Phonological naturalness and phonotactic learning*

Bruce Hayes James White

Department of Linguistics, UCLA

Final version, April 2012

To appear in *Linguistic Inquiry*

Abstract

We investigate whether the patterns of phonotactic well-formedness internalized by language learners are direct reflections of the phonological patterns they encounter, or reflect in addition principles of phonological naturalness. As a research tool we employ the phonotactic learning system of Hayes and Wilson (2008), which carries out an unbiased search of the lexicon for valid phonotactic generalizations. Applying this system to English data, we find that it learns many constraints that seem to be unnatural—they have no evident typological or phonetic basis, yet hold true of the English lexicon.

We tested the status of ten of these constraints in a nonce-probe study, obtaining native-speaker ratings of novel words that violated them. We used 40 such words: 10 violating our unnatural constraints, 10 violating natural constraints assigned comparable weights by the Hayes/Wilson learner, and 20 violation-free forms, each similar to a test form and employed as a control. In our experiment, we found that violations of the natural constraints had a powerful effect on native speaker judgment and violations of the unnatural constraints had at best a weak one. We conclude by assessing a variety of hypotheses intended to explain this disparity, opting ultimately for a learning bias account.

* Author emails: bhayes@humnet.ucla.edu, jameswhite@ucla.edu. Thanks to Colin Wilson for advice, encouragement, and the use of his software, to Patricia Keating for making the recordings, and to our experimental participants for their patience. For helpful input and advice we thank Adam Albright, Robert Daland, Kie Zuraw; talk audiences at the UCLA, University of Alberta, Johns Hopkins University, Cornell University, and two extremely helpful LI reviewers. This study was supported by a grant from the Committee on Research of the UCLA Academic Senate.

A web site with supplementary materials for this article is located at http://www.linguistics.ucla.edu/people/hayes/PhonologicalNaturalness/.

1. Introduction: the problem of unnatural constraints

Our starting point is a classic phonological problem, the origin of phonotactic knowledge (Chomsky and Halle 1965). Speakers can rate novel words of their language, judging them to be fully acceptable (e.g. *blick* [blık]), intermediately acceptable (*bwick* [bwık]), or fully ill-formed (*bnick* [bnık]). Not only their intuitive judgments, but also their behavior reflects such hierarchies of well-formedness, as is shown by experimental evidence from speech production and perception (Massaro and Cohen 1980, Dupoux et al. 1999, Mattys and Jusczyk 2001, Moreton 2002, Berent et al. 2007, Berent et al. 2008, Wilson and Davidson, to appear). Patterns of phonotactic well-formedness are partly language-specific and therefore must, at least to some extent, be learned during the period of phonological acquisition.

The problem of phonotactic learning is particularly suited to the strategy of computational modeling (Hayes 2004, Prince and Tesar 2004, Jarosz 2006, Hayes and Wilson 2008, Pater and Coetzee 2008, Albright 2009, Heinz 2010a, 2010b). A model can be set up to implement a variety of hypotheses about the language faculty as it relates to phonology. Ideally, it should be fed a representative lexicon given in phonetic transcription, approximating the experience of the language-learning child. The model learns a phonotactic grammar, which can then be tested by comparing the well-formedness values it assigns to novel stimuli with well-formedness measures obtained experimentally.

An informative strategy for such work is the "inductive baseline" approach (e.g. Gildea and Jurafsky 1996, Hayes and Wilson 2008). The idea is to start with very simple models, embodying few *a priori* principles, and see where they fail. When augmenting the models with principles of phonological theory produces success instead, we obtain insight into the usefulness of such principles for learning.

A baseline model of this type is proposed by Hayes and Wilson (2008). We review the details of this model below; for present purposes the crucial aspect of the model is that in its core rendition, the *a priori* knowledge that it brings to phonological learning is largely confined to the feature system. This serves as the basis for phonological constraints, which the model constructs by concatenating feature matrices that denote natural classes. For example, the constraint that bans prevocalic lax vowels in English could be stated as *[+syllabic,−tense][+syllabic].

Hayes and Wilson (2008) argue that their model does a fairly good job of locating in some form the generalizations that linguists find when they inspect phonotactic patterns. For example, it finds the principles of featural agreement that govern Shona vowel harmony, and replicates essentially all the phonotactic generalizations proposed by Dixon (1981) in a meticulous phonotactic study of Wargamay, an Australian language. The model is also successful in matching human phonotactic intuitions: it achieves a close match to the experimentally-gathered intuitions on English onset well-formedness collected by Scholes (1966); and more recently has achieved a reasonably good match for the experimental data reported in Albright (2009), Colavin et al. (2010), and Daland et al. (2011).

However, another aspect of the Hayes/Wilson model is potentially far more controversial. When fed with Wargamay data, the model did not learn just the constraints that were needed to recapitulate Dixon's well-motivated analysis; it also learned a number of phonological

constraints that would strike experienced phonologists as unnatural. One example is given in (1), stated first in features then in prose. The symbol "^" may be read "unless".

(1) *A puzzling constraint learned for Wargamay*

a. *[^−sonorant,−anterior][+long][−consonantal]

b. "If a long vowel is followed by a glide, it must be preceded by a palato-alveolar obstruent."

Hayes and Wilson point out two possibilities concerning constraints like (1). One is that they are indeed valid for Wargamay: were it possible to access native-speaker intuitions in this language, forms violating them would be judged as ill-formed to a degree corresponding to the weight of the applicable constraint. Another possibility, however, is that these constraints reflect a defect in the learning model: a constraint could be entirely exception-free in the Wargamay lexicon, yet fail to be implemented by native speakers as part of their phonological grammar. The purpose of this article is to offer evidence from a phonological experiment that bears on which of these two hypotheses is most likely to be correct. For practical reasons we shift our focus to English, where the Hayes/Wilson model also finds unnatural-seeming constraints like (1).

Before starting in, we clarify two terms to be used below.

Phonological constraints are usually defended on two grounds: either typological or phonetic. The typological criterion can be expressed on the basis of Greenbergian implicational universals (the presence of sequences violating the constraint implies the presence of closely similar sequences that do not; Greenberg 1966, 1978). The phonetic criterion is that a constraint should be functionally effective, serving to form a phonological system in which words are easier to articulate or in which possible words are perceptually distinct from one another. We will refer to constraints that satisfy one or the other criterion as *natural*, other constraints as *unnatural*. It is the unnaturalness of constraint (1) that would render it suspect for many phonologists, we think, as a valid constraint of Wargamay phonology.

We will call a constraint *accidentally true* if it holds true of a language's lexicon but experimental investigation indicates that it is not part of the phonotactic knowledge of native speakers. One possible ground for suspecting a constraint of being accidentally true is that it is unnatural. Thus, in these terms, the purpose of our experiment is to test if some of the unnatural constraints learned by the Hayes/Wilson learner are accidentally true.

2. Research background

The problem of unnatural phonotactic constraints is relevant to a current debate in phonology: is phonological learning the result of an unbiased, inductive search for generalizations (see, e.g. Blevins 2004, Ch. 9)? Alternatively, are language learners limited to learning only generalizations that are expressible with a limited, universal set of natural

constraints (see, e.g., Becker, Ketrez, and Nevins 2011)? If the former position is correct, then language learners should in principle be able to access and employ unnatural generalizations. [1]

Experimental evidence bearing on this issue is steadily accumulating. Our reading of this literature is that the evidence is quite mixed and gives no comfort to advocates of either of the two possible extreme positions (all constraints are a priori knowledge/all learning is purely inductive).

2.1 Evidence for learnability of unnatural generalizations

A fairly clear case of an evidently-learnable generalization is the English rule of Velar Softening (Chomsky and Halle 1968:219-221). This rule is argued to be unnatural by Pierrehumbert (2006); notably, it derives [s], rather than the phonetically/typologically expected [tʃ], from /k/. Pierrehumbert's (2006) experiments demonstrate that Velar Softening is surprisingly productive.

Further, processes of phonology or allomorph selection have been shown to apply differentially in a way that is moderated by unnatural factors in the phonological environment. One such case is found in Hungarian vowel harmony (Hayes et al. 2009). The harmony pattern is mostly predictable, but with certain vowel sequences harmony is partly arbitrary, with front or back harmony occurring on a stem-by-stem basis. Corpus data for Hungarian show a statistical skewing, favoring front harmony for stems that ending in bilabial stops, an environment that would not qualify as natural by either phonetic or typological criteria. Hayes et al. find that Hungarian speakers are tacitly aware of this unnatural pattern and several others, respecting them when they apply harmony in words with nonce stems.

Competition between morphological processes is likewise often affected by unnatural factors present in the phonological environment, part of a phenomenon that Albright (2002) calls "islands of reliability." For instance, every English verb stem that ends in a voiceless fricative takes the regular past tense ending. When tested with nonce stems, English speakers show that they particularly prefer regular past tense suffixation (/-d/) for stems of this type (Albright and Hayes 2003).

Finally, experimenters have created novel languages and tested whether unnatural phonological patterns could be learned from them. Often, such studies strengthen their case by comparing the learnability of a particular phonological pattern with its opposite (both cannot be natural). For instance, Onishi, Chambers, and Fisher (2002) compared artificial languages in which {b, k, m, t} were limited to onset and {p, g, n, tʃ} to coda — or vice versa. Both phonotactic systems were learnable by adults, and also by 16.5-month-old infants (Chambers, Onishi, and Fisher 2003). Related work (Dell et al. 2000; Warker et al. 2006, 2008) found similar learning patterns using a speech-error testing paradigm. Adult learners have been able to learn vowel disharmony about as well as (typologically far more common) vowel harmony, in an artificial-language study (Pycha et al. 2003) and in a study where the vowel harmony or

[1] In the inductivist view, the typological patterns that manifest natural principles are attributed instead to diachronic factors: languages change phonologically through phonetic shifts and misperceptions that are sensitive to phonetic or other naturalness principles; see for example Ohala (1981), Blevins (2004).

disharmony rule was used to create a novel "dialect" of French (Skoruppa and Peperkamp 2011). Unnatural consonant alternations (/p, g/ → [ʒ, f] / V ___ V and /ʃ, v/ → [b, k] / V ___ V) were successfully learned by participants in the artificial-language learning study of Peperkamp and Dupoux (2007). Seidl and Buckley (2005) found that 9-month-old infants could learn both phonetically natural patterns and similar unnatural patterns.

2.2 Evidence for the role of naturalness

We next address the opposite possible extreme position, namely that *all* phonological learning is purely inductive, and that naturalness considerations play no role. This strong position is likewise contradicted by evidence in the literature. For instance, Wilson (2006) showed that learners of an artificial language extended a palatalization alternation in one direction (learn forms with /ke/ → [tʃe], test on forms with /ki/ → [tʃi]) but not in the other (learn /ki/ → [tʃi], test /ke/ → [tʃe]); this finding matches both language typology and the predictions of Wilson's phonetic model. Experiments have also shown that participants learn a dependency between the height of two vowels more easily than a dependency between the voicing of a consonant and the height of a vowel (Moreton 2008) and that they learn a directional vowel harmony pattern over a "majority rules" pattern when presented with ambiguous training data (Finley & Badecker 2008, Finley 2008). Other experiments that have supported enhanced learnability for natural phonological patterns are Pater and Tessier (2003), Wilson (2003), Berent et al. (2007), Berent et al. (2008), Peperkamp, Skoruppa, and Dupoux (2006), Berent et al. (2009), and Hayes et al. (2009).

In several of the experiments just cited, the findings support a *bias* effect: the unnatural patterns are learnable but take longer to learn, or yield weaker experimental effects than comparable natural patterns. We will return to the question of bias below.

2.3 Overview

To sum up: at present, purist "all naturalness" and "no naturalness" positions seem ill-supported, but the articulation of a theory explaining how and when naturalness plays a role in phonological learning lies in the future. We hope to contribute to this debate by addressing one particular angle of the problem, one for which the Hayes/Wilson learning model can play a useful role. As noted above, the model learns unnatural-seeming constraints when applied to Wargamay. We have since found the same for English (see below) and believe it would almost certainly find similar constraints when applied to most other languages. Our interest here is not so much the details of the Hayes/Wilson learner but rather its characteristic behavior: through extensive search it discovers phonological "gaps" (unpopulated regions) in a lexical corpus, and the constraints it uses to describe these gaps often appear to be unnatural. The existence of such gaps is in one sense a fact about the lexicon, rather than about the learner itself. The data patterns are there to be discovered; the question is whether native speakers find them.

In what follows, we first review the workings of the Hayes/Wilson phonotactic learner (§3), then discuss how we employed it with English data to discover a variety of unnatural constraints (§4). We then describe how we selected the test words and obtained ratings of them in an experiment (§5). Our main result, that the unnatural constraints have little or no effect on native

speaker intuition, is given in §5.2. We defend our claim against possible confounding effects in §6. In the final section we assess what our results mean for the naturalness debate and suggest how research might proceed from here.

3. The Hayes/Wilson phonotactic learner

In broad outline, the Hayes/Wilson learner forms a space of possible constraints using a feature system given to it in advance. It selects constraints from this set using ranked heuristics, and weights them using the criterion of maximum likelihood. The final grammar learned can assign a likelihood value to any given string, forming a quantitative prediction about phonotactic well-formedness. We elaborate this picture below.

In its simplest form the model uses *SPE*-style phonological representations (Chomsky and Halle 1968) consisting of sequences of feature matrices, and assumes that constraints likewise consist of matrix sequences banning particular sequences of natural classes. A key observation is that although the number of possible feature matrices in any reasonable feature system is extremely large,[2] the number of natural classes these matrices define on a segment inventory is far smaller, typically in the hundreds. This makes it feasible to do exhaustive searching of the class of possible constraints, provided the maximum number of matrices in a constraint is not too high. We assume this number is at least three (e.g. constraints applying to intervocalic consonants are abundant); for discussion see Hayes and Wilson (2008, §4.1.2) and Kager and Pater (forthcoming).

The Hayes/Wilson model iteratively searches for new constraints to add to the grammar, selecting from the full set of possibilities using a hierarchy of heuristics. The top-ranked heuristic is matrix count: unigram constraints are favored over bigram, bigram over trigram. The second-ranked heuristic is accuracy: constraints are preferred if they are violated by very few forms, relative to the number of forms that would be expected to occur given whatever constraints and weighting have been learned so far. The lowest-ranked heuristic is generality; constraints are favored that rule out a larger fraction of the possible strings. Constraint search continues until no further constraints are available that meet the minimal criterion for accuracy, or, where this is convenient, when a user-specified maximum is reached.

As they are selected, constraints are weighted. Intuitively speaking, weights express the relative strength of a constraint; in the completed grammar, higher-weighted constraints play a greater role in lowering the predicted well-formedness of the words that violate them. Weighting follows mathematical procedures, backed by proof, that find the weights that respect the maximum likelihood criterion; i.e. that assign the most probability to the observed data—hence allocating the smallest probability to unobserved data, insofar as this can be done with the available constraints. The result is phonotactic grammars that are restrictive.[3]

[2] Assuming a feature may be +, −, or absent, the number of feature matrices is $3^n - 1$, where n is the number of features. The feature set we use here has 22 features and thus implies about 31 billion matrices.

[3] Following general practice in maxent modeling we employed a Gaussian prior (Goldwater and Johnson 2003, ex.(3)), $\sigma = 1$. The main effect of this is to avoid infinite weights for never-violated constraints.

Both the weight-setting process and the predictions of the learned grammar depend on the basic formula for maxent grammars (Berger, Della Pietra and Della Pietra 1996, Della Pietra, Della Pietra and Lafferty 1997, Goldwater and Johnson 2003), which assigns a probability to a word ω based on its profile of the constraint violations and on the weights of the violated constraints. The computation is summarized in (2):

(2) *Maxent probability computation (Della Pietra et al. 1997, 1)*

$p(\omega) = \exp(-\Sigma_i \lambda_i \chi_i(\omega))/Z$, where $Z = \Sigma_j \exp(-\Sigma_i \lambda_i \chi_i(\omega_j)$

The elements of the formula are as follows:

$p(\omega)$	predicted probability of word ω
$\exp(x)$	e to the power of x
Σ_i	summation across all constraints
λ_i	weight of the ith constraint
$\chi_i(\omega)$	number of times ω violates the ith constraint
Σ_j	summation across all possible words

The end product is a probability value for ω.[4] If one's goal is simply a comparison of different words, it suffices to calculate just the expression $\Sigma_i\lambda_i\chi_i(\omega)$, which can be regarded as a kind of penalty score.

4. Finding candidates for accidentally-true constraints

We began our inquiry by using the Hayes/Wilson learner to search for candidate accidentally-true constraints in English similar to (1) seen above for Wargamay. For this purpose, we needed a training set, ideally as close as possible to what is encountered by the experimental participants during language acquisition. Since our participants were Americans, we used the pronunciations of the Carnegie-Mellon Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). We selected the words that in the CELEX lexical database (Baayen, Piepenbrock and Gulikers 1995) have a frequency of at least one; inspection suggested that this would achieve a reasonably good fit with the words known to our participants. We removed from the list as well as we could all compounds, inflected forms, and forms created by highly transparent processes of morphological derivation, since these tend to have special phonotactic properties—the assumption we made is that simple forms presented to participants as possible words are interpreted as monomorphemic, and that the relevant phonotactics is that of Level I, as defined in the theory of Lexical Phonology (Kiparsky 1982). This assumption is defended in §6.1 below. We also attempted to correct as many errors as possible in the Carnegie-Mellon transcriptions of the words we used, including incorrect markings of primary and secondary stress. Finally, we syllabified our training data following the

A variety of other approaches to constraint weighting are explored and compared in McClelland and Vander Wyk (2006).

[4] Probability is here used in a rather abstract sense: a total probability mass of one is allocated among all possible phonological strings, and the probability of a string is its share in the total.

Maximal Onset Principle (Selkirk, 1982), so that constraints could refer to onset and coda position.[5]

Preliminary exploration indicated that the model was not guaranteed to learn constraints that were obviously natural. Since our interest was in examining constraints that had clear typological and/or phonetic support, we fed 36 such constraints into the grammar in advance, then let it continue until it had learned 160 weighted constraints.[6] The resulting grammar gave reasonably good (not perfect) descriptive performance, ruling out most impossible onset clusters, coda clusters, medial clusters, and we chose it as our base grammar.

4.1 Selecting the relevant constraints

Guided by our knowledge of phonological typology and phonetic naturalness, we picked from the 160 constraints 10 fairly clearly natural constraints and 10 fairly clearly unnatural ones. Five of the 10 natural constraints were manifestations of the well-known Sonority Sequencing Principle (Sievers 1881; Greenberg 1978; Berent et al. 2007). We adopted the feature-based implementation of the Sonority Hierarchy proposed by Clements (1990) and set up constraints that penalize consonant clusters that have less than ideal sequencing for a particular sonority feature. These constraints are listed in (3); for sample violations see the list of experimental stimuli in (7).

(3) *Natural constraints I: Sonority based*

Constraint

a. *[−sonorant][+sonorant] IN CODA
b. *[+consonantal][−consonantal] IN CODA
c. *[−consonantal][+consonantal] IN ONSET
d. *[−continuant][−continuant] IN ONSET[7]
e. *[−continuant][+nasal] IN ONSET

[5] Implementation: we assigned the feature values [+rhyme] and [−rhyme] to consonants in coda and onset position, respectively. The onsets used for maximal onset syllabification, which follow Hayes and Wilson (2008), are posted at the article website. We avoided "exotic" onsets like [km] (*Khmer*), since when maximized they result in implausible syllabifications like *acme* ['æ.kmi]. In the experimental stimuli, we largely avoided questions of syllable division by placing all sequences that violated syllable-based constraints at word edge, where syllabification is unambiguous.

[6] In retrospect this procedure strikes us as having been too cautious. Following a reviewer's suggestion, we reran the learning simulation without using any prior constraints. The resulting grammar ended up including nine out of our ten unnatural constraints ((6)). Most our natural constraints also showed up in recognizable form, either as a notational variant of one of the constraints in (3)-(5) or as a bundle of more complex constraints having similar function to (3)-(5). Like our main test grammar, the no-prior-constraints grammar assigned near-identical (and high) penalties to our natural and unnatural test items and near-zero penalties to our natural control items, It gave penalties to our unnatural controls about one eighth the size of that assigned to the unnatural test items. It thus appears that our results would have been essentially the same had we used the no-prior-constraints grammar. This grammar, and the scores it assigns, may be inspected at the web site for this article.

[7] For this constraint see Morelli (1999), who on the basis of a typological survey suggests a general constraint banning obstruent clusters whose first element is a stop.

where
[−sonorant] = [p t t͡ʃ k b d d͡ʒ g f θ s ʃ h v ð z ʒ]
[+sonorant] = [m n ŋ l ɹ w j]
[+consonantal] = [p t t͡ʃ k b d d͡ʒ g f θ s ʃ h v ð z ʒ l m n ŋ]
[−consonantal] = [ɹ w j]
[−continuant] = [p t t͡ʃ k b d d͡ʒ g]
[+nasal] = [m n ŋ]

Three constraints reflected the common pattern for coda segments to be homorganic with what follows and for onset segments to be heterorganic (see Kager 1999:131; Harris 1983:31-35).

(4) *Natural constraints II: homorganicity/heterorganicity based*

a. *[+labial][+dorsal] IN CODA (heterorganicity in codas)
b. *[+dorsal][+labial] IN CODA (heterorganicity in codas)
c. *[+labial][+labial] IN ONSET (heterorganicity in onsets)

where [+labial] = [p b f v m w], [+dorsal] = [k g ŋ]

The remaining two constraints included a straightforward one ((5a)), the commonplace requirement (Lombardi 1999) that obstruent clusters agree in voicing (in this case, only in one particular direction). The remaining constraint was the only one that we did not design into the grammar, but on reflection seemed a plausible constraint: it forbids the glides [j, w] in syllable coda.[8] This is a plausible restriction for a language like English that has multiple diphthongs; the ban keeps the diphthongs distinct from what would be similar vowel + glide sequences. For the principle of phonological dispersion underlying this view see e.g. Flemming (2004).

(5) *Natural constraints III: other*

a. *[−son,−voice][−son,+voice] Voicing assimilation
b. *[−syllabic, +high] IN CODA Glides in coda, in a diphthongal language

where
[−son,−voice] = [p t t͡ʃ k f θ s ʃ h]
[−son,+voice] = [b d d͡ʒ g v ð z ʒ]
[−syllabic, +high] = [j w]

[8] The Americanist tradition of phonetic transcription (Pullum and Ladusaw 1996:22-24) often depicts the diphthongs of English with glide letters, e.g. [ay, aw] to describe what IPA transcription more accurately notates with vowel symbols ([aɪ, aʊ]). Our recorded tokens of forms violating (5b), given below in (7j), employ true glides with full constriction.

For the unnatural constraints, we combed through the output of the grammar looking for constraints that met several criteria: that they should have, at most, weak typological or phonetic support, that they should have weights similar to those learned for the natural constraints above, and that they should have few or no exceptions in the training data. The constraints are given with prose descriptions in (6).

(6) *10 Unnatural Constraints*

a. *[+round,+high][−consonantal,−sonorant]	*No* [u, ʊ, w] *before* [h]
b. *[+consonantal,−anterior][−sonorant]	[ʃ, ʒ, tʃ, dʒ] *may not precede obstruents*.
c. *[−back][+diphthong]	*No* [j] *before* [aɪ, aʊ, ɔɪ]
d. *[$_{word}$ [−diphthong,+round,+high]	*No word-initial* [u, ʊ]
e. *[+diphthong][+continuant,−anterior]	*No* [aɪ, aʊ, ɔɪ] *before* [ʃ, ʒ]
f. *[+coronal,+continuant,−strident][−sonorant]	*No* [θ,ð] *before obstruents*
g. *[+coronal,+continuant,−strident][−stress,+round]	*No* [θ,ð] *before stressless rounded vowels*
h. *[+diphthong,+round,−back][−anterior]	*No* [ɔɪ] *before* [ʃ, ʒ, tʃ, dʒ]
i. *[+continuant,+voice,−anterior][+stress][−son]	*No* [ʒ] *before stressed vowel + obstruent*
j. *[$_{word}$ [+diphthong,+round][−son,+voice]	*Initial* [aʊ, ɔɪ] *may not precede a voiced obstruent*.

For sample violations, see the list of experimental stimuli in (8) below.

The natural constraints have, in the aggregate, very similar weights to the set of unnatural constraints. The average weight of the natural constraints is 3.79 (range 2.68 – 4.65) and the average weight of the unnatural constraints is 3.96 (range 3.51 – 5.22). In terms of testing the model, the minor discrepancy goes in the direction desired: if it turns out that the unnatural constraints have the weaker effect on native speaker judgment, we do not want to attribute this to the weight difference. Weights for all constraints are given in Appendix A.

4.2 Diachronic origin of unnatural constraints

We digress to address the question of why languages should have any unnatural constraints at all. Some of our constraints have a clear diachronic basis in what could be called "constraint telescoping", analogous to the "rule telescoping" observed by Kenstowicz and Kisseberth (1977, 64-65). The idea is that an originally natural constraint can be obscured by a sequence of natural historical changes while retaining its effects, simply by inertia, in the inherited lexicon. Constraint (6e), banning [aɪ, aʊ, ɔɪ] before [ʃ, ʒ], is one such case. Ignoring the very rare sounds [ʒ] and [ɔɪ] for simplicity, we observe that /ʃ/ originated in English from historical **sk*, and [aɪ] and [aʊ] from historical **iː*, **uː*. Thus (6e) is the historical descendent of a constraint that originally banned long vowels before a consonant cluster, a highly natural pattern. This history is

discussed in detail in Iverson and Salmons (2005), who suggest that for English a synchronic ban on long vowels before /ʃ/ is (in the terms of this article) accidentally true.[9]

Not all of the constraints of (6) have such a clear diachronic origin, and some may indeed be true entirely by accident. Still others may be a blend of diachronically motivated and accidental factors. For (6c), the absence of [jaɪ] has a clear diachronic origin, in that [aɪ] descends from [iː], and bans on [j] before high front vowels are common typologically (Kawasaki 1982, §2.7.2; for English see Jespersen 1909, §58). The lack of [jaʊ], however, may be accidental.

5. Magnitude estimation experiment

We used the constraints described in the preceding section to design the nonce words used in the following word acceptability experiment. We used the magnitude estimation technique, following methods described in Lodge (1981) and Bard, Robertson and Sorace (1996). In this task, participants increase or decrease the magnitude of their response based on the relative increase or decrease in some property of the stimuli. In our case, participants were rating the relative goodness of nonwords as potential words of English. We used number estimation and line drawing as response modalities because these two tasks are easy to implement and their relationship to each other is well understood.

5.1 Method

5.1.1 Participants

Twenty-nine UCLA undergraduate students participated in the experiment for partial course credit. All participants were native speakers of English with no hearing or speech impairments.

5.1.2 Materials

For each constraint in §4.1 (both natural and unnatural) we invented two stimulus pairs. Each pair consisted of a Violating word which, in most cases, violated only the constraint in question, and a Control word, which, in most cases, violated no other constraints. In a couple of cases, the Control word violated a very low-weighted constraint, which was also violated by the Violating word. Aside from the target violation, we tried to make the words as phonotactically bland as possible, and also to avoid strong resemblances to particular existing words. We found that satisfying all of these requirements at once was not easy, and for this reason the pairs were not statistically controlled for resemblance to existing words.

Since there were 10 natural and 10 unnatural constraints, and each constraint was tested with two Violating/Control pairs, there were a total of 80 stimuli: 20 Natural Violating forms, 20 Natural Control forms, 20 Unnatural Violating forms, and 20 Unnatural Control forms. They are listed in (7) and (8) below. We give the forms both in the orthography we employed and in IPA.

[9] In support of this they point out the vulnerability of their constraint to acquiring new counterexamples through borrowing (e.g. *pastiche*, *cartouche*). The more specific constraint we use here, (6e), has been less vulnerable to counterexamples because the likely donor language, French, lacks [aɪ, aʊ, ɔɪ].

(7) *Stimulus pairs for the natural constraints*

	Constraint	*Violating - Control*	*IPA*
a.	*[−son][+son] IN CODA	*kipl - kilp* *canifl - canift*	[ˈkɪpl] - [ˈkɪlp] [kəˈnɪfl] - [kəˈnɪft]
b.	*[+cons][−cons] IN CODA	*tilr - tilse* *shapenr - shapent*	[ˈtɪlɹ] - [ˈtɪls] [ʃəˈpɛnɹ] - [ʃəˈpɛnt]
c.	*[−cons][+cons] IN ONSET	*hlup - plup* *hmit - smit*	[ˈhlʌp] - [ˈplʌp] [ˈhmɪt] - [ˈsmɪt]
d.	*[−cont][−cont] IN ONSET	*cping - sping* *ctice - stice*	[ˈkpɪŋ] - [ˈspɪŋ] [ˈktaɪs] - [ˈstaɪs]
e.	*[−cont][+nasal] IN ONSET	*cnope - clope* *pneck - sneck*	[ˈknoʊp] - [ˈkloʊp] [ˈpnɛk] - [ˈsnɛk]
f.	*[+labial][+dorsal] IN CODA	*trefk - treft* *rufk - ruft*	[ˈtɹɛfk] - [ˈtɹɛft] [ˈɹʌfk] - [ˈɹʌft]
g.	*[+dorsal][+labial] IN CODA	*bikf - bimf* *sadekp - sadect*	[ˈbɪkf] - [ˈbɪmf] [səˈdɛkp] - [səˈdɛkt]
h.	*[+labial][+labial] IN ONSET	*bwell - brell* *pwickon - twickon*	[ˈbwɛl] - [ˈbɹɛl] [ˈpwɪkən] - [ˈtwɪkən]
i.	*[−son,−voice][−son,+voice]	*esger - ezger* *trocdal - troctal*	[ˈɛsgɚ] - [ˈɛzgɚ] [ˈtɹɑkdəl] - [ˈtɹɑktəl]
j.	*[glide] IN CODA	*jouy - jout* *tighw - tibe*	[ˈdʒaʊj] - [ˈdʒaʊt] [ˈtaɪw] - [ˈtaɪb]

(8) *Stimulus pairs for the unnatural constraints*

	Constraint	*Violating - Control*	*IPA*
a.	*[+round,+high][−cons,−son]	*luhallem - laihallem* *tuheim - towheim*	[luˈhæləm] - [leɪˈhæləm] [tuˈheɪm] - [toʊˈheɪm]
b.	*[+cons,−ant][−son]	*ishty - ishmy* *metchter - metchner*	[ˈɪʃti] - [ˈɪʃmi] [ˈmɛtʃtɚ] - [ˈmɛtʃnɚ]
c.	*[−back][+diphthong]	*youse - yoss* *yout - yut*	[ˈjaʊs] - [ˈjɑs] [ˈjaʊt] - [ˈjʌt]
d.	*[$_{\text{word}}$ [−diphthong,+round,+high]	*utrum - otrum* *ooker - ocker*	[ˈutɹəm] - [ˈoʊtɹəm] [ˈʊkɚ] - [ˈɑkɚ]
e.	*[+diphthong][+continuant,−anterior]	*pyshon - pyson* *foushert - fousert*	[ˈpaɪʃən] - [ˈpaɪsən] [ˈfaʊʃɚt] - [ˈfaʊsɚt]

f.	*[+cont,−strident][−sonorant]	*hethker - hethler*	[ˈhɛθkɚ] - [ˈhɛθlɚ]
		muthpy - muspy	[ˈmʌθpi] - [ˈmʌspi]
g.	*[+cont,−strident][−stress,+round]	*potho - pothy*	[ˈpɑθo] - [ˈpɑθi]
		taitho - taithy	[ˈteɪθo] - [ˈteɪθi]
h	*[+diphthong,+round,−back][−anterior]	*noiran - nyron*	[ˈnɔɪɹən] - [ˈnaɪɹən]
		boitcher - boisser	[ˈbɔɪtʃɚ] - [ˈbɔɪsɚ]
i.	*[+cont,+voice,−ant][+stress][−son]	*zhep - zhem*	[ˈʒɛp] - [ˈʒɛm]
		zhod - zhar	[ˈʒɑd] - [ˈʒɑɹ]
j.	*[$_{word}$ [+diphthong,+round][−son,+voice]	*ouzie - oussie*	[ˈaʊzi] - [ˈaʊsi]
		oid - oit	[ˈɔɪd] - [ˈɔɪt]

Our experiment also included a set of filler words, partly as a way of distracting the participants from the fact that the stimuli were paired, and partly in order to provide an independent check on our method. For the fillers, we selected 20 forms each from two earlier phonotactic rating studies: Experiment 5 of Scholes (1966) and Albright (2009). In both cases, the forms selected represented the full range of forms found in each study, from highly well-formed to highly ill-formed. The fillers are listed in Appendix B.

Stimuli were presented auditorily as well as orthographically. To create the auditory stimuli, the nonwords were recorded in a random order by an English-speaking female trained phonetician in a sound booth, assisted by monitoring and feedback from the experimenters. The speaker read from a transcript containing both orthographic and IPA renderings of the words. From this recording, the tokens judged by the experimenters to be the clearest rendering of each intended phonemic sequence were selected, then equalized for volume.[10]

We presented the stimuli both auditorily and orthographically in order to maximize the chance that participants would internalize the intended phonemic representations of the nonwords represented by the IPA transcriptions in (7) and (8). The auditory presentation provided the intended pronunciation in cases where orthography may be ambiguous. However, studies have shown that non-native sequences of sounds may be misperceived by listeners (Dupoux et al. 1999); thus, we chose to provide orthography as well in order to aid participants in parsing the intended sequence of phonemes.

5.1.3 Procedure

The magnitude estimation procedure consisted of three blocks: a calibration block, a number estimation block, and a line drawing block. All participants began with an identical calibration

[10] Discussing how illegal clusters can be rendered by English speakers, Davidson (2007) provides a useful taxonomy: speakers insert a full schwa, a shorter "transitional schwa", or else leave the members of the cluster adjacent. We requested our speaker to avoid inserted schwas of any sort, and spectrographic inspection indicates that indeed no inserted schwas of any sort appear in any of the tokens used in the experiment. We also asked our speaker to avoid rendering sonorants in sonority-reversed clusters as syllabic, and careful listening suggests that the speaker likewise succeeded in this task. The sound files for the experimental tokens may be downloaded from the web site for this article.

phase, following Lodge (1981) and Bard et al. (1996). They were told that they would see multiple lines on the computer screen and that they would be assigning each one a number based on the length of the line. They were shown a horizontal line approximately 35 mm in physical length; this was designated as the reference line and assigned a numerical value of 100. Participants were told to enter numerical values for subsequent lines based on their lengths relative to the reference line; if a line was twice as long as the reference line, they were to enter a number twice as high as 100, and so on. Participants entered numbers using the keyboard and pressed the "next" button to begin the next trial. The reference line was not displayed while the participants were giving their estimations.

After giving numbers for eight lines ranging from 6 to 600 units, participants were given eight numbers of equivalent values (6 to 600) and asked to draw lines. The number 100 was once again used as a reference value. Participants drew horizontal lines by clicking in a rectangular box on the computer screen, dragging the mouse cursor to another part of the box, then releasing the mouse button. If they clicked in the box again, the old line would disappear and a new line could be drawn. When a participant was satisfied with the line in the box, she pressed the "next" button to move on to the next word. An experimenter watched the participants perform the calibration phase to make sure that they understood the task. If a participant was not giving a reasonable response (for example, by entering a number that was less than 100 for a line that was obviously longer than the reference line) then the experimenter would repeat the task instructions to the participant until they were understood. Otherwise, the experimenter gave no further instruction on how to draw lines or give numbers.

After the calibration block had ended, participants were told that they would be performing a similar task but would be rating made-up words. Participants were randomly assigned to perform the number estimations first or to draw lines first.[11] Those who did the number estimation block first were told that they would be entering numbers for made-up words based on how good the words sounded as new words of English. To familiarize the participants with the full range of words they would be looking at, they were given *bzarshk* [ˈbzɑɹʃk] and *kip* [ˈkɪp] as examples of (respectively) strange-sounding and normal-sounding English words. In addition, they were given the word *poik* [ˈpɔɪk] as an example of an intermediate word. All words in the experiment were displayed in English orthography on the screen as well as played through headphones.

The participants were then instructed that *poik* would serve as their reference word and that it should be assigned the number 100. Words that they thought sounded better than *poik* as words of English should be given a number higher than 100 and analogously for words that sounded worse. Participants were encouraged to use a proportional scale, so that for example if they thought a word was twice as good a word of English as *poik,* then they would enter a number twice as high as 100 (200), and similarly would enter 50 for a word that sounded only half as good as *poik*. The rationale for this procedure is that (unlike with ratings scales that use a fixed set of values), participants are free to extend their scale upward or downward when they encounter new items that are unprecedentedly good or bad; it also makes available essentially

[11] Due to a programming error, more participants were given the number estimation task first than were given the line drawing task first. However, post-hoc tests did not find any significant differences between the two groups.

unlimited granularity for their responses, useful when they encounter new words that seem intermediate between two previous words.

The participants completed four practice words before beginning an experimental phase with the 40 fillers and 80 experimental words described above in §5.1.2. Once the "next" button was pressed, the next word appeared on the screen and the sound file was played once automatically. The order of the words was randomized for each participant. The experimenter stayed in the room for the practice trials but left before the participants began the experimental trials.

After completing the number estimation block, the participants were instructed to perform the same task except with line drawing instead of numbering. *Poik* was again used as the reference word, presented with a line of 100 units. If words were twice as good as *poik*, they were instructed to draw a line twice as long, and so on. Participants drew lines for the same set of practice items and experimental items, in a newly-randomized order. This block completed the experiment. Participants who were assigned to perform the line drawing block first completed the same tasks with the same stimuli, but with the blocks in reverse order.

5.2 Results and Discussion

5.2.1 Calibration

Studies using magnitude estimation can be calibrated to assess their validity. We first examine if participants are self-consistent in the training phase described in the previous section: do the lines they draw match up to the numbers they are attempting to match, and vice versa? We can check this by performing a regression analysis, comparing a participant's numerical response to a line of a particular length against the same participant's line length for the same number. This analysis (carried out with log values) yields a strong positive correlation (r = .96). The slope of the regression line is almost exactly 1. This showed that, as in previous work, our participants had no trouble performing the basic magnitude estimation task. In addition, as a group, they neither underestimate nor overestimate in either modality.

We also examined how participants' responses to the nonword items compared across the two modalities. Regression analysis for these values indicated nearly perfect correlation, r = 0.98, and a perfect slope of 1. This indicates that participants were consistent in their nonword ratings across the two modalities. Therefore, we may assume that these values are valid and reliable (for discussion, see Lodge, 1981; Bard et al., 1996).

5.2.2 Replication of Scholes (1966) and Albright (2009)

We found that the mean log ratings of the borrowed fillers correlated strongly with log ratings from Scholes (1966) and Albright (2009), r = 0.90 and r = 0.86, respectively, indicating that our experiment succeeded in eliciting similar phonotactic well-formedness intuitions.

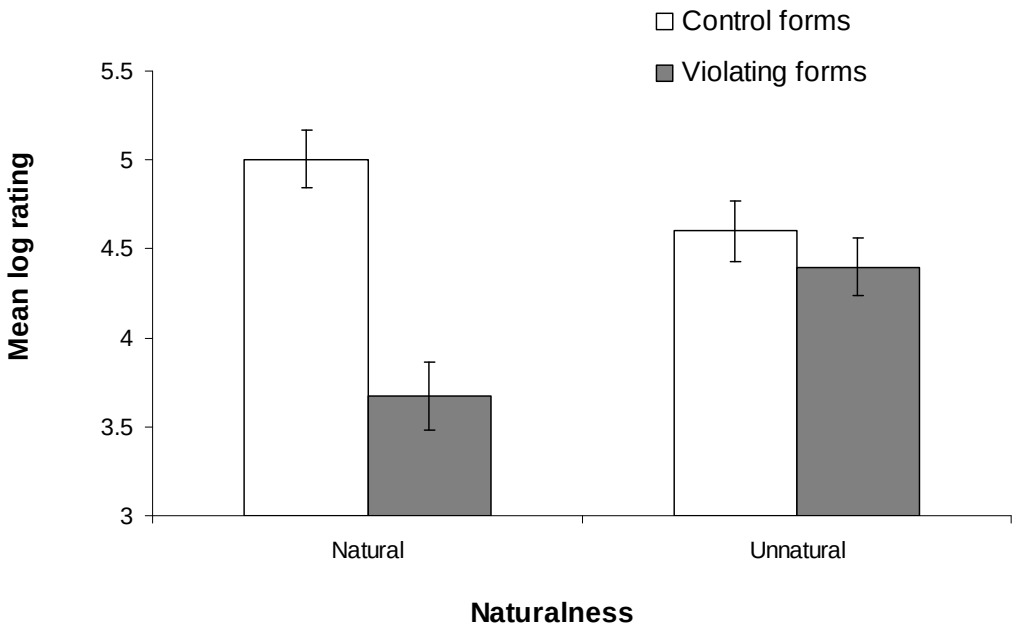
5.2.3 Main results

For the following analyses, data from the line drawing task and the number estimation task, which yielded very similar results, have been collapsed. As a check, we ran all of the analyses on

the line data and numerical data separately, and the results showed the same basic pattern as with the combined data.

Figure 1 shows the mean log ratings for nonwords according to the Naturalness of the constraint being tested (Natural or Unnatural) and to the nonwords' status as Control or Violating forms (error bars represent the standard error of the mean). The figure shows that for Natural constraints, ratings for Violating forms (M = 3.67, SD = 1.02) were much lower than those for Control forms (M = 5.00, SD = 0.87). For the Unnatural constraints, the ratings for Violating forms (M = 4.40, SD = 0.89) were also lower than those for Control forms (M = 4.60, SD = 0.92), but this difference was much smaller — less than a sixth of the difference found for the natural constraints.

Figure 1. Mean log ratings for combined line drawing and number estimation data by Naturalness and Control/Violating status



To evaluate these differences, linear mixed-effects models were created in R (R Development Core Team 2008) using the *lmer*() function of the *lme4* package (Bates, Maechler, and Dai 2008) following Baayen (2008a:ch. 7). As a baseline model, we began with the factors that we were interested in—Naturalness and Control/Violating Status—as fixed effects with an interaction term. Random intercepts were included for Subject and Item because they significantly improved model fit.[12] The results of this model are presented in Table 1. P-values

[12] Following a reviewer's suggestion, we also tried models containing random slopes for subject according to Naturalness and the interaction between Naturalness and Status. These random slopes also significantly improved model fit; however, estimating p-values using MCMC sampling for models with random slopes is not currently implemented in the *languageR* package. The t-values of the model with random slopes were very similar to those in Table 1: Intercept 42.03, Status=Violating −7.30, Naturalness=Unnatural −2.77, and Naturalness/Status interaction 5.03. Moreover, with a large number of degrees of freedom, it can be estimated that a t-value greater than 2 represents a significant value (Baayen 2008a:270). We conclude that adding random slopes to the model does not change the overall pattern or magnitude of significance presented in Table 1.

and 95% confidence intervals (CI) were computed by a Monte Carlo Markov chain (MCMC) sampling method, using the *pvals.fnc*() function of the *languageR* package (Baayen 2008b) with 10,000 samples.

Table 1. Results of mixed-effects model for Naturalness and Control/Violating Status.

Fixed effects

	Estimate	*95% CI*		*t-value*	*p-value*
Intercept	5.00	4.79	5.21	42.85	<0.001
Status = Violating form	−1.33	−1.58	−1.09	−9.49	<0.001
Naturalness = Unnatural	−0.40	−0.65	−0.17	−2.87	0.004
Naturalness = Unnatural & Status = Violating	1.13	0.80	1.48	5.70	<0.001

Random effects

	Standard deviation
Subject (intercept)	0.33
Item (intercept)	0.43
Residual	0.76

Each factor contributed significantly to the model. A potential form begins with the baseline intercept log score of 5.00 (This is in fact the Natural Control mean rating, since Natural Control forms are not further modified by other factors in the model.). The row below *Intercept* indicates that Violating forms had significantly lower ratings in general than Control forms, by about 1.33. In the next row, forms selected for an Unnatural constraint (Control or Violating) received a significantly lower rating than those selected for a Natural constraint by 0.40. Finally, the last row of fixed effects shows that being a Violating form of an Unnatural constraint resulted in a significantly higher rating as compared to forms violating Natural constraint by 1.13. This final factor is the crucial interaction term: it indicates that violating a natural constraint is much worse than violating an unnatural constraint.

To confirm that adding the fixed effects for Naturalness and its associated interaction term improves model fit, the model in Table 1 was compared to an analogous model containing the random effects and a fixed effect for only Control/Violating Status using a log likelihood test, performed with the *anova*() function in R (see Baayen 2008a). The fixed effects for Naturalness and Naturalness x Control/Violating Status significantly improved model fit (negative log likelihoods: *model with naturalness* = −5430 vs. *model without naturalness* = −5446)[13], $\chi^2(2) = 30.18$, $p < .001$. In other words, a model appealing to naturalness fits the data better than a model that treats all of the constraints the same.

5.2.4 Individual constraints

We next examine the results on a constraint-by-constraint basis. We estimate the magnitude of the effect of individual constraints by taking the ratio *log rating of control form/log rating of*

[13] A negative log likelihood closer to 0 indicates a better fit. The fixed effects for naturalness result in a similar increase in log likelihood when random slopes are included in the model.

violator form and average over both data types (line, number) and both word pairs (from (7) and (8)) used to test the constraints. By this measure, with just one exception every natural constraint had a stronger effect on ratings than every unnatural constraint. This is shown in Table 2 below.

Table 2. *Effects of individual constraints*

Constraint	**Status**	**Pairs**	**Effect size**
*[−cont][−cont] IN ONSET	*natural*	*cping/sping, ctice/stice*	1.65
*[glide] IN CODA	*natural*	*jouy/jout, tighw/tibe*	1.56
*[−cons][+cons] IN ONSET	*natural*	*hlup/plup, hmit/smit*	1.51
*[−cont][+nasal] IN ONSET	*natural*	*cnope/clope, pneck/sneck*	1.44
*[+labial][+dorsal] IN CODA	*natural*	*rufk/ruft, trefk/treft*	1.44
*[+dorsal][+labial] IN CODA	*natural*	*bikf/bimf, sadekp/sadect*	1.36
*[+cons][−cons] IN CODA	*natural*	*shapenr/shapent, tilr/tilse*	1.34
*[−sonorant][+sonorant] IN CODA	*natural*	*canifl/canift, kipl/kilp*	1.31
*[+labial][+labial] IN ONSET	*natural*	*bwell/brell, pwickon/twickon*	1.23
*[+cont,−strid][−sonorant]	*unnatural*	*hethker/hethler, muthpy/muspy*	1.14
*[+cont,−strid][−stress,+round]	*unnatural*	*potho/pothy, taitho/taithy*	1.10
*[+diphthong,+round,−back][−ant]	*unnatural*	*boitcher/boisser, noiran/nyron*	1.10
*[+diphthong][+cont,−anterior]	*unnatural*	*foushert/fousert, pyshon/pyson*	1.08
*[$_{word}$ [−diphthong,+round,+high]	*unnatural*	*ooker/ocker, utrum/otrum*	1.03
*[−back][+diphthong]	*unnatural*	*youse/yoss, yout/yut*	1.02
*[$_{word}$ [+diphthong,+round][+voice]	*unnatural*	*oid/oit, ouzie/oussie*	1.02
*[+cont,+voice,−ant][+str][−son]	*unnatural*	*zhep/zhem, zhod/zhar*	1.01
*[+cons,−anterior][−sonorant]	*unnatural*	*ishty/ishmy, metchter/metchner*	0.99
*[−son,−voice][−son,+voice]	*natural*	*esger/ezger, trocdal/troctal*	0.98
*[+round,+high][−cons,−sonorant]	*unnatural*	*luhallem/laihallem, tuhaim/towhaim*	0.97

The exception was *[−son,−voice][−son,+voice], exemplified by *esger* vs. *ezger* and *trocdal* vs. *troctal*. This exception is easily explained: although our speaker was in general able produce our forms with high accuracy, measurement indicates that she did not succeed in producing a voiced closure in either *esger* or *trocdal*; hence what we had intended as phonotactically illegal forms were very close to ordinary (['ɛskɚ] and ['tɹɑktəl]).

5.2.5 Did the unnatural constraints have any effect?

The mixed-effects model establishes that the unnatural constraints did not have as strong an effect as the natural constraints, but the question remains whether the unnatural constraints had any effect at all. To examine this possibility, we created another linear mixed-effects model on a subset of the data containing only the unnatural constraint forms, with Control/Violating Status as a fixed effect and random intercepts for Subject and Item. The model (using *pvals.fnc* as above) found that the small difference between Violating and Control forms, though trending in the right direction, did not reach significance, *Estimate* = −0.20, *t-value* = −1.54, $p = 0.12$. A second version of this experiment using only orthographic forms (not reported here) also found that the Control forms were rated only slightly better, but the difference reached significance in that version. We conclude that the unnatural constraints had, at best, only a small effect on participant ratings.

5.2.6 *Considering additional factors*

To check if factors other than those discussed above played a role in our experiment, we also considered a number of additional variables post-hoc. These included the following: (a) for each form, the weight assigned by the phonotactic learning model to the constraint it violates;[14] (b) the score assigned by Albright's (2009) phonotactic learner;[15] (c) two measures of length: number of syllables and number of segments, and (d) two measures of assessing the simplicity or generality of constraints: number of features in the constraint (roughly following Chomsky and Halle 1968:334), and the proportion of all logically possible *n*-grams that violate each *n*-gram constraint (Hayes and Wilson 2008:394). [16]

Each of these additional factors was examined by adding them to the model in Table 1 one at a time, both with and without interaction terms. The resulting models were compared to the original model using log likelihood tests to determine if adding the additional term(s) would significantly improve model fit.[17]

Only one of these factors resulted in a significant increase to model fit: number of features in the constraint — but the effect was in the wrong direction. That is, violations of constraints with more features had a stronger effect on nonword ratings than constraints with fewer features. This goes against traditional views of generality, in which constraints with fewer features are simpler and simpler constraints are more highly valued. As such, we judge that this effect was most likely an accident. Our factors of interest (i.e., those in Table 1) remain highly significant in the model even when number of features is included, meaning that this additional factor does not confound the main findings of this study.

It is important to keep in mind the following when considering these additional factors: the current study was designed to compare the unnatural constraints to the natural constraints, so we attempted to control other factors as much as possible. As a result, the nonwords varied minimally with respect to these other factors (e.g., constraints were chosen such that their weights were similar and nonwords varied little in their length). Therefore, any effects of these factors (or the lack thereof) are not very meaningful for this experimental design, provided that they do not confound the results of interest. A study intended to test for these additional factors would vary them systematically rather than controlling for them.

6. Possible objections

We consider here various alternative interpretations of our results.

[14] Using the constraint weights in the model is mostly redundant, since they closely match our Violating vs. Control factor. Our interest was whether the small differences in weights among the test stimuli would have a significant effect beyond what would result from the binary factor.

[15] We would like to thank Adam Albright for assistance in computing these scores.

[16] Complexity is worth examining because a considerable body of evidence indicate a bias for simple generalizations in phonological learning; see Moreton and Pater (in prep). For purposes of computing complexity we included the ad hoc features we used for word boundaries and syllabification.

[17] Constraint weights and Albright (2009) scores were centered before running the models by subtracting the mean from each value to reduce collinearity (see Baayen 2008a:276-277).

6.1 The effect of training data

Our training data (§4) were chosen because we felt that they offered the best chance of matching the mental lexicons of our experimental participants. However, it is possible that our training set was inadequate in two ways. First, we excluded many morphologically complex forms, under the assumption that these forms have their own phonotactics and would not affect the judgments of new monomorphemic forms. This assumption may not have been well founded; that is, perhaps complex forms did affect the well-formedness judgments of our participants. Some of our constraints are indeed potentially affected; thus although (6f) *[+cont,−strid][−son] is violation-free in simplex forms, it is violated in past tenses such as *bathed* [beɪðd].

Second, it is possible that we may have underestimated the number of words in the mental lexicons of our participants that are very rare and violate the unnatural constraints: if these are included in the training data, the weights of the unnatural constraints would go down, perhaps explaining the experimental findings. For instance, it seems plausible that many of our participants do not know of *Pushkin* (it is not in our training set) but if they do, and consider it to be an English word, it would produce a slightly lower weight for constraint (6b), *[+cons,−ant][−son].

To test these possibilities, we created three new training sets. The first contained all of the affixed forms that we had previously excluded. For the second, we tried to include all exceptions to our unnatural constraints (such as *Pushkin*) that the participants might plausibly have been familiar with; most of these we found by consulting the full Carnegie-Mellon database. The total number added was 24. The third training set combined the first two, including both the affixed forms and the rare exceptions. We then reweighted the 160-constraint grammar using each of the new training sets. The mean weights for the natural and unnatural constraints using each training set are shown in Table 3.

Table 3. Mean constraint weights for unnatural and natural constraints using the original training set and each of the three modified training sets.

	Original	With affixed forms	With exceptions	With exceptions and affixed forms
Unnatural constraints	3.96	3.84	**3.65**	**3.58**
Natural constraints	3.79	**4.07**	3.79	**4.07**

As the means in Table 3 demonstrate, constraint weights did vary to some extent depending on the training set. The means in bold mark the largest changes. As expected, the weights for the unnatural constraints fell slightly when relevant exceptions were added to the training set. The natural constraints, on the other hand, received higher weights when the affixed forms were added, probably as a result of the larger set of data for which they could "prove their worth" by remaining exceptionless.

However, the changes in mean constraint weight were relatively small for both the unnatural constraints (3.96 → 3.58) and the natural constraints (3.79 → 4.07). Even though the unnatural

constraint weights become smaller than the natural constraint weights, they are still quite similar. It is unlikely that this small difference could explain the large difference in effect between natural and unnatural constraints found in our experiments. In our testing using these weights, constraint weight continued to have no statistically significant effect.

6.2 Have we correctly classified our constraints for naturalness?

It is not easy to establish firmly the naturalness of constraints by either of the criteria laid out in section 1. The reviewers for this article were helpful in offering their input concerning some of our naturalness claims. No one seems to have objected to the classification of our natural constraints as natural, but some of our unnatural constraints may be been prematurely classified as such. Thus, (6f) excludes the non-strident fricatives [θ, ð] before obstruents; the Latin fricative [f], phonetically similar to [θ], was likewise excluded before obstruents.[18] Constraint (6b) forbids palato-alveolars before obstruents; this was a productive phonological constraint in Sanskrit; see Whitney (1889, 72-75). Constraint (6c), forbidding [j] before diphthongs, might be assigned a phonetic rationale: it bans a *high - nonhigh - high* pattern within the syllable, perhaps analogous to the widespread ban on complex (triply-linked) contour tones (Yip 2002, 30).

In light of this, we ran an additional model in which (6b) (6c), and (6f) were recategorized as natural constraints. The effects in the new model were somewhat smaller than in the original model, but the overall pattern remained the same and the crucial interaction term remained highly significant ($p < .001$). Indeed, our main result is fairly robust against further such reclassifications. We experimented with reclassifying as natural not just (6b) (6c), and (6f), but also (6a), (6i), and (6j) — the three unnatural constraints that had the smallest effect size (Table 2), and thus contributed most to our statistical result. Even with just four unnatural constraints still classified as such, the main result remained statistically significant ($p = .046$).

In the long term, finding better ways of assessing phonological naturalness, for example, through typological surveys and modeling, is needed to allow us to pin down the concept of naturalness more precisely.

6.3 How do experimental subjects interpret ill-formed stimuli?

As noted earlier, our experiment faced the problem of how to present to experimental participants phonological sequences that are phonologically ill-formed, given that people sometimes hear phonologically-illegal forms as perceptually similar legal forms. The question arises primarily with our natural violator forms, which, it seems clear, were by far the hardest to hear accurately. We must consider the following scenario: the participants may have perceptually repaired a stimulus (for example, hearing our *hlup* as [flʌp]), but at the same time noticed that the stimulus was a phonetically poor rendition of the perceived phonemic intent.

[18] We confirmed this by searching a Latin electronic corpus; [f] occurs only before vowels and the liquids [l, r].

The very low scores assigned to our natural-violator stimuli might reflect this phonetic factor, rather than the phonological ill-formedness of the phonemic sequences we had intended. [19]

We carried out an informal post hoc test of this hypothesis by asking seven English-speaking undergraduate students who had had one term of phonetic training to transcribe our natural-violator stimuli in IPA notation. Unlike our experimental participants, they listened without the aid of an orthographic form. We found that a number of the stimuli were indeed systematically misheard; the worst-case example was *jouy* [ˈdʒaʊj], heard by all seven consultants as disyllabic [ˈdʒaʊ.i]. For purposes of assessing our main result, we confined our attention to the opposite end of the spectrum: the three forms heard accurately by all seven consultants (*pneck*, *bwell*, and *rufk*) and the three forms heard accurately by six out of seven (*cping*, *sadekp*, and *trefk*).

Redoing the statistical analysis with just these six natural items and their controls, we found that the new model was very similar to the one in Table 1. Most importantly, the interaction effect remained significant: violating an unnatural constraint resulted in a smaller reduction in participant ratings than violating a natural constraint. In fact, the model's estimate of the interaction effect (i.e., *how much* worse is it to violate a natural constraint than to violate an unnatural constraint) actually increased slightly from 1.13 in the original model to 1.38 in the present model. Moreover, the effect cannot be attributed to higher weights for the natural constraints penalizing the accurately-heard forms, because the average weight of these constraints was in fact lower than the average for our natural constraints overall. We conclude that although misperception of stimuli may have occurred in our experiments, it is unlikely to provide an adequate alternative explanation of our results.

6.4 Could the unnatural constraints have been excluded on statistical grounds?

It is possible that the magnitude of a constraint's weight is not a fully accurate reflection of the constraint's importance in accounting for the data. This hypothesis can be checked by carrying out a statistical assessment of a constraint's effect in improving the performance of a grammar. The rationale for doing this is the possibility that language learners might likewise be unconsciously savvy about the effectiveness of constraints, and evaluate them with a procedure analogous to statistical testing.

Pursuing this possibility, we used the likelihood ratio test, which is commonly used to assess models that are in a subset relation.[20] For purposes of testing a constraint, we designate as the *subset model* a grammar (with optimized weights) formed with all of the constraints except the tested one, and the *full model* the grammar (with optimized weights) that uses all the constraints. The likelihood ratio test computes the value $-2 * \log(\text{probability}_{D,\text{full model}})/ (\text{probability}_{D,\text{subset model}})$, where probability_D is the probability that the model assigns to the training data. The distribution of this value can be approximated by a chi-square distribution with one degree of

[19] That detailed phonetic properties of experimental stimuli can strongly affect phonotactic ratings is demonstrated in Wilson and Davidson (to appear).

[20] For a clear description of the test see Pinheiro and Bates (2000:83).

freedom, from which one can determine the probability of the hypothesis that the improvement in accuracy due to including the target constraint could arise by accident.

Using software provided to us by Colin Wilson, we achieved an approximation of this statistic for both our natural and our unnatural constraints.[21] All constraints tested as highly significant by this test; no *p*-value was greater than .007, and many were much smaller. Our conclusion is that the unnatural constraints are as well justified by the lexical data as the natural ones.

7. General discussion

To review, the original impetus for our study was a point made by Hayes and Wilson (2008: sec. 8.5) concerning their Wargamay simulation; namely that in the course of learning the system, their model generated a large set of constraints that are evidently phonologically unnatural. Hayes and Wilson suggested that either (1) language learners are actually very adept at learning such generalizations, so that these constraints would turn out to valid if tested against Wargamay native intuition, or (2) the constraints reveal a defect in the model. Our findings point to the latter conclusion. Colavin, Levy, and Rose (2010), applying the Hayes/Wilson model to a corpus of Amharic roots, obtained a similar result, finding that the model was able to provide only limited improvement over a core model of hand-created constraints. We conjecture that in this case the model was likewise finding unnatural constraints, which are undervalued by Amharic learners.

Our findings do not suffice to identify with certainty where the model is going astray. We consider three possibilities here.

7.1 Naturalness

Our original hypothesis was that natural constraints are learned more easily than unnatural constraints. As we noted earlier, this hypothesis takes two flavors, of which one is that unnatural constraints are simply inaccessible to language learners. We take the extensive evidence reviewed in §2.1 as indicating that this possibility is unlikely; and indeed it is possible that in our own experiment, unnatural constraints did have a modest effect on ratings; see §5.2.5 above.

A more plausible theory is that learners are *biased* to favor natural generalizations, a view suggested by Wilson 2006, Albright 2007, Berent et al. 2007, Finley 2008, Kawahara 2008, Moreton 2008, Finley and Badecker 2009, Hayes et al. 2009, and others. A simple way to check for bias is to examine the output of maxent grammars in which the weights of the unnatural constraints have been "hobbled"; i.e. given a lower weight than would be justified simply by fit to the data. We experimented with this by modifying the grammar described in §4, multiplying the weights of the unnatural constraints by a factor that varied from 0 to 1, and examining the correlation of the resulting scores with the log average participant ratings for each experimental

[21] The approximation was that we could only use the 98 first-learned constraints as our base grammar; memory limitations prevented our testing with the full 160-constraint grammar.

stimulus. The best-fit value of this "hobbling" factor was .33 — a substantial weakening, but not elimination, of the effect of the unnatural constraints.

Language learners must have access to some basis for a learning bias. We think a plausible basis would be phonetically-based phonology (see e.g., Myers 1997; Boersma 1998; Hayes 1999; Steriade 1999, 2001; Côté 2000; Flemming 2001; Hayes, Kirchner and Steriade 2004). Under this approach, language learners evaluating phonotactic generalizations would evaluate them not just for their degree of fit to the learning data, but also for their effectiveness in avoiding articulatory difficulty and in maintaining perceptual distance between contrasting forms in perception.

7.2 Naturalness again: are consonant-vowel generalizations harder to learn?

Moreton (2008) has provided experimental evidence suggesting that phonotactic generalizations that require access to both vowel and consonant identity can in some cases be phonetically natural but nevertheless harder to learn: a bias exists, but it is a general learning bias rather than one based on phonetic naturalness. Becker et al. (2001) suggest that the same principle could also explain their data. As a reviewer pointed out, this explanation might be applicable here. All but one of our natural constraints ((3-5)) evaluate consonant sequences, and all but one of our unnatural constraints ((6)) evaluate consonant-vowel sequences. We have no data that could distinguish whether it is a general learning bias vs. a phonetic one that favors the natural constraints in cases where the two principles are both applicable. We add that the world's phonologies do include a great many cases of vowel-consonant interaction, as in palatalization, nasalization of nasal-adjacent vowels, influence of secondary consonant articulation on vowel quality, and intervocalic lenition, so we think that further research would be needed to establish a learning bias for consonant-vowel patterns more firmly.

7.3 The search heuristics

Another possibility is that the Hayes/Wilson learner might learn fewer (or no) unnatural-seeming constraints if it were modified to use different search heuristics, so that "accidentally true" constraints became less tempting to it. The existing heuristics, reviewed above in §3, favor constraints that are exceptionless or nearly so. Yet exceptionlessness is not necessarily as helpful a criterion as it might seem. For instance, in seeking exceptionlessness the model favored the constraint (6i) *ʒ [+stress][−son], which reduced the impetus to learn a more general ban on pretonic [ʒ] (*ʒ [+stress]); indeed there was no such ban in the 160-constraint grammar we used.[22] If the model were altered to give more priority to generality and less to exceptionlessness, then it might have acquired *ʒ [+stress] first, which would have then devalued (6i) (given it a lower weight). It might even have prevented the selection of (6i) entirely, since the model favors only constraints whose violation counts are below the *expected* value; and learning *ʒ [+stress] would lower the expected value for (6i).

[22] The reason the model included [−sonorant] is that none of the six words in our training set that had pretonic [ʒ] (e.g. *luxuriant*, *regime*, *genre*) included an obstruent after the stressed vowel; thus adding [−sonorant] reduces the number of exceptions from six to zero.

The plausibility of this scenario is increased by the fact that our Unnatural control forms generally received lower ratings than our Natural control forms (see Fig. 1). A possible reason for this is that these forms violate simple constraints unlearned by the model, such as *ʒ [+stress]; these are the constraints that might have preempted the learning of our unnatural constraints had they been learned first.[23]

7.4 For future work

In conclusion, we suggest that the modeling research strategy pursued by Hayes and Wilson could be informative concerning the effectiveness of a naturalness bias approach. As Wilson (2006) showed, a bias can be formalized in maxent through the use of constraint-specific prior terms, which militate against the assignment of high weights to particular constraints, such as (as Wilson suggests) the phonetically natural ones. In this approach, the hobbling of unnatural constraints (as in §7.1) would take place as part of learning system itself, rather than being a post hoc procedure.

Thus, in principle, all the ingredients for exploring the role of natural vs. unnatural constraints in phonotactic learning are at hand. Candidate theories of UG must be formalizable as constraint sets (or systems that construct constraints), and must come with a mechanism for imposing biases, perhaps based on more than one mechanism (here we have considered phonetic naturalness, single-tier status, and simplicity). Such systems could be tested by the method we have used here, comparing model predictions with native speaker ratings of carefully chosen stimulus words. Our hope is that such a program would facilitate progress on the question of naturalness in phonology by making it possible to test specific hypotheses and mechanisms.

[23] A modified learner created by Wilson that appears promising on these lines is reported in Berent, Wilson et al. (2012), Wilson and Davidson (to appear), and Hayes, Wilson and Shisko (forthcoming). This learner uses the principle of "gain" (Della Pietra et al. 1997:1) to select constraints. In a preliminary examination of this modified system, we found that the constraints it selects were indeed more general and less idiosyncratic than those chosen by the Hayes/Wilson (2008) learner; more specifically, it learned none of our ten unnatural constraints. However, it may still be learning *different* ones: for instance, it posited a constraint banning stressed lax vowels before coronal codas as well as a constraint against word-final sonorants. Our testing was tentative, due to memory limitations, and more serious evaluation of the revised model awaits further research.

Appendix A: Mean log experimental ratings by nonword, with constraint and constraint weight.

Phonetic transcriptions of the stimuli are given in (7) and (8).

Naturals					
Violators		Controls		Constraint	Weight
tilr	3.51	*tilse*	4.77	*[+cons][−cons] IN CODA	3.01
shapenr	3.58	*shapent*	4.71		
trefk	3.80	*treft*	5.23	*[+labial][+dorsal] IN CODA	2.68
rufk	3.64	*ruft*	5.47		
bikf	3.23	*bimf*	3.76	*[+dorsal][+labial] IN CODA	3.03
sadekp	3.13	*sadect*	4.88		
esger	4.50	*ezger*	4.20	*[−son,−voice][−son,+voice]	3.14
trocdal	5.12	*troctal*	5.24		
bwell	4.23	*brell*	5.27	*[+labial][+labial] IN ONSET	4.07
pwickon	4.09	*twickon*	4.98		
cnope	3.49	*clope*	5.38	*[−cont][+nasal] IN ONSET	4.21
pneck	3.54	*sneck*	4.76		
hlup	3.56	*plup*	5.06	*[−cons][+cons] IN ONSET	4.27
hmit	3.41	*smit*	5.49		
cping	3.23	*sping*	5.40	*[−cont][−cont] IN ONSET	4.33
ctice	3.29	*stice*	5.33		
kipl	3.77	*kilp*	5.12	*[−son][+son] IN CODA	4.54
canifl	3.58	*canift*	4.51		
jouy	3.69	*jout*	5.20	*[glide] IN CODA	4.65
tighw	3.04	*tibe*	5.29		
Mean	**3.67**		**5.00**		**3.79**

Unnaturals					
Violators		Controls		Constraint	Weight
ouzie	4.23	*oussie*	4.55	*[$_{word}$ [+diphthong,+round][−son,+voice]	3.51
oid	4.28	*oit*	4.09		
pyshon	4.71	*pyson*	5.30	*[+diphthong][+continuant,−anterior]	3.58
foushert	4.37	*fousert*	4.49		
potho	4.68	*pothy*	5.05	*[+cont,−strident][−stress,+round]	3.65
taitho	4.34	*taithy*	4.88		
zhep	3.75	*zhem*	3.76	*[+cont,+voice,−ant][+stress][−son]	3.71
zhod	3.75	*zhar*	3.84		
luhallem	4.21	*laihallem*	4.01	*[+round,+high][−cons,−son]	3.78
tuheim	4.27	*towheim*	4.18		
noiron	4.33	*nyron*	5.26	*[+diphthong,+round,−back][−anterior]	4.01
boitcher	5.02	*boisser*	4.98		
youse	4.57	*yoss*	4.74	*[−back][+diphthong]	4.02
yout	4.70	*yut*	4.69		
hethker	4.45	*hethler*	4.94	*[+cont,−strident][−son]	4.04
muthpy	4.35	*muspy*	5.07		
ishty	3.92	*ishmy*	3.94	*[+cons,−ant][−son]	4.11
metchter	4.71	*metchner*	4.64		
utrum	4.65	*otrum*	4.97	*[$_{word}$ [−diphthong,+round,+high]	5.22
ooker	4.68	*ocker*	4.61		
Mean	**4.40**		**4.60**		**3.96**

Appendix B: Filler items

From Scholes (1966):

stin [ˈstɪn], *smat* [ˈsmæt], *blung* [ˈblʌŋ], *frun* [ˈfɹʌn], *glung* [ˈglʌŋ], *shlurk* [ˈʃlɚk], *skeep* [ˈskip], *vrun* [ˈvrʌn], *srun* [ˈsɹʌn], *vlurk* [ˈvlɚk], *shtin* [ˈʃtɪn], *shnet* [ˈʃnɛt], *zrun* [ˈzɹʌn], *shmat* [ˈʃmæt], *zlurk* [ˈzlɚk], *znet* [ˈznɛt], *fnet* [ˈfnɛt], *zhmat* [ˈʒmæt], *vnet* [ˈvnɛt], *vkeep* [ˈvkip]

From Albright (2009):

wiss [ˈwɪs], *stip* [ˈstɪp], *trisk* [ˈtɹɪsk], *preek* [ˈpɹik], nace [ˈneɪs], *spling* [ˈsplɪŋ], *bize* [ˈbaɪz], *gude* [ˈgud], *drit* [ˈdɹɪt], *skick* [ˈskɪk], *kweed* [ˈkwid], *blig* [ˈblɪg], *gwenge* [ˈgwɛndʒ], *twoo* [ˈtwu], *sfoond* [ˈsfund], *smeerg* [ˈsmirg], *trilb* [ˈtɹɪlb], *ploamf* [ˈploʊmf], *smeenth* [ˈsminθ], *pwudge* [ˈpwʌdʒ]

References

Albright, Adam. 2002. Islands of reliability for regular morphology: Evidence from Italian. *Language* 78:684-709.

Albright, Adam. 2007. Natural classes are not enough: Biased generalization in novel onset clusters. 15th Manchester Phonology Meeting, Manchester UK, May 24–26.

Albright, Adam. 2009. Feature-based generalization as a source of gradient acceptability. *Phonology* 26:9-41.

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

Baayen, R. Harald. 2008a. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, MA: Cambridge University Press.

Baayen, R. Harald. 2008b. languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics." R package version 0.953.

Baayen, R. Harald, Piepenbrock, R., and Gulikers, L. 1995. The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, Penn.: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Bard, Ellen Gurman, Dan Robertson and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72:32-68.

Bates, Douglas, Martin Maechler, and Bin Dai. 2008. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-28. http://lme4.r-forge.r-project.org/.

Becker, Michael, Nihan Ketrez, and Andrew Nevins. 2011. The surfeit of the stimulus: Grammatical biases filter lexical statistics in Turkish voicing deneutralization. *Language* 87:84-125.

Berent, Iris, Donca Steriade, Tracy Lennertz, Vered Vaknin. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104:591-630.

Berent, Iris, Tracy Lennertz, Jongho Jun, Miguel A. Moreno, & Paul Smolensky. 2008. Language universals in human brains. In *Proceedings of the National Academy of Science*, 105:5321-5325.

Berent, Iris, Tracy Lennertz, Paul Smolensky, and Vered Vaknin-Nusbaum. 2009. Listeners' knowledge of phonological universals: Evidence from nasal clusters. *Phonology* 26:75-108.

Berent, Iris, Colin Wilson, Gary F. Marcus, and Douglas K. Bemis. 2012. On the role of variables in phonology: remarks on Hayes and Wilson (2008). *Linguistic Inquiry* 43: 97-119.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22:39–71.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge, MA: Cambridge University Press.

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

Chambers, Kyle. E., Kristine H. Onishi, and Cynthia Fisher. 2003. Infants learn phonotactic regularities from brief auditory experience. *Cognition* 87:B69-B77.

Chomsky, Noam, and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1:97–138.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Clements, George. N. 1990. The role of the sonority cycle in core syllabification. In *Papers in Laboratory Phonology I*, ed. by John Kingston & M. Beckman, 283-333. Cambridge: Cambridge University Press.

Colavin, Rebecca, Roger Levy, and Sharon Rose. 2010. Modeling OCP-place in Amharic with the Maximum Entropy phonotactic learner. To appear in the proceedings volume of the 46th meeting of the Chicago Linguistics Society.

Côté, Marie-Hélène. 2000. Consonant cluster phonotactics: a perceptual approach. Doctoral dissertation, Department of Linguistics, MIT, Cambridge MA.

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis, and Ingrid Norrmann. 2011. Explaining sonority projection effects. *Phonology* 28:197–234.

Davidson, Lisa. 2007. The relationship between the perception of non-native phonotactics and loanword adaptation. *Phonology* 24:261-286.

Dell, Gary S., Kristopher D. Reed, David R. Adams, and Antje S. Meyer. 2000. Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 6:1355-1367.

Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19:380–393.

Dixon, Robert M. W. 1981. Wargamay. In *Handbook of Australian languages, volume II*, ed. by Robert M. W. Dixon and Barry J. Blake, 1–144. Amsterdam: John Benjamins.

Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, and Jacques Mehler. 1999. Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, 25:1568-1578.

Finley, Sara. 2008. Formal and cognitive restrictions on vowel harmony. Doctoral dissertation, Johns Hopkins University, Baltimore, MD.

Finley, Sara, and William Badecker. 2008. Substantive biases for vowel harmony languages. In *Proceedings of the West Coast Conference on Formal Linguistics 27,* ed. by Natasha Abner and Jason Bishop, 168-176. Somerville, MA: Cascadilla Press.

Finley, Sara, and William Badecker. 2009. Artificial language learning and feature-based generalization. *Journal of Memory and Language* 61:423-437.

Flemming, Edward. 2001. Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18: 7-44.

Flemming, Edward. 2004. Contrast and perceptual distinctiveness. In *Phonetically-based phonology,* ed. by Bruce Hayes, Robert Kirchner and Donca Steriade, 232-276. Cambridge, MA: Cambridge University Press.

Gildea, Daniel, and Daniel Jurafsky. 1996. Learning bias and phonological rule induction. *Computational Linguistics* 22:497–530.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Proceedings of the Stockholm Workshop on Variation within Optimality Theory, ed. by Jennifer Spenader, Anders Eriksson, and Osten Dahl, 111–120.

Greenberg, Joseph. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (ed.), *Universals of Language*, ed. by Joseph Greenberg, 73-113. Cambridge, MA: MIT Press.

Greenberg, Joseph H. 1978. Some generalizations concerning initial and final consonant clusters. In *Universals of human language (Vol. 2)*, ed. by Edith A. Moravcsik, 243–279. Stanford, CA: Stanford University Press.

Harris, James. 1983. *Syllable structure and stress in Spanish*. Cambridge, MA: MIT Press.

Hayes, Bruce. 1999. Phonetically-driven phonology: the role of optimality theory and inductive grounding. *Functionalism and formalism in linguistics, Volume I*, ed. by Michael Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatly, 243-285. Amsterdam: John Benjamins.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: The early stages. In *Fixing priorities: Constraints in phonological acquisition*, ed. by René Kager, Joe Pater, and Wim Zonneveld, 158–203. Cambridge, MA: Cambridge University Press.

Hayes, Bruce, Robert Kirchner, and Donca Steriade, eds. 2004. *Phonetically-based phonology*. Cambridge, MA: Cambridge University Press.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379-440.

Hayes, Bruce, Colin Wilson, and Anne Shisko. (forthcoming) Maxent grammars for the metrics of Shakespeare and Milton. Ms., Department of Linguistics, UCLA.

Hayes, Bruce, Kie Zuraw, Peter Siptár and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85:822-863.

Heinz, Jeffrey. 2010a. Learning long-distance phonotactics. *Linguistic Inquiry* 41:623-661.

Heinz, Jeffrey. 2010b. On the role of locality in learning stress patterns. *Phonology* 26: 303–351.

Iverson, Gregory K., and Joseph C. Salmons. 2005. Filling the gap. *Journal of English Linguistics* 33:207-221.

Jarosz, Gaja. 2006. Rich lexicons and restrictive grammars: Maximum likelihood learning in Optimality Theory. Doctoral dissertation, Johns Hopkins University, Baltimore, MD.

Jespersen, Otto. 1909. *A modern English grammar on historical principles. Part I: Sounds and spellings*. London: George Allen & Unwin.

Kager, René. 1999. *Optimality theory*. Cambridge: Cambridge University Press.

Kager, René, and Joe Pater. Forthcoming. Phonotactics as phonology: Knowledge of a complex constraint in Dutch. Ms., Utrecht University and University of Massachusetts, Amherst.

Kawahara, Shigeto. 2008. Phonetic naturalness and unnaturalness in Japanese loanword phonology. *Journal of East Asian Linguistics,* 17: 317-330.

Kawasaki, Haruko. 1982. An acoustical basis for universal constraints on sound sequences. Doctoral dissertation, University of California, Berkeley.

Kenstowicz, Michael, and Charles Kisseberth. 1977. *Topics in phonological theory.* New York: Academic Press.

Kiparsky, Paul. 1982. Lexical phonology and morphology. In *Linguistics in the Morning Calm*, ed. by In-Seok Yang. Seoul.

Lodge, Milton. 1981. *Magnitude scaling: Quantitative measurement of opinions*. Beverly Hills/London: Sage.

Lombardi, Linda. 1999. Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language and Linguistic Theory* 17:267-302.

Massaro, Dominic W, and Michael M. Cohen. 1980. Phonological constraints in speech perception. *Journal of the Acoustical Society of America*, 67, S26.

Mattys, Sven L., and Peter W. Jusczyk. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition* 78:91–121.

McClelland, James L., and Brent C. Vander Wyk. 2006. Graded Constraints on English Word Forms. Ms., Department of Psychology, Stanford University. http://psychology.stanford.edu/~jlm/papers/GCEWFs_2_18_06.pdf

Morelli, Frieda. 1999. The phonotactics and phonology of obstruent clusters in Optimality Theory. Doctoral dissertation, University of Maryland, College Park.

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

Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25:83-127.

Moreton, Elliott and Joe Pater. In preparation. Learning artificial phonology: A review. Ms., University of North Carolina and University of Massachusetts. http://www.unc.edu/~moreton/Papers/MoretonPater.Draft.3.5.pdf.

Myers, Scott. 1997. Expressing phonetic naturalness in phonology. In *Derivations and constraints in phonology*, ed. by Iggy Roca, 125-152. Oxford: Oxford University Press.

Ohala, John. 1981. The listener as a source of sound change. In Carrie S. Masek, Roberta A. Hendrik and Mary Frances Miller (eds.) *Papers from the Parasession on Language and Behavior*. Chicago: Chicago Linguistic Society, 178-203.

Onishi, Kristine H., Kyle E. Chambers, and Cynthia Fisher. 2002. Learning phonotactic constraints from brief auditory experience. *Cognition* 83:B13-B23.

Pater, Joe, and Andries Coetzee. 2008. Weighted constraints and gradient phonotactics in Muna and Arabic. *Natural Language and Linguistic Theory* 26:289-337.

Pater, Joe, and Anne-Michelle Tessier. 2003. Phonotactic knowledge and the acquisition of alternations. In *Proceedings of the 15th International Congress of Phonetic Sciences*, ed. by Maria-Josep Solé´, Daniel Recasens, and Joaquín Romero, 1177–1180. Barcelona: Universitat Autónoma de Barcelona.

Peperkamp, Sharon, and Emmanuel Dupoux. 2007. Learning the mapping from surface to underlying representations in artificial language learning. In *Laboratory Phonology 9*. ed. by Jennifer Cole & José Hualde, 315-338. Berlin: Mouton de Gruyter.

Peperkamp, Sharon, Katrin Skoruppa, and Emmanuel Dupoux. 2006. The role of phonetic naturalness in phonological rule acquisition. In *Proceedings of the 30th Annual Boston University Conference on Language Development,* ed. by David Bamman, Tatiana Magnitskaia, and Colleen Zaller, 464-475. Somerville, MA: Cascadilla Press.

Pierrehumbert, Janet. 2006. The statistical basis of an unnatural alternation. In *Laboratory phonology VIII: varieties of phonological competence*, ed. by Louis Goldstein, D. H. Whalen, and Catherine T. Best, 81-107, Berlin: Mouton de Gruyter.

Pinheiro, José, and Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. Springer.

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Fixing priorities: Constraints in phonological acquisition*, ed. by René Kager, Joe Pater, and Wim Zonneveld, 245–291. Cambridge: Cambridge University Press.

Pullum, Geoffrey K. and William A. Ladusaw. 1996. *Phonetic symbol guide*, 2nd ed. Chicago: University of Chicago Press.

Pycha, Anne, Pawel Nowak, Eurie Shin, and Ryan Shosted. 2003. Phonological rule-learning and its implications for a theory of vowel harmony. In *Proceedings of the West Coast Conference on Formal Linguistics* 22, Gina Garding and Mimu Tsujimura, 423-435. Somerville, MA: Cascadilla Press.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Scholes, Robert. 1966. *Phonotactic grammaticality*. The Hague: Mouton.

Seidl, Amanda and Eugene Buckley. 2005. On the learning of arbitrary phonological rules. *Language Learning and Development* 1:289-316.

Selkirk, Elizabeth O. 1982. The syllable. In The *structure of phonological representations, Part II*, ed. by Harry van der Hulst and Norval Smith, 337-383. Dordrecht: Foris.

Sievers, Eduard. 1881. *Grundzüge der Phonetik*. Breitkopf und Härtel, Leipzig.

Skoruppa, Katrin and Sharon Peperkamp. 2011. Adaptation to novel accents: Feature-based learning of context-sensitive phonological regularities. *Cognitive Science* 35:348-366.

Steriade, Donca. 1999. Alternatives to syllable-based accounts of consonantal phonotactics. In *Proceedings of the 1998 Linguistics and Phonetics Conference*, ed. by Osamu Fujimura, Brian Joseph, and B. Palek, 205–245. Prague: The Karolinum Press.

Steriade, Donca. 2001. Directional asymmetries in place assimilation: A perceptual account. In *The Role of Speech Perception in Phonology*, ed. by Elizabeth Hume and Keith Johnson, 219-250. New York: Academic Press.

Warker, Jill A., Gary S. Dell, Christine A. Whalen, and Samantha Gereg. 2008. Limits on Learning Phonotactic Constraints From Recent Production Experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 34:1289–1295

Warker, Jill A. and Gary S. Dell. 2006. Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32: 387-398.

Whitney, William Dwight. 1889. *Sanskrit grammar*. Cambridge, MA: Harvard University Press.

Wilson, Colin. 2003. Experimental investigation of phonological naturalness. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics*, ed. by Gina Garding and Mimu Tsujimura, 533-546. Somerville, MA: Cascadilla Press.

Wilson, Colin. 2006. Learning phonology with substantive bias: an experimental and computational investigation of velar palatalization. *Cognitive Science* 30:945–982

Wilson, Colin and Lisa Davidson. To appear. Bayesian analysis of non-native cluster production. In *Proceedings of NELS 40*, Cambridge, MA, MIT.

Yip, Moira. 2002. *Tone*. Cambridge: Cambridge University Press.