Grammars leak: Modeling how phonotactic generalizations interact within the grammar

Andrew Martin

Language, Volume 87, Number 4, December 2011, pp. 751-770 (Article)

Published by Linguistic Society of America
DOI: 10.1353/lan.2011.0096

GRAMMARS LEAK: MODELING HOW PHONOTACTIC GENERALIZATIONS INTERACT WITHIN THE GRAMMAR

ANDREW MARTIN

RIKEN Brain Science Institute

I present evidence from Navajo and English that weaker, gradient versions of morpheme-internal phonotactic constraints, such as the ban on geminate consonants in English, hold even across prosodic word boundaries. I argue that these lexical biases are the result of a MAXIMUM ENTROPY phonotactic learning algorithm that maximizes the probability of the learning data, but that also contains a smoothing term that penalizes complex grammars. When this learner attempts to construct a grammar in which some constraints are blind to morphological structure, it underpredicts the frequency of compounds that violate a morpheme-internal phonotactic. I further show how, over time, this learning bias could plausibly lead to the lexical biases seen in Navajo and English.*

Keywords: phonotactic learning, maximum entropy, morphology, language change, Navajo, English

1. INTRODUCTION. A popular explanation for language change attributes linguistic differences across generations to MISLEARNING, that is, failure on the part of one generation to learn exactly the same grammar as that used by the previous generation (e.g. Kiparsky 1968, Lightfoot 1979, 1991, 1999, Clark & Roberts 1993, Hale 1998, Blevins 2004, Hudson Kam & Newport 2005, 2009). In this article I apply this idea to explain, not diachronic change per se, but statistical biases present in the lexicons of Navajo and English. In these languages, compounds license violations of phonotactic constraints that hold within morphemes—for example, in Navajo, multiple sibilants within a root must agree in anteriority, but compounds that combine two roots with disagreeing sibilants are permitted. Likewise, in English, compounds may contain geminate consonants, which are illegal within morphemes. I show that in both languages, compounds that create violations of morpheme-internal generalizations, although legal, are statistically underrepresented.

I argue that these lexical biases are the result of a bias in the phonotactic learning mechanism. Learners of Navajo and English construct a probabilistic grammar on the basis of the input they are exposed to, but the absence of monomorphemes that violate a given phonotactic constraint causes them to acquire grammars that underpredict the frequency of compounds that violate the same constraint, in what might be described as a 'leakage'[1] of a phonotactic generalization from the tautomorphemic domain to the heteromorphemic. Such compounds effectively sound worse to learners than they should, given their frequency. When these same learners go on to form compounds that may in turn become part of the language, this bias will cause them to prefer compounds that obey the constraint, resulting eventually in a lexicon in which compounds that violate the constraint are underrepresented.

* This research was made possible in part by a UCLA Dissertation Year Fellowship, as well as funding from the Centre National de la Recherche Scientifique. I thank Bruce Hayes, Joe Pater, Sharon Peperkamp, Colin Wilson, and Kie Zuraw for much useful discussion. I also thank the anonymous referees for their helpful comments and suggestions.

[1] Although I allude to Sapir's (1921) famous comment in my choice of terminology (and title), I do not use the word 'leak' in the same sense that Sapir did—he was referring to the tendency for linguistic generalizations to have exceptions ('no language is tyrannically consistent'), while I use the word to mean a process by which a generalization in one domain influences a generalization in another domain.

My model of this mislearning process consists of two main components. The first is a distinction between two types of phonotactic constraint: STRUCTURE-SENSITIVE constraints, which take into account morphological structure, and STRUCTURE-BLIND constraints, which ignore morphological structure. Structure-sensitive constraints, which encode generalizations such as 'geminates are not permitted within morphemes', are necessary in order to correctly model the phonotactic differences between monomorphemes and compounds. Structure-blind constraints, by contrast, encode generalizations such as 'geminates are not permitted'. If such a constraint were included in a probabilistic grammar, it could model the relative rarity of geminates in English words overall. Such a grammar would, however, fail to capture the reason that geminates are rare, namely their restriction to a specific context. If in addition to structure-sensitive constraints, learners are also equipped with structure-blind constraints (possibly representing a holdover from very early phonotactic learning), the resulting grammar may be biased against compounds that violate stem-internal generalizations, even when there is no bias in the input data.

The conditions under which this leakage can occur are formally described by the second component of the model, a MAXIMUM ENTROPY (MaxEnt) learning algorithm that assigns probabilities to the space of possible words by constructing a grammar of weighted phonotactic constraints. The algorithm incorporates the idea, prevalent in the machine learning literature, that learning probabilistic generalizations involves a trade-off between accuracy and generality. One consequence of this trade-off is that under certain conditions learners sacrifice accuracy for generality, choosing a grammar that does not model the training data perfectly, but is more general than grammars that are more consistent with the data. Combined with the structure-blind constraints described above, this learning algorithm can account for the mislearning in Navajo and English.

In addition to the learning model, I also propose a model of how newly formed words compete with existing words to be used by speakers, and thereby become part of the language. Assuming that the phonotactic grammar constructed by the learning algorithm influences the creation or adoption of novel words, the model predicts that over several generations, the lexicon used by a speech community will acquire a bias against compounds violating a stem-internal phonotactic constraint, even if the initial lexicon exhibits no such bias. I also show that, assuming that compound formation is influenced by both semantic and phonological factors, the lexical underrepresentation will remain stable across generations—despite certain compound types being dispreferred by speakers, they never completely die out.

The remainder of the article can be broadly divided into three parts. First, I describe the Navajo and English data respectively. I then explicate the maximum entropy learning algorithm and show that, when given certain constraints, the algorithm consistently underpredicts the frequency of compounds that violate a tautomorphemic constraint. Finally, I use a simulation of multigenerational learning to demonstrate that the bias introduced by the learning model can result in a stable lexical pattern in which the dispreferred compounds are underrepresented.

2. NAVAJO.

2.1. NAVAJO SIBILANT HARMONY.

All sibilants in a Navajo root must agree in their specification for the [anterior] feature; thus, a single root can only contain sibilants that are either all anterior or all posterior (Sapir & Hoijer 1967, Kari 1976, McDonough 1991, 2003, Fountain 1998). The two sets of consonants are summarized in Table 1.

[+anterior]	[–anterior]
s	ʃ
z	ʒ
ts^{h}	tʃ^{h}
ts	tʃ
ts’	tʃ’

TABLE 1. Navajo sibilant classes.

Thus, for example, roots like /tʃ’oʒ/ ‘worm’ or /ts’ózí/ ‘slender’ are attested, but */soʃ/ is not a possible Navajo root.

This is not just a cooccurrence restriction on roots—sibilants in affixes must also agree in anteriority with sibilants in the root, resulting in alternations in sibilant-bearing affixes (Sapir & Hoijer 1967). The examples in 1 demonstrate the alternations in prefixed forms (sibilants are in bold).

(1) Examples of sibilant harmony (Fountain 1998)[2]
   a. /ji-**s**-lééʒ/ → [ji-**ʃ**-tɬéé**ʒ**] ‘it was painted’
   b. /ji-**s**-ti**z**/ → [ji-**s**-ti**z**] ‘it was spun’

Typically, assimilation proceeds from the root to the prefixes.

In compounds, however, which contain multiple roots, sibilant harmony does not necessarily apply, meaning that such words can contain disagreeing sibilants.[3]

(2) Exceptions to sibilant harmony in compounds (Young & Morgan 1987)
   a. **tʃ**éí- **ts’**iin ‘rib cage’
      heart bone
   b. **ts^{h}**é- **tʃ**éé**ʔ** ‘amber’
      stone resin

In the next section, I show that compounds in Navajo, although they may violate sibilant harmony, tend to combine roots whose sibilants already agree.

2.2. NAVAJO COMPOUNDS. The data described here is taken from the Young and Morgan (1987) dictionary of Navajo. From this dictionary a list of all compounds containing exactly two sibilants, each sibilant in a different root, was compiled, a total of 140 words—this represents all of the words that could potentially violate sibilant harmony. Because sibilant cooccurrence in compounds is sensitive to distance (Sapir & Hoijer 1967, Martin 2004), the data discussed here is limited to the subset of these words in which the sibilants are in adjacent syllables (there were no cases in which sibilants were in the same syllable, but different roots), a total of ninety-seven words. Representative examples are given in 3.

[2] Navajo examples are given in IPA, with acute accents marking high tones (low tones are unmarked).

[3] A handful of compounds do undergo sibilant harmony, such as ***ts**aa-nééz* ‘mule’, from /tʃaa/ ‘ear’ + /nééz/ ‘long’ (Sapir & Hoijer 1967). I suspect that these words undergo harmony because they have been stored as single units by speakers due to their semantic opacity, but I have included them in the analysis in their underlying (i.e. disagreeing) form, on the assumption that the sibilants disagreed when the compound was originally formed. This is the most conservative approach, since including these words in their harmonized forms would increase the sibilant harmony rate, strengthening the bias I describe in the remainder of the section.

(3) Examples of compounds with two sibilants in adjacent syllables (one per root)

a.	**ts^{h}**ee-	**ts'**iin		'tailbone'
	tail	bone		
b.	k'ii**ʃ**-	**ʒ**in-	ii	'blue beech'
	alder	black	one	
c.	**ts^{h}**é-	**z**éí		'gravel'
	rock	crumbs		

Of these ninety-seven words, twenty-nine (29.9%) contain disagreeing sibilants, violating the stem-internal phonotactic. In order to determine whether this agreement rate significantly differs from chance, I used a Monte Carlo procedure (Kessler 2001) to approximate the distribution of the expected rate. The procedure is described in detail in the following section.

2.3. THE MONTE CARLO TEST FOR SIGNIFICANCE. The Monte Carlo test is performed by randomly recombining the roots that make up the set of compounds in question. The initial root in each compound is combined with another root, pseudo-randomly selected from the same list of compounds (position in the compound is fixed, so that initial roots always remain initial, and final roots remain final). After each such shuffling, the number of disagreeing sibilant pairs under the new permutation is calculated, and the entire process may be repeated as many times as necessary. If the process is repeated sufficiently many times, the result will be a reliable estimate not only of the average expected number of disagreeing sibilants that would occur by chance, but also of the entire distribution of this expected value. With this information, we can determine how likely the actual, attested value is.

The histogram in Figure 1 presents the results of the Monte Carlo procedure on the entire list of Navajo compounds that contain exactly two sibilants in different roots. The *x*-axis represents the number of sibilant pairs (out of a total of ninety-seven) in which

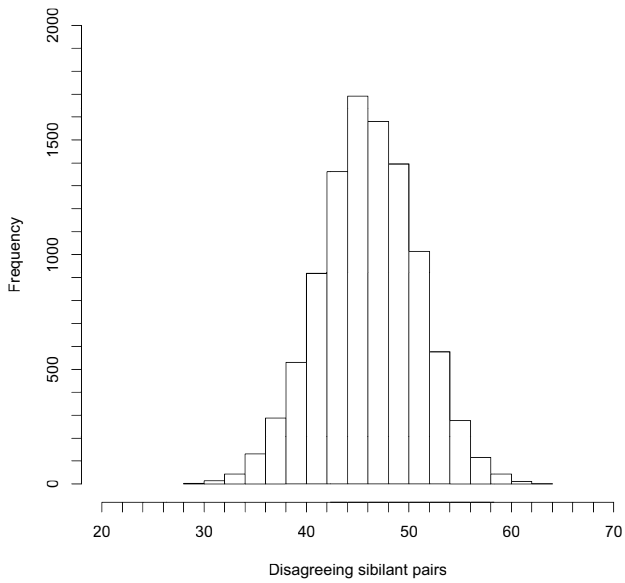

FIGURE 1. Results of Monte Carlo procedure on Navajo compounds.

the sibilants disagreed in anteriority, and the *y*-axis represents the number of iterations (out of 10,000 total)[4] in which a given disagreement rate occurred.

The histogram shows that the values generated in the Monte Carlo test are approximately normally distributed around the mean value of 46.0. The actual number of disagreeing sibilants in the Young and Morgan data, twenty-nine pairs, is extremely unlikely to have arisen by chance—a value this low occurred only three times out of the 10,000 iterations of the Monte Carlo test. From this we can conclude that the actual disagreement rate is significantly ($p < 0.001$) below chance.

In the remainder of this article, Monte Carlo results are summarized by omitting the histogram and simply reporting the 95% confidence interval and the actual value, as in the chart in Figure 2 (which represents the same test reported in Fig. 1). In this and succeeding charts, the triangle indicates the actual value found in the data (in this case, the actual number of Navajo compounds with disagreeing sibilants), while the horizontal bar represents the 95% confidence interval derived from the Monte Carlo test.

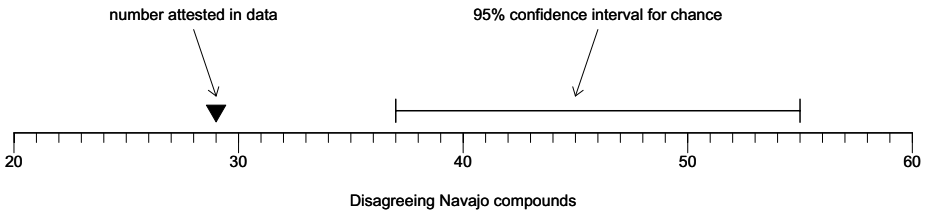

FIGURE 2. Comparing attested Navajo sibilant pairs to Monte Carlo results.

The results of this test show that Navajo compounds tend to obey the sibilant harmony constraint, even though violations of the constraint are permitted, and that this tendency is unlikely to be due to chance. In the next section I present a parallel case from English.

3. ENGLISH. Geminate consonants in English are permitted only across morpheme boundaries (Hammond 1999, Ladefoged 2001, Kaye 2005). Words like *unknown*, *solely*, and *bookcase* are typically pronounced with geminates that have been created by combining morphemes that end and begin with the same consonant. These morphologically created geminates are often called 'fake geminates' (Hayes 1986) to differentiate them from morpheme-internal long consonants—the two types of geminate frequently exhibit different phonological behavior (Payne 2005, Ridouane 2007, Oh & Redford 2009, Pycha 2010). Minimal pairs differing only in consonant length, as in the compounds *carpool* and *carp pool*, may be found in multimorphemic words; in monomorphemic words, however, no such minimal pairs exist—the hypothetical word [hæppi], which would form a minimal pair with existing *happy* [hæpi], is not a possible monomorpheme of English. In the following sections, I show that geminate consonants created by compounding are statistically underrepresented in the lexicon of English.

In order to determine the number of compounds in English that contain geminates, I extracted all of the words marked as noun-noun compounds from the lemmatized version of the CELEX database (Baayen et al. 1993), a total of 4,758 words. Of these, 141 words (3.0%) contain fake geminates—for example, *bus stop*, *hat trick*, *penknife*, *bookkeeper*. The results of a Monte Carlo test on the CELEX compounds are shown in Figure 3.

[4] All of the Monte Carlo tests reported in this article were performed using 10,000 iterations each.

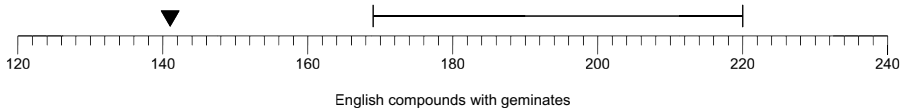

FIGURE 3. Geminates are underrepresented in English compounds.

As the chart makes clear, the number of geminates found in the actual compounds, 141 (3.0%), is significantly lower than expected ($p < 0.001$).

Before accepting that this result tells us something about the compound-formation behavior of English speakers, however, a potential confound must be dealt with. The compounds listed in CELEX were collected by applying an automatic parser to a large text corpus. This raises the possibility that some compounds spelled as separate words (e.g. *sand dune*), might have been misidentified by the parser as separate words. This could bias the results of the Monte Carlo test because of the fact that compounds with geminates are more likely to be spelled with a hyphen or space between the members than as a single word (Sepp 2006). The underrepresentation of geminates could thus be an artifact of the parsing process, combined with people's tendency to spell compounds according to their junctural phonotactics. The effects of the spelling bias are shown in Figure 4, which makes it clear that the underrepresentation is limited to those compounds that are spelled as a single word (this chart depicts the same set of words from CELEX described in Fig. 3, divided according to how they are spelled in the CELEX entry).[5]

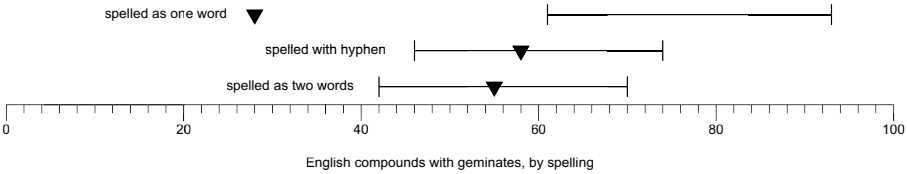

FIGURE 4. Compound spelling is biased by presence of geminate.

To show that the geminate underrepresentation in English is not solely an artifact of the spelling bias, I ran the same Monte Carlo test on the list of compounds compiled by Sepp (2006) from a fourteen-million-word corpus of written American English (see Sepp 2006 for details of the construction of this corpus). Sepp used a part-of-speech tagger and computational parser to extract all potential noun-noun compounds (including any sequence of nouns separated by a space), and then further filtered the list by hand, removing all noncompounds. Because every compound was checked by hand, the likelihood of undercounting compounds spelled with a space is lower than it would be if all parsing were done by algorithm.

Of the 3,222 noun-noun compounds that occur in Sepp's corpus (including both those spelled as single and as separate words), 118 (3.6%) contain false geminates. The

[5] Although the same compound can be spelled different ways by different writers, each compound is listed with a single spelling in CELEX. It is unclear how this spelling was determined. My intuition is that nearly all of the words spelled with hyphens in CELEX would be most often spelled with a space by native speakers (e.g. *space-vehicle*, *rabbit-hutch*, *slot-machine*), a suspicion that is strengthened by the nearly indistinguishable behavior of hyphenated and spaced compounds in Fig. 4. This accords with Sepp's findings that fewer than 5% of the noun-noun compounds in her corpus are spelled with a hyphen more often than either with a space or as one word (many of those are either dvandva compounds (*Clinton-Gore*, *hip-hop*), or involve abbreviations (*op-ed*)).

results of a Monte Carlo test on these compounds, shown in Figure 5, demonstrate that, just as with the CELEX compounds, the actual number of geminates is significantly lower ($p < 0.01$) than the mean expected number of 152.1 (4.7%).

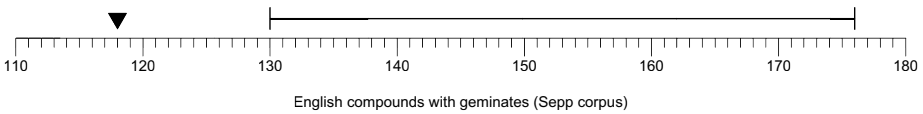

FIGURE 5. Geminates are underrepresented in compounds in Sepp corpus.

Thus, even when the risk of a counting bias is minimized by careful hand-checking, geminates are still underrepresented in compounds overall. This suggests that any orthographic bias that may exist is in addition to a general bias against forming compounds that create geminates.

4. ANALYSIS. The phonotactics in English and Navajo discussed above obey the same generalization: some phonotactic constraint holds within morphemes, and a weaker version of the same constraint holds across morpheme boundaries. Morpheme boundaries, in other words, license violations of phonotactic constraints, but only up to a point. Tautomorphemic and heteromorphemic phonotactic generalizations are thus in some sense entangled rather than computed independently. This property of the phonotactic grammar could simply be stipulated to be part of universal grammar; I argue, however, that this entanglement follows from more general constraints on the human phonotactic learning mechanism. More specifically, I describe a learning algorithm that, when trained on data that exhibits a tautomorphemic phonotactic restriction, cannot help but learn a grammar that encodes a weaker version of the same restriction across morpheme boundaries.

The account that I present of the facts in Navajo and English is thus a theory of linguistic competence. Another possible account of the same facts might appeal instead to universal performance factors. On this view, the tautomorphemic and heteromorphemic phonotactics could both result from the same phonetic pressure. Imagine, for example, that sequences of agreeing sibilants are easier to articulate (or process) than disagreeing sequences. This fact of human physiology could have become phonologized within Navajo roots, resulting in a categorical sibilant harmony constraint, and could also exert a pressure on compound formation, resulting in the underrepresentation of disagreeing compounds. Under this hypothesis, the tautomorphemic constraint does not cause the heteromorphemic constraint; rather, they are both caused by the same underlying factor. A learning bias would thus not need to be invoked to explain the Navajo facts.

Such a performance-only theory, however, runs into empirical problems regarding the universality of these performance factors. If it is true that disagreeing sibilant sequences are universally more difficult than agreeing sequences, and this difficulty exerts a pressure on the contents of the lexicon, then it would be surprising to find many languages in which disagreeing sibilants are overrepresented. English, however, is such a language: /s..ʃ/ sequences are overattested compared to /s..s/ sequences (Berkley 1994, 2000). Gradient similarity AVOIDANCE in consonants, in fact, appears to be a robust crosslinguistic phenomenon and has been shown in a number of languages: Arabic (Frisch et al. 2004), Maltese (Frisch et al. 2004), Italian (Frisch et al. 2004), Muna (Coetzee & Pater 2008), Japanese (Kawahara et al. 2005), and Russian (Padgett 1995). In the most complete study of the phenomenon to date, Pozdniakov and Segerer (2007: 308) find evidence for gradient similarity avoidance in thirty-one different languages,

from which they conclude that it 'is a likely universal property of human language'. If anything, the evidence for universal performance factors appears to point in the opposite direction of the Navajo pattern.

Likewise, the preference for consonant clusters over geminates seen in English is not universal. In Japanese, for example, geminate consonants are permitted within morphemes, while nongeminate clusters (with the exception of homorganic nasal + stop clusters) are not (e.g. /happa/ 'leaf', but */hapta/). Furthermore, nongeminate clusters are repaired when created by the morphology (e.g. /kak/ 'write' + /ta/ 'PAST' → [kaita] 'wrote'; cf. /kat/ 'win' + /ta/ 'PAST' → [katta] 'won'). Luganda is similar to Japanese in that it allows geminates but prohibits other cluster types (Wiltshire 1999). In some languages, like English, geminates are the worst type of cluster; in others, like Japanese, they are the best.

This is not to say that functional, performance-based factors play no role in these phonotactic generalizations. Rather, it is likely that there are multiple functional factors, some of which are at odds with others. Agreeing sibilant pairs and disagreeing pairs may each be considered 'better' along different dimensions, and each language represents a separate compromise among these competing forces. My claim is simply that the correlations between tautomorphemic and heteromorphemic phonotactics in Navajo and English are unlikely to be the result of performance ALONE—the speakers of these languages must learn the language-particular phonotactic generalizations present in morphemes, which they then extend to the formation of complex words. In the remainder of this section I propose a mechanism by which this overgeneralization can occur.

4.1. THE PHONOTACTIC LEARNER. In order to model the gradient constraints observed in the compound data presented above, I assume that speakers make use of a grammar that assigns probabilities to possible words (Hayes & Wilson 2008). On this view, the actual lexicon can be thought of as a finite sample drawn from this probability distribution—the learner's task is to reconstruct the grammar, and therefore the distribution, based on that sample. In the case of English, which I use throughout the rest of the article to illustrate how the learner works, the final grammar should assign very low probabilities to words containing stem-internal geminates, high probabilities to words containing only legal clusters, and intermediate probabilities to words containing geminates across morpheme boundaries.

It would be trivial to construct a probabilistic learner that, fed the lexicon of English, could learn that compounds with geminates are underrepresented. Any algorithm that is capable of simply counting the number of compounds with and without geminates could succeed at this task. My goal, however, is not to produce a learner that correctly learns the patterns in Navajo and English, but to explain why Navajo and English are the way they are. My strategy is therefore to describe a learner that systematically mislearns generalizations across morpheme boundaries—when given input in which the two compound types are equally frequent, it will nonetheless construct a grammar that assigns a lower well-formedness value to compounds that violate the stem-internal phonotactic.

4.2. MAXIMUM ENTROPY GRAMMARS. The maximum entropy formalism has long been a staple of the machine learning literature (Jaynes 1957, Berger et al. 1996, Abney 1997, Della Pietra et al. 1997) and has recently been successfully applied to problems of phonological learning (Goldwater & Johnson 2003, Fischer 2005, Wilson 2006, Jäger 2007, Hayes & Wilson 2008, Hayes et al. 2009). A MaxEnt learning algorithm learns a probability distribution over the members of some set given a sample drawn from that distribution.

A MaxEnt grammar consists of a set of numerically weighted constraints. The constraints that are used in this article ban structures in the output (e.g. 'no geminates within a morpheme') and are equivalent to the markedness constraints used in optimality theory (Prince & Smolensky 2004). Unlike classical optimality theory, in which constraints are strictly ranked, each constraint in a MaxEnt grammar has a weight, represented by a real number, which represents the 'strength' of that constraint. The set of constraints and their weights (which together constitute the grammar) determine a probability for every possible surface form, which is a function of the set of constraints violated by the form and their weights. Specifically, a word's probability is a function of what Hayes and Wilson (2008) call its SCORE, $h(x)$, which is calculated by simply summing the (weight × number of violations) for every constraint in the grammar, as shown in 4.

(4) Definition of score

$$h(x) = \sum_{i=1}^{M} w_i C_i(x)$$

(M = number of constraints; $w_1, w_2, \ldots, w_M$ = constraint weights; x = representation of candidate; $C_i(x)$ = number of violations assigned to x by constraint C_i)

For a grammar with three constraints, for example, C_1 (weight 1.0), C_2 (weight 2.0), and C_3 (weight 3.0), an output form x violating C_1 twice and C_3 once would be assigned a score of $h(x) = (1.0 \times 2) + (2.0 \times 0) + (3.0 \times 1) = 2.0 + 0 + 3.0 = 5.0$.

A word's score is identical to its HARMONY in the HARMONIC GRAMMAR framework (Legendre et al. 1990, Smolensky & Legendre 2006); in a MaxEnt grammar a word's probability is directly related to its score. The equation in 5 describes how scores are mapped to probabilities (Ω represents the set of all possible words).[6]

(5) Determining candidate probability

$$P(x) = \frac{e^{-h(x)}}{\sum_{y \in \Omega} e^{-h(y)}}$$

The goal of the learning algorithm is to reproduce the probability distribution over constraint violations in the learning data. It does this by adjusting the constraint weights so as to maximize the probability of the data—the algorithm thus represents an example of MAXIMUM LIKELIHOOD learning. The probability of the data is calculated by simply multiplying the probabilities of all of the words in the data to arrive at their joint probability, equivalently stated as the sum of the log probabilities of each word, or $\sum_{i=1}^{N} \log P(x_i)$, where N is the number of words encountered during learning.

An algorithm that simply maximizes the probability of the data, however, is prone to OVERFITTING. Because the learner is only given a finite sample of data, a pure maximum likelihood learner will tend to overestimate the probability of items that are in the sample, and underestimate the probability of items that did not happen to occur in the sample. In other words, a probability distribution learned from a finite sample will tend to be skewed in the direction of the observed data.

The standard way to avoid overfitting is to introduce a SMOOTHING TERM into the learning function (Martin et al. 1999). The smoothing term penalizes skewed distribu-

[6] Because the number of possible words (absent any maximum length) is typically infinite in real languages, calculating the probability of any word requires summing the probabilities of an infinite set. The stochastic gradient ascent algorithm used in the simulations in this article avoids this problem by estimating the infinite sum using sample words drawn from the probability distribution defined by the current set of constraint weights (Jäger 2007).

tions and causes the learner to favor more uniform distributions, which ameliorates the tendency to overfit. I use a Gaussian prior over the constraint weights, which prefers that the constraint weights be as uniform as possible. The prior term is subtracted from the likelihood term, resulting in the learning function in 6.

(6) MaxEnt learning function

$$\sum_{i=1}^{N} \log P(x_i) - \sum_{j=1}^{M} \frac{(w_j - \mu_j)^2}{2\sigma_j^2}$$

The Gaussian prior assesses a penalty for constraint weights that deviate from their ideal weights, represented by μ_j. In the implementation of the algorithm I use, μ is set to zero for all constraints, so that the prior penalizes any nonzero weight, with the size of the penalty increasing with the square of the weight. This pressure toward low constraint weights translates into a bias against highly skewed distributions—because the prior term increases with the square of each constraint weight, it prefers grammars with many low-weighted constraints over grammars with a few high-weighted constraints. This means that if multiple constraints are each capable of explaining a given property of the data, the learner will assign all of the constraints low weights rather than choose one and assign it a high weight. This property of the prior will prove crucial in modeling the English and Navajo data.

The learning function thus embodies a trade-off between a pressure to model the data as accurately as possible and a pressure to have as general (i.e. uniform) a grammar as possible. The value of the free parameter σ determines the relative importance of each of these factors. Modeling the connection between tautomorphemic and heteromorphemic phonotactics will rely crucially on this trade-off.

The learning simulations reported in this article utilize the stochastic gradient ascent (SGA) algorithm to assign weights to these constraints.[7] Specifically, I use Jäger's (2007) implementation of the SGA for constraint-based grammars, which estimates (to an arbitrary degree of precision) the maximum of the MaxEnt learning function given in 6. The SGA works as follows: first, the constraints are assigned arbitrary weights (in my simulations, each constraint starts with a weight of zero). Then, input is fed to the algorithm one word at a time. It compares each input word x to a sample word y randomly chosen using the probability distribution determined by the current set of constraint weights. On the basis of this comparison, the current weight for each constraint w_i in the grammar is changed according to the update rule in 7, where β represents the PLASTICITY, a parameter that determines the degree to which constraint weights are perturbed with each incoming learning datum.

(7) SGA update rule

$$w_i = w_i + \beta \cdot (w_i C_i(x) - w_i C_i(y))$$

The update rule serves to change the predicted probabilities in the direction of more closely matching the probabilities observed in the input. The Gaussian prior is implemented in the SGA by decreasing every constraint weight w by an amount $2\alpha w$ after every learning datum (Johnson 2007).[8]

[7] Note that the learning bias I discuss in this article is a property of the learning function itself, and is independent of which algorithm learners actually use to calculate the maximum of that function; the same results can be obtained with the conjugate gradient algorithm used by Hayes and Wilson (2008), for example.

[8] In all of the learning simulations presented here, the data was presented to the learner three times. The plasticity was set to 0.1 for the first presentation, 0.01 for the second, and 0.001 for the third. The Gaussian prior parameter α was set to 0.01 for the first presentation, 0.001 for the second, and 0.0001 for the third. The results reported are averaged over 100 consecutive runs of the algorithm.

4.3. TESTING THE MAXENT LEARNER ON SIMPLIFIED ENGLISH. Let us consider how the learner described in the previous section would handle a schematic representation of the English facts. Imagine a simplified version of English with two salient features: every word is either a monomorpheme or compound, and every word contains exactly one consonant cluster. Each cluster may be nongeminate, which I represent as [tp], or geminate, represented as [pp]. A morpheme boundary intervening between the consonants of a cluster is indicated by '+'. This gives us a set of four codes, listed in 8, with which we can label all of the logically possible words in this simplified English.

(8) Logically possible word types in simplified English
[tp] [pp]
[t+p] [p+p]

Because geminates are illegal within morphemes in English, only three of these word types are attested: [tp], [t+p], and [p+p]. The training data I use for all of the demonstrations of the learner is composed as in Table 2. The ratios of each word type were chosen so that there would be an equal number of monomorphemes and compounds, and an equal number of compounds with geminates and compounds without geminates.

WORD TYPE	RATIO OF WORDS
[tp]	50%
[t+p]	25%
[p+p]	25%

TABLE 2. Training data.

The generalizations formed by a MaxEnt learner given this data will of course depend on the constraints it is given. Let us first consider two possible types of constraint, one structure-blind and the other structure-sensitive, described in Table 3. Note that a plus sign in parentheses indicates an optional morpheme boundary, while the absence of a plus sign (as in *PP) indicates that no morpheme boundary intervenes between the consonants.

STRUCTURE-BLIND CONSTRAINTS	
*P(+)P:	no geminates
*T(+)P:	no nongeminate consonant clusters
STRUCTURE-SENSITIVE CONSTRAINTS	
*PP:	no geminates within a morpheme
*TP:	no nongeminate consonant clusters within a morpheme
*P+P:	no geminates across a morpheme boundary
*T+P:	no nongeminate clusters across a morpheme boundary

TABLE 3. Constraint types.

I propose that human learners incorporate both types of constraints, and that the interaction between them is responsible for the lexical biases in English and Navajo. To understand how this works, it is helpful to first briefly consider how a learner would fare when equipped with only one of the constraint types.

A learner given only structure-sensitive constraints, for example, would be able to capture two generalizations present in the training data—first, that geminates are only legal across compound boundaries, and second, that geminates and nongeminate clusters are equally frequent in compounds. Table 4 lists the weights given to the structure-sensitive constraint set by a learner that is fed the data in Table 2, and the consequent predicted probabilities of the four logically possible words types.

CONSTRAINT	WEIGHT	WORD TYPE	SCORE[9]	PREDICTED PROBABILITY	PROBABILITY IN TRAINING DATA
*PP	4.440	[pp]	4.440	0.004	0.000
*TP	–0.403	[tp]	–0.403	0.480	0.500
*P+P	0.220	[p+p]	0.220	0.258	0.250
*T+P	0.219	[t+p]	0.219	0.258	0.250

TABLE 4. Learning results, structure-sensitive constraints only.

The weight given to *PP in this grammar, 4.440, is much higher than the weights given to other constraints, reflecting the absence of [pp] in the training data. The constraints *P+P and *T+P are given nearly identical weights, meaning that geminates and nongeminates across morpheme boundaries are equally well-formed. Such a grammar accurately models the fact that English speakers find morpheme-internal geminates to be ill-formed, but offers no explanation of the gradient bias against geminates in compounds.

A learner equipped only with structure-blind constraints, however, has the reverse problem, as shown in Table 5.

CONSTRAINT	WEIGHT	WORD TYPE	SCORE	PREDICTED PROBABILITY	PROBABILITY IN TRAINING DATA
*P(+)P	0.496	[pp]	0.496	0.135	0.000
*T(+)P	–0.496	[tp]	–0.496	0.365	0.500
		[p+p]	0.496	0.135	0.250
		[t+p]	–0.496	0.365	0.250

TABLE 5. Learning results, structure-blind constraints only.

This grammar predicts that, in general, words with geminates are less well-formed than words without, but because it is blind to morphological structure, it cannot encode the generalization that geminates are prohibited within morphemes and allowed across morpheme boundaries. Essentially, all this learner knows is that 25% of the clusters occurring in the data are geminates, and 75% are nongeminates.

The structure-blind learner is thus too myopic to serve as a model of actual human language learners. The structure-sensitive learner, by contrast, is in a sense too accurate, in that it does not display a bias that would explain the lexical data in Navajo and English. This bias appears only when the learner uses both types of constraints. The grammar in Table 6 is the result of giving the SGA learning algorithm structure-blind and structure-sensitive constraints, and exposing it to the same data as the previous two learners.

CONSTRAINT	WEIGHT	WORD TYPE	SCORE	PREDICTED PROBABILITY	PROBABILITY IN TRAINING DATA
*P(+)P	0.146	[pp]	4.357	0.004	0.000
*T(+)P	–0.146	[tp]	–0.469	0.483	0.500
*PP	4.211	[p+p]	0.184	0.251	0.250
*TP	–0.323	[t+p]	0.146	0.261	0.250
*P+P	0.038				
*T+P	0.292				

TABLE 6. Learning results, all constraints.

The algorithm assigns a high weight to *PP (4.211), although not as high as the weight assigned to the same constraint by the learner using only structure-sensitive constraints

[9] Remember that higher scores are associated with lower probabilities, by the equation in 5.

(4.440). Furthermore, despite their equal frequencies in the training data, [p+p] compounds are penalized by the grammar more than [t+p] compounds, as can be seen by comparing the scores of the two word types. This is the effect of the Gaussian prior, which is optimized by making the distribution of weights as uniform as possible. Increasing the weight on *P(+)P lowers the probability of [pp] sequences, which allows the weight on *PP to be lower. The price of this more uniform distribution is accuracy in modeling the data; the weight on *P(+)P also slightly lowers the predicted probability of [p+p] sequences, making them appear to the learner to be slightly less frequent than [t+p] sequences.

This shows that a constraint's weight is dependent not just on the properties of the data, but also on the other constraints that are present in the grammar. In this case, *PP gets a higher weight when there is no other constraint that could also explain the absence of [pp] in the data. When the structure-blind constraints are included in the grammar, this generalization is split between two constraints—*PP and *P(+)P—which allows the weight on *PP to be slightly lower. The result is a grammar that evaluates compounds with geminates as more probable than those without geminates, despite the equal probability of the two types of word in the input. The Gaussian prior, in effect, allows probability mass to 'leak' from one type of generalization to another. Although this prior is well motivated on mathematical grounds—without it, the learner will assign infinite weights to constraints that are never violated in the data—it can cause weights to be assigned to constraints that are not strictly necessary to explain the data.

4.4. DISCUSSION. Speakers of English know that geminates are legal across morpheme boundaries but not within morphemes, but they have also encoded in their grammars the fact that geminates are not as frequent as nongeminate clusters. When all of these generalizations are combined in a single grammar, the result is a strong preference for nongeminate clusters within morphemes, and a mild preference for nongeminate clusters across morpheme boundaries.

I have shown that the above scenario can be modeled with a maximum entropy learning algorithm equipped with a Gaussian prior and both structure-sensitive and structure-blind constraints. I have yet to answer a crucial question, however. Why would learners use structure-blind constraints at all, given that structure-sensitive constraints by themselves permit more accurate modeling of the input?

One possibility is that structure-blind constraints represent a holdover from an early stage of phonotactic learning. There is substantial evidence that infants learn a great deal about the phonotactic patterns in their language before they are able to parse the speech stream into morphemes or even words (Peters 1983, Jusczyk 1997). Generalizations formed at this stage are by necessity structure-blind. Of course, once they master morphology, children are presumably able to make use of structure-sensitive constraints. It is possible, though, that the structure-blind constraints they used at the earlier stage remain in the grammar into adulthood (one could imagine, for example, that there is some cost to removing constraints once they are part of the grammar).

Another possibility is that the mechanism responsible for positing constraints is biased toward more general constraints (see Hayes & Wilson 2008 for a proposal in this vein). A constraint that simply bans geminates is more general than a constraint that bans geminates only across morpheme boundaries, and so might be included in the grammar on that basis. Of course, answering this question more definitively would require a more thorough understanding of how language-learning infants go about constructing phonological constraints, a topic on which little research has yet been done.

5. THE EVOLUTION OF THE LEXICON. I have argued that learners consistently underpredict the frequency of compounds that violate a stem-internal phonotactic in their language, resulting in a grammar that assigns such compounds a lower well-formedness value than other compounds. In order to explain why such compounds are underrepresented in Navajo and English, we must further assume that speakers are biased by their phonotactic grammar when they create new compounds (or decide to use novel compounds coined by others). Even if earlier versions of these languages had been unbiased, over time generations of learners would have altered the lexicon of each language, making words that violate stem-internal phonotactics less frequent. However, this model raises another problem: if each generation ends up making the language more biased than the previous generation, given enough time we would expect the underrepresented words to die out completely. Why does English allow geminate consonants in compounds at all, given the rapid turnover in vocabulary typically observed in languages over time?[10]

In order to answer this question, I present the results of a simulation of multigenerational lexical change that incorporates the learning algorithm described in §4. I assume, following Boersma 2007, that lexical change is driven largely by competitions between synonymous lexical items. New words are created or borrowed by speakers, and must compete with existing words that have the same meaning. The tendency of speakers within a speech community to converge on a single way to express a given concept (Lass 1997, Croft 2000, Baronchelli et al. 2006) creates a selection pressure—words that are better at winning these competitions will come to dominate the lexicon (Martin 2007).

The simulation is structured as follows. The speech community is represented by a single agent, who possesses a lexicon and a grammar. Each 'generation' of the simulation is divided into two phases: in the first phase, the agent copies the lexicon of the previous generation's agent, except for the first agent, who begins with an unbiased lexicon containing 2,000 words of type [tp], and 1,000 words each of types [p+p] and [t+p]. Each agent then uses a MaxEnt learning algorithm equipped with the structure-blind and structure-sensitive constraints in Table 3 to learn a grammar using the lexicon as input. In the second phase, the agent is given the option of replacing some of its lexical items with newly generated compounds, with each novel word competing with an existing word. Once these competitions are resolved, the agent's updated lexicon is used as the input for the next generation's agent.

The lexicon-updating phase takes place in two stages. First, a number of potential compounds are generated by the morphology. Throughout the simulation, compounds with and without geminates are equally likely, representing the combining of existing stems based on the semantic needs of speakers. Next, each potential compound is randomly paired with an existing compound in the lexicon, representing a word synonymous with the novel word, and the two words compete (note that the simulation does not model competitions between monomorphemes and other words). The probability that a word will win this competition ($p_{win}(x)$) is proportional to its phonotactic probability ($p(x)$), as shown in 8.

[10] As an example, roughly 85% of the Old English vocabulary is no longer in use (Baugh & Cable 1993), and more than 80% of the Modern English vocabulary consists of words borrowed from other languages (Stockwell & Minkova 2001).

(8) Probability of x winning competition with y

$$P_{win}(x) = \frac{p(x)}{p(x) + p(y)}$$

If a novel word wins a competition, it replaces the existing word; otherwise, the novel word is discarded.

I ran two versions of this simulation for 1,000 generations each. As noted above, the initial agent begins with a lexicon containing 4,000 words. During each generation, 200 new compounds are generated and allowed to compete with 200 randomly chosen existing compounds. In one version of the simulation, the learner is given structure-sensitive and structure-blind constraints, and in the other version the learner is given only structure-sensitive constraints. The resulting frequencies of compounds with geminates for both versions are shown in Figure 6.[11]

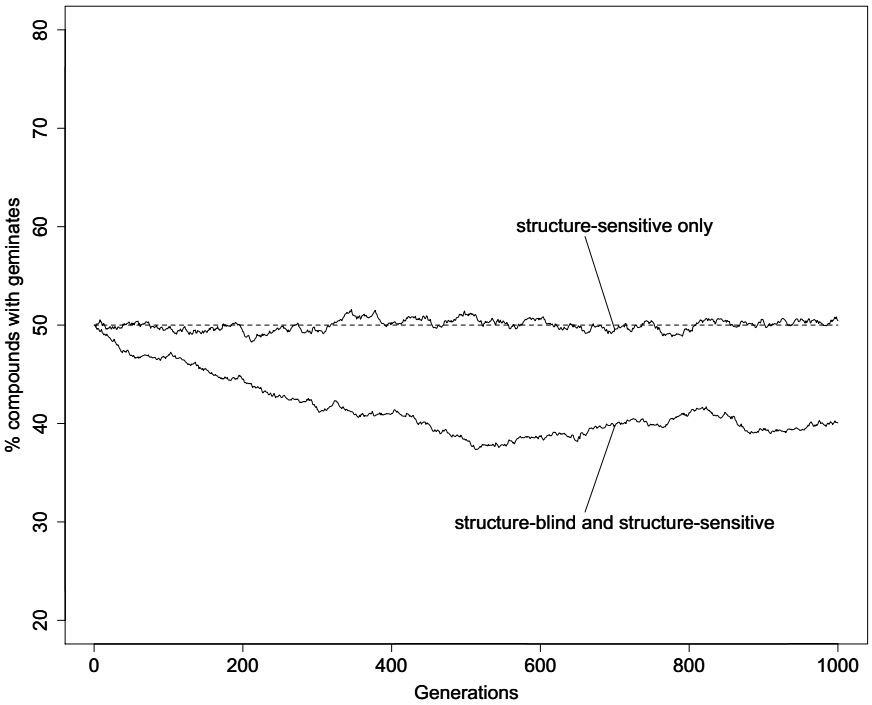

FIGURE 6. Multigenerational learning simulation results.

The graph in Fig. 6 shows that without the structure-blind constraints, the frequency of geminate compounds randomly varies around 50%, the expected value (shown on the graph with a dashed line) given the initial lexicon and the distribution from which replacement words are drawn. When the structure-blind constraints are added, the relative frequency of geminate compounds drops below 50%, but then levels off. With the struc-

[11] The values depicted in Fig. 6 are averaged over ten consecutive runs of each simulation. In order to check that the lexical bias remains stable beyond 1,000 generations, I also performed one run of the simulation for 10,000 generations. With both structure-sensitive and structure-blind constraints, the percentage of compounds with geminates never dropped to zero—between generations 1,000 and 10,000, the mean was 38.5%, the highest occurring percentage was 49.3%, and the lowest 26.8%.

ture-blind constraints, compounds with geminates are consistently underrepresented, but are never completely eliminated from the lexicon.

Intuitively, the stabilization in the lexicon can be understood as resulting from the interaction of three factors, which together determine how the lexical statistics change from generation to generation: the phonotactic well-formedness of each word type, the current frequency of each word type, and the probability distribution from which novel words are drawn. The first of these, the well-formedness of each word type, determines the chances of that type winning a competition with another type. As the [p+p] compounds become rarer, their well-formedness drops, making it harder for them to win subsequent competitions, in a 'poor-get-poorer' feedback loop. Although this would seem to doom these compounds to eventual extinction, low frequency also carries an advantage—less frequent types are correspondingly less likely to be faced with a competing word. In short, as the frequency of any word type decreases, its probability of winning competitions drops, but so does its probability of being forced to compete in the first place.[12] This is why, when structure-blind constraints are added in Fig. 6, [p+p] compounds initially drop rapidly in frequency, and then gradually level off, before eventually reaching an equilibrium.

The other crucial parameter that influences the survival of marked compound types is the probability distribution from which new words are drawn. In the simulations reported above, this is a fixed distribution in which each compound type is equally likely. No matter how rare geminates become in the actual lexicon, new compounds contain geminates half of the time. If this assumption is altered, and novel words are chosen from a distribution reflecting the current lexical frequencies, then stability collapses, and compounds with geminates will eventually be completely eliminated. In other words, if compounding is driven EXCLUSIVELY by phonological factors, then it is indeed a mystery how marked structures survive. But surely this is not the case—morphological operations are at least partly motivated by the semantic needs of speakers, which are presumably blind to phonological considerations.

Thus, the current state of the lexicons of Navajo and English can be seen as a balance between two forces: semantic preferences for certain combinations of morphemes, and phonotactic preferences for certain combinations of sounds. The first drives the English lexicon toward the expected number of geminates, while the second drives the lexicon toward a state with no geminates. The result is a compromise, in which geminates are allowed, but occur at less than the expected rate. Although different languages (or the same language at different times) may enact this compromise to differing degrees, the model predicts that a language in which compounds that violate a categorical stem-internal phonotactic constraint are consistently OVERrepresented would be historically unstable.

6. CONCLUSION.

6.1. SUMMARY. I have argued that a bias in the human phonotactic learner is responsible for a correlation between tautomorphemic and heteromorphemic phonotactics in Navajo and English. In my model this bias results from two factors: a set of structure-blind phonotactic constraints that ignore morphological structure, and a maximum entropy learning algorithm equipped with a smoothing term that penalizes high constraint weights. I have also shown how this learning bias interacts with the creation and selection of new words, resulting in a persistent lexical bias. The biased lexicons in Navajo and English represent a compromise between the needs of the morphology and the needs of the phonology.

[12] I am grateful to Kathryn Pruitt (p.c.) for this insight.

Of course, this is not the only possible analysis of these facts. Hay and Baayen (2005), for example, argue that the distinction between monomorphemes and compounds in the lexicon is itself gradient, a function of the frequencies with which the compound and its component members are used independently. They point out that morphological status along this continuum is influenced by a word's junctural phonotactics, such that a compound containing a geminate is less likely to be lexicalized as a single unit (evidenced by, among other things, the compound acquiring noncompositional semantics). Indeed, the fact that phonotactics bias how English compounds are spelled suggests that something like this is true. This could explain the biases in Navajo and English, as lexicalized compounds would be more likely to appear in a corpus or dictionary.

Unfortunately, the lexical count data presented in this article cannot distinguish between such an account and the phonotactic learning account I have argued for. The two theories could be distinguished, however, in principle. For example, a theory that claims that lexical biases are solely due to lexicalization predicts that a bias against phonotactic violations in compounds emerges only gradually over time, as the lexicalization process applies differentially to words with and without such violations. Thus, evidence that speakers are biased against illegal structures even at the moment of creating a new compound would be evidence against the lexicalization hypothesis. Testing this prediction experimentally is a promising avenue for future research.

6.2. FUTURE DIRECTIONS. The model of lexical change I present here is general in character and could potentially be applied to a wide range of cases in order to determine how these forces interact in a wider range of languages and phenomena. This research would contribute to our understanding of how phonotactic knowledge participates in the shaping of the lexicon, which in turn forms the basis of the next generation's phonotactic knowledge.

Although the cases discussed here involve an interaction between generalizations at different levels of morphological complexity, the learning model also predicts other types of interaction. In a maximum entropy grammar, generalizations stated over constraints that refer to overlapping categories will be interconnected—put simply, when a given structure is underrepresented, similar structures are more likely to also be underrepresented. The model thus makes a very rich set of predictions about not just the generalizations that should be attested in natural languages, but also relations among the generalizations within a single language. Testing these predictions promises to lead to a deeper understanding of the nature of phonotactic learning.

REFERENCES

ABNEY, STEVEN. 1997. Stochastic attribute-value grammars. *Computational Linguistics* 23.597–618.

BAAYEN, R. HARALD; RICHARD PIEPENBROCK; and HEDDERIK VAN RIJN. 1993. The CELEX lexical database. (CD-ROM) Philadelphia: Linguistics Data Consortium, University of Pennsylvania.

BARONCHELLI, ANDREA; MADDALENA FELICI; VITTORIO LORETO; EMANUELE CAGLIOTI; and LUC STEELS. 2006. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment* P06014.

BAUGH, ALBERT, and THOMAS CABLE. 1993. *A history of the English language*. 4th edn. London: Routledge.

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.

BERKLEY, DEBORAH. 1994. The OCP and gradient data. *Studies in the Linguistic Sciences* 24.59–72.

BERKLEY, DEBORAH. 2000. *Gradient OCP effects*. Evanston, IL: Northwestern University dissertation.
BLEVINS, JULIETTE. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.
BOERSMA, PAUL. 2007. The evolution of phonotactic distributions in the lexicon. Talk given at the Workshop on Variation, Gradience and Frequency in Phonology, Stanford University.
CLARK, ROBIN, and IAN ROBERTS. 1993. A computational model of language learnability and language change. *Linguistic Inquiry* 24.299–345.
COETZEE, ANDRIES W., and JOE PATER. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 26.289–337.
CROFT, WILLIAM. 2000. *Explaining language change: An evolutionary approach*. Harlow: Longman.
DELLA PIETRA, STEPHEN A.; VINCENT J. DELLA PIETRA; and JOHN D. LAFFERTY. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.380–93.
FISCHER, MARKUS. 2005. A Robbins-Monro type learning algorithm for an entropy maximizing version of stochastic optimality theory. Berlin: Humboldt University masters thesis.
FOUNTAIN, AMY V. 1998. *An optimality theoretic account of Navajo prefixal syllables*. Tucson: University of Arizona dissertation.
FRISCH, STEFAN A.; JANET B. PIERREHUMBERT; and MICHAEL B. BROE. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22.179–228.
GOLDWATER, SHARON, and MARK JOHNSON. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, ed. by Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–20. Stockholm: Department of Linguistics, Stockholm University.
HALE, MARK. 1998. Diachronic syntax. *Syntax* 1.1–18.
HAMMOND, MICHAEL. 1999. *The phonology of English: A prosodic optimality-theoretic approach*. Oxford: Oxford University Press.
HAY, JENNIFER B., and R. HARALD BAAYEN. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9.342–48.
HAYES, BRUCE P. 1986. Assimilation as spreading in Toba Batak. *Linguistic Inquiry* 17.467–99.
HAYES, BRUCE P., and COLIN WILSON. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440.
HAYES, BRUCE P.; KIE ZURAW; PÉTER SIPTÁR; and ZSUZSA C. LONDE. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85.822–63.
HUDSON KAM, CARLA L., and ELISSA L. NEWPORT. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development* 1.151–95.
HUDSON KAM, CARLA L., and ELISSA L. NEWPORT. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology* 59.30–66.
JÄGER, GERHARD. 2007. Maximum entropy models and stochastic optimality theory. *Architectures, rules, and preferences: Variations on themes by Joan Bresnan*, ed. by Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, 467–79. Stanford, CA: CSLI Publications.
JAYNES, EDWIN T. 1957. Information theory and statistical mechanics. *Physical Review* 106.620–30.
JOHNSON, MARK. 2007. A gentle introduction to maximum entropy models and their friends. Talk given at the Northeastern Computational Phonology Meeting, University of Massachusetts.
JUSCZYK, PETER. 1997. *The discovery of spoken language*. Cambridge, MA: MIT Press.
KARI, JAMES M. 1976. *Navajo verb prefix phonology*. New York: Garland.
KAWAHARA, SHIGETO; HAJIME ONO; and KIYOSHI SUDO. 2005. Consonant co-occurrence restrictions in Yamato Japanese. *Japanese/Korean linguistics 14*, ed. by Timothy J. Vance and Kimberly A. Jones, 7–38. Stanford, CA: CSLI Publications.

KAYE, ALAN S. 2005. Gemination in English. *English Today* 21.43–55.
KESSLER, BRETT. 2001. *The significance of word lists*. Cambridge, MA: MIT Press.
KIPARSKY, PAUL. 1968. Linguistic universals and linguistic change. *Universals in linguistic theory*, ed. by Emmon Bach and Robert T. Harms, 170–202. New York: Holt, Rinehart and Winston.
LADEFOGED, PETER. 2001. *A course in phonetics*. 3rd edn. Fort Worth, TX: Harcourt Brace Jovanovich.
LASS, ROGER. 1997. *Historical linguistics and language change*. Cambridge: Cambridge University Press.
LEGENDRE, GÉRALDINE; YOSHIRO MIYATAL; and PAUL SMOLENSKY. 1990. Harmonic grammar—A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. *Proceedings of the 12th annual conference of the Cognitive Science Society*, 388–95. Hillsdale, NJ: Lawrence Erlbaum.
LIGHTFOOT, DAVID. 1979. *Principles of diachronic syntax*. Cambridge: Cambridge University Press.
LIGHTFOOT, DAVID. 1991. *How to set parameters: Arguments from language change*. Cambridge, MA: MIT Press.
LIGHTFOOT, DAVID. 1999. *The development of language: Acquisition, change, and evolution*. Oxford: Blackwell.
MARTIN, ANDREW. 2007. *The evolving lexicon*. Los Angeles: University of California, Los Angeles dissertation.
MARTIN, ANDREW. 2004. The effects of distance on lexical bias: Sibilant harmony in Navajo compounds. Los Angeles: University of California, Los Angeles masters thesis.
MARTIN, SVEN C.; HERMANN NEY; and JÖRG ZAPLO. 1999. Smoothing methods in maximum entropy language modeling. *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, 545–48.
MCDONOUGH, JOYCE. 1991. On the representation of consonant harmony in Navajo. *West Coast Conference on Formal Linguistics (WCCFL)* 10.319–35.
MCDONOUGH, JOYCE. 2003. *The Navajo sound system*. Dordrecht: Kluwer.
OH, GRACE E., and MELISSA A. REDFORD. 2009. The effect of boundary recoverability on geminate length in English. *Journal of the Acoustical Society of America* 125.2568.
PADGETT, JAYE. 1995. *Stricture in feature geometry*. Stanford, CA: CSLI Publications.
PAYNE, ELINOR M. 2005. Phonetic variation in Italian consonant gemination. *Journal of the International Phonetic Association* 35.153–81.
PETERS, ANN. 1983. *The units of language acquisition*. Cambridge: Cambridge University Press.
POZDNIAKOV, KONSTANTIN, and GUILLAUME SEGERER. 2007. Similar place avoidance: A statistical universal. *Linguistic Typology* 11.307–48.
PRINCE, ALAN, and PAUL SMOLENSKY. 2004. *Optimality theory: Constraint interaction in generative grammar*. Malden, MA: Blackwell.
PYCHA, ANNE. 2010. A test case for the phonetics–phonology interface: Gemination restrictions in Hungarian. *Phonology* 27.119–52.
RIDOUANE, RACHID. 2007. Gemination in Tashlhiyt Berber: An acoustic and articulatory study. *Journal of the International Phonetic Association* 37.119–42.
SAPIR, EDWARD. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace & World.
SAPIR, EDWARD, and HARRY HOIJER. 1967. *The phonology and morphology of the Navaho language*. Berkeley: University of California Press.
SEPP, MARY. 2006. *Phonological constraints and free variation in compounding: A corpus study of English and Estonian noun compounds*. New York: City University of New York dissertation.
SMOLENSKY, PAUL, and GÉRALDINE LEGENDRE. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar*. Cambridge, MA: MIT Press.
STOCKWELL, ROBERT, and DONKA MINKOVA. 2001. *English words: History and structure*. Cambridge: Cambridge University Press.
WILSON, COLIN. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science: A Multidisciplinary Journal* 30.945–82.

WILTSHIRE, CAROLINE. 1999. The conspiracy of Luganda compensatory lengthening. *New dimensions in African linguistics and languages*, ed. by Paul F. A. Kofey, 131–47. Trenton, NJ: Africa World Press.
YOUNG, ROBERT W., and WILLIAM MORGAN. 1987. *The Navajo language: A grammar and colloquial dictionary*. Albuquerque: University of New Mexico Press.

Laboratory for Language Development
RIKEN Brain Science Institute
2-1 Hirosawa, Wako-shi, Saitama 351-0198
Japan
[amartin@brain.riken.jp]

[Received 29 July 2010;
revision invited 12 May 2011;
revision received 24 June 2011;
accepted 30 August 2011]