laryngeal articulation, particularly if that articulation coincides with the onset of the oral constriction. The principal correlates of that articulation are in fact already its coarticulatory effects on the preceding vowel, so there’s little reason to expect them to change or shift off the consonant.

So far the vital role of the listener has remained only implicit in this account (Ohala, 1981). For the contrast to shift to the vowel from the consonant, the listener has to misinterpret the coarticulatory effects of the consonant’s laryngeal articulation as the speaker intending to alter the vowel. The listener may be inclined to do so if other evidence that these effects are properties of the consonant is weak or missing, as when the stop is unreleased or the release is inaudible. Whether the listener misinterprets the effects of coarticulation in this way obviously depends on how often this other evidence is weak or missing, i.e. on speakers’ choices about which variants get used most.
The third complication is perhaps the most intriguing: some of the present-day daughter languages have high tones in stems that end in a glottalic consonant and low tones elsewhere, while others have low tones in such stems and high tones elsewhere. One might be tempted to treat one of these developments as original and the other as a reversal, but in Kingston (in press) I show that it’s actually possible to get both high and low tone directly from different pronunciations of the glottalic consonants. Glottalic consonants are distinguished from non-glottalic ones by a constriction of the glottis that is tight enough to curtail or even cut off air flow through the glottis. The glottis is closed by contracting the interarytenoid and lateral cricoarytenoid muscles while relaxing the posterior cricoarytenoid muscles, and the constriction is tightened by the forceful contraction of the thyroarytenoid muscles, which stiffens the inner bodies of the vocal folds and causes them to press firmly against one another. If this is all the speaker does, the voice quality of adjacent vowels will be creaky and its F0 will be low because the folds’ outer covers remain slack. However, if the speaker also contracts the cricothyroid muscle at the same time, the folds’ outer covers will be stretched and the voice quality in the adjacent vowel will be tense and its F0 high instead.

What determines the speaker’s choice about how to pronounce glottalic consonants, i.e. whether to contract the cricothyroid as well as the thyroarytenoid muscles? In Kingston (1982, 1985), I presented data from Tigrinya and Quiché suggesting that this might be a difference between hyper- vs hypo-articulated pronunciations. Ejectives (= glottalic stops) that raised F0 substantially in adjacent vowels appeared to be hyper-articulated in that their bursts were intense and voice onset was delayed for a long time following the stop release. The glottis remained tightly constricted remained tightly closed during that delay. Ejectives that lowered F0 appeared instead to be hypo-articulated in that their bursts were weak and voice onset was delayed only briefly following the stop release. Once voicing began, its quality following the hypo-articulated ejectives was also noticeably creaky, while that following the hyper-articulated ones was modal or even tense. Since then, data from two other languages, Dakelh (Bird, 2002) and Witsuwit’en (Wright, Hargus, & Davis, 2002), have been reported which show that whether an ejective raises or lowers F0 in adjacent vowels is not predictable from the ejective’s other phonetic properties nor from the extent to which it’s otherwise hyper- or hypo-articulated. These crosslinguistic differences suggest that speakers decide to raise or lower F0 independently of how they otherwise choose to pronounce ejectives.

This account of tonogenesis in Athabaskan relies on the speakers’ choices and listeners’ mistake. The choices speakers made at key points in the history of this sound change are: (1) whether to contract the cricothyroid as well as the thyroarytenoid muscles in pronouncing glottalic consonants, a choice which determines whether high or low tone eventually develops on preceding vowels, (2) not to release stem-final stops.
or to release them inaudibly such that the only the F0 and voice quality of the preceding vowel are reliable correlates of the stops’ laryngeal articulations, and (3) to time the laryngeal articulations of sonorants and fricatives so that they coincide with the onset of the oral constriction, and thus noticeably alter the F0 and voice quality in long as well as short preceding vowels. None of these choices are obligatory, even if they are more typical than alternative choices. The mistake listeners’ make is to interpret the coarticulatory effects of the consonant’s laryngeal articulation on the preceding vowel’s F0 and voice quality as intentional. This mistake is encouraged because speakers fail to release stops or do so inaudibly and because laryngeal articulations are timed to coincide with the onset of the oral constriction in fricatives and sonorants.

3.4.3 Licensing by cue

Let us now use tonogenesis in Athabaskan to evaluate Steriade’s (1999) proposal that contrasts are licensed by cue. The essence of this proposal is that contrasts are maintained in contexts where the cues to their identity are robustly conveyed and neutralized where those cues are reduced, obscured, or absent. The proposal is intended to account for synchronic alternations, e.g. such cases as Lithuanian or Klamath where obstruents contrast for laryngeal articulations – voicing in Lithuanian and aspiration and glottalization in Klamath – before sonorants and neutralize it elsewhere, but it should also account for the sound changes that bring about these alternations. In both Lithuanian and Klamath, laryngeal contrasts in obstruents are maintained before sonorants because cues to those articulations in the consonant release and in the transition to the following sonorant are robust in that context. They are neutralized before other obstruents and word-finally because these cues are absent and those in the transitions from preceding vowels are less robust.

The Athabaskan case appears at first to be quite similar: the contrast between glottalic and non-glottalic stops is maintained at the beginning of stems where the consonants precede a vowel and neutralized, at least as a consonantal contrast, at the ends of the stems, which are often word-final or not followed by a sonorant. Stem-final stops may even have been unreleased when the contrast was lost from them, inducing the listener at that time to think the consonants themselves weren’t different. Moreover, this contrast is maintained in stem-final sonorants because its cues are timed to occur early enough that they are robustly signalled during the transition from the preceding vowel.

The difficulty for the licensing by cue account is that the glottalic:non-glottalic contrast didn’t in fact neutralize in stem-final stops, but instead shifted to a tonal contrast on preceding vowels. How were speakers of Athabaskan languages able to keep morphemes distinct whose stem-final consonants once contrasted in their
laryngeal articulations while speakers of Lithuanian or Klamath failed to do so? If we assume that the phonetic correlates available to act as cues to a particular laryngeal contrast are substantially the same in all languages where that contrast is found, then Lithuanian and Klamath speakers and listeners had at their disposal more or less the same phonetic materials to convey laryngeal contrasts, among them differences in F0 and voice quality on preceding vowels, as Athabaskan speakers. Yet they failed to use them. The solution to this conundrum lies in the idea that speaker’s choose how they are going to pronounce a contrast. That is, they choose which of the available phonetic materials they’re going to use. Licensing-by-cue falls short because it conceives the phonetics as something that happens to speakers, rather than also conceiving them as actively manipulating the phonetics to meet their communicative needs. Contrasts are certainly more robustly signaled in some contexts that others, as laryngeal contrasts in stops are before sonorants, but the phonetic materials are available for speakers to use to increase the robustness with which they’re signaled in other, ostensibly less favorable contexts. The research question then becomes: why do speakers choose to do so in some languages but not others? I strongly suspect that the answer to this question will turn out to be that speakers make this choice when the contrast is lexically or morphological informative and not otherwise. I cannot confirm this suspicion now, but state it here as a challenge to myself (and others). (For further critical evaluation of licensing by cue see Gerfen, 2001; Kingston, 2002.)

3.4.4 Evolutionary phonology
3.4.4.1 Introduction

In Evolutionary Phonology (2004), Blevins argues that many perhaps even most sound changes are phonetically motivated, but that once they’re phonologized, the affected sounds are no longer governed by the sound change’s original phonetic motivation. In particular cases, the outputs of the sound change may appear still to be phonetically motivated, but that’s an illusion prompted by their having undergone no further sound changes in those cases. As noted in §3.3.5 above, Barnes (2002) also promotes severing the outputs of sound change from their phonetic motivation once the sound change has been phonologized. If this is correct, then the alternations and phonotactic restrictions in languages’ synchronic phonologies are arbitrary statements of sound distributions, which are encoded in the representations of and paradigmatic relations between morphemes and words.

Blevins also sweeps out of synchronic phonologies nearly all vestiges of innate constraints on phonological structures and processes, leaving in place only the distinctive features and the constituents of prosodic representations. For Blevins, an important consequence of this purge from synchronic phonology of phonetic motivation and innate constraints is that phonological structures and processes can no longer
optimizing or teleological, not even in the form of optimality theory’s competing optimizations or teleologies.

In this section, I show that the thoroughness of this cleansing is unwarranted, first by presenting evidence that a sound change’s phonetic motivation can remain active in the resulting language’s synchronic phonology after it’s been phonologized, second by showing that restrictions on sounds’ distributions are psychologically and even neurophysiologically active and also distinguishable from statistical generalizations across the lexicon, third by describing a sound change that is optimizing, and finally by demonstrating that language learning is determined by innate predispositions after all.

3.4.4.2 Phonetic persistence

The example of tonogenesis in Athabaskan described in §3.4.2 above is clear evidence that a sound change’s phonetic motivation remains active even after the sound change has been phonologized. Three characteristics of this sound change are relevant to this argument. First, in all the tonal Athabaskan languages, the glottalic:non-glottalic contrast has been lost from stem-final stops and replaced by a tone contrast in preceding vowels. Its replacement is the phonologization of this sound change. Second, distinct tones also develop on the vowels preceding glottalic vs non-glottalic stem-final sonorants and in stems ending in glottal stops vs a vowel, but these tones remain redundant because these stem-final contrasts have not neutralized. Third, in some tonal Athabaskan languages, a low tone developed in stems ending in glottalic consonants and a high tone elsewhere, while in others, the opposite tones developed in the two kinds of stems. Both developments are phonetically possible because the glottis can be constricted tightly by contracting the thyroarytenoid muscles alone, producing creaky voice, low F0, and eventually low tone in preceding vowels, or by also contracting the cricothyroid muscles, producing tense voice, high F0, and eventually high tone in preceding vowels instead.

Now, a property that was glossed over in the account of this sound change above becomes of great importance: in the synchronic phonologies of the tonal Athabaskan languages, the tones that appear in stems ending in glottalic sonorants and glottal stop are always the same as those appear in stems that ended in glottalic stops in the protolanguage, modulo the effects of vowel length on tone development in stop-final stems. If it were once possible to constrict the glottis in such a way as to either lower or raise F0, then it should still be possible to do so. It should therefore also be possible in the subsequent history of a tonal Athabaskan language for its speakers to adopt the pronunciation of glottal constriction that has the opposite effect on F0 and tone in the preceding vowel. The result would be that stems which originally ended in glottalic stops in the protolanguage would have one tone, while those that end today in
glottalic sonorants or glottal stop would have the opposite tone. This has never happened. It hasn’t happened because when the sound change was phonologized, the phonetics of the pronunciation of glottal constriction were, too.\textsuperscript{32} Doing so has constrained glottalic sonorants and glottal stop to be pronounced in the same way throughout the subsequent history of each tonal Athabaskan language as its own glottalic stops were when the sound change was actuated.

3.4.4.3 Active synchronic constraints

In this section, I turn to the evidence that the constraints proposed to account for synchronic sound patterns are psychologically and even neurophysiologically active. This evidence shows that they are also not mere statistical generalizations across the lexicon. These constraints influence the on-line categorization of sounds and the syllabification of segment strings.

Moreton (2002) presented listeners with two sets of stop-sonorant-vowel stimuli. The two sets differed in only the stop, which ranged incrementally from [d] to [g] in the first set and from [g] to [b] in the second. In both sets, the sonorant ranged incrementally between [l] and [w]. Listeners identified the members of the first set as beginning with “gl”, “gw”, “dl”, or “dw” and the members of the second set as beginning with “gl”, “gw”, “bl”, or “bw”. One of the responses to the first, “dl”, is an onset that’s phonotactically prohibited in English, while one of the responses to the second, “bw” is an onset that is statistically very rare but perhaps not prohibited by any constraint. Moreton showed that both [dl] and [bw] have a zero frequency of occurrence as onsets in the 18.5 million words in the London-Lund corpus of written and spoken British English. On statistical grounds alone, then, “dl” and “bw” responses should be equally disfavored. The results were quite different. In responses to the first stimulus set, if listeners identified the stop as “d” they were more than three times less likely to identify the sonorant as “l” than if they identified to stop as “g”. In responses to the second stimulus set, they were actually more than one and a half times more likely to identify the sonorant as “w” if they identified the stop as “b” rather than “g”. These results show that it’s possible to distinguish a zero that’s the result of a phonotactic prohibition from one that’s the consequence of an accidental gap. In a followup experiment in which a vowel was inserted before the stop-sonorant-vowel string in the first stimulus set, the bias against “l” when the stop was identified as “d” disappeared. This result shows that the difference between [dl] and [bw] obtained in the first experiment is not merely a perceptual interaction between the two segments but a consequence of their syllabification. The sequence [d.l] is perfectly acceptable in English if the two segments aren’t both in the onset.

Hallé, Segui, Frauenfelder, & Meunier (1998) report comparable results in
experiments with French listeners, whose language also prohibits [dl] and [tl] in onsets. These listeners frequently mistook [d] and [t] for [g] and [k] before [l]. Hallé, Best, and Bachrach (2003) found that French listeners are much worse at discriminating [d] from [g] and [t] from [k] before [l] than Hebrew listeners, whose language permits [dl] and [tl] in onsets.

Other phonotactic prohibitions have also been shown to influence on-line phoneme categorization (Moreton, 1999; Moreton & Amano, 1999; Coetzee, 2004, in press.) In all these cases, it’s possible to debate the nature of the phonotactic constraint that determines listeners’ responses, but the results of these experiments leave little doubt that their responses are directly governed by quite specific constraints on how sounds may legally combine and not on statistical generalizations across the lexicon. The constraints are specific both to sequences they prohibit and the languages in which they apply.

Results reported in Kirk (2001) show that on-line syllabification is also directly influenced by phonological knowledge, of language-specific allophony, phonotactic constraints, and general constraints on the affiliation of segments to syllables. Kirk used the word-spotting task, a procedure in which positive trials consist of strings of sounds that aren’t themselves words, but which contain words, e.g. the stimulus in a positive trial might be [vukwɔjn], which contains the word wine [wɔjn]. The remainder of the string in such trials, here [vuk], is also not a word. In negative trials, no contiguous subset of the string’s segments is a word. Listeners must determine as quickly and as accurately as possible whether the string contains a word, and if so say it aloud. This task is a fundamentally more naturalistic probe of how listeners parse the signal than either the syllable monitoring task (Mehler, Dommergues, Frauenfelder, & Segui, 1981; Cutler, Mehler, Norris, & Segui, 1986) or having subjects divide words into syllables (Treiman, 1983; Treiman & Danis, 1988; Treiman & Zukowski, 1990) because it involves finding words in longer strings, something listeners do ordinarily when listening to speech.

In demonstrating that English allophony influences on-line syllabification, Kirk presented listeners with two variants of stimuli like the example, one in which the voiceless stop before the word was aspirated and the other in which it was unaspirated, [vukʰwɔjn] vs [vukwɔjn]. Because the embedded word always began with a non-nasal sonorant, English phonotactics permit listeners to syllabify the [kʰ] or [k] with the following segment. Kirk predicted that they would be more likely to syllabify the aspirated [kʰ] with the following sonorant than the unaspirated [k], because voiceless stops in English are consistently aspirated only at the beginnings of syllables. If [kʰ] is syllabified with the sonorant and [k] is not, and syllabification determines where listeners think words begin, then wine should be harder to spot after [kʰ] than [k]. As predicted,
listeners detected wine much slower and less accurately when the preceding voiceless stop was aspirated than unaspirated.

To determine whether phonotactics influenced syllabification independently of allophonics, Kirk pitted the two against one another by comparing the speed and accuracy with which listeners detected a word beginning with [l], e.g. lunch, following aspirated coronal vs non-coronal stops, in such strings as [vit\textsuperscript{th}l\textsuperscript{\~n}t\textsuperscript{\~n}] vs [vik\textsuperscript{th}l\textsuperscript{\~n}t\textsuperscript{\~n}]. On the one hand, the phonotactics prohibit a coronal stop from appearing in an onset before [l], which predicted that lunch should be detected faster and more accurately after [\textsuperscript{th}] than [k\textsuperscript{h}] because the strings would be syllabified differently, [vit\textsuperscript{th}l\textsuperscript{\~n}t\textsuperscript{\~n}] vs [vi.k\textsuperscript{h}l\textsuperscript{\~n}t\textsuperscript{\~n}]. The preceding stops’ aspiration, on the other hand, indicates that they’re in the syllable onset, which predicts that lunch should be detected equally slowly and inaccurately after [k\textsuperscript{h}] as well as [\textsuperscript{th}], because the strings would be syllabified identically, [vi.t\textsuperscript{h}l\textsuperscript{\~n}t\textsuperscript{\~n}] and [vi.k\textsuperscript{h}l\textsuperscript{\~n}t\textsuperscript{\~n}]. The findings showed clearly that phonotactics take priority over allophonics in determining syllabification: listeners detected lunch significantly faster and more accurately after [\textsuperscript{th}] than [k\textsuperscript{h}].

Kirk’s remaining experiments showed that listeners first parse segments into syllables on their way to recognizing words even when neither allophony nor phonotactics dictates a particular syllabification. These experiments show that listeners use constraints requiring that onsets be maximized and that segments affiliate with an adjacent stressed syllable to group segments exhaustively into syllables. Onset maximization was tested by comparing the speed and accuracy with which listeners detected words such as smell vs mad in the strings [vusm\textsuperscript{\~e}l] and [vusm\textsuperscript{\~e}d]. As predicted if listeners maximize onsets, mad was detected slower and less accurately than smell. To ensure that this result reflected onset maximization and not some difference between the targets, Kirk ran another experiment in which listeners had to detect words beginning with [+voice] stops such as bag in the strings [zeb\textsuperscript{\~e}g] vs [zesb\textsuperscript{\~e}g]. As [+voice] stops are typically pronounced voiceless unaspirated at the beginnings of words in English, a preceding [s] could syllabify with them without violating the language’s phonotactics. Again, the results showed that listeners maximize onsets: bag was detected slower and less accurately in [zesb\textsuperscript{\~e}g] than [zeb\textsuperscript{\~e}g].

In the experiments testing the effects of attraction to stress of single intervocalic consonants, a words such as east would be presented in the strings [gw\textsuperscript{\~e}v\textsuperscript{\~i}st] vs [gw\textsuperscript{\~e}v\textsuperscript{\~i}st]. As predicted if [v] is attracted stress, east was detected slower and less accurately in [gw\textsuperscript{\~e}v\textsuperscript{\~i}st] than [gw\textsuperscript{\~e}v\textsuperscript{\~i}st]. The same pattern of results was obtained when the initial stressed syllable contained a tense rather than a lax vowel, i.e. [gw\textsuperscript{\~e}v\textsuperscript{\~i}st] rather than [gw\textsuperscript{\~e}v\textsuperscript{\~i}st]. This finding shows that it was stress rather than the prohibition against lax vowels in open syllables that attracted the consonant into the first syllable in such strings. This result also shows that attraction to stress, i.e. weight to stress, takes priority over a constraint requiring that syllables have an onset.
Using quite different techniques, Dupoux and his colleagues present equally compelling evidence that native language phonotactics determine on-line syllabification, even to the point that they cause listeners to hear segments that aren’t present in the signal (Dupoux, Kakahi, Hirose, Pallier, & Mehler, 1999; Dahaene-Lambertz, Dupoux, & Gout, 2000; Dupoux, Pallier, Kakahi, & Mehler, 2001).

These studies rely on phonotactic differences between French and Japanese in what consonants can cluster together: French but not Japanese permits such clusters as [bz] or [gm]. The phonotactic constraint is, moreover, an active component of Japanese’s grammar in that it forces epenthesis of [u] to break up such clusters in words that Japanese borrows from other languages.

Dupoux, et al. (1999) had French and Japanese listeners respond whether the vowel [u] was present in a continuum of stimuli from [ebuzu] to [ebzo], where the [u] was progressively shortened down to nothing. French listeners’ “yes” responses decreased monotonically from 100 down to 0% as the vowel shortened, while Japanese listeners’ “yes” responses didn’t drop below 70% even for the stimuli from which the entire vowel had been removed. The phonotactic prohibition in Japanese against clusters such as [bz] creates a perceptual illusion: where a vowel must occur, Japanese listeners hear one, even if there’s actually no vowel there at all. Japanese listeners were also much poorer than French listener at speeded discrimination of stimuli in which differed in whether a vowel intervened between two consonants, e.g. [ebuzo] vs [ebzo].

Dupoux, et al. (2001) showed that these effects reflect a phonotactic constraint and not just the facts that the vowel [u] occurs between two consonants in many Japanese words, particularly, in many loan words. Listeners transcribed and made lexical decision judgments for nonword strings containing illegal CC clusters that either have a single lexical neighbor with [u] between the consonants, e.g. the string [sokdo] has the neighbor [sokudo] “speed”, or a single lexical neighbor with some vowel other than [u], e.g. the string [mikdo] has the neighbor [mikado] “emperor”. If the illusion simply reflects the possibility that the string is a word once a vowel has been added to it, then listeners should transcribe [mikdo] with its missing vowel and identify it as a word as readily as [sokdo]. If they instead imagine that an [u] is present because the grammar supplies it as the epenthetic vowel to repair violations of a constraint banning clusters such as [kd], then they should instead identify [sokdo] as a word more often than [mikdo]. They should, moreover, transcribe [mikdo] with [u] between the two consonants as often as [sokdo], even though [mikudo] isn’t a word. In conformity with this alternative prediction, listeners inserted [u] into their transcription of [mikdo] strings nearly as often as into [sokdo], despite the absence of any corresponding word.
[mikudo]. They also identified [mikdo] strings far less often as words than [sokdo] strings, despite the existence of the word *mikado*. Finally, response times in lexical decision to [mikdo] strings were as slow as those to the corresponding nonword strings [mikudo] while those to [sokdo] strings were as fast as those to the corresponding word strings [sokudo]. All these results support the phonotactic constraint explanation and not the lexical similarity explanation. The phonotactic constraint must introduce the illusory vowel before any lexical item is activated because [u] was inserted into the transcriptions of [mikdo] strings nearly as often as into [sokdo] strings, despite the lexical competition from the [a] in [mikado], and [mikdo] was thus identified as a nonword as slowly as [mikudo].

This conclusion is driven home by the behavioral and neurophysiological data reported by Dehaene-Lambertz, et al. (2000). French and Japanese listeners heard four repetitions of strings such as [igumo] or [igmo] followed either by a fifth repetition of the same string or by the string that differed in the presence or absence of the vowel [u] between the two consonants. Listeners labeled the fifth stimulus as the “same” or “different” from the first four as quickly as possible. French listeners responded “different” to different trials far more often than Japanese listeners, 95.1% vs only 8.9%, and responses were significantly slower in different trials compared to same trials for the French but not the Japanese listeners. Like the accuracy data, the absence of any RT difference in the Japanese listeners’ responses suggests they didn’t notice that the fifth stimulus was different in the different trials.

An event-related potential (ERP) obtained in the earliest interval following the moment when the fifth stimulus deviated from the four preceding stimuli on different trials was significantly more negative in voltage than that obtained on same trials for French but not Japanese listeners. Dehaene-Lambertz, et al. interpret this ERP as arising when the brain detects the sensory mismatch between the different stimulus and the sensory memory of the preceding reference stimulus.33 Just as the behavioral response shows that the Japanese listeners seldom consciously heard any difference between [igumo] and [igmo] strings, this early ERP shows that their brains didn’t notice any difference either.

Kirk’s results and those of Dupoux and his colleagues show that on-line syllabification is as strongly influenced by language-specific constraints as is on-line categorization. Like the categorization results, they also cannot be attributed to statistical generalizations across the lexicon. All the knowledge about their languages that the participants in these experiments used in producing these results instead takes the form of psychologically active constraints in their synchronic phonologies.

3.4.4.4 An optimizing sound change
Blevins also claims that sound change isn’t optimizing. Undoubtedly in many cases it isn’t, but in at least some cases it is. The case cited discussed here – the ongoing split in the pronunciation of the diphthong /aɪ/ in southern American English – is particularly interesting because the optimization apparently conflicts with well-grounded phonetic expectations. As we’ll see, there are independent and conflicting phonetic reasons to expect exactly this development.

As documented in Thomas (2000) and Moreton (2004), /aɪ/’s pronunciation is shifting toward a more extreme diphthong – F1 is lower and F2 is higher in its offglide – before voiceless obstruents and toward a monophthong elsewhere. Thomas and Moreton also show that listeners are significantly more likely to identify a following obstruent as voiceless when diphthongization is more extreme.

What is striking about this split is that its direction is exactly opposite what’s expected on phonetic grounds. In the transition to voiceless obstruents, formants are frequently cut off early because the glottis opens and voicing ceases before the oral constriction is complete. Voicing’s continuation into the oral constriction of voiced obstruents permits formants to reach more extreme values at the end of a vowel preceding such consonants. An important consequence is that F1 is typically higher at the end of a vowel before a voiceless than a voiced obstruent; this difference is in fact so reliable that listeners use it to as a cue to the obstruent’s voicing (Parker, Diehl, & Kluender, 1986; Fischer & Ohde, 1990; Kingston & Diehl, 1995). How then can diphthongal offglides become acoustically more extreme precisely in the context where the phonetics leads us to expect they’d become less extreme, while becoming less extreme in the context where the phonetics encourages more extreme acoustic values?

After considering and rejecting a number of alternatives, Moreton offers the following answer to this question: diphthongs are hyper-articulated before voiceless obstruents because voiceless obstruents themselves are hyper-articulated, i.e. produced with more extreme and faster articulations than voiced obstruents, and the consonant’s hyper-articulation spreads to the preceding vowel. The spreading of hyper-articulation compensates for and indeed undoes the cutoff of formant frequencies by the early cessation of voicing before voiceless obstruents. Hyper-articulation is also optimizing in that it not only maintains but exaggerates the voicing contrast.

3.4.4.5 Innateness

Blevins also argues that any aspect of sound patterns that can be learned from exposure to mature speakers should not be analyzed as arising from the learner’s innate endowment. She observes that a pattern or behavior can only analyzed as innate if it emerges when there’s no external stimulus or model, as does vocal babble in
deaf infants. This is surely correct.

Optimality theory’s account of language learning assumes that in the initial ranking all marked constraints outrank all faithfulness constraints, and that the learning of a particular language is a process of demoting specific markedness constraints below competing faithfulness constraints. It seems correct to assume that the initial state of the grammar is innate rather than learned. Therefore, demonstrating that markedness constraints are ranked above faithfulness constraints for infants who are too young to be learning the ambient language’s sound patterns would confirm the hypothesis that this ranking is innate, assuming that markedness constraints do outrank faithfulness constraints in the initial state (see note 37). This result would also presumably demonstrate that the particular constraints observed to be ranked in this order are themselves innate. This is precisely what Davidson, et al. (2004) have done.

Davidson, et al. used the head-turn procedure to determine whether 4.5 month old infants prefer to listen to sequences of the form ... on pa ompa ... or ... on pa onpa ..., where the assimilation in place of the nasal in ompa obeys a markedness constraint, while its failure to assimilate in onpa instead obeys a conflicting faithfulness constraint. The results show that infants at this age significantly prefer to listen to the sequence than obeys the markedness rather than the faithfulness constraint. This preference cannot be a result of these infants’ having learned that nasals assimilate in place to following consonants in the ambient language, because 6 month old infants don’t significantly prefer to listen to sequences that conform to the ambient language’s phonotactics. Such preferences don’t emerge before 9 months (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Jusczyk, Luce, & Charles-Luce, 1994; Jusczyk, Houston, & Newsome, 1999; Mattys & Jusczyk, 2001). This result confirms the prediction that markedness constraints are ranked above competing faithfulness constraints in the initial state as well as the hypothesis that this ranking and the constraints themselves are innate.

Certainly, this initial ranking isn’t preserved in the grammar of the ambient language that is eventually learned; it can’t be. Nonetheless, in the most parsimonious account, that grammar is composed of the same constraints as govern infants’ responses at 4.5 months, and no more markedness constraints have been demoted below faithfulness constraints than is necessary to account for the surface distributions of sounds in that language.

3.4.4.6 Summary

In this section, I have presented positive evidence in support of five claims rejected in Blevins (2004): (1) a sound change’s phonetic motivation remains active
after it’s been phonologized, (2) constraints attributed to synchronic phonologies are psychologically and neurophysiologically active, (3) these constraints are not mere statistical generalizations across the lexicon, (4) sound changes can be optimizing, and (5) the constraints of which synchronic phonologies are composed and their initial ranking are innate. This evidence points to a rich, on-going interchange between the phonetics and phonology and to a contentful synchronic phonology.

4 Implementation

4.1 Introduction

In this concluding section of the chapter, I discuss three examples of how the phonetics implements phonological representations. In each case, the way in which a representation is implemented reflexively provides an explanation for properties of the representation itself. The first example concerns how tones are aligned with an utterance’s segments, what’s known as “tune-text alignment”. In at least some languages, the nature of the tune-text alignment confirms that the hypothesis that intonational contrasts are best represented as a string of tones, while in others it suggests that they are better analyzed as rises and falls. In the second example, I take up the question of what it means phonetically for a sound to be phonologically marked vs unmarked. In many analyses, the coronal place of articulation is treated as unmarked compared to the marked labial and dorsal places. The evidence shows that the unmarked coronal place is more variable in its articulation than the marked non-coronal places, and listeners perceive coronal place when acoustic evidence for its presence is weak or even altogether absent. These behaviors may explain why coronals assimilate in place of articulation to adjacent non-coronals while non-coronals don’t assimilate. It is often been proposed that phonetics manipulates gradients, while phonology instead manipulates categories. This distinction is the central issue in the third example where it is extended to differences in how phonetic and phonological constraints are prioritized. The evidence presented there suggests that one phonetic constraint isn’t categorically ranked above or below another the way phonological constraints are, but that the constraint that has higher priority is weighted more heavily in evaluating a possible output’s well-formedness. Weight is inherently gradient rather than categorical.

4.2 Tune-text alignment

Perhaps most influential work on phonetic implementation has been the studies of intonation initiated by Pierrehumbert (1980). The essential claim in this approach to the analysis of intonation is simple: the variation in F0 across the course of an utterance can be represented phonologically as a string of tones, in many analyses just two, a high (H) and a low (L) tone. These tones specify F0 targets which are to be reached at
particular moments in the utterance, i.e. on the segments with which the tones are aligned. This alignment of tones is the “tune” with the segments. The text is the topic of this section.

A number of recent studies, of Greek (Arvaniti, Ladd, & Mennen, 1998, 2000), English (Ladd, Faulkner, Faulkner, & Schepman, 1999; Ladd & Schepman, 2003; Dilley, Ladd, & Schepman, 2005), Dutch (Ladd, Mennen, & Schepman, 2000), and German (Atterer & Ladd, 2004), have shown that the first of two tones in a bitonal pitch accent is aligned to the beginning of the syllable bearing that accent and the second is aligned to the end of the syllable (English, Dutch, German) or to the beginning of the next syllable (Greek). These alignments are demonstrated by the fact that the timing of the F0 targets that realize these tones varies little relative to these moments in the segmental string as the phonological quantity and/or phonetic duration of the syllable is varied. As quantity and/or duration increase, the interval between the two targets stretches out, and as they decrease, it instead compresses. The slope of F0 change between the two targets varies inversely with the interval between them.

This description of the results of these studies abstracts away from consistent differences between languages and dialects in how early or late the F0 targets are relative to their alignment points. In particular, both tones are aligned later in German than English or Dutch and later in Southern than Northern German (Atterer & Ladd, 2004). Atterer & Ladd argue that invariant timing of F0 targets relative to particular segments in the accented syllable is evidence of the phonological alignment of the tones to these segments, and the subtler differences in just when an F0 target occurs relative to the segment it’s aligned with instead reflect language-specific phonetic implementation.

As emphasized by Ladd, et al. (2000), the very finding that F0 targets are aligned consistently to particular segments in the accented syllable supports the phonological analysis of the intonation contour as a string of tones. If the phonological representation instead specified that F0 rise or fall on particular syllables, there is no reason why the beginnings or ends of those rises and falls should consistently occur at the same times relative to the syllable’s segments.

The timing of F0 targets in the Mandarin rising tone apparently reflect yet another pattern of alignment (Xu, 1998). The final high F0 target is consistently aligned with the end of the syllable or with the C-V boundary at the beginning the next syllable. The initial low F0 target (or the beginning of the rise) and the moment when the slope of the rise is steepest both occur later when the syllable is longer. That is, the initial target, too, is aligned with the end of the syllable. Perhaps, the initial F0 target is displaced by the late realization of the final target of the preceding syllable’s tone.
Unfortunately, F0 targets are not aligned with particular segments in all languages. Erikson & Alstermark (1972) and Bannert & Bredvad-Jensen (1975, 1977) show that in most Swedish dialects, the F0 contours that realize the bitonal pitch accents are truncated rather than compressed when the vowel on which they're realized is shortened, although compression is observed in a few dialects. This finding is compatible with an account in which the first tone's F0 target is aligned with the beginning of the accent syllable but not the second tone's. Following the first target, F0 rises or falls at a more or less constant rate toward the second F0 target, which would be reached at a more or less fixed time after the first. If the vowel is long enough, the second target is reached, but if it's shortened, F0 falls short of the second target, in proportion to the shortening of the vowel.

Besides the majority of Swedish dialects, truncation has also been reported for Danish (Grønnum, 1989), Palermo Italian (Grice, 1995), Hungarian (Ladd, 1996), for a different pitch accent in German than the one studied by Atterer & Ladd (Grabe, 1998a,b), and in the varieties of English spoken in Leeds and Belfast but not those spoken in Cambridge and Newcastle (Grabe, Post, Nolan, & Farrar, 2000).

As did Ladd, et al. (2000), Grabe, et al. (2000) emphasize the crucial role that tune-text alignment plays in arguing that intonation contours are represented phonologically as strings of tones. It is harder to argue for at least the second of these tones if its target isn’t reached as a result of truncation. The observed contours might equally well be analyzed as the realization of a tone followed by an instruction to rise if that tone is L or to fall if it’s instead H. Other phonetic evidence must be sought to argue for the second tone’s presence in the phonological representation in such cases.

4.3 The phonetics of markedness

In heterosyllabic sequences of a coronal stop followed by a non-coronal stop in English, e.g. [t.k] or [d.g], the coronal articulation is typically briefer, it may be substantially reduced, even to the point where the tongue tip and blade don’t reach the alveolar ridge, and it is often fully overlapped by the following non-coronal articulation (Nolan, 1992; Byrd, 1996). For some speakers, coronal stops in this context assimilate completely to the following non-coronal, in some or all tokens (Ellis & Hardcastle, 2002). When the order of the places of articulation are reversed, the non-coronal isn’t shortened, reduced, nor overlapped nearly as much.

This articulatory asymmetry is matched by a corresponding perceptual one. Gaskell & Marslen-Wilson (1996) report the results of a cross-modal priming task, in which an assimilated pronunciation of a coronal stop, e.g \textit{lea[m]} bacon, sped up recognition that a simultaneous visual probe \textit{lean} is a word just as much as did the
unassimilated pronunciation, *lea[n] bacon* (Gaskell & Marslen-Wilson, 1996). These results were obtained despite the fact that the coronal stops were deliberated pronounced at the same place of articulation as the following consonant (cf. Gow, 2002). However, an assimilated pronunciation of a non-coronal stop, e.g. *la[r]e goat*, slowed recognition that the visual probe *lame* was a word significantly compared to the unassimilated pronunciation, *la[m]e goat*. Monitoring for the phoneme beginning the second word, e.g. the /b/ in *lean bacon*, is also facilitated by an assimilated pronunciation of the the preceding coronal, whether assimilation is full (Gaskell & Marslen-Wilson, 1998) or only partial (Gow, 2003). These phoneme monitoring results suggest that the listener parses the non-coronal place information off the assimilated consonant and attributes it to the following non-coronal.

But how or why do listeners parse place information in this way? Both Gaskell & Marslen-Wilson and Gow argue that they do so because they know the articulatory facts described at the beginning of this section: coronal stops are extensively overlapped by and even assimilate to following non-coronals. The coronal stops in Gow’s stimuli were only partially and not fully assimilated, so his results could arise from listeners’ actually hearing coronal as well as non-coronal place information in the affected consonant. But as noted, the coronal stops in Gaskell & Marslen-Wilson’s stimuli were fully assimilated, so their listeners weren’t responding to phonetic evidence of a coronal articulation in the signal. Instead, when listeners hear a non-word ending in a non-coronal consonant before a homorganic consonant and that non-word becomes a word when its final non-coronal is replaced by a coronal, they infer that the non-coronal place information belongs to the following consonant and that the intended consonant is coronal. They don’t infer another non-coronal because they have no comparable experience of non-coronals being extensively overlapped by or assimilating to the place of articulation of the following consonant. This interpretation is supported by Gaskell & Marslen-Wilson’s finding that an assimilated coronal that isn’t homorganic with a following non-coronal, e.g. *lea[m] goat*, neither primes recognition of the visual probe *lean* nor facilitates detection of the initial /g/ in the following word. The inferences are blocked in this “non-viable” assimilation because the non-coronal place of the [m] cannot be parsed onto the following [g].

Coenen, Zwitserlood, & Bölte (2001) report cross-modal priming experiments run with German listeners in which the procedures and results closely resemble those reported by Gaskell & Marslen-Wilson (1996). Lahiri & Reetz (2002) also report the results of cross-modal priming experiments with German listeners, but the procedures differed in two respects from those used by Gaskell & Marslen-Wilson or Coenen, et al. First, semantic rather than form priming was used. Far more important is the fact the primes were presented in isolation, without any following word whose initial consonant might be an assimilation trigger. In the first experiment, the auditory primes were
isolated words either ending a coronal such as *Bahn* “railway” or ending in a non-c coronal such as *Lärm* “noise”, and non-words made by replacing the final coronal with a non-coronal or vice versa, *Bahm* vs *Lärn*. Both *Bahn* and *Bahm* primed recognition of the related visual probe *Zug* “train” but only *Lärm* primed *Krach* “bang, racket”. This result can’t be attributed to the listeners’ actually parsing the non-coronal place information at the end of *Bahm* onto a following homorganic consonant because there was no following consonant. However, perhaps listeners were still able to separate the labial place information from *Bahn* because *Bahn* is sometimes pronounced *Bahm* in front of a word beginning with a bilabial consonant. This alternative is ruled out by the second experiment, where the manipulated consonants were medial rather than final, and therefore in a context where they never assimilate to a following consonant.

Auditory primes were words with medial coronal or non-coronal consonants, e.g. *Düne* “dune” or *Schramme* “a scratch”, and corresponding non-words made by substituting a non-coronal for a coronal or vice versa, *Düme* or *Schranne*. Both *Düne* and *Düme* primed recognition of the related visual probe *Sand* “sand” but only *Schramme* primed *Kratzer* “a scratch”. These results rather definitively rule out the inferential parsing account proposed by Gaskell & Marslen-Wilson or Gow.

How then are they and the earlier results to be interpreted? Lahiri & Reetz’s interpretation relies on coronal place not being specified phonologically, while the labial and dorsal places are specified. When there is phonetic evidence in the signal for a non-coronal place, as in *Bahm* or *Düme*, this evidence doesn’t mismatch the stored forms of the words *Bahn* or *Düne*, because the /n/ in these words isn’t specified for place, and these words are activated. Because *Bahm* and *Düme* aren’t words, this evidence for non-coronal place also doesn’t activate any competing words. Phonetic evidence of coronal place as in *Lärm* or *Schranne*, however, does mismatch the phonological specification for labial place in the words *Lärn* or *Schramme*, which inhibits their activation.

The proposal here to leave the unmarked members of contrasts unspecified phonological is a quite general one, which Lahiri & Reetz have implemented in a model for automatic speech recognition known as FUL (for “fully underspecified lexicon”). FUL is remarkably robust, in that it can successfully recognize connected speech in noisy environments from multiple talkers without training. It is also a model of humans’ recognition of speech that makes quite precise and therefore testable predictions about their behavior.

The one result that this interpretation doesn’t handle easily is the failure of the non-viable assimilation in [lim##got] to prime *lean*. The phonetic evidence for the labial place in the [m] wouldn’t mismatch the missing place specification of the /n/ in this string any more than in an isolated word. However, in an earlier cross-modal priming cross-
model priming study with German listeners in which the auditory primes were followed by another word whose initial consonant could be an assimilation trigger (1995), Lahiri obtained priming for non-viably as well as viably-assimilated coronals, i.e. Bahm primed Zug even when the following word didn't begin with a labial consonant. This result indicates that viability needs to be reexamined.

The articulatory data reviewed at the beginning of this section shows that the unmarked member of a contrast may vary substantially more in its pronunciation than the marked member(s). The perceptual data indicate that listeners can readily tolerate the phonetic effects of the unmarked member’s variation, either because they’ve had long experience of it or because the unmarked member is actually not specified phonologically and the variation creates no mismatch between phonetic evidence and the phonological specification.

4.4 Categories and gradients


In this section, I discuss a more recent use, by Zsiga (2000) in her analysis of differences between English and Russian in the extent to which the coronal gesture for an [s] is overlapped by the palatal gesture of a following [j] across word boundaries. The extent of overlap was assessed acoustically by comparing the spectral distribution of fricative noise at the beginning, middle, and end of the fricative in the [s# #j] sequences with the noise distribution at comparable points in [s] and [j]. If [j]’s palatal articulation overlaps [s]’s, then the center of gravity of the noise distribution should shift to lower frequencies such as those in [j].

Lower, more [j]-like noise distributions were observed in the English speakers’ pronunciations of [s# #j] sequences, but with the exception of just a few tokens only at the end of the fricative. The center of gravity of the noise distributions at beginning and middle of the fricatives remained high and [s]-like in the great majority of tokens. These acoustic data agree with palatographic data reported in Zsiga (1995), which showed a shift from coronal to palatal contact only at the end of an [s] preceding [j]. Both the acoustic and articulatory data can be interpreted as evidence that starting after the middle of the fricative, the [s]’s coronal articulation gradually blends with the following [j]’s palatal articulation. This blending produces an articulation that’s midway between these two articulations at the end of the fricative.
The acoustics of the Russian speakers’ pronunciations, however, showed no
evidence of articulatory overlap between the two articulations: the center of gravity of
the noise distribution was just as high and [s]-like at the end of the fricative as earlier.
The noise distribution in the Russian speakers’ pronunciations of [s] before [j] also
differs from that in their pronunciations of palatalized [sj], where there are strong peaks
at both high and low frequencies. The coronal and palatal articulations are
simultaneous but not blended: both coronal and palatal articulations are produced, not
an articulation halfway between them.

Zsiga proposes that English and Russian differ in the relative priorities of
phonetic constraints requiring the speaker to achieve particular articulatory targets
specified in the phonological representation. These constraints resemble faithfulness
constraints in requiring that a property of the input, in this case, the surface phonological
representation, be realized unchanged in the output, the actual pronunciation. For
English speakers, the requirement to maintain the coronal constriction specified by /s/
gradually gives way over the last half of the fricative to the requirement to reach the
following palatal constriction. The result is a progressive blending of the two
constrictions that by the end of the fricative produces a constriction midway between the
alveolar ridge and the palate. For Russian speakers, however, the requirement to
maintain the coronal constriction remains a higher priority all the way to the end of the
fricative constriction in [s# #j] sequences, as well as in palatalized [sj], where the
coronal articulation is maintained despite complete overlap with the palatal articulation.\(^{38}\)
Because the coronal constraint’s priority doesn’t change in Russian even when the
coronal articulation is completely overlapped by the palatal articulation, while its priority
diminishes gradually as a result of overlap in English, Zsiga argues that these phonetic
constraints are weighted continuously with respect to one another and not ranked
categorically. This difference in how priority conflicts are resolved is the basis for her
treating the phonetic constraint evaluation as autonomous from and following
phonological constraint evaluation.

This sequential model is quite different from that advocated by Steriade (1999) or
Flemming (2004, submitted) in which phonetic constraints are integrated among and
even supplant phonological constraints, and where the phonetic constraints are also
strictly, i.e. categorically ranked. Their models do not, as far as I know, try to account
for phonetic detail to the extent that Zsiga’s proposal does, but there seems to be no
formal barrier to their doing so. In the example of tune-text alignment, the differences
between languages pointed to opposite and irreconcilable analyses. Here, there simply
aren’t enough facts yet to decide whether phonological and phonetic constraint
evaluation are a single, integrated process or instead sequential.

Summary and concluding remarks
In the three sections of this chapter, I have tried to show how distinctive features might be defined, how phonological patterns might be explained, and how phonological representations might be implemented.

The essential problem that has to be solved in defining distinctive features is that their articulations and acoustics vary enormously. This variability can be largely eliminated by moving away from the details of particulation phonetic realizations, either toward the articulatory plan for the utterance embodied in articulatory phonology’s gestures or toward the auditory effects of the signal’s acoustic properties. Evidence was reviewed that pointed to the second move as the right one, i.e. that the speaker’s goal is to produce particular auditory effects in the listener.

Explaining phonological patterns is difficult because they are typically determined by more than one phonetic constraint, as well as by phonological constraints, and these constraints may conflict with one another. The eventual explanation is a description of the resolution of this conflict. It is largely because phonetic explanations are complex in this way that I think no bright line can be drawn between the phonetic and phonological components of a grammar. It’s interesting to note in this connection that many of those who advocate such bright lines (e.g. Hale & Reiss, 2000; Blevins, 2004) also reject phonological models in which the surface phonological representation corresponding to a particular underlying representation is selected by applying well-formedness constraints in parallel to all possible surface representations, as in optimality theory. The replacement of serial derivation by parallel evaluation removes the barrier to phonetic constraints being interpersed among and interacting with phonological constraints. (Zsiga’s, 2000, proposal, as described in §4.4, is an obvious exception to this generalization.)

The problem in trying to understand phonetic implementation is actually very similar to that arising in attempts to explain phonological patterns in phonetic terms: phonetic constraints not only regulate how a phonological representation can be realized but also determine at least some of its properties. These properties of the phonological representation emerge out of its implementation in much the same way that the distinctive features emerge out of the solution to the variability problem.
References


Cohn, A. C. (1993). Nasalisation in English: Phonology or phonetics, Phonology, 10, 43-81.

Crosswhite, K. (2004). Vowel reduction, B. Hayes, R. Kirchner, & D. Steriade (eds.), *Phonetically Based Phonology*, (pp. 191-231), Cambridge, UK: Cambridge University Press.


Flemming, E. (2004). Contrast and perceptual distinctiveness, B. Hayes, R. Kirchner,
& D. Steriade (eds.), *Phonetically Based Phonology*, (pp. 232-276), Cambridge, UK: Cambridge University Press.


Ohala, J. J. (1983a). The origin of sound patterns in vocal tract constraints, P. F. MacNeilage (ed.), *The Production of Speech*, (pp. 189-216), New York: Springer-
Verlag.


* Acknowledgments. I could not have written this chapter without feedback and stimulation I received from the students in the seminar I taught on the phonetics-phonology interface at the University of Massachusetts, Amherst in the fall of 2004, Kathryn Flack, Shigeto Kawahara, Anne-Michelle Tessier, and Matthew Wolf. I have also made great use of the work done by some of the students who I’ve worked very closely with here at the University of Massachusetts, Amherst, Andries Coetzee, Cecilia Kirk, and Elliott Moreton, without which I would have been gravelled for lack of evidence in at many crucial points in the development of my arguments. Finally, I must thank Paul de Lacy for inviting me to write this chapter in the first place; his doing so forced me to take up a task that I had been putting off for some time: working out just what I think the interface is (and isn’t) between phonetics and phonology. In short, I owe many of the virtues of the present chapter, such as they are, to others, particularly to my students. As is customary, I keep the faults, of which there are many, for myself.
1. I have omitted many technical details in discussing the various cases presented in this chapter when doing so would not impair understanding or when they may be easily found in the sources that I cite.

2. In articulatory phonology, this collection of gestures is actually the phonological representation of the utterance, which doesn’t have a more abstract representation in terms of distinctive feature values. Browman & Goldstein make this move to avoid the very difficult problem of translating between an utterance’s linguistic representation as a sequence of discrete cognitive categories and its physical realization as continuous and overlapping actions having spatial and temporal extents. Translation is so difficult because categories and actions are so thoroughly incommensurate (Fowler, Rubin, Remez, & Turvey, 1980). Moreover, if the phonological and phonetic representations differ to this extent, it’s hard to see how either could constrain the other, yet there is considerable evidence that they do (Browman & Goldstein, 1995).

3. Keating’s (1990, 1996) window model of coarticulation achieves a similar result by specifying spatial and temporal ranges for particular articulators’ movements. The limits on these ranges ensure that the speaker reaches the goal, but their width permits individual articulators’ movements to vary in extent depending on context.

4. Holt, Lotto, & Kluender (1991) that the perceptual interaction between F0 onset frequency and voice onset delay in stops arises from learners’ having experienced their reliable covariation and not from any auditory similarity between the two properties.

5. A syllable is heavy if its rime contains a long vowel and/or at least one coda consonant but light if its rime contains just a short vowel. If each of these constituents of the rime counts as a mora, a syllable is heavy if its rime consists of two (or more)
moras, and light if it consists of just one.

6. In note 5, syllable weight or quantity was defined in terms of the number of segments or moras in the rime, not the actual duration of the rime. Nonetheless, a rime containing a long vowel or a short vowel and a coda consonant, i.e. one consisting of two moras, does last longer than one containing just a short vowel, i.e. one consisting of just a single mora. Stressed syllables in a quantity-sensitive language could therefore be distinguished from unstressed ones either by having one more mora or by lasting longer; either the phonological or the phonetic comparison yields the same results. The preference then to group a shorter unstressed syllable with a following longer stressed syllable into an iamb in such languages apparently follows naturally from the preference to group sounds that alternate in duration into short-long pairs. In a quantity-insensitive language, syllable rimes may still differ in mora count and duration but these differences aren’t recognized as differences in prominence in such languages. Stressed syllables must be distinguished from unstressed ones by other properties. Intensity is the most likely candidate, as fundamental frequency is very likely to be determined by lexical tone or pitch accent contrasts and/or intonation and thus unavailable to serve as the principal correlate of stress (for relevant discussion see Beckman, 1986). If intensity differences are used to distinguish stressed from unstressed syllables in quantity-insensitive languages, then the preference for pairing a softer unstressed syllable is paired with a preceding louder stressed syllable into a trochee also follows naturally from the preference for grouping sounds alternating in intensity into loud-soft pairs.

7. They also presented the listeners with tone sequences alternating in frequency. Frequency alternation isn’t relevant to evaluating the phonetic explanation for the
iambic-trochaic law, so these results won’t be discussed here.

8. Pulmonic stops are those where the flow of the air stream is initiated by contraction of the lungs, while glottalic stops are those where the glottis is closed or constricted, isolating the oral cavity aerodynamically from the subglottal cavity and the flow of the air stream is initiated by contracting or expanding the oral cavity.

9. Dental and alveolar stops are lumped together because languages differ in whether their coronal stops are dental or alveolar and rarely contrast stops at these two places of articulation and because coronal stops aren’t explicitly identified as dental or alveolar for many languages in UPSID.

10. $X^2$ analyses would be used for contingency tables with just two dimensions.

11. The vowel space has a fixed acoustic volume because when articulators move past certain limits, they impede air flow enough that what was a vowel becomes a fricative. This limit on tongue body raising toward the palate and tongue root retraction into the pharynx restricts F1 to being no less than 250-300 Hz and no more than 800-900 Hz in vowels. F2's minima and maxima are similarly limited, to a range of roughly 500-2200 Hz, by how far front or back the tongue body can be moved and how much the mouth opening can be constricted by lip rounding without also constricting the vocal tract too much. Of course, these limits vary between speakers, being lower in those whose vocal tracts are larger and higher in those whose vocal tracts are smaller, but the limits don’t go away. Nonetheless, there is a critical degree of constriction that will impede air flow enough to turn a vowel into a fricative in a vocal tract of any size.

12. A very different approach to predicting the contents of vowel inventories is presented by de Boer (2000). It appears to be somewhat more successful at predicting
which arrangements of vowels will be preferred in inventories of a given size than the
approach considered here.
13. These include voicelessness, breathiness, laryngealization, velar stricture,
pharyngealization, and a property called “extra-shortness”, which is used in UPSID to
describe vowels that are shorter than ordinary short vowels.
14. 51 languages in the sample contrast long with short vowels: 14 of them distinguish
more qualities among the long than the short vowels, 14 distinguish the same number of
qualities in long as short vowels, and the remaining 23 distinguish more short than long
vowel qualities.
15. Sometimes, a language lacks a peripheral vowel at a particular height and
backness, but has the central vowel, e.g. the language has /u/ but no /u/ or has /e/ but
no /a/. In such cases, the otherwise central vowel is treated as peripheral in the tallies
below. Thus, a vowel quality is only central if it contrasts in a language with a minimally
different peripheral vowel.
16. These patterns and their frequencies closely resemble those in Table I in Schwartz,
et al. (1997a), which is unsurprising given that they analyzed an earlier, smaller version
of UPSID consisting of 317 languages (Maddieson, 1984).
17. This way of describing the facts is not meant to suggest that one or two central or
peripheral vowels were literally added to an earlier 5+0 inventory at some time in these
languages’ histories. Instead, it’s intended to describe how vowel patterns would change
or remain the same when additional vowels are present in an inventory.
18. This generalization holds if /a/ is treated as a front vowel in the Qawasqar inventory
/a, a, o/. It is at least more front than /o/.
19. This, too, is plainly a kind of functional explanation, but one concerned with the making maximal use of the available resources for contrast between messages rather than with the distinctness of messages.

20. Because half the values in the range are below the median and half are above, all the values are on average closer to the median than to any other value, and it is less affected by extreme values than the mean. The median thus estimates the distribution of values better than the mean.

21. The average excess of oral vowels is 2.45, but as language cannot have a fraction of a vowel, the average excess is better described as between 2 and 3 vowels.

22. Here “mid” encompasses the range from lower to higher mid, “high” includes high and lowered high, and “low” includes low and raised low.

23. These findings agree those reported in other studies, which report that a higher vowel will sound more nasalized than a lower vowel for a given degree of coupling of the nasal to the oral cavity (House & Stevens, 1956; Lubker, 1968; Ohala, 1975; Abramson, Nye, Henderson, & Marshall, 1981; Benguerel & Lafargue, 1981; Stevens, Fant, & Hawkins, 1987; Maeda, 1993, cf. Lintz & Sherman, 1961; Massengill & Bryson, 1967; Bream, 1968; Ali, Gallagher, Goldstein, & Daniloff, 1971).

24. Barnes (2002) restricts the reduction pattern shown in (1c) to eastern varieties of Bulgarian alone; Crosswhite does not. This disagreement isn’t material to the discussion here.

25. Crosswhite (2004) cites two languages, Standard Slovene and northeastern dialects of Brazilian Portuguese, in which reduction lowers higher mid vowel to a lower mid one. In standard descriptions of Standard Slovene, lower mid /ɛ, ə/ are described
as being pronounced [ɛ, ɔ] in unstressed as well as stressed syllables, while higher mid /e, o/ are lowered to [ɛ, ɔ] in unstressed syllables. However, Crosswhite cites formant frequency measurements reported by Lehiste (1961) which show that the unstressed realizations of all these vowels are actually intermediate between higher and lower mid qualities. No comparable reanalysis is given for the Brazilian Portuguese case.

26. Petterson & Wood (1987a,b) present cineradiographic evidence from Bulgarian showing that it is the jaw rather than the tongue that undershoots its target in unstressed syllables. The tongue remains lower for the non-high vowels [e, o, a] than for [i, u, ə], but the jaw is higher, close to its position in [i, u, ə]. The jaw is apparently high enough to lower F1 enough that unstressed [e, o, a] sound like the vowels just above them. See also Westbury & Keating (1980) for evidence that vowels with lower jaw positions last longer.

27. Brazilian Portuguese doesn’t distinguish higher from lower mid nasal vowels.

28. Although these choices are not conscious, speakers must still be viewed as making choices because there is nothing about the circumstances or contexts in which they hypo- or hyper-articulate or choose one allophone rather than another that obligates them to pronounce the sounds in those ways rather than others. See Blevins (2004) for a proposal that speakers choose from an array of alternative pronunciations.

29. The spirantization of stops in specific phonological and morphological conditions is a very old process in Athabaskan. It must antedate tonogenesis because its reflexes are observable in the conservative non-tonal Alaskan and Pacific Coast languages as well as the tonal languages (Leer, 1979, 1999).

30. Steriade (1993) discusses other phonological consequences of the acoustic
difference between a stop’s onset and release compared to the acoustic similarity of these events in fricatives and sonorants.

31. In Kingston (2004, see also Kingston & Solnit, 1989; Solnit & Kingston, 1989), I show that apparent tone reversals of this kind are in fact widespread and also occur when the historical sources of the tones are an earlier contrast between voiced and voiceless obstruents or between aspirated and unaspirated consonants – the latter include sonorants as well as obstruents. In these cases, too, it may be possible to pronounce the consonants such that they either raise or lower F0.

32. Blevins might appeal to structural analogy as the source of this uniformity. It’s influence could be exerted through paradigmatic alternations in the verbs in these languages, but it’s hard to see how it could be extended to the nouns where few if any helpful alternations occur. This is not to say that analogy has played no role in Athabaska tonogenesis, but its role is limited to the extension of tonogenesis to other morphemes than stems (Kingston, in press).

33. This ERP’s timing and cortical topography closely resembles the mismatch negativity (MMN) obtained whenever the brain detects that the current stimulus is auditorily different from the immediately preceding one.

34. Moreton reports that the offglides of other complex nuclei /ɔɪ, ɑʊ, ɛɪ/ are also more extreme before voiceless obstruents than elsewhere, but these nuclei are not (yet?) becoming monophthongs in southern American English when they don’t occur before voiceless obstruents.

35. See Smolensky (1996) and Davidson, Jusczyk, & Smolensky (2004) for arguments why markedness constraints must outrank faithfulness constraints in the initial state of
the grammar. These arguments depend on the fundamental assumption in optimality theory that the base of inputs to the grammar is itself unconstrained or “rich” in optimality theory’s terminology. If the base can be shown not to be rich, then the argument developed in this section will have to be reassessed.

36. See also Liberman & Prince (1977) for an even earlier analysis of a small piece of English intonation along essentially the same lines. The formalization and coverage in that paper were not sufficient to stimulate the explosion of work that followed Pierrehumbert (1980) three years later. Pierrehumbert & Beckman (1988) is an important landmark in the subsequent development of Pierrehumbert’s original proposals, and Ladd (1996) and Jun (2005) provide comprehensive reviews of the state-of-the-art in work along these lines.

37. Phonetic implementation also determines the frequency with which these targets are realized. This determination is exceeding complex and will not be discussed here. For detailed investigations, see Liberman & Pierrehumbert (1984), Pierrehumbert & Beckman (1988), and Ladd (1996), among many others.

38. Zsiga notes that Russian speakers may wish to avoid any blending in [s# #j] because they must keep /s/, /ʃ/, and /sʲ/ distinct.