

Running Head: SUBSTANTIVE BIAS AND VELAR PALATALIZATION

Manuscript submitted for publication (May 2005):
please do not quote, cite, or distribute without permission

Learning Phonology with Substantive Bias:
An Experimental and Computational Study of Velar Palatalization

Colin Wilson
Department of Linguistics
UCLA

Address for Correspondence:

Colin Wilson
Department of Linguistics
3125 Campbell Hall
Los Angeles, CA 90095 USA
Phone: (310) 991-3706
Email: colin@humnet.ucla.edu

Abstract

There is an active debate within the field of phonology concerning the cognitive status of substantive phonetic factors such as ease of articulation and perceptual distinctiveness. A new framework is proposed in which substance functions as a bias, or prior, on phonological learning. Two experiments tested this framework with a method in which participants are provided highly impoverished evidence of a new phonological pattern, and then tested on how they extend this pattern to novel contexts and novel sounds. Participants were found to generalize velar palatalization (e.g., the change from [k] as in *cap* to [tʃ] as in *cheap*) in a way that accords with linguistic typology, and that is predicted by a cognitive bias in favor of changes that relate perceptually similar sounds. Palatalization was extended from the mid front vowel context (i.e., before [e] as in *cape*) to the high front vowel context (i.e., before [i] as in *keep*), but not vice versa. The key explanatory notion of perceptual similarity is quantified with a psychological model of categorization, and the substantively biased framework is formalized as a conditional random field. Implications of these results for the debate on substance, theories of phonological generalization, and the formalization of similarity are discussed.

Keywords: Phonology; Phonetics; Language game; Inductive bias; Conditional random field

1 Introduction

With the introduction of theories of grammar that are based on violable constraints, and Optimality Theory ('OT'; Prince & Smolensky, 1993/2005) in particular, has come a renewed interest in the *substantive* factors that shape human languages. As originally defined (Chomsky, 1965), 'substance' refers to the system of categories that figure in the mental representation of linguistic knowledge. For example, the claim that the sounds of all languages are mentally represented with a particular set of distinctive features (e.g., [voice]), and that these features have universal articulatory and acoustic content, is a claim about substance. In the field of generative phonology, which studies knowledge of linguistic sound systems, substance is now used in a more broad sense to refer to any aspect of grammar that has its basis in the physical properties of speech. These properties include articulatory inertias, aerodynamic pressures, and degrees of articulatory salience and distinctiveness.

Recent work has emphasized the importance of acoustic/auditory/perceptual properties, an area that was previously somewhat neglected (but cf. Ohala, 1981, 1992; Lindblom, 1986, Stevens & Keyser, 1989). By studying the speech signal, as shaped by the vocal tract and processed by the auditory and perceptual systems, one gains a deeper understanding of several aspects of sound systems, including the inventories and distributions of sounds in the languages of the world (Beckman, 1999; Flemming, 2002; Gilkerson, to appear; Kawasaki-Fukumori, 1992; Kochetov, 2002; Ohala, 1992; Padgett, 2004; Steriade, 2001ab; Zhang, 2001), the characteristic changes that sounds undergo in particular phonological contexts (Cho & McQueen, in preparation; Côté 2000, 2004; Jun, 1995; Steriade, 2001ab; Wilson, 2001), lexical stress systems (Hayes, 1995; Gordon, 2004; Peperkamp, 2004), the perception and production of structures that do not occur in the native language (Davidson, 2003, to appear; Dupoux et al., 1999), and the extension of native-language phonological patterns to borrowed words (Fleischhacker, 2001; Kang, 2004; Kenstowicz, 2003; Zuraw, 2005). Two recent volumes (Hayes et al., 2004; Hume & Johnson, 2001) testify to both the dramatic advances that have been made in integrating perception into phonology and the pivotal role that OT has played in the formalization of the resulting theories.¹

In spite of these empirical and theoretical developments, there is no consensus on the central question of what role substance plays in grammar. Do the phonological grammars that speakers acquire have a significant substantive component? That is, do the cognitive computations that support phonological behavior make reference to knowledge of perceptual similarity, degree of articulatory difficulty, and, broadly speaking, other phonetic aspects of speech? Two opposing answers are given in the recent literature.

According to the framework known as *phonetically based phonology* (e.g., Hayes et al., 2004), phonological cognition is rich in substance. Speakers have detailed knowledge of articulatory and perceptual properties, and their grammatical systems make reference to that knowledge. Within OT, this takes the form of violable constraints that ban articulatorily difficult sounds and sound sequences, and that require sounds to appear in phonological environments that facilitate their perception.

An alternative framework, known as *evolutionary phonology* (e.g., Blevins, 2004; Blevins

¹Significant advances have also been made in integrating articulatory information into phonology (Davidson, 2003, to appear; Gafos, 1999, 2002; Hall, 2003; Hayes, 1999; Kirchner, 2000, 2001), but these developments are not of direct relevance for this paper.

& Garrett, 2004), claims that the evidence cited in support of phonetically based phonology is also consistent with an account in which substantive factors influence diachrony (the development of language over time) but not synchronic phonologies (the computational systems of speakers at a given point in time). An additional point in support of this alternative is that, as has long been known, phonological patterns without apparent phonetic motivation are attested in the languages of the world (Anderson, 1974, 1981, 1985; Buckley, 2000; Chomsky & Halle, 1968; Hyman, 2001). Recent work has also established that such patterns are found in child language (Buckley, 2003) and are not distinguished from more substantively-motivated patterns by infants in certain experimental conditions (Seidl & Buckley, in press).

The goal of this paper is to develop and support a modified version of phonetically based phonology—one that avoids the problems just mentioned—that I refer to as *substantively biased phonology* (see also Steriade, 2001c; Wilson, 2003). In this framework, knowledge of substance acts as a bias (or *prior*) that favors phonological patterns that accord with phonetics. But the bias is not so strong that it excludes phonetically-unmotivated patterns from being acquired or productively applied. The main empirical claim of substantively biased phonology is not that all phonological systems must be phonetically natural, but rather that a bias in favor of substantively motivated phonology will emerge when speakers extend patterns from impoverished input data.²

I focus on one specific type of phonological pattern, referred to throughout as *velar palatalization* and introduced in Section 2; a simple example of the pattern would be the change of pronunciation from *keep* ([kip]) to *cheap* ([tʃip]). The formal development of substantively biased phonology in Section 3 makes use of mathematical methods from the theory of categorization (Luce, 1963; Nosofsky, 1986; Shepard, 1987) and conditional random fields (Lafferty et al., 2001). Perhaps the most original contribution of the paper is an experimental paradigm, dubbed the *poverty of the stimulus method* (PSM), that tests for substantive bias by requiring participants to generalize new phonological patterns based on extremely limited exposure. For example, in one condition of Experiment 1 participants were exposed to instances of velar palatalization before the mid front vowel [e], and were then tested on whether they would generalize the change to new words containing the same vowel and, of most interest, to words containing the high front vowel [i]. Results from both Experiment 1 (Section 4) and Experiment 2 (Section 5) support substantively biased phonology over a formally-equivalent but unbiased alternative, thus defusing the arguments from theoretical simplicity that have been advanced in favor of evolutionary phonology (Blevins, 2004; Hale & Reiss, 2000; Ohala, 1992, 1995). These results have additional consequences for theories of phonological generalization and similarity, as I discuss in Section 6.

2 Background on velar palatalization

For the purposes of this paper, ‘velar palatalization’ refers to the change from a velar stop consonant, voiceless [k] (as in *cap*) or voiced [g] (as in *gap*), to the corresponding palatoalveolar affricate, voiceless [tʃ] (as in *cheap*) or voiced [dʒ] (as in *jeep*), respectively. We will examine velar palatalization before three vowels: the high front vowel [i] (as in *keep*), the mid front [e] (as

²I take the absolute limits on human phonologies to be set by formal properties of the type investigated within OT by Albro (2005), Eisner (1997), and Riggle (2004) and within rule-based phonology by, for example, Reiss (2003). See Frank (2004) for general discussion of formal complexity in grammar.

in *cape*), and the low back vowel [ɑ] (as in *cop*). Simple examples appear in Table 1.³

Table 1 about here

Velar palatalization was selected as the focus of this paper because the articulatory, acoustic, perceptual, and phonological properties of velars and palatoalveolars have been studied extensively. In order to establish the substantive basis for the experiments and modeling efforts that appear later, I now summarize the relevant findings.

Articulation. It is well-known that in many languages the velar stop consonants [k] and [g] are articulated further forward on the palate when they appear immediately before front vowels such as [i] and [e] than when they appear immediately before back vowels such as [ɑ] (Butcher & Tabain, 2004; Keating & Lahiri, 1993; Ladefoged, 2001). Keating & Lahiri (1993) review X-ray and other articulatory evidence of this fronting effect in English and other languages. They conclude that “[t]he more front the vowel, the more front the velar” (p. 89) holds in all of the languages for which data was available. A more recent study by Butcher & Tabain (2004), which investigates several Australian Aboriginal languages as well as Australian English, comes to essentially the same conclusion based on static palatography data (although Butcher & Tabain suggest that their data supports only a binary distinction between non-back and back vowel contexts).

Fronting is relevant here because it makes the articulation of velar stops more similar to that of palatoalveolar affricates. Keating & Lahiri (1993) speculate that there may be additional points of articulatory similarity, especially before the high front vowel [i], but as far as I know these have not been investigated by any subsequent study.

Acoustics. The articulatory similarity of velars and palatoalveolars before front vowels gives rise to an acoustic similarity. As discussed by Keating & Lahiri (1993), the main peak in the spectrum of a consonant release (the brief period of time after the articulatory constriction of the consonant ends) is due to “a front cavity resonance whose frequency value largely depends on the following vowel” (p. 96). Velars before more front vowels have smaller resonant cavities, and therefore higher-frequency peaks, as demonstrated by acoustic measurements in Butcher & Tabain (2004), Guion (1996, 1998), Keating & Lahiri (1993), and many references cited therein. For example, Guion’s (1996, 1998) investigation of American English found that the peak spectral frequency of a velar release is directly proportional to the frontness of the following vowel. The peak is higher before [i] than before [e], and higher before [e] than before [ɑ]. (Note that though [i] and [e] are both phonologically front, [i] is phonetically further front than [e].)

Guion also measured the peak spectral frequencies in the corresponding regions of [tʃ] and [dʒ]. The results show that the affricates have peaks that are approximately constant across vowel contexts, and high relative to those of the velars. It follows that velar stops before more front vowels are more acoustically similar to palatoalveolar affricates, at least with respect to the peak spectral frequency measure.

As was noted in the case of articulation, there are likely to be additional acoustic properties

³Square brackets, as in [ki], indicate broad phonetic transcription. For convenience, the vowel [eɪ] (as in *cape*) is transcribed throughout as [e]. The vowel in *cop*, which I transcribe throughout as back [ɑ], may be closer to central [ɐ] for some speakers. The paper does not require knowledge of any distinctive features beyond [voice], which distinguishes sounds such as [k] and [g], and the features given in Table 1. If desired, most of this section could be skipped on a first reading; the main points are summarized at the end.

that are shared by palatoalveolars and velars before front vowels. For example, the length of frication and aspiration at the release of a velar stop has been found to be proportional to the frontness of the following vowel (see references cited in Guion, 1996, 1998).

Perception. Experiments reported in Guion (1996, 1998) establish further that velars and palatoalveolars are more perceptually similar—more likely to be confused by listeners—before more front vowels. In one of the experiments, native English speakers performed forced-choice identification of consonant-vowel stimuli that were composed of [k, tʃ, g, dʒ] followed by [i, a, u]. The stimuli were excised from faster-speech recordings of English words and truncated so that the duration of the vowel was 100ms. They were played to participants both without masking noise and with white masking noise at a signal-to-noise ratio of +2 dB. Very few identification errors were found in the absence of noise (95% correct responses). In contrast, there were many errors in the presence of noise (69% correct responses), and the error patterns are largely understandable in terms of the articulatory and acoustic evidence reviewed above.

Table 2 reproduces a portion of the confusion matrix data published in Guion (1998: p.35, Table 5). Note that the design of the experiment did not allow participants to report vowel misidentifications, therefore the cells corresponding to such errors have been left blank.

Table 2 about here

As can be seen from the table, the rate at which [ki] is misidentified as [tʃi] is higher than the rate at which [ka] is misidentified as [tʃa]. Similarly, [gi] was misidentified as [dʒi] more often than [ga] was misidentified as [dʒa], though the overall error rate for [g] is lower than that for [k]. Note that the confusion rates are asymmetric; for example, [ki] was identified as [tʃi] 3.5 times more frequently than [tʃi] was identified as [ki]. Asymmetric confusion is not of central interest for this paper, but it does have some consequences, discussed in Section 3, for formal modeling of the confusion data; see Ohala (1997), Plauché et al. (1997), and, more generally, Tversky (1977) for further discussion of asymmetric confusability. Errors in which voicing was confused (e.g., [ki] misidentified as [gi]) were relatively rare, a finding that replicates many other speech perception experiments (e.g., Benkí, 2002), and such errors will not be considered further in this paper.

An earlier study of consonant perception by Winitz et al. (1972) found a high rate of [ki] > [ti] errors in a forced-choice identification task with [p t k] as the possible response options (the stimuli were consonant release bursts excised from their contexts). One could speculate that the listeners in Winitz et al.'s study also misperceived [ki] as something closer to [tʃi], and selected [t] as the available response that was most faithful to their perception.

Phonology. As originally observed by Ohala (1992) and expanded upon by Guion (1996, 1998), there is a striking relationship between the phonetic and perceptual facts reviewed above and two implicational laws that govern velar palatalization (recall Table 1). These laws were revealed by surveys of genetically diverse languages that either have velar palatalization as part of their phonological systems, or that have undergone a velar palatalization sound change during their diachronic development (Bhat, 1978; Chen, 1972, 1973; Guion, 1996, 1998; Neeld, 1973).

The first law is that palatalization before more back vowels asymmetrically implies palatalization before more front vowels. For example, if a language palatalizes velars before the back vowel [a] ([ka] → [tʃa] and [ga] → [dʒa]), then it is also expected to palatalize velars before the front vowels [i] and [e] ([ki] → [tʃi], [gi] → [dʒi], etc.), but not necessarily vice versa. Similarly,

palatalization before mid [e] implies palatalization before high [i] (recall that [i] is phonetically more front than [e]), but not vice versa.

The second law is that palatalization of voiced velars asymmetrically implies palatalization of voiceless velars. In other words, if palatalization applies to voiced [g] in a given vowel context, then it is also expected to apply to voiceless [k] in the same context, but not necessarily vice versa.

Comparing these statements about phonological systems with the confusion matrix in Table 2, we see that the two laws can be given a unified explanation in terms of perceptual similarity (Guion, 1996, 1998; Ohala, 1992). The greater the perceptual similarity of a velar stop and palatoalveolar affricate in a given context (as measured by rate of confusion in noise), the greater the expectation that velar palatalization will apply in that context (in the specific, implicational sense of ‘expectation’ defined by the laws).

Also relevant is the finding that in the lexicons of many languages velar stops co-occur with front vowels, in particular [i], less often than would be expected by chance (Maddieson & Precoda, 1992). This is an instance of a well-known generalization about phonological typology: the same forces that drive changes in some languages (e.g., [ki] → [tʃi]) are visible in the static distribution of sounds in other languages (e.g., relative rarity of [ki]). In the present case, we can trace both types of pattern back to the same relation of perceptual similarity.

Summary. The study of velar palatalization presents us with a near-perfect correlation between substance and phonological patterning. Velar stops and palatoalveolar affricates are more articulatorily, acoustically, and perceptually similar before front vowels (e.g., more similar before [i] than before [e], and more similar before [e] than before [ɑ]), and front vowels condition velar palatalization more strongly in attested phonological systems (i.e., palatalization before a front vowel asymmetrically implies palatalization before a less front vowel that is otherwise identical). Similarity is greater overall for the voiceless stops and affricates than for the voiced ones, and voiceless stops undergo palatalization more easily (i.e., palatalization of voiced velar stops asymmetrically implies palatalization of voiceless velar stops). What is the proper theoretical understanding of this correlation?

In the framework of phonetically based phonology, such findings have been taken to reveal a cognitive principle that privileges alternations between perceptually similar sounds (Steriade, 2001ab; see also Côté, 2000, 2004; Jun, 1995; Wilson, 2001; Zuraw, 2005; and see Chen, 1972, 1973 for a foundational proposal in the same spirit). Thus, for example, the principle favors the change [k] → [tʃ] before [i] over the same change before [ɑ], precisely because the terms related by the change are more similar in the former context than in the latter. According to this view, the observed laws on velar palatalization derive from mental structures (such as rules or rankings of violable constraints) that are in turn shaped by substance.

In contrast, evolutionary phonology takes such correlations to be evidence of the role that substance plays in diachronic change, not in the mental grammars of speakers (Blevins 2004, to appear; Blevins & Garrett, 2004; this is also the view expressed explicitly by Ohala, 1992, 1995 and appears to be the one held by Guion 1996, 1998). Velar palatalization applies more strongly in contexts where velars and palatoalveolars are more similar, according to this view, because those are exactly the contexts in which learners of one generation are most likely to misperceive the velar stops of the previous generation as palatalized. Phonological rules or constraint rankings are symbolic reifications of such misperception patterns (and other types of interpretation/reanalysis that are claimed to be characteristic of language acquisition); they obey implicational laws only

because the underlying error patterns are lawful.

It is unlikely that traditional linguistic description and analysis, though they remain of vital importance to the field as a whole, are sufficient to resolve this particular controversy. The proponents of phonetically based phonology have not been deterred by the fact that many substantively-motivated implicational laws—including those governing velar palatalization (Chen 1972, 1973)—are known to have a small number of exceptions, just as the proponents of evolutionary phonology have not been swayed by the high level of explicitness achieved within the other framework.

The rest of this paper presents two alternative techniques, one computational and one experimental, that are aimed at resolving this impasse. In the next section, I show that the new framework of substantively biased phonology—and in particular the cognitive principle that favors changes involving perceptually similar sounds—can be formalized in a quantitatively precise way. As noted in the introduction, by using substance as a bias rather than an absolute restriction on phonological systems, the framework avoids the incorrect or implausible predictions of the strongest version of phonetically based phonology (e.g., that a child exposed to a language in which velars palatalize only before the low back vowel [ɑ] would somehow fail to acquire this pattern). The formalism easily accommodates different degrees of bias, including none, and therefore allows competing analyses to be compared on a level playing field. This comparison is worked out for the results of the experiments reported in Sections 4 and 5, which involve briefly exposing participants to “language games” involving velar palatalization and testing their generalization of those games. The experiments reveal that participants generalize in a way that accords with the first implicational law discussed above (i.e., palatalization before more back vowels implies palatalization before more front vowels), a result that supports substantively biased phonology over evolutionary and other emergentist alternatives (de Boer, 2001; Kirchner, 2004; Redford et al., 2001).

3 Substantively biased phonology

In this section, I introduce substantively biased phonology in two parts. In the first part, I combine acoustic and confusion-matrix data with the generalized context model of classification (‘GCM’; Nosofsky 1986) to evaluate the perceptual similarity of velar stops and palatoalveolar affricates across vowel contexts. The result is a quantitative version of the *P(erceptual)-map* of Steriade (2001abc), which represents speakers’ knowledge of similarity across phonological contexts. In the second part, I introduce conditional random fields (‘CRF’; Lafferty, et al. 2001), which are a special case of more general maximum entropy / log-linear models, and show how the similarity values derived earlier in the section can function as a source of substantive bias (or prior) on CRF learning. I also discuss qualitative properties of the CRF learning mechanism, in preparation for modeling the experimental results in Sections 4 and 5.

3.1 Quantifying perceptual similarity

The GCM is defined by three equations that relate similarity on measurable stimulus dimensions (e.g., peak spectral frequency) to confusability under identification. The first equation states that the distance d_{ij} between two points x_i and x_j in the space defined by the stimulus dimensions is a weighted function of the difference between x_i and x_j on each dimension.

$$(1) \quad d_{ij} = c \left[\sum_{k=1}^N w_k |x_{ik} - x_{jk}|^r \right]^{1/r}$$

The index k runs over the stimulus dimensions (e.g., x_{ik} is the value of stimulus x_i on dimension k). Three dimensions were used in the simulations presented here: a binary-valued voicing dimension (0 = voiceless, 1 = voiced), a binary-valued vowel dimension (0 = [i], 1 = [a]), and a real-valued peak spectral frequency dimension (see Section 2 for discussion of this measure). Clearly it is the third dimension that is of central interest; the other two were included in order to allow a single model to be fit to a confusion matrix.

Each stimulus dimension has an attention weight w_k . The weights on all dimensions are constrained to be non-negative and to sum to unity ($\sum_{k=1}^N w_k = 1$). There is also a scale parameter c , constrained to be non-negative, that relates to the overall level of discriminability among elements of the stimulus space (larger c corresponds to greater ‘stretching’ of the space). The r parameter, which controls how the three stimulus dimensions interact with one another, was set to 2 (corresponding to the Euclidean distance metric). This setting reflects the assumption that at least some of the stimulus dimensions are *integral* (that is, not perceived separately from one another; Nosofsky 1986). And indeed, Benkí (1998) has established that place of articulation (here, velar vs. palatoalveolar) and voicing (voiceless vs. voiced) are perceived in a non-separable fashion. (Similar results were obtained with $r = 1$, which corresponds to the city-block metric that is appropriate when stimulus dimensions are assumed to be perceptually separable.)

The second GCM equation expresses the well-known finding that perceptual similarity η_{ij} between two points x_i and x_j falls exponentially as the distance between the points increases (Nosofsky 1984, 1986, Shepard 1957, 1987).

$$(2) \quad \eta_{ij} = \exp(-d_{ij})$$

Nosofsky (1986) gives a more general version of this equation, in which the distance d_{ij} is raised to a power p within the exponential, but the special case given in Eqn. (2) was found to be sufficient for present purposes.

The final equation projects perceptual similarities onto predicted confusion rates according to the Luce choice rule (Luce 1963, Shepard 1957). The probability of response x_j given stimulus x_i is a function of the perceived similarity of x_i and x_j , relative to the perceived similarity of x_i and all of the possible responses.⁴

$$(3) \quad P(\text{response} = x_j | \text{stimulus} = x_i) = \frac{b_j \eta_{ij}}{\sum_{k=1}^n b_k \eta_{ik}}$$

This equation presupposes that every stimulus x_k has an associated response bias b_j (all of the response biases are required to be non-negative and to sum to unity: $\sum_{k=1}^n b_k = 1$). In the present

⁴I abuse notation by identifying a stimulus category with one of its members. Because all of the stimulus-dimension values employed here were averages over multiple tokens, some mixing of type/token levels is unavoidable.

case, the response bias parameters allow the model to capture the finding, noted in Section 2, that velar/palatoalveolar confusion rates are asymmetric (e.g., [k] is misidentified as [tʃ] much more often than [tʃ] is misidentified as [k]). This is probably undesirable as an ultimate account of the asymmetry — among other considerations, the relative frequencies of velars and palatoalveolars in words of English might suggest a response bias in the opposite direction — but it does provide a provisional solution that is compatible with the underlying symmetry assumptions of the GCM.

Given the values that a set of items take on the stimulus dimensions, and a confusion matrix over the same items, perceptual similarities can be inferred from Eqns. (1), (2), and (3) with the maximum likelihood ('ML') method (Nosofsky, 1996; Nosofsky & Zaki, 2002; see Myung, 2003 for a general presentation). The likelihood equation used here was the one given in Nosofsky & Zaki (2002, p. 930); optimization was performed with the `optim` method of the R statistical package (R Core Development Team 2005). The free parameters were the similarities ($\{\eta_{ij}\}$), attention weights ($\{w_k\}$), response biases ($\{b_j\}$), and the scale (c).

The confusion matrix given in Table 2 above and stimulus values for tokens of [ki, tʃi, ka, tʃa, gi, dʒi, ga, dʒa] were entered into the model. The values for the voicing and vowel dimensions were dummy-coded, as already described. The values for the peak spectral frequency dimension were taken from average data published in Guion (1996). Similar results were obtained with peak spectral frequencies that were measured from the stimulus items of the experiments reported in Sections 4 and 5. All spectral frequencies were converted to the auditory Bark scale with Traunmüller's approximation $(26.81/(1 + (1960/f)) - .53$; Traunmüller, 1990). The resulting predicted confusion matrix was qualitatively similar to the observed matrix. The ML estimate of the parameters had a log likelihood of -64.6, and the Kullback Leibler distance (Cover & Thomas 1991) between the observed confusion proportions and the predicted confusion probabilities was 1.44 (the minimum possible value is 0).

Because Guion's (1998) confusion matrix contains information for the vowels [i] and [a], but not [e], only the perceptual similarities of [ki]/[tʃi], [gi]/[dʒi], [ka]/[tʃa], and [ga]/[dʒa] could be directly assessed. These (unitless) values are given in Table 3. Note that the values are the perceptual similarities $\{\eta_{ij}\}$ multiplied by the appropriate response bias terms $\{b_j\}$; as discussed above, response biases are employed here to capture the fact that confusion rates are asymmetric within a pair. The perceptual similarities of the remaining pairs [ke]/[tʃe] and [ge]/[dʒe] were determined by interpolation from the ML fit. The peak spectral frequencies for velars and palatoalveolars before [e] were taken from Guion's (1996, 1998) data, and the response biases for [tʃe] and [dʒe] were set equal to those for [tʃi] and [dʒi], respectively. The resulting values, marked with italicization, are also given in Table 3. (As before, similar values were obtained using measurements from the stimuli in the present experiments.)

Table 3 about here

Notice that, as expected from the confusion data and the distribution of velar palatalization in natural languages, voiceless velars and palatoalveolars are more similar overall than the corresponding voiced sounds, and within a voicing category similarity decreases with vowel frontness (i.e., from high front [i] to mid front [e] to low back [a]). We will return to these values at the end of the next subsection.

3.2 Conditional random fields for phonology

Lafferty et al. (2001) introduce a general framework, referred to as *conditional random fields* (CRF), and apply it to the problem of labeling sequences (see also Gregory & Altun, 2004; McCallum 2003; Roark et al., 2004; Sha & Pereira, 2003; among others). Many attested phenomena in phonology can be considered as types of labeling, therefore applying CRF models to phonology is a promising direction for research. For example, consider a grammar that would be standardly described as mapping the hypothetical input sequence /kinə/ to the output sequence [tʃinə] (where [ə] is the final vowel in *rhumba*). This grammar can also be thought of as assigning an output label to each sound in the input: /k/:[tʃ], /i/:[i], /n/:[n], /ə/:[ə]. Indeed, a much richer labeling system known as Correspondence Theory (McCarthy & Prince, 1999), which allows transposition and multiple labeling, has become standard in work on phonology within OT.⁵

In the most general terms, a CRF defines a probability distribution over a set of output random variables \mathbf{y} given values for a set of input random variables \mathbf{x} . Each output variable takes on a value in the finite set \mathcal{Y} . In the present setting, we identify the input variables \mathbf{x} with the sequence of sounds in one phonological form (the input) and the output variables \mathbf{y} with the sequence of sounds in a possibly different phonological form (the output). The set \mathcal{Y} is the set of all possible phonological segments, possibly expanded to include a special symbol representing deletion.

The defining structural property of CRF models is that the relationships among the output variables are described by an undirected graph. There is a one-to-one correspondence between the output variables and the vertices of the graph, and an output variable y_i can probabilistically depend on another output variable y_j , given the input variables \mathbf{x} , only if the corresponding vertices are connected by an edge in the graph. In other words, the output variables satisfy the Markov property, when conditioned on the input variables, with respect to the graph underlying the CRF. A simple graphical structure, discussed by Lafferty et al. (2001) and sufficient for present purposes, arranges the output vertices into a chain. Given an input sequence $\mathbf{x} = \langle x_1, x_2, \dots, x_n \rangle$, there is a corresponding output sequence $\mathbf{y} = \langle y_1, y_2, \dots, y_n \rangle$. Each output variable in the sequence corresponds to a vertex in the graph, and edges between vertices represent adjacency in the output sequence (i.e., there is an edge for each pair (y_i, y_{i+1}) , $1 \leq i \leq n - 1$). In this setting, the input-output mapping of /kinə/ to [tʃinə] can be written with coindexation: /k₁i₂n₃ə₄/. [tʃ₁i₂n₃ə₄].

As demonstrated by Hammersley & Clifford (1971), the joint probability distribution that a CRF defines over the output variables is equivalent to the Gibbs distribution (see also Geman & Geman, 1984; Smolensky, 1986):

$$(4) \quad P(\mathbf{y}|\mathbf{x}) = Z_{\mathbf{x}}^{-1} \prod_{c \in \mathcal{C}} \exp\left(\sum_{k=1}^K \lambda_k f_k(\mathbf{x}_c, \mathbf{y}_c)\right)$$

where \mathcal{C} is the set of cliques in the graph, $\exp(\sum_{k=1}^K \lambda_k f_k(\mathbf{x}_c, \mathbf{y}_c))$ is the potential on clique c

⁵The main limitation of the CRF model as an approach to phonology is that it cannot accommodate epenthesis (insertion of sounds) without an a-priori bound on the number of epenthetic segments. Goldwater & Johnson (2003) present a more general maximum-entropy model that does not have this limitation, but do not discuss how the probability distribution over the resulting (infinite) set of possible labelings/outputs is approximated. In current research, I am applying standard Markov Chain Monte Carlo methods to this problem.

if \varnothing one segment

(discussed further below), and Z_x is the partition function with respect to input x :

$$(5) \quad Z_x = \sum_{y'} \prod_{c \in \mathcal{C}} \exp\left(\sum_{k=1}^K \lambda_k f_k(\mathbf{x}_c, \mathbf{y}'_c)\right)$$

Eqns. (4) and (5) define the probability of the output y , given the input x , by comparing y to all possible outputs y' for the same input. This is the probabilistic analogue of the OT claim that the grammatical output is selected by competition among all possible candidate outputs.

The clique potentials $\exp(\sum_{k=1}^K \lambda_k f_k(\mathbf{x}_c, \mathbf{y}_c))$ have a close relationship to the notion of *harmony* in OT—and an even closer relationship to harmony in Harmony Theory ('HT'; Smolensky, 1986; Smolensky & Legendre, 2005)—therefore I will refer to them with the term *CRF-harmony*. CRF-harmony is defined in terms of a set of functions $\{f_k\}$, each of which evaluates input/output pairs.

In standard CRF terminology, the f_k functions are referred to as *features*, but we will think of them as *constraints* like those in OT. Each f_k is a function from input-output mappings (more precisely, cliques in the input-output mapping) to the non-negative integers; this conception of constraints as functions from candidates to violation levels is familiar from Eisner (1997), Samek-Lodovici & Prince (199), and others. Each constraint has a real-valued weight λ_k , which we will take throughout to be non-positive, thereby expressing the central OT tenet that constraint violations *decrease* harmony—that is, in the probabilistic setting, more violations imply lower probability. The entire set of weights for K constraints will be denoted by Λ .

To assess the CRF-harmony of any given pair (x, y) , we do essentially this. Find the number of violations that the pair incurs on each constraint ($f_k(x, y)$). Multiply each violation score by the corresponding weight ($\lambda_k f_k(x, y)$). Sum the weighted violations ($\sum_{k=1}^K \lambda_k f_k(x, y)$). And finally raise the natural number e to that sum (written $\exp \sum_{k=1}^K \lambda_k f_k(x, y)$). Technically this must be done separately for each clique in the graph, with the values for the cliques multiplied together as shown in Eqns. (4) and (5). But suffice it to say that, with the simple graph structure assumed above (i.e., a chain graph), the result is the same as long as the constraints satisfy certain locality conditions. Among the allowable constraint types are those that assess violations for pairs of adjacent segments in the output (e.g., [k] followed immediately by [i]) and those that assess violations for single elements of the input/output mapping (e.g., /k/:[tj]).

One main difference between the CRF model and OT lies in the way that constraint violations are combined into a harmony score. CRF-harmony is an exponential function of the weighted sum of constraint violations, much as in HT: a pair (x, y) is more harmonic than another (x, y') iff the value of the CRF-harmony is greater for the former than for the latter. OT-harmony, on the other hand, is determined by lexicographic comparison of constraint violations: a pair (x, y) is more harmonic than another (x, y') iff the highest-ranked constraint that distinguishes between the two pairs prefers the latter (Prince & Smolensky, 1993/2005). (OT rankings could also be expressed with real-valued weights but only the ordering of the weights would be relevant.)

There were two motivations in the present context for adopting a CRF rather than an OT approach. First, CRFs generate probability distributions over candidate outputs, and therefore hold the promise of yielding precise quantitative matches to the stochastic behavior of the participants in the experiments reported later in the paper. Second, there is a provably (asymptotically) correct algorithm that converges on the globally optimal CRF weights given a body of training data

(Lafferty et al. 2001; see Boyd & Vandenberghe, 2004 for the general picture). No current instantiation of OT has both of these advantages. While the original formulation of the theory by Prince & Smolensky (1993/2004) has a correct and convergent ranking algorithm (Tesar & Smolensky, 1998; 2004), it does not readily generate probability distributions. The alternative formulation known as stochastic OT (Boersma 1998; Boersma & Hayes 2001) is explicitly probabilistic, but the learning algorithm supplied with it has no correctness or convergence proofs and is known to fail to converge in practice (Bruce Hayes, p.c.). The same learning problem holds, as far as I know, for other varieties of OT that define probability distributions.⁶

CRF learning in this paper was performed by minimizing the objective function in Eqn. (6) (Goldwater & Johnson, 2003; Lafferty et al., 2001; McCallum, 2003), which assumes a body of training data D that consists of N input-output pairs ($D = \{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \dots, (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})\}$).

$$(6) \quad L_{\Lambda} = \left[- \sum_{j=1}^N \log P_{\Lambda}(\mathbf{y}^{(j)} | \mathbf{x}^{(j)}) \right] - \left[\sum_{k=1}^K \frac{(\lambda_k - \mu_k)^2}{2\sigma_k^2} \right]$$

This equation defines the likelihood of the weights (L_{Λ}) as a function of two bracketed terms, which have interpretations that are familiar from the theory of induction (e.g., Grünwald et al., 2005; Smolensky, 1996). The first term is the negative log probability, given the weights Λ , of the outputs in the data $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(N)}\}$ given the inputs $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}\}$. Minimizing this term is equivalent to finding the weights that maximize the probability of the observed outputs given the corresponding inputs. The second bracketed term is a Gaussian prior, or regularizer, on the weights (Chen & Rosenfeld, 1999). For each weight λ_k the regularizer specifies a target value (μ_k) and imposes a penalty for deviating from that value. An important relationship in the following will be that **smaller values of σ_k yield greater penalties for deviating from μ_k** . As we will see, substantive bias can be injected into the CRF model by assigning different σ values to different constraints.

3.2.1 Constraints on velar palatalization

A particular application of the CRF model of phonology is characterized by a specific set of constraints. For the purposes of analyzing the experimental results in Sections 4 and 5, I have found it sufficient to adopt a relatively small set of a-priori constraints on velar palatalization. (The strategy of assuming a known constraint set, rather than inducing constraints from the data, is familiar from OT but not standard in work on random fields. **One goal of future research is to develop a model that is able to induce both the constraints and their weights.**)

The constraints fall into two classes, as is standard within OT phonology. The first, *Faithfulness* class contains constraints that are violated when an input variable x_i and the corresponding output variable y_i have different values. For empirical reasons discussed later, I assume that the velar stops k and g are subject to two different Faithfulness constraints, each one violated by velar palatalization. $F(k)$ is violated when input /k/ corresponds to output [tʃ]; $F(g)$ is violated when input /g/ corresponds to output [dʒ]. I assume further that all other input-output disparities run afoul

⁶Recent results of Lin (2004) provide a solution to this problem for stochastic OT in particular, but I have not had the opportunity to apply Lin's methods to the present problem.

of an inviolable faithfulness constraint. This is not a realistic assumption for all of phonology, of course, but it accords with the design and results of the present experiments.

The second, *Markedness* class contains the constraints shown in Table 4. Each of these constraints has the form $*\alpha\phi$, where α is one of the velar stops ([k] or [g]) and ϕ is either a single vowel ([i], [e], or [a]) or a class of vowels. ‘V’ stands for the class of all vowels; the other classes can be derived from Table 1. The Markedness constraints are violated by velar stops—and satisfied by palatoalveolar affricates—that appear immediately before the designated vowels in the output. (The other information in the table is explained in the following subsection.)

Table 4 about here

With the constraints in hand, we can now distill the analysis of velar palatalization down to essentials. Given an input form that begins with a velar stop (e.g., /k₁i₂n₃ə₄/), the inviolable Faithfulness constraint eliminates all but two of the logically-possible candidate outputs: the fully-faithful candidate (e.g., [k₁i₂n₃ə₄]) and the candidate that is identical to the input except that the velar has been replaced with a palatoalveolar of the same [voice] specification (e.g., [tʃ₁i₂n₃ə₄]). The faithful candidate satisfies $F(\alpha)$, where α is the initial velar in the input, but it violates one or more of the Markedness constraints (e.g., [k₁i₂n₃ə₄] violates *ki, *kV_[-low], and *kV). Conversely, the velar palatalization candidate violates $F(\alpha)$, but satisfies the Markedness constraints completely.

In OT, the relative ranking of these Markedness and Faithfulness constraints would determine a unique output for each input. For example, if $F(k)$ were to dominate all three of *ki, *kV_[-low], and *kV, then velar palatalization could not apply to the /k/ in /kinə/; the grammatical output would be the faithful one. In contrast, the numerical weights on the constraints in the CRF model do not determine an absolute winner. Instead, they define a probability distribution over the two candidates. This distribution is described by Eqn. (7), which is the special case of Eqns. (4) and (5) when there are exactly two candidates. I have made the substitution $H(\mathbf{x}, \mathbf{y}) = \prod_{c \in \mathcal{C}} \exp(\sum_{k=1}^K \lambda_k f_k(\mathbf{x}_c, \mathbf{y}_c))$ in order to bring out the essence of the competition.

$$(7) \quad P(\mathbf{y}^{pal} | \mathbf{x}) = \frac{H(\mathbf{x}, \mathbf{y}^{pal})}{H(\mathbf{x}, \mathbf{y}^{pal}) + H(\mathbf{x}, \mathbf{y}^{faith})}$$

(making more negative)

Recalling that stronger constraints have weights that are further below 0, we see that decreasing the weight of Faithfulness relative to Markedness makes the palatalization output \mathbf{y}^{pal} less probable. Conversely, decreasing the weight of Markedness relative to Faithfulness makes the palatalization output \mathbf{y}^{pal} more probable.

3.2.2 Biased instantiation

I now bring together the two strands of this section to complete the formulation of substantively biased phonology. The type of bias studied here, due to Steriade (2001ab) and others, is the proposed cognitive preference for changes involving sounds that are more perceptually similar. For example, the bias should assign a lower cost to the change [ki] → [tʃi] than to the change [ka] → [tʃa].

The key idea is to impose a systematic relationship between similarity values (as in Table 3) and the σ parameters for the Markedness constraints in the CRF. Recall that the Markedness constraints are the ones that force phonological changes, in this case by eliminating velar stops before certain vowels. Recall also that the smaller the σ_k for a given constraint f_k , the more tightly the prior/regularizer holds the constraint's weight to the target value μ_k . If Markedness constraints that force alternations among *less* perceptually similar sounds are assumed to be subject to *greater* pressure to remain at their target weights, then we are close to embodying the desired bias.

Specifically, I propose that the σ_k for a Markedness constraint f_k is determined according to the following steps. First find all of the changes that can be forced by f_k (here, changes are of the type velar stop \rightarrow palatoalveolar affricate in a particular vowel context). Among these, find the change that relates the sounds that are *least* perceptually similar, in the given context. Suppose that $b_j \eta_{ij}$ is the perceptual similarity, multiplied by the response bias of the outcome, for the sounds involved in that change. Set σ_k equal to that value. In short, the prior σ of a Markedness constraint is equal to the perceptual similarity of the sounds in the greatest change that is motivated by the constraint. The columns labeled 'Biased' in Table 4 give the σ s for the Markedness constraints assumed here.

The target weight value μ for each constraint must also be specified, and these values depend on the particular human population whose phonological learning we are trying to model. For adult native speakers of English, a language that does not have a productive process of velar palatalization, one natural possibility is that all of the Markedness have a target weight $\mu_M = 0$ whereas all of the Faithfulness constraint have a target weight that is substantially more negative. In the simulations reported here, I used the values $\mu_F = -10$ and $\sigma_F = 10^{-2}$ for all of the Faithfulness constraints. The latter value is important only insofar as it gives the Faithfulness weights greater overall freedom of movement than the Markedness weights.⁷

To aid understanding of the qualitative properties of learning and generalization in the CRF model of phonology, Fig. 1 shows the forces that apply to the weights when, starting from the adult state, the model is exposed to training data in which a velar stop α (i.e., [k] or [g]) undergoes palatalization in context K (i.e., before one of the three vowels [i], [e], or [a]).

Figure 1 about here

Handwritten note: \downarrow ? w/ weights < 0 ?

The force labeled D shown in the figure is due to the training data. It pulls the weight of $M(\alpha/K)$ upward (away from 0) and the weight of $F(\alpha)$ downward (toward 0). Given sufficient training data, the system will learn that α palatalizes in context K . What else the system learns depends on the relative size of the prior forces $\sigma_{M(\alpha/K)}$ and $\sigma_{F(\alpha)}$.

Handwritten note: confusion w/ [k]?

Case I. If $\sigma_{M(\alpha/K)}$ and $\sigma_{F(\alpha)}$ are of comparable size, then the system learns nothing beyond palatalization of α in K . The weight of $F(\alpha)$ lowers at roughly the same rate that $M(\alpha/K)$ rises, with the consequence that at the end of learning $F(\alpha)$'s weight is still above the weights of the other Markedness constraints. Thus those constraints remain too weak, relative to Faithfulness, to cause any other type of palatalization.

Case II. Things work out differently if $\sigma_{M(\alpha/K)}$ is substantially smaller than $\sigma_{F(\alpha)}$. The greater prior on the Markedness constraint prevents its weight from being displaced too far from

⁷Another possibility, not yet explored, would be to set the adult μ_M and μ_F values in a way that models the relative frequencies of velar stops and palatoalveolar affricates in the lexicon of English. The proper settings of μ_M and μ_F for a child acquiring her first language are discussed in a companion paper.

The key idea is to impose a systematic relationship between similarity values (as in Table 3) and the σ parameters for the Markedness constraints in the CRF. Recall that the Markedness constraints are the ones that force phonological changes, in this case by eliminating velar stops before certain vowels. Recall also that the greater the σ_k for a given constraint f_k , the more tightly the prior/regularizer holds the constraint's weight to the target value μ_k . If Markedness constraints that force alternations among *less* perceptually similar sounds are assumed to be subject to *greater* pressure to remain at their target weights, then we are close to embodying the desired bias.

Specifically, I propose that the σ_k for a Markedness constraint f_k is determined according to the following steps. First find all of the changes that can be forced by f_k (here, changes are of the type velar stop \rightarrow palatoalveolar affricate in a particular vowel context). Among these, find the change that relates the sounds that are *least* perceptually similar, in the given context. Suppose that $b_j\eta_{ij}$ is the perceptual similarity, multiplied by the response bias of the outcome, for the sounds involved in that change. Finally, set σ_k equal to the inverse of that value, $(b_j\eta_{ij})^{-1}$. In short, the prior σ of a Markedness constraint is determined by the perceptual 'cost' of the greatest change that is motivated by the constraint. The columns labeled 'Biased' in Table 4 give the σ s for the Markedness constraints assumed here.

The target weight value μ for each constraint must also be specified, and these values depend on the particular human population whose phonological learning we are trying to model. For adult native speakers of English, a language that does not have a productive process of velar palatalization, one natural possibility is that all of the Markedness have a target weight $\mu_M = 0$ whereas all of the Faithfulness constraint have a target weight that is substantially more negative. In the simulations reported here, I used the values $\mu_F = -10$ and $\sigma_F = 10$ for all of the Faithfulness constraints. The latter value is important only insofar as it gives the Faithfulness weights greater overall freedom of movement than the Markedness weights.⁷

To aid understanding of the qualitative properties of learning and generalization in the CRF model of phonology, Fig. 1 shows the forces that apply to the weights when, starting from the adult state, the model is exposed to training data in which a velar stop α (i.e., [k] or [g]) undergoes palatalization in context K (i.e., before one of the three vowels [i], [e], or [a]).

Figure 1 about here

The force labeled D shown in the figure is due to the training data. It pulls the weight of $M(\alpha/K)$ upward (away from 0) and the weight of $F(\alpha)$ downward (toward 0). Given sufficient training data, the system will learn that α palatalizes in context K . What else the system learns depends on the relative size of the prior forces $\sigma_{M(\alpha/K)}$ and $\sigma_{F(\alpha)}$.

Case I. If $\sigma_{M(\alpha/K)}$ and $\sigma_{F(\alpha)}$ are of comparable size, then the system learns nothing beyond palatalization of α in K . The weight of $F(\alpha)$ lowers at roughly the same rate that $M(\alpha/K)$ rises, with the consequence that at the end of learning $F(\alpha)$'s weight is still above the weights of the other Markedness constraints. Thus those constraints remain too weak, relative to Faithfulness, to cause any other type of palatalization.

Case II. Things work out differently if $\sigma_{M(\alpha/K)}$ is substantially greater than $\sigma_{F(\alpha)}$. The

⁷Another possibility, not yet explored, would be to set the adult μ_M and μ_F values in a way that models the relative frequencies of velar stops and palatoalveolar affricates in the lexicon of English. The proper settings of μ_M and μ_F for a child acquiring her first language are discussed in a companion paper.

greater prior on the Markedness constraint prevents its weight from being displaced too far from μ_M , therefore the weight of the Faithfulness constraint must compensate by descending further. If the $F(\alpha)$ weight lowers so far that it becomes roughly equivalent to the weight of $M(\alpha/K')$, then the system will to some extent generalize palatalization of α from the context K to the new context K' , even though no examples of palatalization in K' appeared in the training data. I will refer to this type of behavior as *generalization on the context*.

The link between perceptual similarity and σ values tells us when each case will apply. For example, if the system is exposed to palatalization of [k] before [i], we expect Case I behavior. The prior forces on $M(ki)$ (= *ki) and $F(k)$ are approximately the same, therefore no substantial generalization should result. On the other hand, if the system is exposed to palatalization of [g] before [e], then Case II behavior is expected. The prior force on $M(ge)$ (= *ge) is substantially greater than that on $F(g)$, therefore a small degree of generalization should be found. Note that the prediction is generalization of [g] palatalization to two environments—both [i] and [a]—because $F(g)$ will descend within range of both *gi and *ga. Generalization to the [i] context would be predicted under any sensible implementation of substantive bias; generalization to the [a] context is a more subtle consequence of the current implementation, one we will see to be borne out in Experiment 1 (see Section 4).

The system makes a further prediction, namely that what I will refer to as *generalization on the target* should not occur. This type of generalization would involve extending a palatalization process that applies to one velar stop α (e.g., [k]) to another velar stop β (e.g., [g]). Such generalization is impossible in the current system because the Faithfulness constraints that apply to the two velar stops are distinct. This prediction is also borne out in both experiments reported below.

To summarize, the substantively biased model of phonology developed above makes detailed quantitative predictions about patterns of learning and generalization. The predictions can to a certain extent be subject to qualitative analysis by considering the various forces that act on constraint weights during learning. The predictions concern types of generalization that should be found and, of equal importance, types that should not. The predictions are asymmetric, mirroring the asymmetries of substance, and follow in a non-trivial way from the representations and computations of the model.

3.2.3 Unbiased instantiation

As a formalism, the CRF model of phonology is equally compatible with a prior that is not substantively biased. In Sections 4 and 5 I compare the biased instantiation described above with an unbiased instantiation in which the σ values for all constraints, both Markedness and Faithfulness, are equal. Table 4 gives the values assumed in the unbiased version.

Lack of bias in the model leads to absence of asymmetry in the predictions. The unbiased instantiation learns any velar palatalization pattern just as easily as any other, and predicts that the pattern in the training data will be generalized to new words but not to new contexts or targets. We turn now to the experimental evidence against this more empiricist theory of phonological learning. (An alternative implementation, one that would not fare better on the experimental results, would predict an equal rate of generalization for all types of training data. The fundamental point is that only by referring to substance can we correctly predict when generalization occurs and when it does not.)

4 Experiment 1: Testing generalization on the context

Language games are naturally-occurring phenomena in which the pronunciation of words are altered in systematic ways, and that often have the purpose of disguising speech or indicating group membership (Bagemihl, 1995). An important inspiration for the current experiments comes from McCarthy (1981), which shows that games found in nature shed light on the mechanisms by which learners generalize from impoverished input. Previous studies that have used experimentally-constructed language games include Pierrehumbert & Nair (1995) and Treiman (1993).

The experimental method employed here is fairly straightforward and directly motivated by the model introduced in Section 3. Participants are first presented with spoken examples of a novel language game. For example, one group of participants in the present experiment heard examples such as [kenə] ... [tʃenə] and [gɛpə] ... [dʒɛpə], which illustrate palatalization of velar stops before the mid vowel [e] (‘...’ indicates a short pause; all stimuli were nonwords). Importantly, the same participants were not presented with any examples in which the velar stops [k g] appeared before the high front vowel [i]. Then, in the second part of the experiment, participants were tested on whether they would apply velar palatalization before [e], in both previously-heard and new nonwords, and before the vowel [i]. I refer to the latter as the *novel* context. If participants exposed to examples such as [kenə] ... [tʃenə] extend the pattern to the novel context as in [kiwə] ... [tʃiwə], but not vice versa, this will provide strong evidence for the substantively biased system developed in Section 3.

This method deliberately withholds crucial information—in this case, whether palatalization applies in the novel context—and thereby forces participants to rely upon their ability to generalize from limited exposure to a new phonological pattern. I therefore refer to it as the *poverty of the stimulus method* (PSM). The issue of whether natural-language input is highly impoverished remains a contentious one (Blevins, 2004; Idsardi, 2005; Pullum & Scholz, 2002), but there can be no debate about the degree of impoverishment in a PSM experiment (although of course adult participants will bring a wealth of knowledge of their native language to bear on the task). This method is therefore exactly the right one to test claims about mechanisms of learning and generalization such as those posited in substantively biased phonology.

The present experiment tested for generalization on the context.

4.1 Methods

4.1.1 Stimuli

The stimuli were pairs of $C_1V_1C_2V_2$ nonwords (where ‘C’ stands for consonant and ‘V’ for vowel). Lexical stress was always on the initial syllable, and the final vowel (V_2) was always schwa ([ə], as in *rhumba*). Within a pair, the first vowel (V_1) and the second consonant (C_2) were held constant. V_1 was drawn from the set [i e ə]. C_2 came from [p b k g m n f v θ ð s z tʃ dʒ l r w], which is a sizable subset of the English consonants. (The sound [θ] is as in *think*) and [ð] is as in *that*. With the exception of the palatoalveolar affricates [tʃ] and [dʒ], which have already been discussed, all of the other consonants were pronounced as expected from English orthography.)

In the first member of a pair of nonwords, the initial consonant (C_1) was drawn from the set [p b k g]. Items that began with [k] or [g] (i.e., one of the two velar stops) are referred to as

critical items. Items that began with [p] or [b] (i.e., one of the two labial stops) are referred to as *fillers*. The possible initial consonants ([p b k g]) and possible first vowels ([i e a]) were fully crossed in the stimulus set. For each C₁V₁ combination, a phonetically balanced set of following C₂s was selected from the specified inventory of second consonants. This resulted in a set of 82 total nonwords that served as the first members of stimulus pairs.

The second member of a stimulus pair was either phonologically identical to the first member, or differed from it by the application of velar palatalization to the initial consonant (i.e., [k] → [tʃ] or [g] → [dʒ]). No change was ever applied to items beginning with [p b]. Application of palatalization always resulted in a nonword.

The stimuli were recorded by a phonetically trained native American English speaker who was naive to the purpose of the experiment. Recordings were conducted in the soundbooth of the UCLA Phonetics Lab. All stimuli were spoken in the standard frame “Say ____ again” with no pauses between words. The first and second members of each pair were recorded separately, even when they were phonologically identical across all conditions. Individual stimulus items were excised from the recordings and their amplitudes were normalized.

A complete list of the stimuli for Experiment 1 appears in Appendix A.

4.1.2 Procedure

There were two conditions (High, Mid), with four experimental phases in each condition (Practice, Exposure, Break, Testing). During the experiment participants were seated in front of a desktop computer in a sound attenuated booth in the UCLA Phonetics Lab. Stimuli were played through two speakers at the sides of the computer; speaker volume was constant for all participants. Stimulus presentation was controlled by PsyScope (Cohen et al. 1993), with timing performed by the PsyScope Button Box.

At the beginning of the experiment, participants were told that the computer would teach them a new language game by presenting them with spoken examples. They were told that a language game could be thought of as a way of pronouncing certain words, and that to play the game they would first listen to a word that the computer said and then give a spoken response. Participants were told that all of the words in the experiment were made-up and not intended to be words of English or any other language. They were not given any additional information about the experimental stimuli, and the instructions included no reference to rules, constraints, patterns, or generalizations. The procedure for trials in the Practice and Exposure phases were as follows:

1. A trial began with a 2s period of silence during which the computer screen was blank.
2. A rectangle containing the text “I say . . .” appeared on the left side of the screen.
3. 250ms later, the first member of a stimulus pair was played from the speakers.
4. There was a 1s interstimulus interval (ISI) that began at the end of the stimulus. The text box remained on the screen during the ISI.
5. The first text box was removed from the screen and a rectangle containing the text “You say . . .” appeared on the right side of the screen.
6. 250ms later, the second member of a stimulus pair was played from the speakers.

7. The participant repeated the second member of the stimulus pair (i.e., repeated the word corresponding to his/her response). Participants were directed to repeat this word in the instructions, with the explanation that doing so would help them to learn the game.
8. A trial ended when the participant pressed the spacebar on the computer keyboard.

There were two practice trials, one in which the members of the stimulus pair were phonologically identical ([bələ] ... [bələ]), and one that illustrated velar palatalization (High: [gibə] ... [dʒibə]; Mid: [gefə] ... [dʒefə]).⁸ The Practice phase was followed by 32 Exposure trials, as schematized in Table 5. The trials in the Exposure phase were divided into 4 blocks of 8 trials each. A block contained 2 examples of velar palatalization (one each of [k] → [tʃ] and [g] → [dʒ]), 2 examples of velars that did not palatalize (one instance each of [k] and [g] before [ɑ]), and 4 fillers. The order of the blocks and the order of items within blocks were randomized across participants. The stimulus sets for the High and Mid conditions differed only in the items that illustrated velar palatalization. No stimulus was repeated during the first part of the experiment (i.e., the Practice and Exposure phases).

Table 5 about here

After the Exposure phase, there was a 2min break during which participants worked on pencil-and-paper math problems. The problems were designed to be of moderate difficulty (multiplication of two 3-digit numbers) and were identical across all participants. Participants were informed that the problems were designed to occupy their minds during the break, but would not play any other role in the experiment. The computer played a brief tone to signal the conclusion of the 2min period and the beginning of the Testing phase.

The instructions at the beginning of the experiment made the participants aware that there would be a Testing phase, and directed them to play the game in this phase by using their intuition based on the examples that they had heard in the first part of the experiment. A screen of instructions at the beginning of the Testing phase reiterated these directions.⁹

The procedure for the testing trials was identical to that of the practice and exposure trials, except that steps (6) and (7) were replaced with (6'):

- 6' After the rectangle containing the text “You say ...” appeared on the screen, the participant generated a response to the word that the computer had played in step (3).

There were 80 testing trials. The stimuli consisted of the full set of 82 original nonwords that were constructed for the experiment (i.e., the first member of each stimulus pair), with the 2 critical items used for the Practice trials removed. The testing list was thus exactly the same for both conditions. The trials were distributed as schematized in Table 6 and presented in an order that was randomized for each participant without blocking.

⁸The fact that the practice items illustrated palatalization of voiced [g] only was a deliberate design feature, as an earlier experiment (described in a companion paper) had used practice items that illustrated palatalization only of voiceless [k]. The nature of the practice items has a measurable effect on participants' behavior, as noted below.

⁹Note that the word “testing” did not appear in any of the instructions. Rather, the Testing phase was referred to as simply the “second part” of the experiment, and participants were told that they would “play the game with the computer” in that part. Participants were also assured that their responses would not be judged as correct or incorrect.

Table 6 about here

In the *kaCV*, *gaCV*, and filler categories, half of the testing items were identical to stimuli that had been presented during the Exposure phase; these were identical for both conditions. Thus, for example, there were three testing items of the type *baCV* that all participants heard in the Exposure phase of the experiment, and three testing items of the same type that were novel for all participants. In addition, half of the *kiCV* and *giCV* testing items were identical to exposure items for the High group, just as half of the *keCV* and *geCV* testing items were identical to exposure items for the Mid group. All of the *keCV* and *geCV* testing items were novel for participants in the High condition, just as all of the *kiCV* and *giCV* testing items were novel for Mid participants.

Participants' responses in the Practice, Exposure, and Training phases were recorded with a Sony Portable MiniDisc Recorder MZ-B100 and a Sony ECM-44B Electret Condenser Microphone with a tie-clip. Though the Exposure and Testing phases were self-paced, they had quite similar durations across participants. The Exposure phase lasted approximately 4min and the Testing phase lasted approximately 7.5min.

4.1.3 *Participants*

Twenty-two native American English speaking undergraduate students at UCLA participated in the experiment. Participants were randomly assigned to the two experimental conditions, with the restriction that there be an equal number in each condition. They were paid a nominal fee or received a small amount of extra credit in an introductory course.

4.2 Results and analysis

The recorded responses in the Practice, Exposure, and Testing phases were transcribed by a phonetically-trained native American English speaker (not the author). There were few errors or unexpected responses in the Practice or Exposure phases, which required the participants to simply repeat words that were played by the computer.

The vast majority of the responses in the Testing phase could be classified into two categories: no-change (the participant responded with the same nonword that was produced by the computer) and palatalized (the participant responded with the same nonword except that the initial consonant was replaced by a palatoalveolar affricate). Palatalization was applied very infrequently to the labial stops [p] and [b]; only 5 responses (less than 1% of all total responses) were of this type. The statistical analysis below focuses on the rate of palatalization of critical testing items (i.e., items that began with [ki], [ke], [ka], [gi], [ge], or [ga]). Fig. 2 displays the palatalization rate for each type of critical item in each condition; Table 7 gives the means and standard errors.

Figure 2 about here**Table 7 about here**

The central issue addressed by this experiment is whether participants exposed to palatalization before the vowel [i] (High condition) and participants exposed to palatalization before the

vowel [e] (Mid condition) will show different patterns of generalization. Recall that velar stops and palatoalveolar affricates are more perceptually similar before [i] than before [e], and that palatalization before [e] asymmetrically implies palatalization before [i] in most attested languages. If participants have a system of substantively-biased generalization of the kind presented in Section 3, then we expect more generalization of palatalization in the Mid condition than in the High condition. On the other hand, if participants do not have such a system or cannot access it—if they do not approach the problem of learning a new phonological pattern with the implicit knowledge that alternations among more perceptually similar sounds are favored—then there is no particular expectation of greater generalization in one condition than in the other.

The correct way to test for an asymmetric generalization pattern is to look for an interaction between experimental condition and vowel context. The *exposure* context is the vowel that conditioned velar palatalization in the Exposure phase: [i] for the High group, [e] for the Mid group. The *novel* context is the other front vowel, the one that did not occur after velars in the Exposure phase: [e] for the High group, [i] for the Mid group. (I return below to the issue of generalization before the low back vowel [ɑ].)

A repeated-measures ANOVA with participant as a random factor was performed on the proportion of palatalization responses computed for each participant in each of the two contexts, with responses broken down by consonant category (voiceless [k] vs voiced [g]). The between-participants factor was condition (High vs Mid). There were two within-participants factors: consonant ([k] vs [g]) and vowel context (exposure vs novel). The main effect of condition was not significant ($F < 1$), suggesting that the different types of exposure to velar palatalization did not induce different overall rates of palatalization. There was a significant main effect of consonant ($F(1, 20) = 8.0, p < .05, MS_e = .07$), a significant main effect of vowel context ($F(1, 20) = 8.3, p < .01, MS_e = .08$), and a marginally significant interaction between condition and consonant ($F(1, 20) = 4.2, p < .06, MS_e = .07$). The crucial interaction between condition and vowel context was significant ($F(1, 20) = 8.3, p < .01, MS_e = .08$) and supported by planned post-hoc paired t -tests. Participants in the High condition palatalized velars at a significantly higher rate before the exposure vowel [i] than before the novel vowel [e] (mean of the differences: .35, $t(10) = 3.0, p < .05$). But participants in the Mid condition applied palatalization at a statistically indistinguishable rate before the exposure vowel [e] and the novel vowel [i] (mean of the differences: 0, $t(10) = 0, p = 1$).¹⁰

The ability of the biased and unbiased instantiations of the conditional random field model (Section 3) to capture this asymmetric pattern of generalization was tested by fitting the models to the aggregate data for each group.¹¹ Both instantiations of the model had a single free parameter

¹⁰The two-way interaction between consonant and vowel context, and the three-way interaction among condition, consonant, and vowel context were both non-significant ($F_s < 1$). Because of issues of non-normality that arise when proportional data are analyzed with ANOVA, the statistics reported in the text were also performed under the arcsin transformation $\sin^{-1}(\sqrt{x})$ of the proportions. The pattern of statistical significance did not change, except that the interaction between condition and consonant reached significance at the $\alpha = .05$ level ($F(1, 20) = 4.6, p < .05, MS_e = .15$). The interaction between condition and vowel context remained significant ($F(1, 20) = 9.6, p < .01, MS_e = .19$).

¹¹At this point it would be customary in psycholinguistic studies to perform the same statistical analysis with items as a random factor. Such an analysis would test the hypothesis that the effects found in the by-participants ANOVA are uniform across items (see Clark 1973 for general discussion). However, there is little reason, in this or many other experiments on language, to believe that such a hypothesis could be valid. With a small stimulus set, it is likely

D , which scales the size of the training data (i.e., determines the magnitude of the ‘D’ force in Fig. 1) relative to the prior. This parameter is necessary because it is not known how the number of exposure trials in the experiment corresponds to degree of processing in the psychological system. Each item in the exposure list was assigned a weight of 1; practice items were assigned a weight of 10, reflecting the fact that they were presented at the beginning of the experiment and in relative isolation from other items. The total body of training data for the models was obtained by multiplying the weight of each item by D . These details aside, the models were exposed to exactly the same stimuli as the human participants.

Table 8 gives the correlations between the observed velar palatalization rates and the best-fitting predictions of the biased and unbiased models. Also included are correlations for variants of the models in which each example that underwent palatalization in the exposure phase was encoded as a word-specific constraint (or ‘memory’). Figs. 3 and 4 display the correlations between the substantively biased model, with memory constraints, and the data for the High and Mid groups, respectively. Memory contributed little to the predictions of either instantiation of the model, therefore the figures are similar when the word-specific constraints are removed.

Table 8 about here

Figure 3 about here

Figure 4 about here

4.3 Discussion

The results of this experiment support the substantively-biased model over the unbiased model, especially with respect to the Mid condition. Participants generalized velar palatalization from the mid vowel [e] to the high vowel [i], but did so much less in the opposite direction, a result that is in line with the findings from language typology that were reviewed in Section 2 and that is explained within the framework of substantively-biased phonology through the incorporation of perceptual similarity into the priors on constraint weights. The substantively-biased model yields detailed qualitative and quantitative fits to the pattern of behavioral data: the asymmetry between [i] and [e]; the extension of palatalization to the [a] context (in spite of the fact that velars did not palatalize before [a] in the exposure items); and the overall higher rate of palatalization of [g] (a finding that can be traced to the practice items, which only instantiated [g] palatalization). The model without bias fits the data much more poorly, explaining about 45% less of the variance in the Mid condition.

There was one qualitative feature of the results that was not predicted by the substantively-biased model: namely, the relatively high rate at which palatalization was extended to [ki] in the

that the idiosyncratic properties of some particular items (e.g., their phonotactic probability, or similarity to existing words, or degrees of similarity to other stimulus items) will substantially affect participants’ behavior. Fortunately, the argument for substance does not depend on the hypothesis that all items of a particular type were treated identically. Though we wish to establish a general claim about a population of human learners, making the by-participants analysis a sensible one, we do not desire or need to make the claim that all nonwords beginning with a particular CV sequence are identical, even for the limited purpose of predicting velar palatalization.

Mid condition. This is likely due to a defect in the similarity values that were entered into the model (see Table 3) rather than in the model itself (thanks to Matt Gordon for this suggestion). Recall that the similarity of [k] and [tʃ] before [e] was estimated by interpolation. The present findings suggest that this value is too high (i.e., the consonants are being treated as too similar), a possibility that could be tested in a perception experiment of the kind reported in Guion (1996, 1998).

To put these results in a broader context, we return to the debate between phonetically-based phonology and evolutionary phonology (see Section 1). One of the central claims made within evolutionary phonology and related frameworks is that typological asymmetries, such as the implicational laws observed to govern velar palatalization, need not be attributed to cognitive asymmetries; mechanisms by which languages change over time provide an alternative explanation. A possible response to this claim, fine as far as it goes, is that very little work in this vein has been formalized to a degree that allows falsifiability (cf. de Boer, 2001; Redford et al., 2001). A more positive response by the proponents of substance is to seek out new types of evidence that cannot plausibly be accounted for with the evolutionary mechanisms of misperception, reinterpretation, self-organization, and the like. The PSM experiments and analyses presented here were conducted in that spirit. By demonstrating that participants generalize from a brief period of exposure in the way predicted by a formal, substantively-biased learning model—not in the way predicted by a formally identical model that lacks substantive bias—the present results shift the debate from speculation over the source of typological distribution to experimental investigation of human learning (see also Pater & Tessier, 2003; Pycha et al., 2003; Zhang & Lai, in progress; Zuraw, 2005; Wilson, 2003).

5 Experiment 2: Testing generalization on the focus

As noted in Section 3, both the substantively biased and unbiased instantiations of the conditional random field model predict that palatalization of one velar stop should not be generalized to the other velar stop. This prediction follows from the assumption that the stops are subject to distinct Faithfulness constraints, $F(k)$ and $F(g)$. The purpose of the present experiment was to test this prediction and to provide an independent set of data on which to test the claims of substantively biased phonology.

5.1 Methods

5.1.1 Stimuli

The nonword recording used in this experiment were the same as those in Experiment 1.

5.1.2 Procedure

The experiment had two conditions (Voiceless, Voiced), with four phases in each condition (Practice, Exposure, Break, Testing). The equipment and procedures were identical to those in Experiment 1, except with respect to the stimulus lists that were presented to participants in the Practice and Exposure phases.

There were three practice trials: one in which the members of the stimulus pair were phonologically identical ([bələ] ... [bələ]), and two in which the members of the stimulus pair were re-

lated by velar palatalization (Voiceless: [kiwə] ... [tʃiwə] and [kenə] ... [tʃenə]; Voiced: [gipə] ... [dʒipə] and [gefə] ... [dʒefə]).

During the Exposure phase there were 34 trials, as schematized in Table 9. The trials were grouped into 4 blocks. Each block contained 2 examples of velar palatalization ([k] → [tʃ] or [g] → [dʒ] before [i] and [e]), 1 or 2 examples of velars that did not palatalize ([k] or [g] before [ɑ]), and 4 or 5 fillers. The blocks that contained 4 fillers also included one example in which velar palatalization applied to the novel voicing category. In other words, participants in the Voiceless condition heard exactly two examples of palatalization of [g] (one before [i] and one before [e]), and participants in the Voiced condition heard exactly two examples of palatalization of [k] (one before [i] and one before [e]). These items were included in order to encourage generalization during testing — a manipulation that was not successful, as we will see. The order of the blocks and the order of items within blocks were randomized across participants.

Table 9 about here

The Testing phase contained 80 trials, as schematized in Table 6 (see Section 3). The testing list was exactly the same for both conditions, and was randomized for each participant without blocking.

5.1.3 Participants

Twenty-two native American English speaking undergraduate students at UCLA participated in the experiment. Participants were randomly assigned to the two experimental conditions, with the restriction that there be an equal number in each condition. They were paid a nominal fee or received a small amount of extra credit in an introductory course. None of the participants in this experiment had participated in Experiment 1.

5.2 Results and analysis

The recorded response in the Practice, Exposure, and Testing phases were transcribed by a phonetically-trained native American English speaker (the author). As in Experiment 1, almost all of the responses in the Practice and Exposure phases consisted of errorless repetitions, and the great majority of the responses in the Testing phase could be classified as no-change or palatalized. Palatalization was applied very infrequently to the labial stops ([p] and [b]); only 2 responses (less than 1% of all total responses) were of this type. The following statistical analysis therefore focuses on the rate of palatalization of critical testing items. Fig. 5 displays the palatalization rate for each type of critical item in each condition; Table 10 gives the means and standard errors.

Figure 5 about here

Table 10 about here

A repeated-measures ANOVA with participant as a random factor was performed on the proportion of palatalization responses computed for each subject and each critical consonant, with responses broken down by vowel context. The between-participants factor was condition (Voice-

less vs Voiced). There were two within-participant factors: consonant (exposure vs novel) and vowel context (high front [i] vs mid front [e]). The main effect of condition was not significant ($F < 1$), suggesting that the two exposure conditions did not lead to different overall rates of palatalization. There was also a significant main effect of consonant ($F(1, 20) = 10.5, p < .01, MS_e = .12$). All other main effects and interactions were non-significant. In particular, there was no significant interaction between condition and consonant ($F < 1$), suggesting that participants do not extend velar palatalization from [g] to [k] at a higher rate than the (relatively low) rate at which they extend the change from [k] to [g]. Both generalization rates are low relative to that observed in Experiment 1.¹²

The biased and unbiased instantiations of the model were fit to the averaged experimental data following exactly the same procedure described for Experiment 1. Table 11 gives the correlations between the observed velar palatalization rates and the best-fitting predictions of the biased and unbiased models. Also included are correlations for variants of the models in which each example that underwent palatalization in the exposure phase was encoded as a word-specific constraint (or ‘memory’). The biased model significantly outperforms the unbiased model with respect to the Voiced condition—the condition in which the participants extended palatalization to the [a] context most strongly. In that condition, the biased model explained approximately 10% more of the variance than the unbiased model. Figs. 6 and 7 display the correlations between the substantively biased model, with memory constraints, and the data for the Voiceless and Voiced groups, respectively.

Table 11 about here

Figure 6 about here

Figure 7 about here

5.3 Discussion

The results of this experiment support the prediction that palatalization is not generalized from a velar stop with one specification for [voice] to a velar stop with a different [voice] specification. They also provide additional evidence in favor of the substantively-biased model, which predicts the detailed pattern of velar application better than the unbiased model.

The lack of generalization on the target converges with results of Goldrick (2004), who also found little generalization between the two velar stops [k g] in a quite different experimental paradigm. Absence of generalization—and ultimately the existence of two distinct Faithfulness constraints, F(k) and F(g)—may itself have a perceptual explanation. It is a well-known finding of speech perception experiments that the [voice] specification of a stop is perceptually robust, much more so its place of articulation (e.g., Benkí, 2002). This line of explanation may also account for another finding of Goldrick (2004), namely that generalization between voiceless and

¹²The statistics reported in the text were repeated with arcsin-transformed proportions ($\sin^{-1} \sqrt{x}$). The pattern of statistical significance did not change. The main effect of consonant (exposure vs novel) was significant ($F(1, 20) = 11.7, p < .01, MS_e = .22$) and there was no significant interaction of condition and consonant ($F < 1$).

voiced fricatives ([f v]) does occur. For aerodynamic reasons, the [voice] distinction is likely to be weaker for fricatives than for stops; this perhaps gives rise to an identification of their Faithfulness constraints.

Kie Zuraw points out another source of converging evidence, this time from loanword phonology. It has been observed that, whereas phonological patterns in the native language are typically extended to novel *contexts* in borrowed words, extension to novel *segments* is more rare. Again, we might expect a nuanced version of this generalization according to which a non-native sound is subject to native phonology in proportion to how strongly it perceptually resembles native sounds.

The present findings do, however, appear to be incompatible with one of the typological implications discussed in Section 2. Recall that palatalization of [g] asymmetrically implies palatalization of [k] in the languages of the world. This could have lead us to expect generalization in the same direction, for essentially the same reason that we expected (asymmetric) generalization on the context in Experiment 1.

This apparent tension can be resolved by considering an important difference between the present experiments and how velar palatalization is likely to arise in natural languages. The experiments presented palatalization as an instantaneous, categorical change from a velar stop to a palatoalveolar affricate. But natural velar palatalization likely develops in a series of smaller steps. I make the minimal assumption that the first step is strong coarticulation between front vowels and all preceding velar stops (a state of affairs that could be transcribed roughly as [kⁱ, gⁱ]). The typological rarity or non-existence of palatalization of voiced velars only could then follow from the hypothesis that learners are unlikely to reinterpret a heavily coarticulated [gⁱ] as [dʒi] without also reinterpreting a heavily coarticulated [kⁱ]—which is perceptually more similar to the corresponding palatoalveolar affricate—as [tʃi]. In short, I take the explanation for the typological implication to be of the kind championed in evolutionary phonology: palatalization of only voiced velars is a possible sound pattern, but unlikely to arise in nature. Adopting this explanation does not, as I have shown, prevent us from also investigating cognitive biases that make reference to the same underlying substantive factors.

6 Conclusions

The main issue addressed in this paper was whether human learners have a system of cognitive biases, rooted in knowledge of phonetic substance, that shapes the way in which they learn and generalize from phonological data. I began by reviewing the articulatory, acoustic, perceptual, and typological properties of velar palatalization, and then (following work by Ohala and Guion) focused on perception as the central substantive factor. A general model of categorization adopted from work in psychology was used to quantify the perceptual similarity of velars and palatoalveolars in three vowel contexts. The resulting similarity values function as a prior, one that favors changes involving more similar sounds, in the proposed framework of substantively biased phonology. The framework was made fully explicit with conditional random fields. Two PSM experiments and accompanying modeling results revealed novel, detailed patterns of generalization—and lack of generalization—that support the biased model over a formally matched unbiased model.

In addition to their implications for the debate over substance, the present findings have consequence for the theories of generalization and similarity. First, phonological learning cannot

proceed exclusively by minimal (or ‘least general’) generalization (Albright & Hayes, 2003; Pierrehumbert & Nair, 1995), because such a mechanism could not explain the observed patterns in which velar palatalization is extended to a novel context (Experiments 1 and 2). The same problem holds for exemplar-based theories of phonological generalization (Daelemans et al., 2003; Kirchner, 2005).¹³ I would tentatively suggest that both types of theory are valid only when the evidence available to the learner is abundant, thereby allowing for fine-grained comparison of the predictive value of specific stimulus properties. The current investigation is targeted at the opposite extreme—closer to the original empirical motivation for generative grammar—in which the learner’s input is highly impoverished. (Minimal generalization / exemplar theories are of course compatible with the apparent lack of generalization on the focus in Experiment 2, but I have given an alternative explanation for that finding in terms of Faithfulness.)

Second, the predictions of the substantively biased model depend crucially on a notion of similarity that is context-sensitive. This contrasts sharply with recent research in which much more coarse-grained similarity metrics are applied to the problem of predicting various aspects of lexical and phonological behavior (Bailey & Hahn, 2001, 2005; Frisch et al., 2004; Hahn & Bailey, to appear; Luce, 1986). The large amounts of unexplained variance in the important studies of Bailey & Hahn (2001, 2005) and Hahn & Bailey (to appear) in particular suggest that judgments of wordlikeness and word similarity cannot be successfully modeled unless contextual effects on sound perception are not taken into account.

I should also note some limitations of the present paper and directions for future research. For reasons of space, I have not been able to consider several additional alternative explanations of the experimental data, most notably those that would draw upon the participants’ knowledge of English. Such alternatives are considered and shown to be inadequate in a companion paper. I have already noted that the interpolated similarity value for the pair [ki]/[tʃi] is likely too high, and suggested another study that could test this possibility. There are two other rather open-ended directions for research. The first would systematically vary the amounts and types of exposure in the PSM paradigm to further test the quantitative predictions of substantively biased phonology. The second would apply the paradigm to other putatively substantively-motivated phonological patterns, dozens of which appear in the literature (e.g., Hayes et al. 2004).

I conclude with a final remark on the general perspective advanced here. In the foundational work of generative phonology, Chomsky & Halle (1968) set a goal of defining a notational system in which well-attested, substantively-motivated phonological patterns have concise descriptions. The framework of substantively biased phonology continues this line of research, with the difference that the preference for certain patterns is expressed as a prior on constraint weights rather than as a set of notational conventions. Like Chomsky & Halle (1968), I claim that the bias is a component of cognition that is important for phonological learning and generalization. Also like Chomsky & Halle (1968), I do not take the bias to be so strong that it excludes disfavored patterns. The experimental and computational methods applied here allow such claims to be investigated in

¹³Tests of the TiMBL exemplar-based model (Daelemans et al., 2003) have yielded poor fits to the experimental results. While the model gave a reasonable account of the behavior of the participants in the High (Experiment 1) and Voiceless (Experiment 2) conditions ($r = .90$ and $r = .51$ for the critical items, respectively), it could not account for the behavior of participants in the other two conditions, Mid (Experiment 1) and Voiced (Experiment 2) ($r = .08$ and $r = -.21$ for the critical items, respectively). The latter two conditions were the ones that gave rise to the greatest generalization beyond the exposure data, and for that reason the results appear to lie beyond the reach of this model.

unprecedented detail, providing a potentially vast body of new data and theoretical insights on the nature of phonological learning.

Acknowledgments [to be added]

References

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

Albro, D. M. (2005). Computational Optimality Theory and the Phonological System of Malagasy. Ph.D. thesis, University of California, Los Angeles.

Anderson, S. R. (1974). *The Organization of Phonology*. New York: Academic Press.

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

Anderson, S. R. (1985). *Phonology in the Twentieth Century: Theories of Rules and Theories of Representations*. Chicago: University of Chicago Press.

Bagemihl, B. (1995). Language Games and Related Areas. In J. A. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 697-712). Cambridge, MA: Blackwell.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44, 568-591.

Bailey, T. M., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, 52, 339-362.

Beckman, J. (1999). *Positional Faithfulness: An Optimality Theoretic Treatment of Phonological Asymmetries*. New York: Garland.

Benkí, J. R. (1998). *Evidence for Phonological Categories from Speech Perception*. Ph.D. thesis, University of Massachusetts, Amherst. Distributed by GLSA.

Benkí, J. R. (2002). Analysis of English nonsense syllable recognition in noise. *Phonetica*, 60, 129-157.

Bhat, D. (1978). A general study of palatalization. In J. Greenberg (Ed.), *Universals of human language*, vol. 3 (pp. 47-92). Stanford: Stanford University Press.

Blevins, J. (2004). *Evolutionary Phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Blevins, J. (to appear). Phonetic explanations for recurrent sound patterns: diachronic or synchronic? In C. Cairns & E. Raimy (Eds.), *Phonological Theory: Representations and Architecture*. Cambridge, MA: MIT Press.

Blevins, J., & Garrett, A. (2004). The evolution of metathesis. In B. Hayes et al. (2004) (pp. 117-156).

Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32, 45-86.

Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Buckley, E. (2000). On the naturalness of unnatural rules. In *Proceedings from the Second Workshop on American Indigenous Languages. UCSB Working Papers in Linguistics (Vol. 9)*.

Buckley, E. (2003). Children's unnatural phonology. In *Proceedings of the Berkeley Linguistics Society 29* (pp. 523-534).

Butcher, A., & Tabain, M. (2004). On the back of the tongue: dorsal sounds in Australian languages. *Phonetica*, 61, 22-52.

Chen, M. (1972). On the formal expression of natural rules in phonology. *Journal of Linguistics*, 9, 209-383.

- Chen, M. (1973). Predictive power in phonological description. *Lingua*, 32, 173-191.
- Chen, S. F., & Rosenfeld, R. (1999). A Gaussian prior for smoothing maximum entropy models. Technical Report CMU-CS-99-108.
- Cho, T., & McQueen, J. (in preparation). Mapping phonologically altered speech onto the lexicon: The case of consonant cluster simplification in Korean. Ms., Max Planck Institute for Psycholinguistics.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. Cambridge, MA: MIT Press.
- Cohen J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, 25, 2, 257-271.
- Côté, M.-H. (2000). *Consonant Cluster Phonotactics: A Perceptual Approach*. Ph.D. thesis, MIT.
- Côté, M.-H. (2004). Syntagmatic distinctness in consonant deletion. *Phonology*, 21, 1, 1-41.
- Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York: John Wiley & Sons.
- Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2003). TiMBL: Tilburg Memory Based Learner, version 5.0, Reference Guide. ILK Technical Report 03-10, available from <http://ilk.uvt.nl/downloads/pub/papers/ilk0310.pdf>.
- Davidson, L. (2003). *The Atoms of Phonological Representation: Gestures, Coordination and Perceptual Features in Consonant Cluster Phonotactics*. Ph.D. thesis, Johns Hopkins University.
- Davidson, L. (to appear). Phonology, phonetics, or frequency: influences on the production of non-native sequences. *Journal of Phonetics*.
- de Boer, Bart. (2001). *The Origins of Vowel Systems*. Oxford: Oxford University Press.
- Dupoux, E., Kakehi, H., Hiroshi, Y., Pallier, C. , & Mehler, J. (1999). Epenthetic vowels in Japanese: a perceptual illusion. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 6, 1568-1578.
- Eisner, J. (1997). Efficient generation in primitive Optimality Theory. In *Proceedings of the 35th Annual Meeting of the ACL* (pp. 313-320).
- Flemming, E. (2001). Scaler and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, 18, 1, 7-44.
- Flemming, E. (2002). *Auditory Representations*. New York: Routledge.
- Fleischhacker, H. (2001). Cluster-dependent epenthesis asymmetries. In A. Albright & T. Cho (Eds.), *UCLA Working Papers in Linguistics 7: Papers in Phonology 5* (pp. 71-116).
- Frank, R. (2004). Restricting grammatical complexity. *Cognitive Science*, 28, 669-697.
- Frisch, S. A., Pierrehumbert, J. B., & Broe, M. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, 22, 1, 179-228.
- Gafos, D. (1999). *The Articulatory Basis of Locality in Phonology*. New York: Garland.
- Gafos, D. (2002). A grammar of gestural coordination. *Natural Language & Linguistic Theory*, 20, 2, 269-337.
- Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- Gilkerson, J. (to appear). Categorical perception of natural and unnatural categories: evidence for

innate category boundaries. *UCLA Working Papers in Linguistics: Language Development and Breakdown 2*.

- Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language*, 51, 586-603.
- Goldwater, S., & Johnson, M. (2003). Learning OT Constraint Rankings Using a Maximum Entropy Model. In *Proceedings of the Workshop on Variation within Optimality Theory*, Stockholm University.
- Gordon, M. (2004). Syllable weight. In Hayes et al. 2004 (pp. 277-312).
- Gregory, M. L., & Altun, Y. (2004). Using Conditional Random Fields to Predict Pitch Accents in Conversational Speech. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL)*, Barcelona (pp. 677-683).
- Grünwald, P. D., Myung, I. J., & Pitt, M. A. (2005). *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press/Bradford Books.
- Guion, S. G. (1996). *Velar Palatalization: Coarticulation, Perception and Sound Change*. Ph.D. thesis, University of Texas, Austin.
- Guion, S. G. (1998). The role of perception in the sound change of velar palatalization. *Phonetica*, 55, 18-52.
- Hahn, U., & Bailey, T. M. (to appear). What makes words sound similar? *Cognition*.
- Hale, M., & Reiss, C. (2000). Substance abuse and “disfunctionalism”: current trends in phonology. *Linguistic Inquiry*, 31, 1, 157-169.
- Hall, N. (2003). *Gestures and Segments: Vowel Intrusion as Overlap*. Ph.D. thesis, University of Massachusetts, Amherst.
- Hayes, B. (1995). *Metrical Stress Theory*. Chicago: University of Chicago Press.
- Hayes, B. (1999). Phonetically-driven phonology: the role of Optimality Theory and inductive grounding. In M. Darnell, E. Moravcsik, M. Noonan, F. Newmeyer, & K. Wheatley (Eds.), *Functionalism and formalism in linguistics, volume I: general papers* (pp. 243-285). Amsterdam: John Benjamins.
- Hayes, B. Kircher, R., Steriade, D. (2004). *Phonetically Based Phonology*. Cambridge: Cambridge University Press.
- Hume, E., & K. Johnson. (2001). *The role of speech perception in phonology*. San Diego: Academic Press.
- Hyman, L. M. (2001). The limits of phonetic determinism in phonology: *NC revisited. In E. Hume & K. Johnson (2001) (pp. 141-185).
- Idsardi, W. J. (2005). Poverty of the Stimulus Arguments in Phonology. Ms., University of Delaware.
- International Phonetic Association. (1996). Reproduction of the International Phonetic Alphabet (Revised to 1993, Updated to 1996). <http://www2.arts.gla.ac.uk/IPA/ipachart.html>.
- Jun, J. (1995). *Perceptual and Articulatory Factors in Place Assimilation: An Optimality Theoretic Approach*. Ph.D. thesis, University of California, Los Angeles.
- Kang, Y. (2004). Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean. *Phonology*, 20, 2, 173-218.
- Kawasaki-Fukumori, H. (1992). An acoustical basis for universal phonotactic constraints. *Language and Speech*, 35, 1-2, 73-86.
- Keating, P., & Lahiri, A. (1993). Fronted velars, palatalized velars, and palatals. *Phonetica*, 50,

73-101.

Kenstowicz, M. (2003). Salience and similarity in loanword adaptation: a case study from Fijian. Ms., MIT.

Kingston, J. (1994). Change and stability in the contrasts conveyed by consonant releases. In P. A. Keating (Ed.), *Phonology Structure and Phonetic Form: Papers in Laboratory Phonology III* (pp. 354-361). Cambridge: Cambridge University Press.

Kirchner, R. (2000). Geminate inalterability and lenition. *Language*, 76, 509-545.

Kirchner, R. (2001). *An Effort Approach to Consonant Lenition*. New York: Routledge.

Kirchner, R. (2005). Exemplar-based phonology and the time problem: a new representational technique. Presented at LabPhon 9, UIUC.

Kochetov, A. (2002). *Production, Perception, and Emergent Phonotactic Patterns: A Case of Contrastive Palatalization*. London: Routledge.

Ladefoged, P. (2001). *A Course in Phonetics* (3rd ed.). San Diego: Harcourt Brace.

Lafferty, J., McCallum, M., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. ICML*.

Lin, Y. (2004). Learning Stochastic OT Grammars with a Gibbs Sampler. Ms., University of California, Los Angeles.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. Ohala & J. Jaeger (Eds.), *Experimental Phonology* (pp. 13-44). Orlando: Academic Press.

Luce, P. D. (1986). *Neighborhoods of Words in the Mental Lexicon*. Ph.D. thesis, Indiana University, Bloomington, IA.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103-190). New York: Wiley.

McCallum, A. (2003). Efficiently inducing features of conditional random fields. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

McCarthy, J. (1981). The role of the evaluation metric in the acquisition of phonology. In C. L. Baker, & J. McCarthy (Eds.), *The Logical Problem of Language Acquisition* (pp. 218-248). Cambridge, MA: MIT Press.

McCarthy, J. (1982). Prosodic structure and expletive infixation. *Language*, 58, 574-590.

McCarthy, J., & Prince, A. (1999). Faithfulness and identity in Prosodic Morphology. In R. Kager, H. van der Hulst, & W. Zonneveld (Eds.), *The Prosody-Morphology Interface* (pp. 218-309). Cambridge: Cambridge University Press.

Maddieson, I., & Precoda, K. (1992). Syllable structure and phonetic models. *Phonology*, 9, 1, 45-60.

Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition*, 84, 55-71.

Moreton, E., Feng, G., & Smith, J. L. (2005). Syllabification, sonority, and perception: new data from a language game. Presented at the 41st Regional Meeting of the Chicago Linguistic Society, Chicago, Illinois, April 7-9.

Myers, S. (2000). Boundary disputes: The distinction between phonetic and phonological sound patterns". In N. Burton-Roberts, P. Carr, & G. Docherty (Eds.), *Phonological Knowledge: Conceptual and empirical issues*. Oxford: Oxford University Press.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47, 90-100.

- Neeld, R. (1973). Remarks on palatalization. In *Working papers in Linguistics No. 14: Studies in phonology and methodology*. Columbus, OH: Department of Linguistics, The Ohio State University.
- Nevins, A., & Vaux, B. (2003). Underdetermination in language games: dialects of Pig Latin. Presented at the Linguistic Society of America (LSA) Annual Meeting, Atlanta, GA.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 1, 39-57.
- Nosofsky, R. M., & S. R. Zaki. (2002). Exemplar and prototype models revisited: response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 5, 924-940.
- Ohala, J. J. (1981). The listener as the source of sound change. In C. S. Masek, R. A. Hendrick, M. F. Miller (Eds.), *Papers from the parasession on language and behavior* (pp. 178-203). Chicago: Chicago Linguistic Society.
- Ohala, J. J. (1990). The phonetics and phonology of aspects of assimilation. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology (Vol. 1)* (pp. 258-275). Cambridge: Cambridge University Press.
- Ohala, J. J. (1992). What's cognitive, what's not, in sound change. In Kellermann, G., & Morrissey, M. (Eds.), *Diachrony within synchrony: language history and cognition. Duisberger Arbeiten zur Sprach- und Kulturwissenschaft 14* (pp. 309-355). Frankfurt: Peter Lang.
- Ohala, J. J. (1995). Experimental phonology. In J. A. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 713-722). Cambridge, MA: Blackwell.
- Ohala, J. J. (1997). Comparison of speech sounds: distance vs. cost metrics. In S. Kiritani, H. Hirose, & H. Fujisaki (Eds.), *Speech Production and Language: In honor of Osamu Fujimura* (pp. 261-270). Berlin: Mouton de Gruyter.
- Padgett, J. (2004). Russian vowel reduction and dispersion theory. In *Phonological studies 7* (pp. 81-96). Tokyo: Kaitakusha.
- Pater, J., & Tessier, A.-M. (2003). Phonotactic Knowledge and the Acquisition of Alternations. In M.J. Sol, D. Recasens, & J. Romero (Eds.) *Proceedings of the 15th International Congress on Phonetic Sciences, Barcelona* (pp. 1777-1180).
- Peperkamp, S. (2004). Lexical exceptions in stress systems: arguments from early language acquisition and adult speech perception. *Language*, 80, 98-126.
- Pierrehumbert, J. B. (submitted). An unnatural process. *Laboratory Phonology 8*. Berlin: Mouton de Gruyter.
- Pierrehumbert, J. B., & Nair, R. (1995). Word games and syllable structure. *Language and Speech*, 38, 78-114.
- Plauché, M. C., Delogu, C., & Ohala, J. J. (1997). Asymmetries in consonant confusion. In *Proceedings of EuroSpeech '97 (Vol. 4)* (2187-2190).
- Prince, A., & Smolensky, P. (1993/2004). *Optimality Theory: Constraint Interaction in generative grammar*. Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, April 1993. Cambridge, MA: Blackwell.
- Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19, 9-50.
- Pvcha, A., Nowak, P., Shin, E., & Shosted, R. (2003). Phonological rule-learning and its implica-

tions for a theory of vowel harmony. In G. Garding & M. Tsujimura (Eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics* (pp. 533-546). Somerville, MA: Cascadilla.

- R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Raaijmakers, J. G., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with “the language-as-fixed-effect fallacy”: common misconceptions and alternative solutions. *Journal of Memory and Language*, 41, 416-426.
- Redford, M. A., Chen, C. C., & Miikkulainen, R. (2001). Constrained Emergence of Universals and Variation in Syllable Structure. *Language and Speech*, 44, 27-56.
- Reiss, C. (2003). Quantification in Structural Descriptions: Attested and Unattested Patterns. *The Linguistic Review*, 20, xxx-xxx.
- Riggle, Jason. (2004). Generation, Recognition, and Learning in Finite State Optimality Theory. Ph.D. thesis, University of California, Los Angeles.
- Roark, B., Saraclar, M., Collins, M., & Johnson, M. (2004). Discriminative Language Modeling with Conditional Random Fields and the Perceptron Algorithm. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona.
- Samek-Lodovici, V., & Prince, A. (1999). Optima. *Rutgers Optimality Archive* 363.
- Schwartz, B. D. (1986). The epistemological status of second language acquisition. *Second Language Research*, 2, 120-159.
- Seidl, A., & Buckley, E. (in press). On the learning of arbitrary phonological rules. *Language Learning and Development*.
- Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of Human Language Technology, NAACL*.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.
- Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1, pp. 194-281). Cambridge, MA: MIT Press/Bradford Books.
- Smolensky, P. (1996). Overview: statistical perspectives on neural networks. In Smolensky, P., Mozer, M. C., & Rumelhart, D. E. (Eds.), *Mathematical Perspectives on Neural Networks* (pp. 453-496). Mahwah, NJ: Lawrence Erlbaum.
- Smolensky, P., & Legendre, G. (2005). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammars* (two volumes). Cambridge, MA: MIT Press.
- Steriade, D. (2001a). Directional asymmetries in place assimilation: A perceptual account. In Hume & Johnson (2001), pp. 219-250.
- Steriade, D. (2001b). The phonology of perceptibility effects: the P-map and its consequences for constraint organization. Ms., MIT.
- Steriade, D. (2001c). What to expect from a phonological analysis. Presented at the Linguistic Society of America (LSA) Annual Meeting, Washington, DC.
- Stevens, K. N. & Keyser, S. J. (1989). Primary features and their enhancement in consonants. *Language*, 65, 81-106

- Tesar, B., & Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, 29, 229-268.
- Tesar, B., & Smolensky, P. (2000). *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.
- Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88, 97-100.
- Treiman, R. (1983). The structure of spoken syllables: evidence from novel word games. *Cognition*, 15, 49-74.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.
- Wilson, C. (2001). Consonant cluster neutralisation and targeted constraints. *Phonology*, 18, 1, 147-197.
- Wilson, C. (2003). Experimental Investigation of Phonological Naturalness. In G. Garding & M. Tsujimura (Eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics* (pp. 533-546). Somerville, MA: Cascadilla.
- Winitz, H., Scheib, M. E., & Reeds, J. A. (1971). Identification of Stops and Vowels for the Burst Portion of /p, t, k/ Isolated from Conversational Speech. *The Journal of the Acoustical Society of America*, 51, 1309-1317.
- Wright, R. (1996). *Consonant Clusters and Cue Preservation in Tsou*. Ph.D. thesis, University of California, Los Angeles.
- Zhang, J. (2001). *The Effects of Duration and Sonority on Contour Tone Distribution—Typological Survey and Formal Analysis*. Ph.D. thesis, University of California, Los Angeles.
- Zhang, J., & Lai, Y. (in progress). Psychological differences between natural and unnatural tone sandhi in Mandarin Chinese. Ms., University of Kansas.
- Zuraw, K. (2005). Knowledge of consonant clusters: corpus and survey evidence from Tagalog. Ms., University of California, Los Angeles.

Appendix A: Stimuli for Experiments 1 and 2

'kitfə	'gibə	'pidʒə	'bilə
'kigə	'gimə	'piθə	'bipə
'kirə	'gipə	'pivə	'bizhə
'kiwə	'girə	'pibə	'biðə
'kifə	'gisə	'pilə	'bishə
'kimə	'gitfə	'pizhə	'bizə
'kinə	'gikə	'pekə	'beðə
'kisə	'givə	'pevə	'begə
'ketfə	'giwə	'pezə	'benə
'kegə	'gefə	'pebə	'bedʒə
'kenə	'geðə	'peðə	'bevə
'kewə	'gemə	'pesə	'bezhə
'kedʒə	'gepə	'patfə	'balə
'kemə	'gerə	'pagə	'bashə
'kerə	'getfə	'parə	'bazə
'kezə	'gekə	'pafə	'batfə
'kapə	'gevə	'padʒə	'bashə
'kaθə	'gewə	'pavə	'bavə
'kavə	'gafə		
'kaðə	'gakə		
'kagə	'garə		
'kazhə	'gadʒə		
	'gapə		
	'gawə		

Tables

Table 1. Examples of velar palatalization in three vowel contexts.

[-voice] velar				[+voice] velar			
Vowel context		Example		Vowel context		Example	
[i]	[+high, -low, -back]	[ki]	→ [tʃi]	[i]	[+high, -low, -back]	[gi]	→ [dʒi]
[e]	[-high, -low, -back]	[ke]	→ [tʃe]	[e]	[-high, -low, -back]	[ge]	→ [dʒe]
[a]	[-high, +low, +back]	[ka]	→ [tʃa]	[a]	[-high, +low, +back]	[ga]	→ [dʒa]

Table 2. Confusions of velars and palatoalveolars (adapted from Guion 1998).

Stimulus	Response							
	[ki]	[tʃi]	[gi]	[dʒi]	[ka]	[tʃa]	[ga]	[dʒa]
[ki]	43	35	10	12				
[tʃi]	10	85	0	5				
[gi]	4	4	71	21				
[dʒi]	9	28	12	51				
[ka]					84	13	3	0
[tʃa]					10	87	0	3
[ga]					4	0	87	9
[dʒa]					2	23	10	65

Table 3. Maximum likelihood estimates of perceptual similarities in three vowel contexts.

[ki]/[tʃi]	[ke]/[tʃe]	[ka]/[tʃa]	[gi]/[dʒi]	[ge]/[dʒe]	[ga]/[dʒa]
9.23^{-1}	12.68^{-1}	88.72^{-1}	21.13^{-1}	40.60^{-1}	126.93^{-1}

Note. i/j denotes $b_j\eta_{ij}$. Values in italics are interpolated.

Table 4. Markedness constraints on palatalization. Note negative exponents of σ^2 terms.

Constraint	Prior values				Constraint	Prior values			
	Biased		Unbiased			Biased		Unbiased	
	μ	σ^2	μ	σ^2		μ	σ^2	μ	σ^2
*ki	0.0	9.23^{-2}	0.0	10^{-2}	*gi	0.0	21.13^{-2}	0.0	10^{-2}
*ke	0.0	12.68^{-2}	0.0	10^{-2}	*ge	0.0	40.60^{-2}	0.0	10^{-2}
*ka	0.0	88.72^{-2}	0.0	10^{-2}	*ga	0.0	126.93^{-2}	0.0	10^{-2}
*kV _[-low]	0.0	12.68^{-2}	0.0	10^{-2}	*gV _[-low]	0.0	40.60^{-2}	0.0	10^{-2}
*kV _[-high]	0.0	88.72^{-2}	0.0	10^{-2}	*gV _[-high]	0.0	126.93^{-2}	0.0	10^{-2}
*kV	0.0	88.72^{-2}	0.0	10^{-2}	*gV	0.0	126.93^{-2}	0.0	10^{-2}

Tables

Table 1. Examples of velar palatalization in three vowel contexts.

[-voice] velar				[+voice] velar			
Vowel context		Example		Vowel context		Example	
[i]	[+high, -low, -back]	[ki]	→ [tʃi]	[i]	[+high, -low, -back]	[gi]	→ [dʒi]
[e]	[-high, -low, -back]	[ke]	→ [tʃe]	[e]	[-high, -low, -back]	[ge]	→ [dʒe]
[a]	[-high, +low, +back]	[ka]	→ [tʃa]	[a]	[-high, +low, +back]	[ga]	→ [dʒa]

Table 2. Confusions of velars and palatoalveolars (adapted from Guion 1998).

Stimulus	Response							
	[ki]	[tʃi]	[gi]	[dʒi]	[ka]	[tʃa]	[ga]	[dʒa]
[ki]	43	35	10	12				
[tʃi]	10	85	0	5				
[gi]	4	4	71	21				
[dʒi]	9	28	12	51				
[ka]					84	13	3	0
[tʃa]					10	87	0	3
[ga]					4	0	87	9
[dʒa]					2	23	10	65

Table 3. Maximum likelihood estimates of perceptual similarities in three vowel contexts.

[ki]/[tʃi]	[ke]/[tʃe]	[ka]/[tʃa]	[gi]/[dʒi]	[ge]/[dʒe]	[ga]/[dʒa]
9.23^{-1}	12.68^{-1}	88.72^{-1}	21.13^{-1}	40.60^{-1}	126.93^{-1}

Note. i/j denotes $b_j\eta_{ij}$. Values in italics are interpolated.

Table 4. Markedness constraints on palatalization.

Constraint	Prior values				Constraint	Prior values			
	Biased		Unbiased			Biased		Unbiased	
	μ	σ^2	μ	σ^2		μ	σ^2	μ	σ^2
*ki	0.0	9.23^2	0.0	10^2	*gi	0.0	21.13^2	0.0	10^2
*ke	0.0	12.68^2	0.0	10^2	*ge	0.0	40.60^2	0.0	10^2
*ka	0.0	88.72^2	0.0	10^2	*ga	0.0	126.93^2	0.0	10^2
*kV _[-low]	0.0	12.68^2	0.0	10^2	*gV _[-low]	0.0	40.60^2	0.0	10^2
*kV _[-high]	0.0	88.72^2	0.0	10^2	*gV _[-high]	0.0	126.93^2	0.0	10^2
*kV	0.0	88.72^2	0.0	10^2	*gV	0.0	126.93^2	0.0	10^2

Table 5. Exposure trials for the two conditions in Experiment 1.

Condition	Trial type (number)							
High	kiCV	...	tʃiCV	(4)	giCV	...	dʒiCV	(4)
Mid	keCV	...	tʃeCV	(4)	geCV	...	dʒeCV	(4)
Both	kaCV	...	kaCV	(3)	gaCV	...	gaCV	(3)
	piCV	...	piCV	(3)	biCV	...	biCV	(3)
	peCV	...	peCV	(3)	beCV	...	beCV	(3)
	paCV	...	paCV	(3)	baCV	...	baCV	(3)

Table 6. Testing trials for the two conditions in Experiment 1.

Critical trial type (number)				Filler trial type (number)			
kiCV ...	(8)	giCV ...	(8)	piCV ...	(6)	biCV ...	(6)
keCV ...	(8)	geCV ...	(8)	peCV ...	(6)	beCV ...	(6)
kaCV ...	(6)	gaCV ...	(6)	paCV ...	(6)	baCV ...	(6)

Table 7. Mean observed and predicted rates of velar palatalization for critical item types in Experiment 1. Values in parentheses are standard errors. The predictions are those of the substantively biased + memory model.

Condition		kiCV	keCV	kaCV	giCV	geCV	gaCV
High	Observed	.44 (.10)	.13 (.06)	.05 (.03)	.52 (.10)	.14 (.06)	.14 (.08)
	Predicted	.39	.04	.03	.54	.19	.13
Mid	Observed	.20 (.07)	.19 (.10)	.15 (.09)	.48 (.10)	.49 (.12)	.39 (.13)
	Predicted	.08	.15	.07	.40	.58	.29

Table 8. Correlations (r) between observed and predicted rates of palatalization in Experiment 1. Values in parentheses are % variance explained (r^2).

Condition	Model	All items	Critical items
High	Substantively biased	.910 (.828)	.870 (.757)
	Substantively biased + memory	.917 (.841)	.878 (.771)
	Unbiased	.913 (.834)	.871 (.759)
	Unbiased + memory	.916 (.839)	.877 (.769)
Mid	Substantively biased	.859 (.738)	.758 (.575)
	Substantively biased + memory	.851 (.724)	.787 (.619)
	Unbiased	.550 (.303)	.396 (.157)
	Unbiased + memory	.554 (.307)	.425 (.181)

Table 9. Exposure trials for the two conditions in Experiment 2.

Condition	Trial type (number)							
Voiceless	kiCV	...	tʃiCV	(4)	keCV	...	tʃiCV	(4)
	giCV	...	dʒiCV	(1)	geCV	...	dʒeCV	(1)
Voiced	giCV	...	dʒiCV	(4)	geCV	...	dʒeCV	(4)
	kiCV	...	tʃiCV	(1)	keCV	...	tʃeCV	(1)
Both	kaCV	...	kaCV	(3)	gaCV	...	gaCV	(3)
	piCV	...	piCV	(3)	biCV	...	biCV	(3)
	peCV	...	peCV	(3)	beCV	...	beCV	(3)
	paCV	...	paCV	(3)	baCV	...	baCV	(3)

Table 10. Mean observed and predicted rates of velar palatalization for critical item types in Experiment 2. Values in parentheses are standard errors. The predictions are those of the substantively biased + memory model.

Condition		kiCV	keCV	kaCV	giCV	geCV	gaCV
Voiceless	Observed	.39 (.10)	.36 (.10)	.12 (.09)	.14 (.06)	.11 (.06)	.09 (.05)
	Predicted	.38	.38	.03	.33	.22	.06
Voiced	Observed	.26 (.11)	.20 (.09)	.00 (.00)	.50 (.13)	.44 (.10)	.23 (.07)
	Predicted	.23	.20	.00	.47	.47	.25

Table 11. Correlations (r) between observed and predicted rates of palatalization in Experiment 2. Values in parentheses are % variance explained (r^2).

Condition	Model	All items	Critical items
Voiceless	Substantively biased	.807 (.651)	.689 (.475)
	Substantively biased + memory	.817 (.667)	.707 (.500)
	Unbiased	.800 (.640)	.684 (.468)
	Unbiased + memory	.811 (.658)	.701 (.491)
Voiced	Substantively biased	.920 (.846)	.832 (.692)
	Substantively biased + memory	.911 (.830)	.813 (.661)
	Unbiased	.875 (.766)	.753 (.567)
	Unbiased + memory	.871 (.759)	.751 (.564)

Figures

Figure 1. Mechanism of learning and generalization in the conditional random field model.

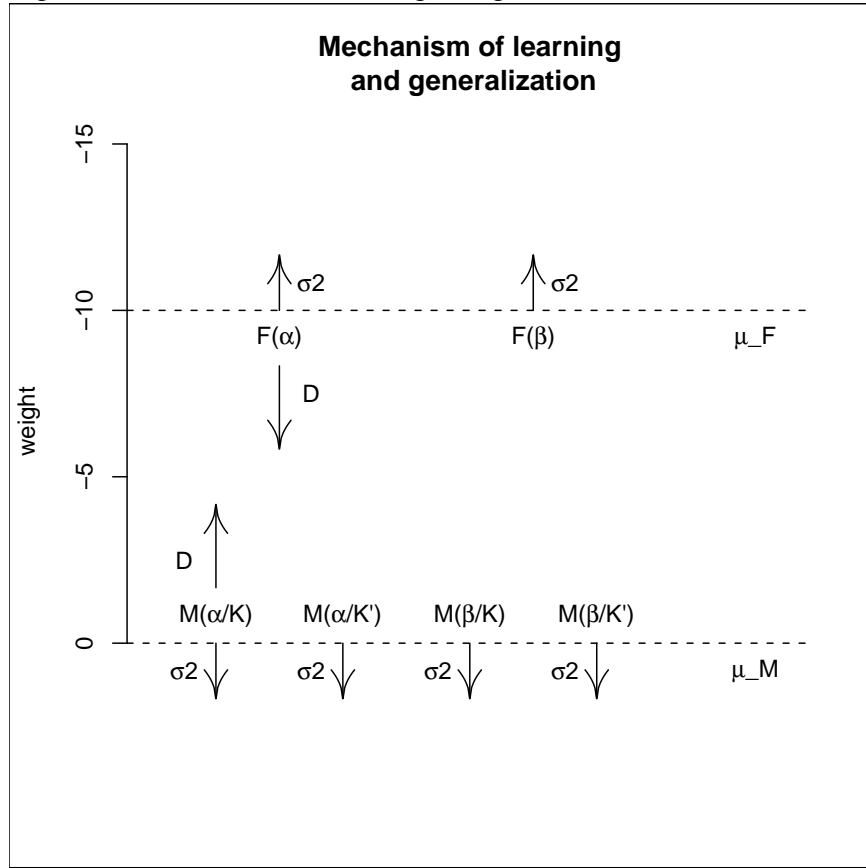
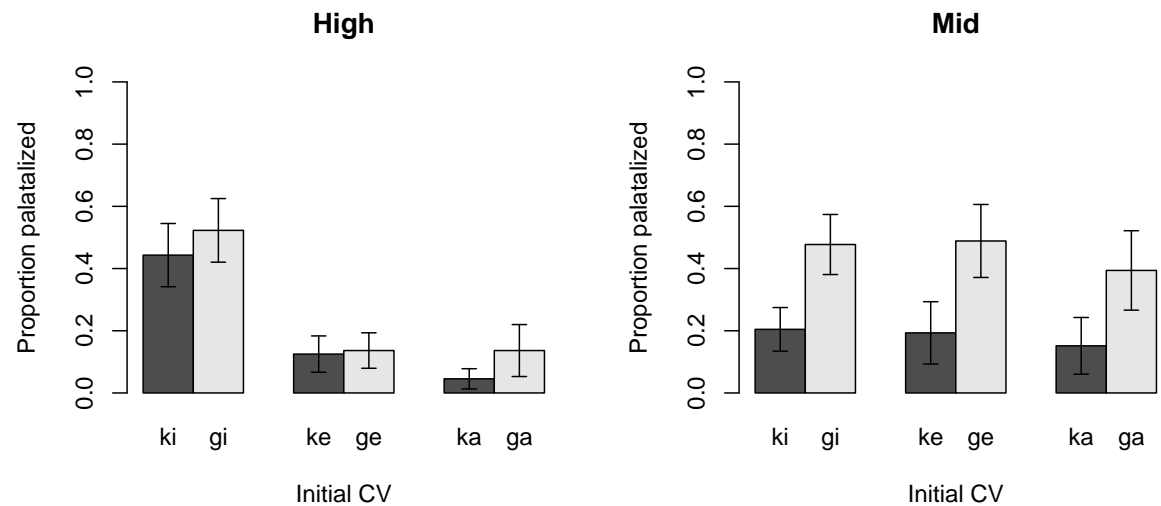


Figure 2. Results of Experiment 1 by condition.



Note. Error bars represent standard error of the mean.

Figure 3. Observed and predicted rates of velar palatalization by item in Experiment 1, High condition. Predicted values are those of the substantively biased + memory model.

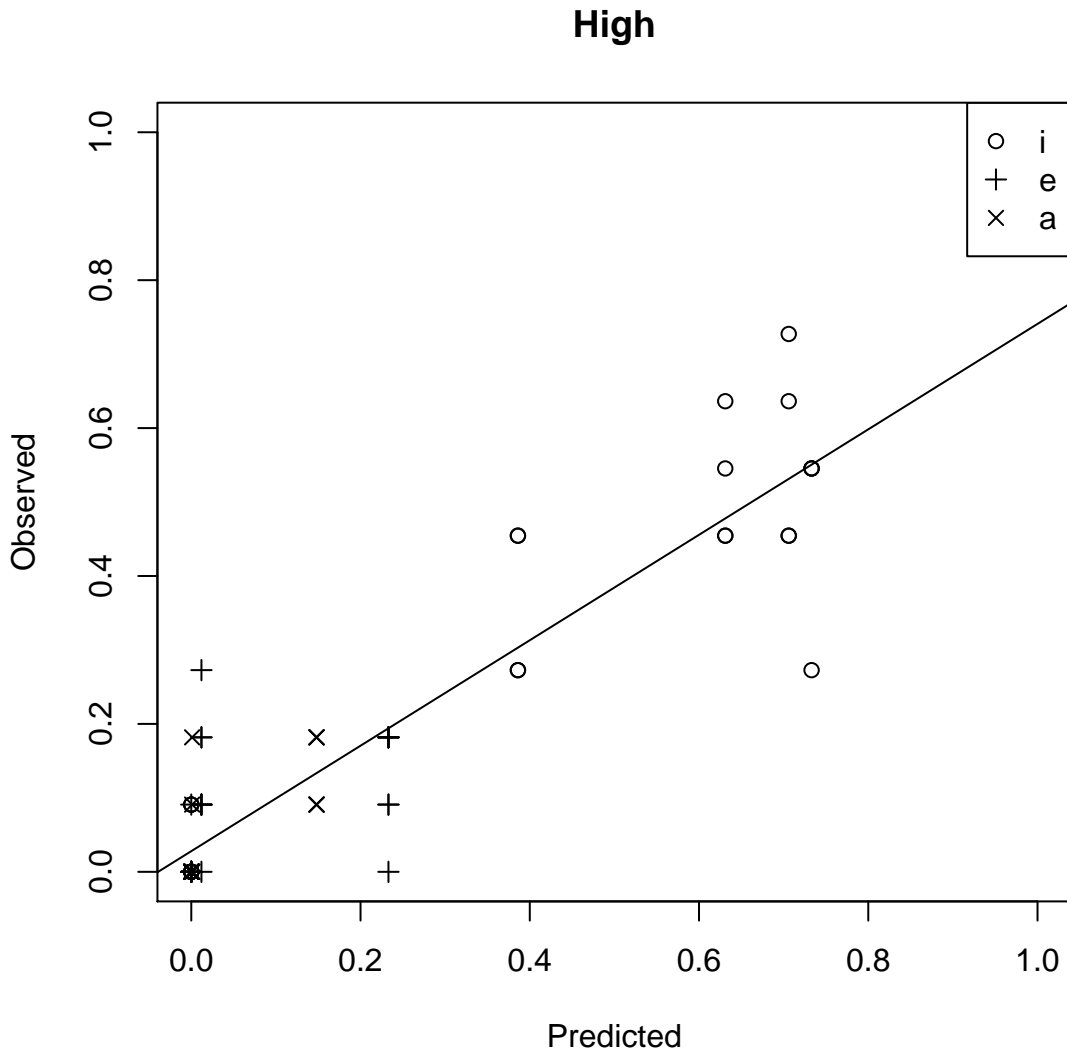


Figure 4. Observed and predicted rates of velar palatalization by item in Experiment 1, Mid condition. Predicted values are those of the substantively biased + memory model.

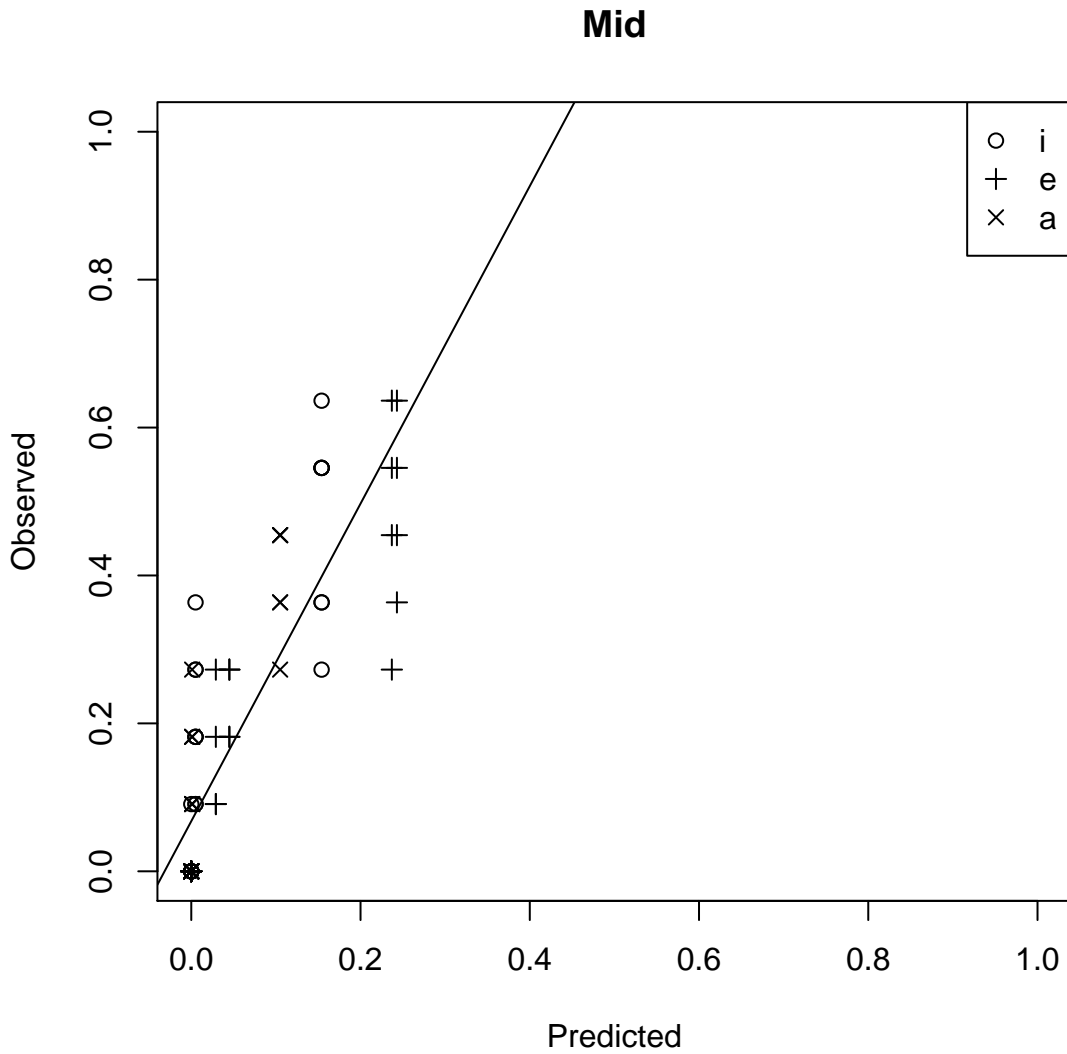
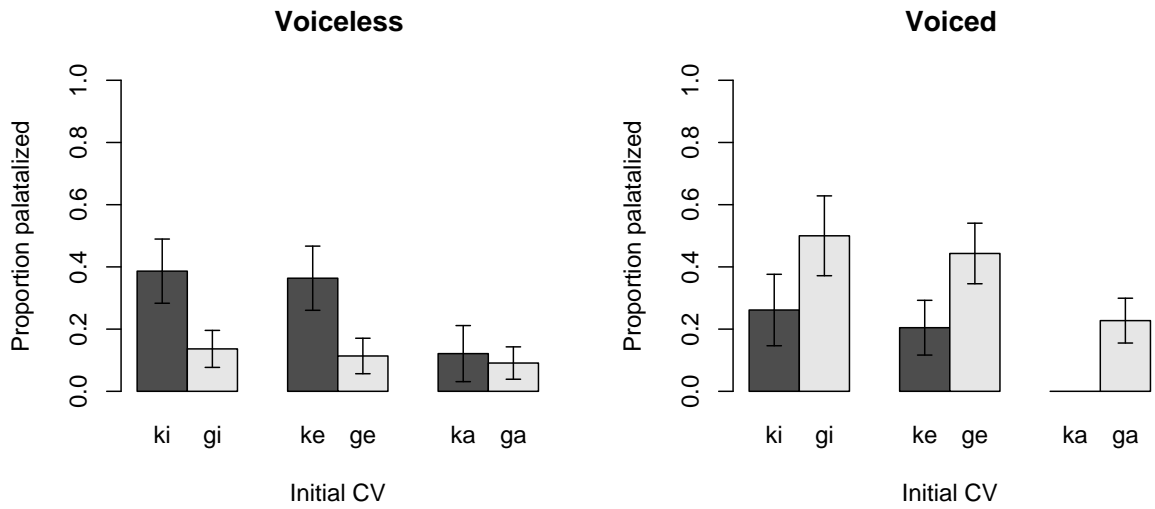


Figure 5. Results of Experiment 2 by condition.



Note. Error bars represent standard error of the mean.

Figure 6. Observed and predicted rates of velar palatalization in Experiment 2, Voiceless condition. Predicted values are those of the substantively biased + memory model.

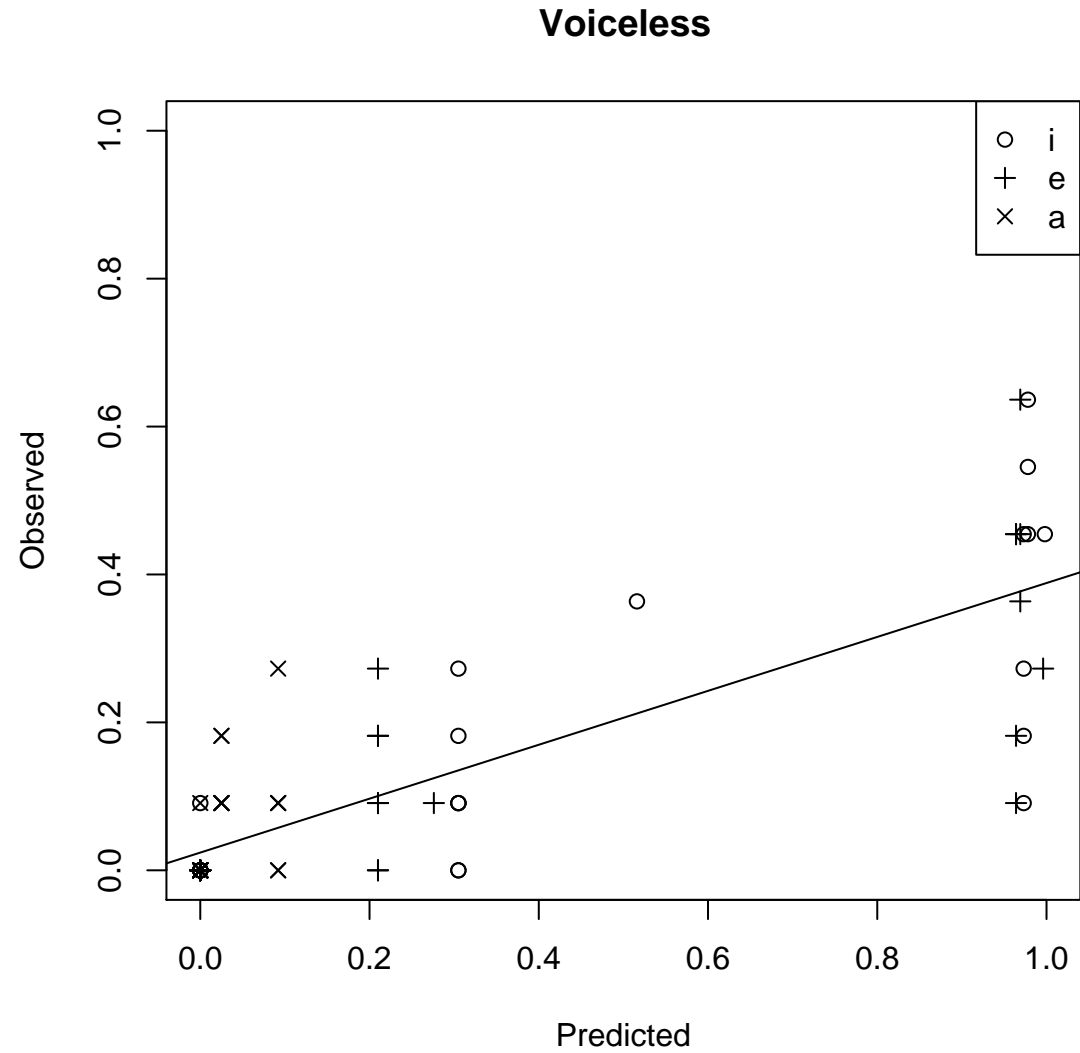


Figure 7. Observed and predicted rates of velar palatalization in Experiment 2, Voiced condition. Predicted values are those of the substantively biased + memory model.

