Cognitive Science 30 (2006) 945–982

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

Colin Wilson

Department of Linguistics, UCLA

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 acts as a bias, or prior, on phonological learning. Two experiments tested this framework with a method in which participants are first 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 *keep* to [t͡ʃ] 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. Velar 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 (Prince & Smolensky, 1993/2004) 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

Correspondence should be addressed to Colin Wilson, Department of Linguistics, UCLA, 3125 Campbell Hall, Los Angeles, CA 90095. E-mail: colin@humnet.ucla.edu

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 broader 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 auditory salience and distinctiveness.

Recent work has emphasized the importance of acoustic, auditory, and perceptual properties, an area that was previously somewhat neglected (but cf. Lindblom, 1986, Ohala, 1981, 1992; Stevens & Keyser, 1989). By studying the speech signal, as shaped by the vocal tract and processed by the auditory and perceptual systems, phonologists have gained 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, 2005; Kawasaki-Fukumori, 1992; Kochetov, 2002; Ohala, 1992; Padgett, 2004; Steriade, 2001a, 2001b; Zhang, 2001), the characteristic changes that sounds undergo in particular phonological contexts (Cho & McQueen, 2006; Côté, 2000, 2004; Jun, 1995; Steriade, 2001a, 2001b; Wilson, 2001), lexical stress systems (Gordon, 2004; Hayes, 1995; Peperkamp, 2004), the perception and production of structures that do not occur in the native language (Davidson, 2003, 2006; Dupoux, Kakehi, Hirasi, Pallier, & Mehler, 1999), and the extension of native-language phonological patterns to borrowed words (Fleischhacker, 2001; Kang, 2004; Kenstowicz, 2003; Zuraw, 2005). Two recent volumes (Hayes, Kircher, & Steriade, 2004; Hume & Johnson, 2001) testify both to the dramatic advances that have been made in integrating perception into phonology and to the pivotal role that optimality theory 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 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 optimality theory, this knowledge 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 & 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). Proponents of this framework have called attention to the fact that phonological patterns without apparent phonetic motivation can be found in natural languages (Anderson, 1974, 1981, 1985; Buckley, 1999; Chomsky & Halle, 1968; Hyman, 2001; Yu, 2004; but cf. Zsiga, Gouskova, & Tlale, 2006). 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 under certain experimental conditions (Seidl & Buckley, 2005).

The goal of this article is to develop and support a modified version of phonetically based phonology 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 phonetic naturalness. The bias is not so strong that it excludes phonetically unmotivated patterns from being acquired or productively applied.[2] Therefore, the main claim of substantively biased phonology is not that all phonological systems are phonetically natural, or that the language learner has difficulty acquiring phonetically unnatural patterns under all conditions. The main claim is instead that the learner is predisposed toward patterns that are phonetically natural. In this article, I show that this naturalness bias can be quantified and embedded within a formal theory of the learner, and that the bias is revealed under experimental conditions in which the learner is required to generalize from highly impoverished input data.

I focus on one specific type of phonological pattern, referred to throughout as *velar palatalization,* introduced in section 2; a simple example of the pattern would be the change of pronunciation from *keep* ([kip]) to *cheap* ([t͡ʃip]). 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, McCallum, & Pereira, 2001). Perhaps the most original contribution of the article 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 limited exposure. For example, in one condition of Experiment 1 (section 4) 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]. The results from both Experiment 1 and Experiment 2 (section 5) support substantively biased phonology. The biased model predicts the detailed quantitative patterns observed in the experiments better than the unbiased model, a finding that defuses the arguments from theoretical simplicity that have been advanced in favor of evolutionary phonology (Blevins, 2004; Hale & Reiss, 2000; Ohala, 1992, 1995). The 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 article, *velar palatalization* refers to the change from a velar stop consonant, voiceless [k] (as in *keep*) or voiced [g] (as in *gear*), to the corresponding palatoalveolar affricate, voiceless [t͡ʃ] (as in *cheap*) or voiced [d͡ʒ] (as in *jeer*), respectively. We examine velar palatalization before three vowels: the high front vowel [i] (as in *keep*), the midfront [e] (as in *cape*), and the low back vowel [ɑ] (as in *cop*). Simple examples appear in Table 1.[3]

Velar palatalization was selected as the focus of this article because the articulatory, acoustic, perceptual, and phonological properties of velars and palatoalveolar affricates have been studied extensively. To establish the substantive basis for the experiments and modeling that appear later, I now summarize the relevant findings.

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] → [t͡ʃi]	[i] [+high, ≈low, ≈back]	[gi] → [d͡ʒi]
[e] [≈high, ≈low, ≈back]	[ke] → [t͡ʃe]	[e] [≈high, ≈low, ≈back]	[ge] → [d͡ʒe]
[ɑ] [≈high, +low, +back]	[kɑ] → [t͡ʃɑ]	[ɑ] [≈high, +low, +back]	[gɑ] → [d͡ʒɑ]

2.1. 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 and 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 were available. A more recent study by Butcher and 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 and Tabain suggested that their data support only a binary distinction between nonback and back vowel contexts).

Fronting is relevant here because it makes the articulation of velar stops more similar to that of palatoalveolar affricates. Keating and 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 the relevant phonetic properties have not been investigated by any subsequent study.

2.2. Acoustics

The articulatory similarity of velars and palatoalveolars before front vowels gives rise to an acoustic similarity. As discussed by Keating and 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 and Tabain (2004), Guion (1996, 1998), Keating and 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 although [i] and [e] are both front vowels phonologically, [i] is phonetically further front than [e].)

Guion also measured the peak spectral frequencies in the corresponding regions of [t͡ʃ] and [d͡ʒ]. 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 in the case of articulation, there are likely to be additional acoustic properties that are shared by palatoalveolar affricates 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). All other things being equal, stops with longer frication or aspiration will be more acoustically similar to affricates.

2.3. Perception

Experiments reported in Guion (1996, 1998) establish further that velar stops and palatoalveolar affricates are more perceptually similar 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, t͡ʃ, g, d͡ʒ] followed by [i, ɑ, u]. The stimuli were excised from faster speech recordings of English words and truncated so that the duration of the vowel was 100 msec. 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 already reviewed.

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, so the cells corresponding to such errors have been left blank.

As can be seen from Table 2, the rate at which [ki] is misidentified as [t͡ʃi] is higher than the rate at which [kɑ] is misidentified as [t͡ʃɑ]. Similarly, [gi] was misidentified as [d͡ʒi] more often than [gɑ] was misidentified as [d͡ʒɑ], although the overall error rate for [g] is lower than that for [k].[4] 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 are not considered further in this article.

Table 2
Confusions of velars and palatoalveolars

	Response							
Stimulus	[ki]	[t͡ʃi]	[gi]	[d͡ʒi]	[kɑ]	[t͡ʃɑ]	[gɑ]	[d͡ʒɑ]
[ki]	43	35	10	12				
[t͡ʃi]	10	85	0	5				
[gi]	4	4	71	21				
[d͡ʒi]	9	28	12	51				
[kɑ]					84	13	3	0
[t͡ʃɑ]					10	87	0	3
[gɑ]					4	0	87	9
[d͡ʒɑ]					2	23	10	65

Note. From "The Role of Perception in the Sound Change of Velar Palatalization," by S. G. Guion, 1998, Phonetica, 55, pp. 18–52. Copyright 1998 by S. Karger AG, Basel. Adapted with permission.

An earlier study of consonant perception by Winitz, Scheib, and Reeds (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 [t͡ʃi], and selected [t] as the available response that was most faithful to their perception.

2.4. Phonology

As originally observed by Ohala (1992) and expanded on by Guion (1996, 1998), there is a striking relation between the phonetic and perceptual facts discussed earlier 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 [ɑ] ([kɑ] → [t͡ʃa] and [gɑ] → [d͡ʒɑ]), then it is also expected to palatalize velars before the front vowels [i] and [e] ([ki] → [t͡ʃi], [gi] → [d͡ʒi], 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). If a velar stop and its corresponding palatoalveolar affricate are more similar, as measured by confusion rate, in context C than in context C′, then velar palatalization in C′ asymmetrically implies velar palatalization in C.

A related finding is that, in the lexicons of many languages, velar stops cooccur with front vowels, and in particular [i], less often than would be expected by change (Maddieson & Precoda, 1992). This finding is relevant because it illustrates a well-known property of phonological typology, namely that the same forces that lead to changes in some languages (e.g., [ki] → [t͡ʃi]) are visible in the static distribution of sounds in other languages (e.g., relative rarity of [ki]). In this case, we can trace both dynamic and static patterns back to the same relation of perceptual similarity.

2.5. Summary

The study of velar palatalization presents us with a near-perfect correlation between phonetics 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]). Front vowels condition

velar palatalization more strongly in attested phonological systems (i.e., palatalization before a front vowel asymetrically 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).

In the framework of phonetically based phonology, such correlations have been taken to reveal a cognitive principle that privileges alternations between perceptually similar sounds (Steriade, 2001a, 2001b; 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] → [t͡ʃ] before [i], relative to 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 phonetic 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, in press; 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 velar stops and palatoalveolar affricates 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 or 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, although 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 article 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 made quantitatively precise. 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). Substantively biased phonology nevertheless makes falsifiable predictions about how novel phonological patterns will be generalized from impoverished input. The experiments in sections 4 and 5, which involve briefly exposing participants to "language games" involving velar palatalization and then measuring the degree to which they generalize the games to new phonological contexts, were designed to test some of these predictions. The experiments reveal

that participants generalize in a way that accords with the first implicational law discussed earlier (i.e., palatalization before more back vowels implies palatalization before more front vowels), a finding that supports substantively biased phonology over evolutionary and other emergentist alternatives (de Boer, 2001; Kirchner, 2005; Redford, Chen, & Miikkulainen, 2001).

3. Substantively biased phonology

In this section, I introduce the framework of 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 (2001a, 2001b, 2001c), 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 or 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 psychological similarity on the dimensions that define the stimuli 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.

$$d_{ij} = c[\sum_{m=1}^{M} w_m \mid x_{im} - x_{jm} \mid^r]^{1/r} \tag{1}$$

The index m runs over the stimulus dimensions (e.g., x_{im} is the psychological value of stimulus x_i on dimension m). 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 = [ɑ]), and a real-valued peak spectral frequency dimension. Clearly it is the third dimension that is of central interest; as discussed later, values on this dimension were transformed to the Bark scale to better match the perceptual (as opposed to acoustic) relations among the stimuli (see Johnson, 1997). The other two dimensions were included to allow the model to be fit to the confusion matrix in Table 2.

Each stimulus dimension has an attention weight w_m. The weights on all dimensions are constrained to be nonnegative and to sum to 1 $\left(\sum_{m=1}^{M} w_m = 1\right)$. There is also a scale parameter c, constrained to be nonnegative, 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* (i.e., not perceived separately from one another; Nosofsky 1986). Indeed, Benkí (1998) established that place of articulation (here, velar vs. palatoalveolar) and voicing (voiceless vs. voiced) are perceived in a nonseparable 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 psychological distance between the points increases (Nosofsky, 1984, 1986; Shepard, 1957, 1987).

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

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; however, the special case given in Equation 2 (corresponding to $p = 1$) was found to be sufficient for this study.

The final equation projects perceptual similarities onto predicted confusion rates according to the Luce choice rule (R. D. 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.[5]

$$P(response = x_j \mid stimulus = x_i) = \frac{b_j \eta_{ij}}{\sum_{n=1}^{N} b_n \eta_{in}} \tag{3}$$

This equation presupposes that every stimulus x_n has an associated response bias b_n (all of the response biases are required to be nonnegative and to sum to one: $\left(\sum_{n=1}^{N} b_n = 1\right)$. In this case, the response bias parameters allow the model to capture the finding, noted in section 2, that the confusion rates for velar stops and palatoalveolar affricates are asymmetric (e.g., [k] is misidentified as [t͡ʃ] much more often than [t͡ʃ] 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 (see Nosofsky, 1991, for general discussion).

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 Equations 1, 2, and 3 with the maximum likelihood (ML) method (Nosofsky, 1986; Nosofsky & Zaki, 2002; see also Myung, 2003, for a general presentation). The likelihood equation used here was the one given in Nosofsky and 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 attention weights ($\{w_k\}$), response biases ($\{b_j\}$), and the scale (c).

The confusion matrix given in Table 2 and stimulus values for tokens of [ki, t͡ʃi, kɑ, t͡ʃa, gi, d͡ʒi , gɑ, d͡ʒɑ] 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

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

[ki]/[t͡ʃi]	[ke]/[t͡ʃe]	[kɑ]/[t͡ʃɑ]	[gi]/[d͡ʒi]	[ge]/[d͡ʒe]	[gɑ]/[d͡ʒɑ]
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.

reported in sections 4 and 5. All spectral frequencies were converted from Hz to the Bark scale, which better matches auditory and perceptual similarity relations for stimuli, like the present ones, that have significant high-frequency components; conversions were performed 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 negative log likelihood of 64.6, and the Kullback–Leibler divergence (Cover & Thomas, 1991) between the observed confusion proportions and the predicted confusion probabilities was 1.44.

Because Guion's (1998) confusion matrix contains information for the vowels [i] and [ɑ], but not [e], only the perceptual similarities of [ki] / [t͡ʃi], [gi] / [d͡ʒi], [kɑ] / [t͡ʃɑ], and [gɑ] / [d͡ʒɑ] 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 earlier, 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] / [t͡ʃe] and [ge] / [d͡ʒe] were determined by interpolation from the ML fit. The peak spectral frequencies for velars and palatoalveolars before [e] were taken from Guion (1996, 1998), and the response biases for [t͡ʃe] and [d͡ʒe] were set equal to those for [t͡ʃi] and [d͡ʒi], 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 these experiments.)

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 midfront [e] to low back [ɑ]). We return to these values at the end of the next subsection.

3.2. Conditional random fields for phonology

Lafferty et al. (2001) introduced a general framework, referred to as CRF, and applied it to the problem of labeling sequences (see also Gregory & Altun, 2004; McCallum 2003; Roark, Saraclar, Collins, & Johnson, 2004; Sha & Pereira, 2003, among others). Many 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 [t͡ʃinə] (where [ə] is the final vowel in *rhumb*a). This grammar can also be thought of as assigning an output label to each sound in the input: /k/:[t͡ʃ], /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 optimality theory.[6]

In the most general terms, a CRF defines a probability distribution over a set of output random variables y given values for a set of input random variables x. Each output variable takes on a value in the finite set $\mathcal{Y}$. In this setting, we identify the input variables x with the sequence of sounds in one phonological form (the input) and the output variables 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 relations 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 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 purposes here, arranges the output vertices into a chain. Given an input sequence $x = (x_1, x_2, \ldots, x_n)$, there is a corresponding output sequence $y = (y_1, y_2, \ldots, y_n)$. 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}), $1 \le i \le n-1$). In this setting, the input–output mapping of /kinə/ to [t͡ʃinə] can be written with coindexation: /$k_1i_2n_3ə_4$/, [t͡$ʃ_1i_2n_3ə_4$].

As demonstrated by Hammersley and 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):

$$P(\mathbf{y} \mid \mathbf{x}) = Z_{\mathbf{x}}^{-1} \exp\left(-\sum_{m=1}^{M} \lambda_m f_m(\mathbf{x}, \mathbf{y})\right) \tag{4}$$

where $\exp\left(-\sum_{m=1}^{M} \lambda_m f_m(\mathrm{x}, \mathrm{y})\right)$ is the potential (discussed further later)[7] and Z_X is the partition function with respect to input x:

$$Z_{\mathbf{x}} = \sum_{\mathbf{y}'} \exp\left(-\sum_{m=1}^{M} \lambda_m f_m(\mathbf{x}, \mathbf{y}')\right) \tag{5}$$

Equations (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 analog of the optimality theory claim that the grammatical output is selected by competition among all possible candidate outputs.

The potentials $\exp\left(-\sum_{m=1}^{M} \lambda_m f_m(\mathrm{x}, \mathrm{y})\right)$ have a close relation to the notion of Harmony in optimality theory—and an even closer relation to Harmony in harmony theory (HT; Smolensky, 1986; Smolensky & Legendre, 2005)—therefore I refer to them with the term *CRF-harmony*. CRF-harmony is defined in terms of a set of functions $\{f_m\}$, each of which evaluates input–output pairs.

In standard CRF terminology, the f_m functions are referred to as *features,* but we think of them as constraints like those in optimality theory. Each f_m is a function from input–output mappings (more precisely, cliques in the input–output mapping) to the nonnegative integers; this conception of constraints as functions from candidates to violation levels is familiar from Eisner (1997), Samek-Lodovici and Prince (1999), and others. A constraint has a real-valued weight λ_m, which we take throughout to be nonnegative, thereby expressing the central optimality theory tenet that constraint violations decrease harmony—that is, in the probabilistic setting, more violations imply lower probability. The entire set of weights for M constraints is 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_m(x, y)). Multiply each violation score by the corresponding weight ($\lambda_m f_m$(x, y)). Sum the weighted violations $\left(-\sum_{m=1}^{M} \lambda_m f_m(\mathrm{x}, \mathrm{y})\right)$. Finally, raise e (the base of the natural logarithm) to the negative of that sum, written as $\exp\left(-\sum_{m=1}^{M} \lambda_m f_m(\mathrm{x}, \mathrm{y})\right)$.

One main difference between the CRF model and optimality theory 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′) if and only if the value of the CRF-harmony is greater for the former than for the latter. Optimality theory harmony, on the other hand, is determined by lexicographic comparison of constraint violations: A pair (x, y) is more harmonic than another (x, y′) if and only if the highest ranked constraint that distinguishes between the two pairs prefers the latter (Prince & Smolensky, 1993/2004). (Optimality theory rankings could also be expressed with real-valued weights, but only the ordering of the weights would be relevant.)

There were two motivations in this context for adopting a CRF rather than an optimality theory 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 article. Second, globally optimal CRF weights can be learned from a body of training data using standard gradient-based optimization algorithms (Lafferty et al., 2001; see Boyd & Vandenberghe, 2004, for the general picture). No current instantiation of optimality theory has both of these advantages. Although the original formulation of the theory by Prince and Smolensky (1993/2004) has a correct and convergent ranking algorithm (Tesar & Smolensky, 1998, 2000), it does not readily generate probability distributions. The alternative formulation known as stochastic optimality theory (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 (B. Hayes, personal communication, July 1, 2005). The same learning problem holds, as far as I know, for other varieties of optimality theory that define probability distributions.[8]

Learning in this article was performed by minimizing the objective function in Equation 6 (Goldwater & Johnson, 2003; Lafferty et al., 2001; McCallum, 2003) with gradient-based optimization. The learning process is batch rather than online; that is, the learner is assumed to be given an entire sequence D of training data, where D consists of N input–output pairs (i.e., $D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\}$).

$$L_\Lambda = [-\sum_{n=1}^{N} \log P_\Lambda(\mathbf{y}^{(n)} \mid \mathbf{x}^{(n)})] + [\sum_{m=1}^{M} \frac{(\lambda_m - \mu_m)^2)}{2\sigma_m^2}] \tag{6}$$

Equation 6 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, Myung, & Pitt, 2005; Smolensky, 1996). The first term is the negative log probability, given the weights Λ, of the outputs in the data $\{y^{(1)}, y^{(1)}, \ldots y^{(N)}\}$ given the inputs $x^{(1)}, x^{(1)}, \ldots x^{(N)}$ under the assumption that the input–output pairs are independent of one another. 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_m the regularizer specifies a target value (μ_m) and imposes a penalty for deviating from that value. An important relation in the following will be that smaller values of σ_m yield greater penalties for deviating from μ_m. 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 optimality theory but not common practice in work on CRFs. 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 optimality theory 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 [t͡ʃ]; F(g) is violated when input /g/ corresponds to output [d͡ʒ] . I assume further that all other input–output disparities run afoul of an inviolable Faithfulness constraint. This is not a realistic assumption for all of phonology, but it accords with the design and results of the experiments in sections 4 and 5.

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 [ɑ]) 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 section.)

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., /$k_1i_2n_3ə_4$/), an assumed inviolable Faithfulness constraint eliminates all but two of the logically possible candidate outputs: the fully faithful candidate (e.g., [$k_1i_2n_3ə_4$]) 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., [$t͡ʃ_1i_2n_3ə_4$], where output $[t͡ʃ]_1$ corresponds to input /$k/_1$). 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., [$k_1i_2n_3ə_4$] violates *ki, *$kV_{[-low]}$, and *kV). Conversely, the velar palatalization candidate violates $F(\alpha)$, but satisfies the Markedness constraints completely.

Table 4
Markedness constraints on palatalization

	Prior Values			
	Biased		Unbiased	
Constraint	μ	σ^2	μ	σ^2
*ki	0.0	9.23^{-2}	0.0	10^{-2}
*ke	0.0	12.68^{-2}	0.0	10^{-2}
*kɑ	0.0	88.72^{-2}	0.0	10^{-2}
*$kV_{[-low]}$	0.0	12.68^{-2}	0.0	10^{-2}
*$kV_{[-high]}$	0.0	88.72^{-2}	0.0	10^{-2}
*kV	0.0	88.72^{-2}	0.0	10^{-2}
*gi	0.0	21.13^{-2}	0.0	10^{-2}
*ge	0.0	40.60^{-2}	0.0	10^{-2}
*gɑ	0.0	126.93^{-2}	0.0	10^{-2}
*$gV_{[-low]}$	0.0	40.60^{-2}	0.0	10^{-2}
*$gV_{[-high]}$	0.0	126.93^{-2}	0.0	10^{-2}
*gV	0.0	126.93^{-2}	0.0	10^{-2}

In optimality theory, 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 Equation 7, which is the special case of Equations 4 and 5 when there are exactly two candidates. I have made the substitution $H(\mathrm{x},\mathrm{y}) = \exp\left(-\sum_{m=1}^{M} \lambda_m f_m(\mathrm{x},\mathrm{y})\right)$ to bring out the essence of the competition.

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

Recalling that stronger constraints have weights that are further above 0, we see that increasing the weight of Faithfulness relative to Markedness makes the palatalization output y^{pal} less probable. Conversely, increasing the weight of Markedness relative to Faithfulness makes the palatalization output 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 (2001a, 2001b) 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] → [t͡ʃi] than to the change [kɑ] → [t͡ʃɑ].

The key idea is to impose a systematic relation between similarity values (as in Table 3) and the σ parameters for the Markedness constraints in the CRF. Recall that the Markedness con-

straints are the ones that force phonological changes, in this case by eliminating velar stops before certain vowels. Recall also that the smaller the σ_m for a given constraint f_m, the more tightly the prior or regularizer holds the constraint's weight to the target value μm_m. 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_m for a Markedness constraint f_m is determined according to the following steps. First find all of the changes that can be forced by f_m. Among these changes, find the one 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_m 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 values for the Markedness constraints assumed here.[9]

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 constraints have a target weight $\mu_M = 0$, whereas all of the Faithfulness constraints have a target weight that is substantially greater. In the simulations reported here, I use 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.[10]

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

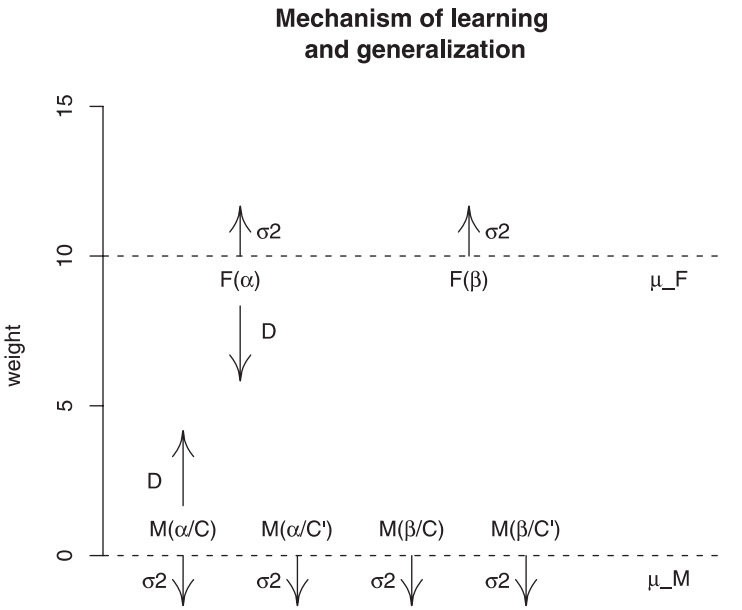

Fig. 1. Mechanism of learning and generalization in the conditional random field model.

adult state, the model is exposed to training data in which a velar stop α (i.e., [k] or [g]) undergoes palatalization in context *C* (i.e., before one of the three vowels [i], [e], or [ɑ]). The force labeled *D* shown in Fig. 1 is due to the training data (recall Equation 6). It pulls the weight of M(α/*C*) upward (away from 0) and the weight of F(α) 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 $\sigma_{M(\alpha/C)}$ and $\sigma_{F(\alpha)}$.

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

3.2.2.2. Case II. Things work out differently if $\sigma_{M(\alpha/C)}$ is substantially smaller than $\sigma_{F(\alpha)}$. The 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*(α) weight lowers so far that it becomes roughly equivalent to the target weight μ of some other constraint M(α/*C*′), then the system will to some extent generalize palatalization of α from the context *C* to the new context *C*′, even though no examples of palatalization in *C*′ appeared in the training data. I refer to this type of behavior as generalization on the context.

The link between perceptual similarity and σ values tells us when to expect each case. 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 some degree of generalization on the context should be found. Note that the precise prediction is generalization of [g] palatalization to two environments—both [i] and [ɑ]—because *F*(g) will descend within range of both *gi and *gɑ. Generalization to the [i] context would be predicted under any sensible implementation of substantive bias; generalization to the [ɑ] context is a more subtle consequence of the current implementation, one we see borne out in Experiment 1 (see section 4).

Because of the analytic decision to distinguish the Faithfulness constraints *F*(k) and *F*(g) (as opposed to collapsing them into a single constraint *F*(velar-stop)), the system makes a further prediction: namely, that what I refer to as generalization on the focus 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. Both experiments reported later support this prediction; I discuss other converging evidence and potential theoretical explanations for the distinct faithfulness constraints in section 5.3.

To summarize, the substantively biased model of phonology just developed 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, mir-

roring the asymmetries of substance, and follow in a nontrivial way from the representations and computations of the model.

The following is a brief review of the components of substantively biased phonology that have already been introduced:

- Perceptual similarity among speech sounds is quantified with the GCM, which uses the psychological values of the stimuli along specified dimensions, along with data from perceptual confusions, to estimate similarities ($\{\eta_{ij}\}$) and response biases ($\{b_j\}$). Biased similarities ($\{b_j\ \eta_{ij}\}$) account for asymmetric relations among the sounds.
- The grammar takes the form of a CRF that assigns a probability to each output under consideration, given the input. The CRF is in turn defined in terms of a set of constraints ($\{f_m\}$) and nonnegative weights ($\{\lambda_m\}$). The constraints are essentially as in optimality theory, but numerical weighting yields a different type of constraint interaction, one that is much closer to HG.
- Each constraint f_m in the grammar has an associated target weight μ_m and a standard deviation σ_m. In the current application, μ_m values are appropriate for the native language of the adult participants in Experiments 1 and 2 (i.e, $\mu_{\text{Markedness}} < \mu_{\text{Faithfulness}}$ because the native language does not have velar palatalization in any context). The σ_m values constitute the substantive bias, which is plausibly common to both adult and child learners. Taking advantage of the relation that smaller σ_m implies greater penalty for deviating from μ_m, the value of σ_m has been set to the biased perceptual similarity of the greatest change that can be forced by f_m. Within the system of interacting constraints, this quantifies the idea that changes involving more perceptually distant sounds are dispreferred.

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 already with an unbiased instantiation in which the α 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. The experimental results presented next provide evidence against this more empiricist theory of phonological learning.

4. Experiment 1: Testing generalization on the context

Language games are naturally occurring phenomena that involve altering the pronunciation of words 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. Related work on natural language games appears in Barlow (2001) and Nevins and Vaux (2003). Previous studies that have

used experimentally constructed language games include McCarthy (1982), Moreton, Feng, and Smith (2005), Pierrehumbert and Nair (1995), and Treiman (1983).

The experimental method employed here is fairly straightforward and directly motivated by the model introduced in section 3. In the first part of the experiment, participants are presented with spoken examples of a novel language game. For example, participants in the mid condition of this experiment heard examples such as [kenə] … [t͡ʃenə] and [gepə] … [d͡ʒepə], which illustrate palatalization of velar stops before the midvowel [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]. Participants in the high condition heard examples illustrating palatalization of velar stops before the high vowel [i], but were not presented with any examples in which the velar stops appeared before [e].

In the second part of the experiment, participants are presented with words containing the velar stops in three vowel contexts ([i e ɑ]). For each participant, exactly one of the vowels [i e] conditioned velar palatalization in the first part of the experiment; I refer to this vowel as the *exposure* context. The other member of [i e] did not appear after a velar stop in the first part of the experiment; I refer to this vowel as the *novel* context. The relative rates of velar palatalization in the exposure and novel contexts constitute the dependent measure of interest. If participants exposed to examples such as [kenə] … [t͡ʃenə] extend velar palatalization to words such as [kinə], but participants exposed to examples such as [kinə] … [t͡ʃenə] do not extend the change to words such as [kenə], this will support 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 on 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). However, there can be no debate about the degree of impoverishment in a PSM experiment (although, of course, the adult participants will bring other knowledge, not provided by the experiment, 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.

This 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 ɑ]. C_2 came from [p b k g m n f v θ ð s z t͡ʃ d͡ʒ 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 [t͡ʃ] and [d͡ʒ], 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 ɑ]) were fully crossed in the stimulus set. For each C_1V_1 combination, a phonetically balanced set of following C_2s 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] → [t͡ʃ] or [g] → [d͡ʒ]). 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 the Appendix.

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, MacWhinney, Flatt, & Provist, 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 2-sec 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. 250 msec later the first member of a stimulus pair was played from the speakers.
4. There was a 1-sec 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. 250 msec 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 or 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 ([bɑlə] … [bɑlə]), and one that illustrated velar palatalization (High: [gibə] … [d͡ʒibə]; Mid: [gefə] … [d͡ʒefə]).[11] The practice phase was followed by 32 exposure trials, as schematized in Table 5. The trials in the exposure phase were divided into four blocks of eight trials each. A block contained two examples of velar palatalization (one each of [k] → [t͡ʃ] and [g] → [d͡ʒ]), two examples of velars that did not palatalize (one instance each of [k] and [g] before [ɑ]), and four 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).

After the exposure phase, there was a 2-min break during which participants worked on pencil-and-paper math problems. The problems were designed to be of moderate difficulty (multiplication of two three-digit numbers) and were identical across all participants. Participants were informed that the problems were designed to occupy their time 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 2-min 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.[12]

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 two critical items used for the practice trials removed. The testing list was thus exactly the same for

Table 5
Exposure trials for the two conditions in Experiment 1

Condition	Trial Type (number)
High	kiCV … t͡ʃiCV (4) giCV … d͡ʒiCV (4)
Mid	keCV … t͡ʃeCV (4) geCV … d͡ʒeCV (4)
Both	kɑCV … kɑCV (3) gɑCV … gɑCV (3)
	piCV … piCV (3) biCV … biCV (3)
	peCV … peCV (3) beCV … beCV (3)
	pACV … pɑCV (3) bɑCV … bɑCV (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)
kɑCV … (6) gɑCV … (6)	pɑCV … (6) bɑCV … (6)

both conditions. The trials were distributed as schematized in Table 6 and presented in an order that was randomized for each participant without blocking.

In the kɑCV, gɑCV, and filler categories (pɑCV, bɑCV), 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 bɑCV 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. Although the exposure and testing phases were self-paced, they had quite similar durations across participants. The exposure phase lasted approximately 4 min and the testing phase lasted approximately 7.5 min.

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 five responses (less than 1% of all total responses) were of this type. The following statistical analysis focuses on the rate of palataliza-

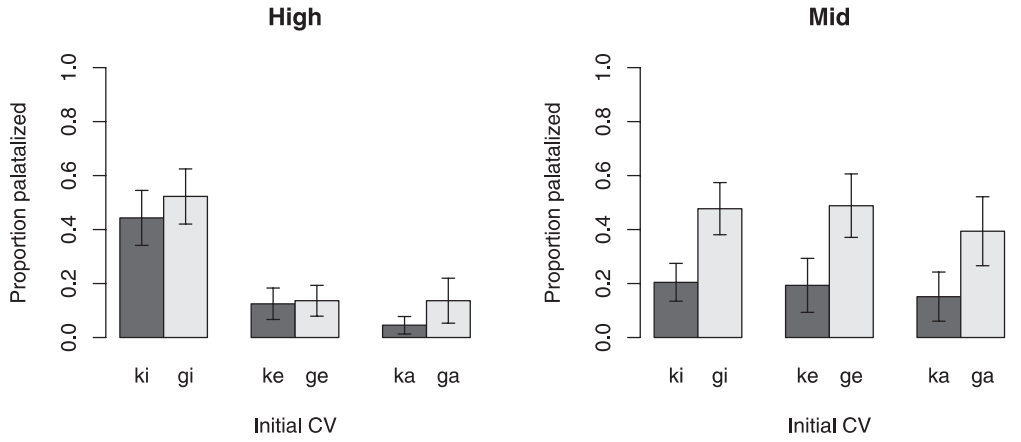

Fig. 2. Results of Experiment 1 by condition. *Note.* Error bars represent standard error of the mean.

tion of critical testing items (i.e., items that began with [ki], [ke], [kɑ], [gi], [ge], or [gɑ]). Fig. 2 displays the palatalization rate for each type of critical item in each condition. Table 7 gives the means and standard errors.

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 for a particular participant is the vowel that conditioned velar palatalization in the Exposure phase: [i] for participants in the High group, [e] for those in 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 later to the issue of generalization before the low vowel [ɑ].)

Table 7
Mean observed rates of velar palatalization for critical item types in Experiment 1

Condition	kiCV	keCV	kɑCV	giCV	geCV	gɑCV
High	.44 (.10)	.13 (.06)	.05 (.03)	.52 (.10)	.14 (.06)	.14 (.08)
Mid	.20 (.07)	.19 (.10)	.15 (.09)	.48 (.10)	.49 (.12)	.39 (.13)

Note. Values in parentheses are standard errors.

A repeated-measures analysis of variance (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, $t(10) = 3.0$, $p < .05$. However, 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, $t(10) = 0$, $p = 1$.[13]

In qualitative terms, the participants in the Mid condition, but not those in the High condition, generalized velar palatalization from the exposure context to the novel context. This effect is quite reliable across participants. In the High condition, 8 out of 11 participants exhibited a greater rate of palatalization before [i] than before [e], with the difference between the palatalization rates before the two vowels ranging from 1.0 to –.13 across participants. However, only 5 out of 11 participants in the Mid condition showed greater rates of palatalization before [e] than before [i], and the difference between the palatalization rates before the two vowels fell within a much smaller range across participants in this condition (.13 to –.19).

The ability of the biased and unbiased instantiations of the CRF model (section 3) to capture this asymmetric pattern of generalization was tested by fitting the models to the aggregate data for each group.[14] Both instantiations of the model had a single free parameter 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.

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

Table 8
Correlations (r) between observed and predicted rates of palatalization in Experiment 1

Condition	Model	All Items	Critical Items
High	Substantively biased	.910 (.83)	.870 (.76)
	Unbiased	.913 (.83)	.871 (.76)
Mid	Substantively biased	.859 (.74)	.758 (.58)
	Unbiased	.550 (.30)	.396 (.16)

Note. Values in parentheses are percentage variance explained (r^2).

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 [ɑ] context (in spite of the fact that velars did not palatalize before [ɑ] 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 44% less of the total variance, and 42% less of the variance for the critical items, in the mid condition.

There was one qualitative feature of the results that was not predicted by the substantively biased model: the relatively high rate at which palatalization was extended to [ki] in the 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 M. Gordon, personal communication, January 20, 2004, for this suggestion). Recall that the similarity of [k] and [t͡ʃ] before [e] was estimated by interpolation. These 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 future 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 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 an otherwise identical model that lacks substantive bias—the results reported here shift the debate from speculation over the source of typological distribution to experimental investigation of human learning (see also Pater & Tessier, 2003; Pycha, Nowak, Shin, & Shasted, 2003; Wilson, 2003; Zhang & Lai, 2005; Zuraw, 2005).

5. Experiment 2: Testing generalization on the focus

As noted in section 3, both the substantively biased and unbiased instantiations of the CRF 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 this 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 recordings 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 ([bɑlə] … [bɑlə]), and two in which the members of the stimulus pair were related by velar palatalization (Voiceless: [kiwə] … [t͡ʃiwə] and [kenə] … [t͡ʃenə]; Voiced: [gipə] … [d͡ʒipə] and [gefə] … [d͡ʒefə]).

During the exposure phase there were 34 trials, as schematized in Table 9. The trials were grouped into four blocks. Each block contained two examples of velar palatalization ([k] → [t͡ʃ] or [g] → [d͡ʒ] before [i] and [e]), one or two examples of velars that did not palatalize ([k] or [g] before [ɑ]), and four or five fillers. The blocks that contained four fillers also included one example in which velar palatalization applied to the novel voicing category; that is, participants in the Voiceless condition heard exactly two examples of palatalization of voiced [g] (one before [i] and one before [e]), and participants in the Voiced condition heard exactly two examples of palatalization of voiceless [k] (one before [i] and one before [e]). These items were included to encourage generalization during testing—a manipulation that was not suc-

Table 9
Exposure trials for the two conditions in Experiment 2

Condition	Trial Type (Number)
Voiceless	kiCV … t͡ʃiCV (4) keCV … t͡ʃiCV (4)
	giCV … d͡ʒiCV (1) geCV … d͡ʒeCV (1)
Voiced	giCV … d͡ʒiCV (4) geCV … d͡ʒeCV (4)
	kiCV … t͡ʃiCV (1) keCV … t͡ʃeCV (1)
Both	kACV … kɑCV (3) gɑCV … gɑCV (3)
	piCV … piCV (3) biCV … biCV (3)
	peCV … peCV (3) beCV … beCV (3)
	pACV … pɑCV (3) bɑCV … bɑCV (3)

cessful, as we will see. The order of the blocks and the order of items within blocks were randomized across participants.

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 responses 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 two 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. 3 displays the palatalization rate for each type of critical item in each condition. Table 10 gives the means and standard errors.

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

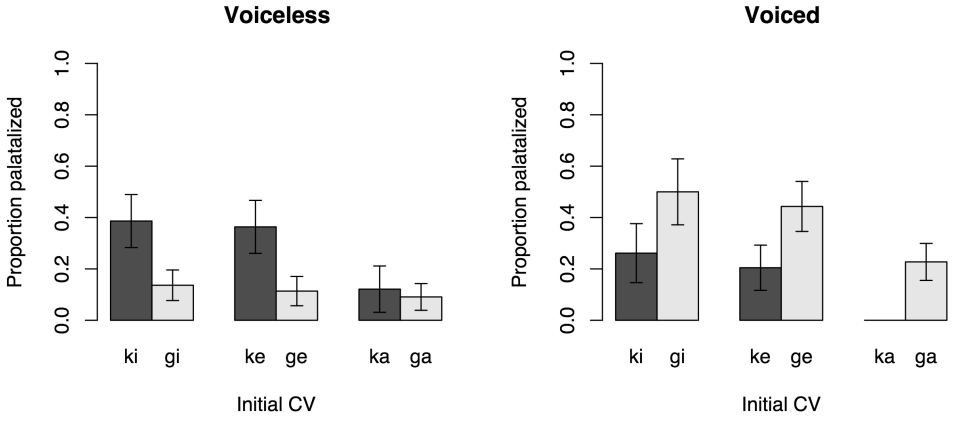

Fig. 3. Results of Experiment 2 by condition. *Note.* Error bars represent standard error of the mean.

Table 10
Mean observed rates of velar palatalization for critical item types in Experiment 2

Condition	kiCV	keCV	kɑCV	giCV	geCV	gɑCV
Voiceless	.39 (.10)	.36 (.10)	.12 (.09)	.14 (.06)	.11 (.06)	.09 (.05)
Voiced	.26 (.11)	.20 (.09)	.00 (.00)	.50 (.13)	.44 (.10)	.23 (.07)

Note. Values in parentheses are standard errors.

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 nonsignificant. 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.[15]

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. The biased model significantly outperforms the unbiased model with respect to the voiced condition—the condition in which the participants extended palatalization to the [ɑ] context most strongly. In that condition, the biased model explained approximately 8% more of the total variance, and 12% more of the variance for the critical items, than the unbiased model.

5.3. Discussion

The results of this experiment support the prediction that velar 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 palatalization better than the unbiased model.

The lack of generalization on the focus 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 find-

Table 11
Correlations (r) between observed and predicted rates of palatalization in Experiment 2

Condition	Model	All Items	Critical Items
Voiceless	Substantively biased	.807 (.65)	.689 (.48)
	Unbiased	.800 (.64)	.684 (.47)
Voiced	Substantively biased	.920 (.85)	.832 (.69)
	Unbiased	.875 (.77)	.753 (.57)

Note. Values in parentheses are percentage variance explained.

ing of speech perception experiments that the [voice] specification of a stop is perceptually robust, much more so than 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 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.

K. Zuraw (personal communication, May 21, 2005) 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 rare. Again, we might expect a nuanced version of this generalization, according to which a nonnative sound is subject to native phonology in proportion to how strongly it perceptually resembles native sounds, to be consistent with the entire body of data.

These 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 led 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 these 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. However, 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 [kʲi, gʲi]). The typological rarity or nonexistence of palatalization of voiced velars only could then follow from the hypothesis that learners are unlikely to reinterpret a heavily coarticulated [gʲi] as [d͡ʒi] without also reinterpreting a heavily coarticulated [kʲi]—which is perceptually more similar to the corresponding palatoalveolar affricate—as [t͡ʃi]. 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 article 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 CRF. Two PSM experi-

ments 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, these 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, Zavrel, van der Sloot, & van den Bosch, 2003; Kirchner, 2005).[16] 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. This 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 and 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, Pierrehumbert, & Broe, 2004; Hahn & Bailey, 2005; P. A. Luce, 1986). The large amounts of unexplained variance in the important studies of Bailey and Hahn (2001, 2005) and Hahn and Bailey (2005) in particular suggest that judgments of word likeness and word similarity cannot be successfully modeled unless contextual effects on sound perception are taken into account.

I should also note some limitations of this article 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 on the participants' knowledge of English. Such alternatives are considered and shown to be inadequate in a companion article. I have already noted that the interpolated similarity value for the pair [ke]/[t͡ʃe] is likely too high, and suggested another study that could test this possibility. There are three 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). The third, and most crucial, would investigate the relation between infant acquisition of phonology and adult phonological learning of the kind studied here.

I conclude with a final remark on the general perspective advanced here. In the foundational work of generative phonology, Chomsky and 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 and Halle, I claim that the bias is a component of cognition that is important for phonological learning and generaliza-

tion. Also like Chomsky and Halle, I do not take the bias to be so strong that it excludes unfavored patterns. The experimental and computational methods applied here allow such claims to be investigated in unprecedented detail, providing a potentially vast body of new data and theoretical insights on the nature of phonological learning.

Notes

1. Significant advances have also been made in integrating articulatory information into phonology (Davidson, 2003, 2006; Gafos, 1999, 2002, this issue; Hall, 2003; Hayes, 1999; Kirchner, 2000, 2001).
2. I take the absolute limits on human phonologies to be not substantive, but rather set by formal properties of the type investigated within optimality theory 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.
3. Square brackets, as in [ki], indicate broad phonetic transcription in IPA (International Phonetic Association, 1999). Note that [t͡ʃ] and [d͡ʒ] are considered to be single consonants (not sequences of two consonants), as indicated by the top ligature. For expository convenience, the diphthong [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 many speakers. The article 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.
4. The confusion rates for these sounds are asymmetric; for example, [ki] was identified as [t͡ʃi] 3.5 times more frequently than [t͡ʃi] was identified as [ki]. Such asymmetries are commonly found in identification experiments (Nosofsky, 1991; Ohala, 1997; Plauché, Delogu, & Ohala, 1997; Tversky, 1977). Although not of central interest for this article, the observed asymmetries do have some consequences, discussed in section 3, for formal modeling of the confusion data.
5. I 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 and token levels is unavoidable.
6. The 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 and Johnson (2003) presented 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 or outputs is approximated. In current research, I am using standard Markov Chain Monte Carlo methods to address this problem.
7. Technically, there is a potential for every clique in the undirected graph, and the potentials are multiplied together in the equation for $P(y|x)$. However, the equation in the text is sufficient for the chain-graph structure assumed here.

8. Recent results of Lin (2004) and Wilson (2006) provide a solution to this problem for stochastic optimality theory in particular, but I have not had the opportunity to apply these methods to this problem.
9. For the purposes of this article, only changes of the type velar stop → palatoaleveolar affricate are considered. This limitation is appropriate for the experiments reported later, and reflects the inviolable Faithfulness constraint against other changes that were assumed earlier. However, according to standard assumptions in optimality theory, a given Markedness constraint can in general be satisfied by many different types of change. To apply the current model in the context of multiple changes, it would perhaps be possible to relate the σ_m of constraint f_m to the average over the biased perceptual similarities of all of the changes that could be forced by f_m. Alternatively, it is possible to revise optimality theory in a way that limits each Markedness constraint to one particular change (Wilson, 2001), in which case the definition of σ_m in the text would always be sufficient.
10. Another 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 his or her first language are discussed in a companion article.
11. The fact that the practice items illustrated palatalization of voiced [g] only was a deliberate design feature, as an earlier experiment (described in a companion article) 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 later.
12. Note 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.
13. The two-way interaction between consonant and vowel context, and the three-way interaction among condition, consonant, and vowel context were both nonsignificant (Fs < 1). Because of issues of nonnormality that arise when proportional data are analyzed with an 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, $F(1, 20) = 9.6$, $p < .01$, $MS_e = .19$.
14. At 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 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. Although 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.

15. The 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 ($F < 1$).
16. Tests of the TiMBL exemplar-based model (Daelemans et al., 2003) have yielded poor fits to the experimental results. Although 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.

Acknowledgments

This article is dedicated to Paul Smolensky, whose enthusiasm for the richness of linguistic patterns reflects the formal elegance that he sees in them. "What I Learned From Paul" would be the title of a very long manuscript; I hope this shorter contribution is true to some of his lessons. For useful comments and discussion I would like to thank Eric Bakovic, Luigi Burzio, Bernard Comrie, Lisa Davidson, Adamantios Gafos, Bruce Hayes, Matt Goldrick, Robert Goldstone, Matt Gordon, Géraldine Legendre, Kie Zuraw, the participants of seminars at NYU, UCLA, and UCSB, and an anonymous *Cognitive Science* reviewer. Special thanks go to Bill Badecker for crucial methodological advice, and to Jill Gilkerson for assistance in coding the data.

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.* Unpublished doctoral dissertaion, University of California, Los Angeles.

Anderson, S. R. (1974). *The organization of phonology.* New York: Academic.

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.

Barlow, J. A. (2001). Individual differences in the production of initial consonant sequences in Pig Latin. *Lingua, 11,* 667–696.

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*. Unpublished doctoral dissertation, University of Massachusetts, Amherst, MA.

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, CA: Stanford University Press.

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

Blevins, J. (in press). 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, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 117–156). Cambridge, UK: Cambridge University Press.

Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives.* Hague, The Netherlands: 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, UK: Cambridge University Press.

Buckley, E. (1999). On the naturalness of unnatural rules. In *Santa Barbara Papers in Linguistics, Vol. 9: Proceedings from the Second Workshop on American Indigenous Languages* (pp. 16–29). Santa Barbara: UCSB Linguistics Deprtment.

Buckley, E. (2003). Children's unnatural phonology. *Proceedings of the Berkeley Linguistics Society, 29,* 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* (Tech. Rep. No. CMU-CS-99-108). Computer Science Department, Carnegie Mellon University.

Cho, T., & McQueen, J. (2006). *Mapping phonologically altered speech onto the lexicon: The case of consonant cluster simplification in Korean.* Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

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.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provist, J. (1993). Psyscope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25,* 257–271.

Côté, M.-H. (2000). *Consonant cluster phonotactics: A perceptual approach*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Côté, M.-H. (2004). Syntagmatic distinctness in consonant deletion. *Phonology, 21,* 1–41.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York: Wiley.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2003). *Timbl: Tilburg memory based learner, version 5.0: Reference guide* (Tech. Rep.). Tilburg, The Netherlands: ILK.

Davidson, L. (2003). *The atoms of phonological representation: Gestures, coordination and perceptual features in consonant cluster phonotactics.* Unpublished doctoral dissertation, Johns Hopkins University, Baltimore.

Davidson, L. (2006). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics, 34*(1), 104–137.

de Boer, B. (2001). *The origins of vowel systems.* Oxford, UK: Oxford University Press.
Dupoux, E., Kakehi, H., Hirosi, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1568–1578.
Eisner, J. (1997). Efficient generation in primitive optimality theory. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL)* (pp. 313–320). Ann Arbor, MI: Association of Computational Linguistics.
Fleischhacker, H. (2001). Cluster-dependent epenthesis asymmetries. In A. Albright & T. Cho (Eds.), *UCLA working papers in linguistics 7: Papers in phonology* (Vol. 5, pp. 71–116). Los Angeles: UCLA, Linguistics Department.
Flemming, E. (2002). *Auditory representations in phonology.* New York: Routledge.
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 and Linguistic Theory, 22,* 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 and Linguistic Theory, 20,* 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. (2005). Categorical perception of natural and unnatural categories: Evidence for innate category boundaries. In R. Okabe & K. Nielsen (Eds.), *UCLA Working Papers in Linguistics, 13: Papers in psycholinguistics* (Vol. 2, pp. 34–58). Los Angeles: UCLA, Linguistics Department.
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, Sweden: Stockholm University.
Gordon, M. (2004). Syllable weight. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 277–312). Cambridge, UK: Cambridge University Press.
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)* (pp. 677–683). Ann Arbor, MI: Association for Computational Linguistics.
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.* Unpublished doctoral dissertation, 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. (2005). What makes words sound similar? *Cognition, 97,* 227–267.
Hale, M., & Reiss, C. (2000). Substance abuse and "disfunctionalism": Current trends in phonology. *Linguistic Inquiry, 31,* 157–169.
Hall, N. (2003). *Gestures and segments: Vowel intrusion as overlap.* Unpublished doctoral dissertation, University of Massachusetts, Amherst, MA.
Hammersley, J. M., & Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished.
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: Vol. I. General papers* (pp. 243–285). Amsterdam: John Benjamins.
Hayes, B., Kircher, R., & Steriade, D. (2004). *Phonetically based phonology.* Cambridge, UK: Cambridge University Press.
Hume, E., & Johnson, K. (2001). *The role of speech perception in phonology.* San Diego, CA: Academic.
Hyman, L. M. (2001). The limits of phonetic determinism in phonology: *NC revisited. In E. Hume & K. Johnson (Eds.), *The role of speech perception in phonology* (pp. 141–185). San Diego, CA: Academic.

Idsardi, W. J. (2005). *Poverty of the stimulus arguments in phonology.* Unpublished manuscript, University of Delaware, Newark, DE.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge, UK: Cambridge University Press.

Johnson, K. (1997). *Acoustic and auditory phonetics.* Malden, MA: Blackwell.

Jun, J. (1995). *Percepetual and articulatory factors in place assimilation: An optimality theoretic approach.* Unpublished doctoral dissertation, University of California, Los Angeles.

Kang, Y. (2004). Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean. *Phonology, 20,* 173–218.

Kawasaki-Fukumori, H. (1992). An acoustical basis for universal phonotactic constraints. *Language and Speech, 35,* 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. Unpublished manuscript, Massachusetts Institute of Technology, Cambridge, MA.

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.* Unpublished manuscript, University of Alberta, Edmonton.

Kochetov, A. (2002). *Production, perception, and emergent phonotactic patterns: A case of contrastive palatalization.* London: Routledge.

Ladefoged, P. (2001). *A course in phonetics.* San Diego, CA: Harcourt Brace.

Lafferty, J., Pereira, F., & McCallum, M. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *International Conference on Machine Learning (ICML'01)* (pp. 282–289). San Francisco, CA: Morgan Kaufmann.

Lin, Y. (2004). *Learning stochastic OT grammars with a gibbs sampler.* Unpublished manuscript, University of Arizona, Tucson.

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

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. Unpublished doctoral dissertation, Indiana University, Bloomington, IN.

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.

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

McCallum, A. (2003). Efficiently inducing features of conditional random fields. *UAI '03, Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence, August 7–10, 2003, Acapulco, Mexico* (pp. 403–441). San Francisco, CA: Morgan Kaufmann.

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). New York: Cambridge University Press.

Moreton, E., Feng, G., & Smith, J. L. (2005, April). *Syllabification, sonority, and perception: New data from a language game.* Paper presented at the 41st Regional Meeting of the Chicago Linguistic Society, Chicago.

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, 14: Studies in phonology and methodology.* Columbus: Ohio State University, Department of Linguistics.

Nevins, A., & Vaux, B. (2003, January). *Underdetermination in language games: Dialects of Pig Latin.* Paper presented at the Linguistic Society of America (LSA) Annual Meeting, Atlanta, GA.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology, 23,* 94–140.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 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).

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, UK: Cambridge University Press.

Ohala, J. J. (1992). What's cognitive, what's not, in sound change. In G. Kellermann & M. Morrissey (Eds.), *Diachrony within synchrony: Language history and cognition. Duisberger Arbeiten zur Sprach- und Kulturwissenschaft 14* (pp. 309–355). Frankfurt, Germany: Peter Lang.

Ohala, J. J. (1995). Experimental phonology. In J. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 713–722). Oxford, UK: Blackwell.

Ohala, J. J. (1997). Comparison of speech sounds: Distance vs. cost metrics. In S. Kiritani & 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. *Phonological Studies, 7,* 81–96.

Pater, J., & Tessier, A.-M. (2003). Phonotactic knowledge and the acquisition of alternations. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress on Phonetic Sciences* (pp. 1777–1780). Dordrecht: Foris.

Peperkamp, S. (2004). Lexical exceptions in stress systems: Arguments from early language acquisition and adult speech perception. *Language, 80,* 98–126.

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, pp. 2187–2190).

Prince, A., & Smolensky, P. (2004). *Optimality theory: Constraint interaction in generative grammar.* Cambridge, MA: Blackwell. Technical Report CU–CS–696–93, University of Colorado at Boulder, Department of Computer Science, and Technical Report TR–2, Rutgers University, Rutgers Center for Cognitive Science, New Brunswick, NJ, April 1993. Rutgers Optimality Archive 537 version, 2002.

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19,* 9–50.

Pycha, A., Nowak, P., Shin, E., & Shosted, R. (2003). Phonological rule-learning and its implications 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 Core Development Team. (2005). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

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,* 305–338.

Riggle, J. (2004). *Generation, recognition, and learning in finite state optimality theory.* Unpublished doctoral dissertation, 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)* (pp. 507–514). Ann Arbor, MI: Association for Computational Linguistics.

Samek-Lodovici, V., & Prince, A. (1999). *Optima*. Unpublished manuscript, University of London, UK, and Rutgers University, New Brunswick, NJ. [Available on Rutgers Optimality Archive 537 version, 2002]

Seidl, A., & Buckley, E. (2005). On the learning of arbitrary phonological rules. *Language Learning and Development, 1,* 289–316.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Lingusitics on Human Language Technology–Volume 1* (pp. 134–141). Edmonton, Canada: North American Chapter of the Association for Computational Lingusitics.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22,* 325–345.

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 (Eds.), *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 P. Smolensky, M. C. Mozer, & D. E. Rumelhart (Eds.), *Mathematical perspectives on meural networks* (pp. 453–496). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Smolensky, P., & Legendre, G. (2005). *The harmonic mind: From neural computation to optimality-theoretic grammars.* Cambridge, MA: MIT Press.

Steriade, D. (2001a). Directional asymmetries in place assimilation: A perceptual account. In E. Hume & K. Johnson (Eds.), *The role of speech perception in phonology* (pp. 219–250). San Diego, CA: Academic.

Steriade, D. (2001b). *The phonology of perceptibility effects: The P-map and its consequences for constraint organization.* Unpublished manuscript, Massachusetts Institute of Technology, Cambridge, MA.

Steriade, D. (2001c, January). *What to expect from a phonological analysis.* Paper 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.

Traunmüuller, 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,* 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). Sommerville, MA: Cascadilla.

Wilson, C. (2006). *Stochastic ranking as probabilistic choice.* Unpublished manuscript, University of California, Los Angeles.

Winitz, H., Scheib, M. E., & Reeds, J. A. (1972). Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *Journal of the Acoustical Society of America, 51,* 1309–1317.

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

Zhang, J. (2001). *The effects of duration and sonority on contour tone distribution—Typological survey and formal analysis*. Unpublished doctoral dissertation, University of California, Los Angeles.

Zhang, J., & Lai, Y. (2005). *Testing the role of phonetic naturalness in Mandarin tone sandhi.* Unpublished manuscript, University of Kansas.

Zsiga, E., Gouskova, M., & Tlale, O. (2006). On the status of voiced stops in Tswana: Against *ND. To appear in *Proceedings of the North East Linguistic Society 36.* Amherst, MA: GLSA.

Zuraw, K. (2005). *Knowledge of consonant clusters: Corpus and survey evidence from Tagalog.* Unpublished manuscript, University of California, Los Angeles.

Appendix: Stimuli for Experiments 1 and 2

ˈkit͡ʃə
ˈkigə
ˈkirə
ˈkiwə
ˈkifə
ˈkimə
ˈkinə
ˈkisə
ˈket͡ʃə
ˈkegə
ˈkenə
ˈkewə
ˈked͡ʒə
ˈkemə
ˈkerə
ˈkezə
ˈkɑpə
ˈkɑθə
ˈkɑvə
ˈkɑðə
ˈkɑgə
ˈkɑʒə

ˈgibə
ˈgimə
ˈgipə
ˈgirə
ˈgisə
ˈgit͡ʃə
ˈgikə
ˈgivə
ˈgiwə
ˈgefə
ˈgeðə
ˈgemə
ˈgepə
ˈgerə
ˈget͡ʃə
ˈgekə
ˈgevə
ˈgewə
ˈgɑfə
ˈgɑkə
ˈgɑrə
ˈgɑd͡ʒə
ˈgɑpə
ˈgɑwə

ˈpid͡ʒə
ˈpiθə
ˈpivə
ˈpibə
ˈpilə
ˈpiʒə
ˈpekə
ˈpevə
ˈpezə
ˈpebə
ˈpeðə
ˈpesə
ˈpɑt͡ʃə
ˈpɑgə
ˈpɑrə
ˈpɑfə
ˈpɑd͡ʒə
ˈpɑvə

ˈbilə
ˈbipə
ˈbiʒə
ˈbiðə
ˈbiʃə
ˈbizə
ˈbeðə
ˈbegə
ˈbenə
ˈbed͡ʒə
ˈbevə
ˈbeʒə
ˈbɑlə
ˈbɑSə
ˈbɑzə
ˈbɑt͡ʃə
ˈbɑʃə
ˈbɑvə