Nat Lang Linguist Theory
DOI 10.1007/s11049-012-9179-z

Frequency biases in phonological variation

Andries W. Coetzee · Shigeto Kawahara

Received: 24 August 2010 / Accepted: 23 May 2011

Abstract In the past two decades, variation has received a lot of attention in mainstream generative phonology, and several different models have been developed to account for variable phonological phenomena. However, all existing generative models of phonological variation account for the overall rate at which some process applies in a corpus, and therefore implicitly assume that all words are affected equally by a variable process. In this paper, we show that this is not the case. Many variable phenomena are more likely to apply to frequent than to infrequent words. A model that accounts perfectly for the overall rate of application of some variable process therefore does not necessarily account very well for the actual application of the process to individual words. We illustrate this with two examples, English *t/d*-deletion and Japanese geminate devoicing. We then augment one existing generative model (noisy Harmonic Grammar) to incorporate the contribution of usage frequency to the application of variable processes. In this model, the influence of frequency is incorporated by scaling the weights of faithfulness constraints up or down for words of different frequencies. This augmented model accounts significantly better for variation than existing generative models.

Keywords Variation · Usage frequency · Harmonic Grammar · *t/d*-deletion · Japanese geminate devoicing

Electronic supplementary material The online version of this article (doi:10.1007/s11049-012-9179-z) contains supplementary material, which is available to authorized users.

A.W. Coetzee (✉)
Department of Linguistics, University of Michigan, 440 Lorch Hall, 611 Tappan Street, Ann Arbor, MI 48109, USA
e-mail: coetzee@umich.edu

S. Kawahara
Department of Linguistics, Rutgers University, 18 Seminary Pl., New Brunswick, NJ 08901-1108, USA
e-mail: kawahara@rci.rutgers.edu

Published online: 22 September 2012

1 Introduction

1.1 The changing prospects of variation

Although the existence of phonological variation has been acknowledged since the early years of generative phonology (Postal 1966:185, 1968:14–15), variation received relatively little attention in mainstream generative phonology during the first 25 years of the history of this field. To the extent that variation was acknowledged, it was usually relegated to the late stages of phonology or to phonetic implementation, and was hence not considered a part of the core of phonological grammar. In Lexical Phonology, for instance, it was assumed that lexical rules apply obligatorily while "postlexical rules can be optional and subject to variation" (Kaisse and Shaw 1985:6; see also Kiparsky 1985:86).

This low valuation of variation in mainstream generative phonology contrasts with how it was viewed in the Labovian variationist tradition. This research tradition, spearheaded by Labov's work in the late 1960's (Labov 1966, 1969, etc.), developed concurrently with mainstream generative phonology, but had little impact on this field. In this approach, variation is central to grammar rather than an accidental property that applies only on the edges of grammar. In fact, Labov (2004:6) claims that variation is "the central problem of linguistics".

In the past 15 years, the prospects of variation in generative phonology have changed dramatically. It now occupies a central place in the study of phonology, and to some extent dictates the architecture of phonological grammar. A clear indication of this change is how variation has been treated in handbooks of phonological theory. The first edition of the Blackwell *Handbook of Phonological Theory* (Goldsmith 1995), which reflects the situation in generative phonology at the beginning of the 1990's, does not even contain the word "variation" in its subject index. In contrast, every handbook since contains a chapter dedicated to variation (Anttila 2002b, 2007; Coetzee 2012; Coetzee and Pater 2011; Guy 2011). Similarly, several articles on variation have appeared in theoretical, generatively-oriented journals over the past decade (Anttila 2002a, 2006; Anttila et al. 2008; Boersma and Hayes 2001; Coetzee 2006; etc.).

This same period has seen the development of several versions of current generative phonological grammar intended to deal with variation. These models have all been developed in some version of a constraint-based grammar, be that classic discrete Optimality Theory (OT) (Anttila 1997, 2002a, 2006, 2007; Anttila et al. 2008; Bane 2011, to appear; Coetzee 2004, 2006, 2009c; Kiparsky 1993; Reynolds 1994), stochastic OT (Boersma 1997; Boersma and Hayes 2001), or noisy Harmonic Grammar (HG) (Coetzee 2009a; Coetzee and Pater 2011; Jesney 2007).[1]

In fact, variation has become so important that the ability of a grammatical model to account for variation is now often used as one of the measures of the model's sufficiency. Anttila (2002b:211) claims that an adequate theory of phonology should account for the "locus of variation" (where variation is observed and where it is not),

[1] Noisy HG was first implemented by Paul Boersma in *Praat* (Boersma and Weenink 2009) as early as 2006.

and the "degrees of variation" (the frequency of different variants). Using these two criteria as a measure of success, most of the models mentioned above have been very successful. All of these models have formal mechanisms that can account for the locus of variation. With the exception of Coetzee's 2004/2006 model, these models also predict the degrees of variation. In fact, they have all been shown to be relatively successful in modeling the frequency with which different variants are observed for a range of variable phenomena.

In spite of the obvious progress that has been made in accounting for phonological variation, much work still remains. All of the existing generative models mentioned above are purely grammatical models that do not incorporate the influence of non-grammatical factors on variation. Decades of research in variationist sociolinguistics and more recent investigations of large speech corpora, however, have shown that variation is influenced by many factors in addition to grammar. In this paper, we take the next logical step in accounting for phonological variation by developing an extension of one of the existing generative models of phonological variation (noisy HG) that allows both grammatical and non-grammatical factors to impact the pattern of observed variation.

1.2 Non-grammatical influences on variation

One of the persistent results of the variationist research tradition is that variation is influenced, in addition to grammatical factors, by many non-grammatical factors. In fact, reviewing this tradition, Bayley (2002:118) identified "the principle of multiple causes" as one of the four core principles of this tradition. These non-grammatical factors include speech genre (word lists, informal conversations, read speech, etc.), discourse situation, age, sex or educational background of the speaker, etc.

Although mainstream generative phonology has adopted the variationist tradition's higher valuation of variation over the past decade, mainstream approaches have focused nearly exclusively on the grammatical factors that impact variation. Existing generative models make no formal allowance for the influence of other factors. Yet, the variationist tradition has established that phonological variation is influenced by many factors in addition to grammar. The next step, then, is to augment generative models so that they can account for both the grammatical and non-grammatical factors that influence variation. This idea is not original to us. Boersma and Hayes (2001) mention this explicitly with regard to their stochastic OT model of variation, and suggest a way in which their model could be augmented to incorporate some non-grammatical aspects of variation. This paper follows up in more detail on their suggestion (although we will develop our model in noisy HG rather than their stochastic OT).

1.3 Usage frequency as a non-grammatical influence on variation

As mentioned above, many non-grammatical factors influence the application of variable phonological processes. In this paper, we focus on usage frequency—i.e., the observation that some variable processes apply at different rates to words that differ in frequency. Our selection of usage frequency is one of convenience: since frequency

is already quantitative, it is straightforward to incorporate it into a quantitative model of variation. We also acknowledge that usage frequency would not be considered external to the grammar in all grammatical models. In fact, in several recent models of grammar, grammar can be described as structured memory encoding of frequency—see the usage-based and exemplar models of grammar, for instance (Bybee 2001, 2006, 2007; Gahl and Yu 2006, and papers therein; Pierrehumbert 2001; etc.). In the generative tradition, however, usage frequency is not encoded in the grammar—generative models do not treat two words differently merely because they differ in their usage frequencies. In this paper, we subscribe to the standard generative assumption, and we will hence treat usage frequency as external to the grammar. See also Sect. 5.1 for further discussion.

Some variable phonological processes (typically reduction or simplification processes, though see Sect. 5.2) are more likely to affect words with higher than lower usage frequency. For example, Bybee reports that the schwa in frequent *memory* is more likely to delete than the schwa in the nearly identical, but infrequent, *mammary* (Hooper 1976; see also Bybee 2000:68).

This correlation between frequency and simplification processes is widespread and has been reported for many different phonological processes. For instance, the variable deletion of word-final *t/d* from consonant clusters in English is more likely to apply to frequent than infrequent words—i.e., more deletion from frequent *just* than infrequent *jest* (Bybee 2000:69–70, 2002; Coetzee 2009a:272–273, 2009c; Lacoste 2008:187–207). The same process also applies in Dutch, where the correlation between frequency and the probability of deletion also holds (Goeman 1999:182; Phillips 2006:65). (See Sect. 3 for a more detailed discussion of *t/d*-deletion.) A similar correlation of usage frequency and variation has also been illustrated for flapping in American English (Patterson and Connine 2001), word-medial *t*-deletion in English (Raymond et al. 2006), word-final *s*-lenition in Spanish (File-Muriel 2010), *l*-vocalization in American English (Lin et al. 2011), and for geminate devoicing in Japanese loans (on which there is more in Sect. 4; see Kawahara 2011a, 2011b). See Phillips (2006) for a recent review of many more similar examples.

A model of variation that incorporates only grammatical influences on variation cannot capture the influence of factors like usage frequency. As a concrete illustration, we include Fig. 1, which represents the rate of *t/d*-deletion from word-final clusters in English for a selection of words from the *Buckeye Corpus* (Pitt et al. 2007), plotted against the log frequency of the words, as measured in CELEX (Baayen et al. 1995).[2] (See Sect. 3.1 on the details of how these data were extracted from the *Buckeye Corpus*.) The three panels show the rate of deletion before consonant-initial words (*west bank*), vowel-initial words (*west end*), and before pause (*west.*). The broken horizontal lines show the overall deletion rate in each context—i.e., deletion rate based on token counts. In existing grammatical models of variation, these are the variation

[2]Throughout this paper, all logarithmic transformations use a base of 10. For instance, the word *and* has a CELEX frequency of 514,946, and hence a log frequency of $\log_{10}(514{,}946) = 5.71$. In the Buckeye Corpus, *and* appears in pre-vocalic context 3,273 times, and in 2,966 of these occurrences its final /d/ was deleted. In this context, *and* therefore shows a deletion rate of 90.6 %. In the middle panel of Fig. 1, the data point that appears in the upper right-hand corner of the graph therefore corresponds to *and* at a log frequency of 5.71 on the x-axis, and at a deletion rate of 90.6 % on the y-axis.

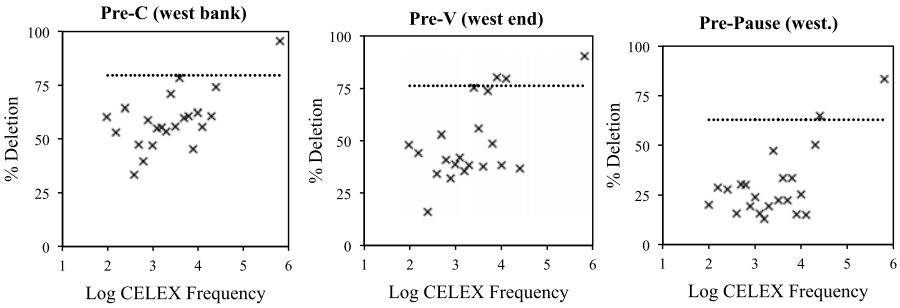

Fig. 1 Relation between deletion rate and frequency in the *Buckeye Corpus*

frequencies that are modeled. These rates capture the difference between the different grammatical contexts (most deletion pre-consonantally, then pre-vocalically, and least deletion pre-pausally). However the actual, observed rates deviate quite drastically from the overall rates, especially for words of lower frequency. To account not only for the grammatical influences on variation, but also for the influence of usage frequency, existing grammatical models would need to be augmented in some way. In the rest of this paper, we augment one of the existing grammatical models of variation (noisy HG). We add an extra parameter, incorporating usage frequency into the noisy HG model of variation, and show that this augmented model accounts significantly better for the deletion rates of words with different usage frequencies.

Although we treat frequency as if it is a standalone property of a word, it is actually only one subpart of the larger concept of predictability. A word's predictability depends on many factors in addition to its frequency, as has been documented by many studies in speech processing and production. A word is, for instance, primed by other words to which it is semantically (McNamara 2005; etc.) or phonologically (Goldinger et al. 1992; etc.) related, or by repetition (Versace and Nevers 2003; etc.). On the other hand, a word is inhibited (i.e., becomes less predictable) if it inhabits a dense lexical neighborhood (Luce and Pisoni 1986; Vitevitch and Luce 1998, 1999; etc.). Many studies have documented that factors such as these influence speech production, with the general result being that less predictable words (words that are inhibited or less strongly primed) tend to be produced more slowly, and with more effort or clarity (Baese-Berk and Goldrick 2009; Bell et al. 2009; Gahl 2008; Jurafsky et al. 2001; Scarborough 2004, 2010; etc.). Similar results have also been reported in the literature on "Uniform Information Density" (Frank and Jaeger 2008; Jaeger 2010; etc.), which shows that speakers have a tendency to spread out information equally across an utterance. Since more predictable words carry less information, speakers tend to reduce these words. Ultimately, it would be necessary to determine an overall measure of the predictability of a word that includes contributions from all of these aspects. Our focus on usage frequency is only an initial step.

2 Noisy Harmonic Grammar with weight scaling

We develop our model in a noisy version of Harmonic Grammar (Pater 2009; Smolensky and Legendre 2006). HG is a constraint-based theory that is closely re-

lated to OT (Prince and Smolensky 1993, 2004) and is, in fact, an historical predecessor of OT (Goldsmith 1993; Legendre et al. 1990). The main difference between HG and OT is that HG works with weighted rather than ranked constraints. Noisy HG is a stochastic implementation of HG, similar to the noisy implementation of OT, known as stochastic OT (Boersma 1997; Boersma and Hayes 2001). Noisy HG and stochastic OT are closely related; we could have developed our model in this paper just as successfully in stochastic OT rather than noisy HG (see Coetzee and Pater 2011 for evidence that noisy HG and stochastic OT account for most variable phenomena equally well). In the rest of this section, we first show how noisy HG accounts for variation, and then how we will augment this model to incorporate the influence of frequency on variation.

2.1 Noisy Harmonic Grammar

HG, like OT, is a constraint-based theory of grammar. The main difference between HG and OT is that OT relies on constraint ranking, and HG on constraint weighting. This difference is illustrated in the tableaux in (2) using the familiar OT constraints in (1). These tableaux represent the grammar of a language that does not allow tautosyllabic consonant clusters, and that repairs such clusters via deletion. In the HG tableau, *w*(CON) stands for the weight of constraint CON.

(1)

MAX	Assign one violation mark for every segment in the input that lacks a correspondent in the output (no deletion). (McCarthy and Prince 1995:371)
DEP	Assign one violation mark for every segment in the output that lacks a correspondent in the input (no epenthesis). (McCarthy and Prince 1995:371)
*COMPLEX	Assign one violation for every tautosyllabic consonant cluster. (Prince and Smolensky 1993:96)

(2) a. Optimality Theory: DEP ≫ *COMPLEX ≫ MAX

/lʊst/	DEP	*COMPLEX	MAX
lʊst		*!	
☞ lʊs			*
lʊs.ti	*!		

b. Harmonic Grammar: *w*(DEP) > *w*(*COMPLEX) > *w*(MAX)

/lʊst/	5 DEP	1.5 *COMPLEX	1 MAX	H
lʊst		−1		−1.5
☞ lʊs			−1	−1
lʊs.ti	−1			−5

In HG, each constraint is weighted, and these weights are indicated with Arabic numerals above the constraint names in HG tableaux.[3] Constraint violations are marked with negative whole numbers rather than asterisks. A *harmony score* H is calculated for every candidate, using the formula in (3)—i.e., by taking the product of the weight of each constraint and the violation index of the candidate, and summing these products. These H-scores are indicated in the last column of the tableau. The H-score of the first candidate in (2b), for instance, is calculated as follows: The weight of DEP (= 5) is multiplied by the violation index of the candidate in terms of DEP (zero, since this candidate does not violate DEP). The weight of *COMPLEX (= 1.5) is then multiplied with the violation index of the candidate for *COMPLEX (−1), giving −1.5. Similarly, the weight of MAX (= 1) is multiplied with the violation index of the candidate (zero again). Finally, these products are summed, giving an H-score of −1.5 for this candidate. Since H-scores are negative, the candidate with the H-score closest to zero wins.

(3) $$H(cand) = \sum_{i=1}^{n} w_i C_i(cand)$$

Where w_i is the weight of constraint C_i, and $C_i(cand)$ is the number of times that candidate *cand* violates C_i, expressed as a negative integer.

The version of HG illustrated above is not noisy HG, and cannot generate variation—given these constraints and weights, the grammar will always map /lʊst/ onto [lʊs]. However, HG has an implementation known as "noisy HG" that can generate variable outputs (Coetzee 2009a; Coetzee and Pater 2011; Jesney 2007). Noisy HG is closely related to stochastic OT (Boersma 1997; Boersma and Hayes 2001). In stochastic OT, constraint ranking is along a continuous scale, rather than a discrete scale as in classic OT. Every time that the grammar is used, the ranking of each constraint is perturbed by a negative or positive noise value (randomly selected from a normal distribution with a mean of zero). Because of this noisy evaluation, the relative ranking between two constraints can differ from one occasion to the next, resulting in variation. Noisy HG shares with stochastic OT this noisy evaluation procedure. The only difference is that, in noisy HG, the weights of constraints rather than their rankings are perturbed by random noise. If the weights of two conflicting constraints are close enough, the noisy evaluation can result in their relative weights flipping around between evaluation occasions, potentially causing variation.

In (4), the HG tableau from (2) is repeated, this time with noise. In these tableaux, *w* stands for the weight of a constraint and *nz* for the noise added to a constraint at the specific evaluation occasion. The effective weight of constraints (the sum of *w* and *nz*) is given in parentheses after the constraint names. In the first tableau, the weight of *COMPLEX is adjusted down by the addition of noise at −0.4, and the weight of MAX is adjusted upward by a positive noise value of 0.2. The effect is that violation of *COMPLEX is now less serious than the violation of MAX, so that the

[3] In noisy HG, the weights of the constraints are determined by a gradual learning algorithm, closely related to the learning algorithm developed by Boersma and Hayes (2001) for their stochastic OT model. For more on this, see Sect. 3.2.2.

faithful candidate has the highest H-score, and is selected as the output. In the second tableau, the weight of *COMPLEX is adjusted upward and that of MAX downward, so that the deletion candidate has the highest H-score and is selected as the output. These tableaux show how the same grammar (the same constraints with the same weights) can select different outputs on different evaluation occasions because of the addition of noise to the evaluation. An updated version of the formula used to calculate H-scores that include noise is given in (5).

(4) a. Faithful candidate optimal

	w	*nz*	*w*	*nz*	*w*	*nz*	
	5	–0.7	1.5	–0.4	1	0.2	
/lʊst/	DEP (4.3)		*COMPLEX (1.1)		MAX (1.2)		H
☞ lʊst			–1				–1.1
lʊs					–1		–1.2
lʊs.ti	–1						–4.3

b. Deletion candidate optimal

	w	*nz*	*w*	*nz*	*w*	*nz*	
	5	–0.8	1.5	0.1	1	–0.1	
/lʊst/	DEP (4.2)		*COMPLEX (1.6)		MAX (0.9)		H
lʊst			–1				–1.6
☞ lʊs					–1		–0.9
lʊs.ti	–1						–4.2

(5) $$H(cand) = \sum_{i=1}^{n}(w_i + nz_i)C_i(cand)$$

Where w_i is the weight of constraint C_i, nz_i the noise associated with constraint C_i at this evaluation occasion, and $C_i(cand)$ is the number of times that *cand* violates C_i, expressed as a negative integer.

Several authors have shown have shown that this model of phonological variation can account for a variety of variable phenomena (Coetzee 2009a; Coetzee and Pater 2011; Jesney 2007). Coetzee and Pater (2011), in particular, show that it performs at least as well as stochastic OT. This model, however, still treats all words exactly the same. There is no place in the formula in (5) where any factor such as usage frequency can impact the H-score of a candidate. In the next section, we augment this model to allow for factors such as usage frequency to impact the H-score of a candidate.

2.2 Weight scaling

We need a model that can account for the fact that more frequent words are more likely to be treated unfaithfully. This correlation can be captured by scaling the weight of faithfulness constraints down for frequent words and up for infrequent words. Violating a faithfulness constraint will then contribute less to the H-score of a frequent word, resulting in unfaithfulness being more likely, while it will contribute more to the H-score of an infrequent word, resulting in faithfulness being more likely.

There are precedents for this idea in the literature. Van Oostendorp (1997) and Itô and Mester (2001), for instance, suggested that the higher likelihood of faithfulness in more formal speech registers can be captured by ranking faithfulness constraints higher in formal speech situations—an idea that echoes the concept of "carefulness weights" in Lindblom's Hyper- and Hypoarticulation theory of speech production (Lindblom 1990). Boersma and Hayes (2001: Appendix C) similarly suggest scaling the ranking values of constraints to account for different rates of unfaithfulness observed with different speech registers.

By adding such weight scaling to the model, two words that differ in usage frequency may be evaluated differently in the same grammatical context. Continuing with the example from the previous section, assume that /lʊst/ and /nʊst/ differ in frequency such that /lʊst/ is frequent and /nʊst/ infrequent. For the sake of the illustration, assume that /lʊst/ will be associated with a weight scaling factor of −1, and /nʊst/ with a factor of +1. The weight of faithfulness constraints will be scaled down by one unit in the evaluation of /lʊst/, and up by one unit in the evaluation of /nʊst/. The tableaux in (6) show how this addition of scaling factors affects the evaluation of these words. In these tableaux, the same grammatical settings (the same constraint weights and noise values) are used. All that differs is the scaling factors associated with the faithfulness constraints (marked by *sf* in the tableaux). The result is that frequent /lʊst/ is mapped onto its unfaithful candidate [lʊs], while infrequent /nʊst/ is mapped onto its faithful candidate [nʊst]. An updated version of the H-score formula that incorporates the scaling factor is given in (7).

(6) a. Evaluating frequent /lʊst/, with $sf = -1$

	w	*nz*	*sf*	*w*	*nz*	*w*	*nz*	*sf*	
	5	0.7	−1	1.5	0.1	1	0.2	−1	
/lʊst/	DEP (4.7)			*COMPLEX (1.6)		MAX (0.2)			H
lʊst				−1					−1.6
☞ lʊs						−1			−0.2
lʊs.ti	−1								−4.7

b. Evaluating infrequent /nʊst/, with $sf = +1$

	w	*nz*	*sf*	*w*	*nz*	*w*	*nz*	*sf*	
	5	0.7	1	1.5	0.1	1	0.2	1	
/nʊst/	DEP (6.7)			*COMPLEX (1.6)		MAX (2.2)			H
☞ nʊst				−1					−1.6
nʊs						−1			−2.2
nʊs.ti	−1								−6.7

(7) $$\mathsf{H}(cand) = \sum_{i=1}^{n}(w_i + nz_i)M_i(cand) + \sum_{j=1}^{m}(w_j + nz_j + sf)F_j(cand)$$

Where M_i is the i-th markedness constraint, w_i the weight associated with M_i, nz_i the noise associated with M_i at this evaluation occasion, and $M_i(cand)$ the number of times that *cand* violates M_i (expressed as a negative integer); and where F_j is the j-th faithfulness constraint, w_j the weight associated with F_j, nz_j the noise associated with F_j at this evaluation occasion,

and $F_j(cand)$ the number of times that *cand* violates F_i (expressed as a negative integer); and where *sf* is the scaling factor associated with the specific word being evaluated.

In this model, only faithfulness constraints have scaling factors. The same effect could also be achieved by scaling markedness weights, or even by scaling the weights of both markedness and faithfulness constraints. In fact, Boersma and Hayes (Boersma and Hayes 2001: Appendix C) propose scaling the ranking values of both markedness and faithfulness constraints to incorporate style effects into their stochastic OT model. Although there are subtle differences in the variation patterns predicted by these different options, any of these options could have accounted equally well for the data that we discuss in this paper. We return to this issue briefly in Sect. 5.2, but leave the question of the difference between these options for future research.

2.3 A linking function between frequency and scaling factors

The final part of our model is a linking function between frequency and scaling factors: Given a word of some frequency, what is the scaling factor that should be used in evaluating this word? This problem could be approached from two different directions. One possibility is that the mapping between frequency and scaling factors has to be learned on a language-by-language basis. The language learner will then have to take note of how words that are equivalent in their phonological properties but differ in frequency are treated differently by the grammar. From this information, he/she will deduce a function that best maps from frequency to scaling factors. Since the linking function is then determined on a language-particular basis, we would not necessarily expect to see universal tendencies in how frequency maps to scaling factors. See Coetzee (2009a) for an implementation of this kind of approach.

A different possibility is that there is some universal linking function that applies similarly to all languages. The expectation would then be that frequency has the same basic influence in all languages. Given the large amount of evidence that frequency has the same basic influence in all languages (More frequent words are more likely to undergo reduction processes—see the references above in Sect. 1.3 and the discussion in Sect. 5.2 below.), we pursue the second option—that is, that the same basic linking function applies in all languages. In this paper, we illustrate how such a universal mechanism accounts well for two different variable phenomena in two unrelated languages (*t/d*-deletion in English, and geminate devoicing in Japanese).

We propose that every word is associated with a distribution function, whose shape is determined by the frequency of the word. These functions are modeled as instantiations of the *beta* distribution (Gupta and Nadarajah 2004), and the scaling factor associated with a word is read off its distribution function.[4] The formula of the *beta* distribution is given in (8). In addition to its argument x, the distribution has three parameters. ρ specifies the range of the function as spanning from $-\rho$ to ρ. α and β are shape parameters that determine the skewness of the distribution. When $\alpha = \beta$,

[4]See later in this section on why we use the *beta* distribution rather than a more well-known distribution such as the normal distribution.

the distribution is symmetric around zero. When $\alpha > \beta$, it is left-skewed, and when $\alpha < \beta$, it is right-skewed. Additionally, the larger the difference between α and β, the more severe the skewness of the distribution is.

$$f(x, \alpha, \beta, \rho) = \rho \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 x^{\alpha-1}(1-x)dx} \tag{8}$$

Frequent words must have a negative, and infrequent words a positive scaling factor. But what counts as "frequent" or "infrequent"? A reference frequency has to be established such that words that appear more frequently than this reference frequency will be treated as frequent, and words that appear less frequently will be treated as infrequent. There are several ways in which such a reference point can be established. The average or median frequency of all the words in the lexicon could be used, for instance. We explored several different options, and settled on the one that resulted in the best fit of our model to the data. Specifically, the reference frequency is set in such a way that (at least) half of the tokens in the corpus are being treated as frequent, and (at most) half as infrequent. The exact way in which we determine the reference frequency is stated in (9).

(9) Let N be the total number of tokens in the corpus.

i. Order the words in the corpus in terms of frequency.
ii. Determine the point on this ordering such that at least $N/2$ of all the tokens are above this point.
iii. Determine the log frequency of the word just above this point, and the word just below this point.
iv. Let the reference frequency be halfway between these two log frequencies.

We illustrate how this algorithm works with an example. In Sect. 3, we work with a corpus of *t/d*-deletion examples, extracted from the *Buckeye Corpus* (Pitt et al. 2007). The corpus contains 16,460 tokens. Ordering the tokens according to their CELEX frequencies (Baayen et al. 1995), the word *and* occupies the topmost position. It also accounts for more than half of the tokens in the corpus (*and* appears 8,827 times in our corpus). The reference point is halfway between the log CELEX frequency of *and* and the log CELEX frequency of *just*, the next most frequent word in our corpus. For reasons that we explain in Sect. 3, we grouped words together into larger log groups. *Just* was placed into the 4.4 log frequency group, while *and* went into the 5.8 log frequency group. The midpoint between these two is 5.1, and this value serves as the reference point in our modeling of the data in our *t/d*-deletion corpus.

Having established the reference frequency, the values of the shape parameters (α and β) of the *beta* distribution associated with each word, as well as the scaling factor associated with each word, can now be determined. Specifically, we propose that α is set equal to the log reference frequency, and β to the log frequency of the specific word. The α-parameter therefore represents the reference frequency (i.e., neither frequent nor infrequent). The β-parameter represents the frequency of a specific word. For a word that appears less often than the reference frequency (so that $\alpha > \beta$), the distribution will be left-skewed and hence have a positive mode—see the distribution for *interrupt* in Fig. 2. We propose that the mode is used as the frequency scaling

factor associated with a specific word. For a word that appears less frequently than the reference frequency, the scaling factor will therefore be positive. The weight of faithfulness constraints will be scaled up in the evaluation of such a word, so that the word will resist an unfaithful mapping more strongly. On the other hand, for a word that appears more often than the reference frequency (so that $\alpha < \beta$), the distribution will be right-skewed, and the mode thus negative—see the function for *and* in Fig. 2. The scaling factor of such a frequent word (the mode of the *beta* distribution) will be negative, diminishing the contribution of faithfulness constraints in evaluating the word, resulting in a higher likelihood of an unfaithful mapping. The table in (10) summarizes the effect of the values of α and β on the skewness of the *beta* distribution, and the effect that this has on the mode of the distribution (and the scaling factors in the model that we propose here).

(10) Determining the values of α, β, and the scaling factor associated with each word

Shape parameters				
Reference frequency		Frequency of specific word	Skewness	Mode (= scaling factor)
α	<	β	Right	negative
α	=	β	Symmetrical	zero
α	>	β	Left	positive

The last parameter to set is the range parameter ρ. ρ does not influence the shape of the *beta* distribution, but only its range. In particular, it specifies the minimum and maximum value of the function on the x-axis: The higher the value of ρ, the higher the absolute value of the mode. The higher ρ is, the higher the scaling factors will be. And the higher the scaling factors, the more influence the frequency of words can have on their evaluation. ρ therefore determines how much frequency is allowed to influence how the grammar functions. We propose that the value of ρ be fit to the data—i.e., for every corpus, the value of ρ that results in the best fit between the model and the data is used.[5]

In (11), we give examples of the parameter values and the modes for three words from our *t/d*-deletion corpus. *And* is used as an example of a frequent word. *And*'s distribution function is right-skewed, so that the mode of this function, and hence *and*'s scaling factor, is negative. *Interrupt* and *weekend* both appear less frequently than the reference frequency, and both serve as examples of infrequent words. Their distributions are left-skewed, so that their modes are positive, and the scaling factors associated with these two words are also positive. Although both *interrupt* and *weekend* are infrequent, they differ in frequency. *Interrupt* has a CELEX log frequency of

[5]We also leave open the possibility that the value of ρ can vary across different speech styles. A larger value for ρ results in a larger range for the *beta* distribution, and hence in modes that deviate more from zero. Since the mode of the *beta* distribution is used as the scaling factor in the evaluation of some word, a larger ρ (and hence more extreme mode and scaling factor) will increase the influence that frequency can have on the determination of H-scores. It is therefore possible that the value of ρ may fluctuate to account for speech situations in which frequency has a bigger or smaller impact. We do not explore this possibility further in this paper, however.

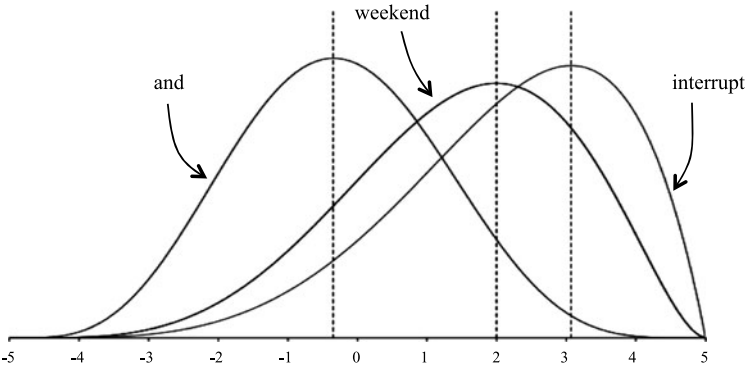

Fig. 2 *Beta* distributions for words from (11) with $\rho = 5$. Vertical broken lines mark the modes for the distributions, and hence the scaling factors associated with these words

1.98 and *weekend* has one of 2.76. In the distribution function associated with *interrupt*, the difference between the values of α and β is hence larger than in *weekend* ($\alpha = 5.1$, $\beta = 1.98$ vs. $\alpha = 5.1$, $\beta = 2.76$). We include both of these words to show that the larger the difference between α and β, the more skewed the distribution, and hence the more extreme the mode of the distribution. The more the frequency of a word (represented by β) differs from the reference frequency (represented by α), the more its scaling factor will differ from zero. Faithfulness will hence be scaled down more for more frequent words, and up more for less frequent words. The table also gives the modes for these distributions at three different values of ρ. Note how a change in ρ influences only the absolute value of the modes, and not their signs. In Fig. 2 we show the shape of the distribution functions for these tokens when $\rho = 5$ (the value that we use for ρ in Sect. 3).[6]

(11) Examples of scaling factors in the *t/d*-deletion corpus (see Sect. 3.3)

	Shape parameters			Mode (scaling factor)		
	α	β				
Word	(reference frequency)	(log frequency of word)	Skew	ρ = 1	ρ = 5	ρ = 10
and	5.1	5.71	Right	−0.07	−0.35	−0.70
weekend	5.1	2.76	Left	0.40	2.00	4.00
interrupt	5.1	1.98	Left	0.61	3.07	6.14

In principle, scaling factors could be deduced from a more well-known distribution such as the normal distribution. Our selection of the *beta* rather than the normal distribution is motivated by the fact that the *beta* distribution has a finite range (specified by ρ), while the normal distribution has an infinite range. The finite range of the *beta* distribution places an absolute limit on the influence that non-grammatical factors such as frequency can have via weight scaling. If scaling factors were taken from the

[6] An Excel file for the calculation of the *beta* distribution's mode under different settings of the three parameters is available from http://www.quantitativeskills.com/sisa/rojo/distribs.htm. In this file, the range parameter ρ is represented by A and B, with $A = -\rho$ and $B = \rho$. The shape parameter α is represented by p, and β by q.

normal distribution with its infinite range, there would be no principled limit on the extent to which non-grammatical factors could influence the application of variation. See Sect. 5.1 for more detailed discussion.

3 English *t/d*-deletion

Word-final *t/d* variably deletes from consonant clusters in English, so that a word like *west* can be pronounced as [wɛst] or [wɛs]. This deletion process has been described in detail for countless dialects of English (see Coetzee 2004: Chap. 5 for a review), and even for languages other than English (on Dutch, see Goeman 1999; Goeman and van Reenen 1985; Schouten 1982, 1984). Since this process has been studied so extensively, the factors (both grammatical and non-grammatical) that influence its application are reasonably well understood. We begin this section by first reviewing some of the grammatical and non-grammatical factors that are known to influence this process, focusing on those aspects for which we will provide an account. We then develop a purely grammatical account in the noisy HG framework. Once the grammatical account has been established, we augment it to account for the influence of usage frequency according to the method described above in Sects. 2.2 and 2.3.

3.1 Grammatical and non-grammatical influences

We first review evidence that this process is influenced by the same kinds of grammatical considerations as those that influence "ordinary" non-variable phonological rules. Echoing an idea that has been present throughout the variationist research tradition for nearly four decades, Anttila (1997:44) takes this fact to be a motivation for expecting phonological grammar to account for at least part of variation: "... if variation preferences are based on phonological variables, then it seems reasonable to expect phonology to make sense of them."

In a summary of the grammatical factors that influence *t/d*-deletion, Labov (1989) includes the following: (i) *Stress*: *t/d* is more likely to delete from an unstressed syllable (*cúbist*) than a stressed syllable (*insíst*); (ii) *Cluster size*: Deletion is more likely from tri-consonantal (*tanked* [tæŋkt]) than from bi-consonantal clusters (*tacked* [tækt]); (iii) *Similarity to preceding segment*: Deletion is more likely after consonants that share more features with *t/d* than consonants that share fewer features—there is more deletion from *kissed*, where [s] shares place (coronal) and sonorancy (non-sonorant) with the following [t] than from *seemed*, where [m] shares no major features with the following [d]; (iv) *Morphology*: *t/d* that functions as the past tense suffix of a regular past tense verb (*missed*) is less likely to delete than *t/d* that functions as the past tense suffix in a semi-weak verb (*kept*), which is less likely to delete than *t/d* that is part of a morphological root (*mist*).

Another grammatical factor that influences *t/d*-deletion is the context that follows the word-final *t/d*. We use this factor as an example of a grammatical factor in the rest of this section, and will therefore discuss it in more detail. In every dialect of English for which *t/d*-deletion has been studied, it has been found that deletion is most likely if the next word begins with a consonant (*west bank*). Dialects diverge on

whether a following vowel-initial word (*west end*) or a pause (*west.*) results in more deletion. The table in (12) contains a sample of the data available on the influence of the following context.[7] The data on all but Columbus English are taken from the literature, with references given in footnote 8.

(12) Percent *t/d*-deletion in different English dialects in pre-consonantal, pre-vocalic, and pre-pausal contexts.[8]

Relative deletion rate		Pre-C *west bank*	Pre-Pause *west.*	Pre-V *west end*
Pre-C > Pre-Pause > Pre-V	AAVE (Washington, DC)	76	73	29
	Jamaican English	85	71	63
	Tejano English	62	46	25
	Trinidadian English	81	31	21
Pre-C > Pre-V > Pre-Pause	Chicano English	62	37	45
	Columbus English	80	63	76

The data on Columbus English were extracted from the *Buckeye Corpus* (Pitt et al. 2007). This is a corpus of conversational speech collected from 40 lifelong residents of Columbus, Ohio. All of the speech was both orthographically and phonetically transcribed. In order to compile a list of words from the corpus to which *t/d*-deletion could apply, we extracted all words that end orthographically in *-Ct* or *-Cd* (where *C* stands for any consonant). Since *t/d* that corresponds to the past tense suffix is consistently treated differently (see discussion above), and since our focus is on the influence of the phonological context, we excluded words with this suffix. The principle by which we selected tokens from the corpus already excluded past tense forms that end orthographically in *-ed*. We manually removed the semi-weak past tense forms, such as *kept*. We also removed a few other classes of words. First, due to the difficulty of determining whether word-final *t/d* has been realized before a word that starts with [t] or [d], we removed all such tokens from the list. Secondly, we removed words that end orthographically in *-rt/-rd* or *-lt/-ld*. These tokens showed unexpectedly low deletion rates in the corpus. In these tokens, *r* and *l* were often phonetically realized as coloring on the preceding vowel rather than as a separate consonant, so that *-rt/-rd* and *-lt/-ld* words often do not actually end in consonant clusters phonologically (Guy and Boberg 1997). Lastly, we removed words such as *thought* and *could*, that end orthographically but not phonologically in *-Ct/-Cd*. This whole procedure left a list of 16,460 tokens, representing 459 different words. The phonetic transcription in the corpus for each of the token words was then consulted, and each token was coded as either "*t/d* deleted" or "*t/d* retained".[9] Each token was also classified as

[7] These data are simplified with regard to the pre-consonantal context. Labov (1989) and Guy (1991), among others, show that *t/d*-deletion rates are different before consonants of different types. We follow the practice in the vast majority of the *t/d*-deletion literature of lumping all of the consonants together.

[8] Sources: AAVE (Fasold 1972), Jamaican (Patrick 1992), Tejano (Bayley 1995), Trinidad (Kang 1994), Chicano (Santa Ana 1991).

[9] A token was coded as "*t/d* deleted" if no segment was transcribed for the underlying *t/d*. In the *Buckeye Corpus*, underlying *t/d* was transcribed with several different surface realizations, including faithful realizations [t] or [d], glottalized realizations [t^{ʔ}] or [d^{ʔ}], flap [ɾ], etc. All tokens transcribed with one of these

pre-consonantal, pre-vocalic, or pre-pausal based on the context in which the token appeared in the corpus.[10,11]

Several non-grammatical factors that influence the application of *t/d*-deletion have also been documented in the variationist literature. For example, biographical factors, such as the age, sex, or ethnicity of the speaker, have been shown to influence application of the process. Additionally, speech register also influences the deletion rate, with less formal registers associated with higher deletion rates. Browman and Goldstein (1990), for instance, found little evidence of *t/d*-deletion in the reading of a word list, but they did find evidence for the process in a more casual conversational speech style. Mitterer and Ernestus (2006) studied the analogous process in Dutch in two speech corpora. One corpus consisted of read speech (literally, novels read on tape for the blind)—i.e., a rather formal speech register. The other corpus consisted of recordings of casual speech. They found evidence of deletion in both corpora, but at very different rates (8 % for the read speech vs. 45 % for the casual speech).

The non-grammatical factor on which we focus is usage frequency, and we therefore report on it in more detail. As we already showed in Sect. 1.3, phonological processes such as *t/d*-deletion usually apply at higher rates to words of higher frequency—i.e., there is more deletion from frequent *just* than from phonologically similar but infrequent *jest*. Bybee (2000:69–70), for instance, analyzes Santa Ana's 1991 corpus of Chicano English, and finds a deletion rate of 54.4 % in high frequency words compared to 34.4 % for low frequency words.[12] Phillips (2006:65) shows that frequency has the same influence in the analogous process in Dutch.

In order to investigate the influence of frequency on *t/d*-deletion in the *Buckeye Corpus*, we determined the frequency of each of the words that we selected from this corpus in CELEX (Baayen et al. 1995), and then transformed these counts by

realizations were coded as "*t/d* retained". Since the corpus contains no articulatory data, deletion is defined here as the absence of any acoustic evidence of *t/d*. An actually articulated *t/d* might not have any acoustic realization when it is articulated before a labial consonant. If the labial closure of the following consonant is made before the release of the *t/d*, the potential acoustic effect of the coronal release is masked by the labial closure, and hence becomes inaudible (Browman and Goldstein 1990). The actual articulatory *t/d*-deletion rate before consonants may therefore be somewhat lower than the acoustic rate reported here. As a check of the potential influence that this acoustic masking could have on our data, we counted the number of tokens in our pre-consonantal category followed by labial and non-labial consonants. We found that more than 80 % of the pre-consonantal tokens appear before non-labial consonants.

[10]The coding conventions in the *Buckeye Corpus* do not actually include a category for pauses. We coded as pre-pausal the following tokens: (i) tokens where the corpus indicates that silence followed an utterance; (ii) tokens where the corpus indicates that an utterance was followed by the interviewer speaking, and where it was clear from the context that the interviewer did not interrupt the interviewee mid-utterance; (iii) utterances followed by some kind of non-speech vocalization noise, and where the context made it clear that this vocalization noise did not occur mid-utterance.

[11]The corpus of *t/d*-words that we used is available as "supplementary material" on the Springer link for this article, or from the first author upon request.

[12]Bybee (2001) and Jurafsky et al. (2001:252–255) show that mere lexical usage frequency does not capture the full influence of frequency. Just as important, and in some instances maybe even more important, is frequency of use within a specific syntagmatic context. That is, the [t] in *best* may delete more often from a more frequent phrase such as *best friend* than from a less frequent phrase such as *best fruit*. Although an adequate account of phonological variation will ultimately have to incorporate all relevant types of frequency influences (and all other relevant influences), we will focus only on lexical usage frequency in this article.

taking their logarithms (with base 10).[13,14] Because the *Buckeye Corpus* is relatively small, words with a low CELEX frequency count appear infrequently in the corpus. (In fact, several words appear only once.) It is consequently not possible to calculate reliable deletion rates for individual words, and we therefore divided the words into frequency bins before calculating deletion rates (cf. also Bybee 2000:69–70; Lacoste 2008:188–189). Most of the frequency bins spanned 0.1 intervals on the log-transformed frequency values. If some bin contained fewer than 50 tokens, we combined it with one of its adjacent bins so that a few bins spanned a wider range than 0.1. In total, 23 frequency bins were created ranging in log-transformed frequency from (0 to 2.0) up to (5.7 to 5.8).[15] The deletion rate in each of the three contexts (pre-vowel, pre-consonant, pre-pause) was then calculated for each frequency bin. This procedure gives a data set where deletion rates in each of the contexts can be plotted against frequency to look for a correlation, as in Fig. 3. This figure shows a positive correlation between frequency and deletion rate in all three contexts. In fact, the correlation is significant in all three contexts (Pre-C: $r^2 = 0.46$, $p < 0.01$; Pre-V: $r^2 = 0.39$, $p < 0.01$; Pre-Pause: $r^2 = 0.43$, $p < 0.01$).[16]

In the next section, we first develop an account for the influence of the following phonological context on *t/d*-deletion in Columbus English, as given in (12). In doing this, we will abstract away from the influence of usage frequency, shown in Fig. 3.

[13]Since log of zero is undefined, a constant of one was added to all frequencies before they were log-transformed.

[14]One could raise some concerns about using CELEX to measure usage frequency. First, CELEX is a British corpus, and usage frequency may differ between CELEX and the American speakers included in the *Buckeye Corpus*. Second, although CELEX includes some spoken sources, the majority of the frequency counts in CELEX come from written texts. Usage frequency may be different between spoken and written language.

A possibly more accurate measure of the usage frequency of words for the speakers who contributed to the *Buckeye Corpus* would be the *Buckeye Corpus* itself—i.e., just counting the frequency with which each token appears in the corpus. However, since the *Buckeye Corpus* is comparatively small, it does not differentiate well between words with low usage frequencies—many words appear only once in the corpus. Facing the same problem with regard to the *Buckeye Corpus* and CELEX, Raymond et al. (2006) showed that CELEX and *Buckeye* frequencies are highly correlated ($r = 0.82$). In fact, using CELEX for frequency counts, even when dealing with American English, is standard practice in the field (Albright 2009; Coetzee 2005, 2008). We therefore follow the standard practice, using CELEX for frequency counts in our study.

[15]The decision to use 23 frequency bins is to some extent arbitrary. A finer-grained division into more bins could potentially give a more detailed picture of how usage frequency interacts with deletion. However, relying on more bins also results in some bins containing too few data points to reliably calculate deletion rates. There is a trade-off between the reliability of the deletion rate for each frequency bin and the fine-grainedness with which the frequency range is sampled. We decided to use bins that contain at least 50 tokens each, resulting in the 23 bins used here.

[16]On each of the three graphs, there is one data point with an extremely high log frequency, just below 6. This data point corresponds to the word *and*, which accounts for more than half of all the tokens in our corpus. If this data point is removed, the positive correlation between frequency and deletion rate remains, even if it is less strong (Pre-C: $r^2 = 0.21$, $p < 0.05$; Pre-V: $r^2 = 0.22$, $p < 0.05$; Pre-Pause: $r^2 = 0.14$, $p < 0.11$). Due to the fact that extremely high frequency words such as *and* show much higher deletion rates, these words are often excluded from the data sets used in variationist sociolinguistic studies of *t/d*-deletion (Patrick 1992:172). By including frequency as a factor in our model, we do not have to exclude frequent words. Their seemingly anomalous behavior is no longer anomalous, but rather expected given the model that we develop.

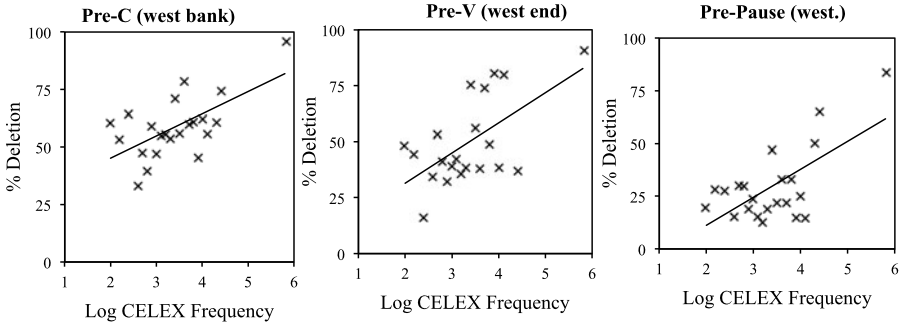

Fig. 3 The relation between frequency and deletion rate in Columbus English in Pre-C, Pre-V and Pre–Pause contexts. The *x*-axis represents log-transformed CELEX frequencies. Deletion rate is plotted on the *y*-axis

Once this grammatical account is in place, we will augment it to incorporate the influence of frequency.

3.2 A grammatical account

In this section we develop a noisy HG account for the overall deletion rates observed in Columbus English, as shown in the table in (12). For similar accounts of the other data from this table, see Coetzee and Pater (2011).

3.2.1 Constraints

The constraints that we use are given in (13). The two contextual faithfulness constraints are in the spirit of Steriade's "licensing by cue" constraints—i.e., they protect segments from deletion in contexts where the cues for their perception are saliently licensed (Steriade 1999, 2001; Côté 2004).

(13)

*CT]$_{\text{Word}}$	Assign one violation mark for every word that ends in the sequence [-Ct] or [-Cd].[17]
MAX	Assign one violation mark for every input segment lacking an output correspondent (no deletion). (McCarthy and Prince 1995:371)
MAX-PRE-V	Assign one violation mark for each segment that appears in pre-vocalic context in the input, and that does not have a correspondent in the output (no deletion before a vowel). (Côté 2004:22)

[17]This constraint is a special version of the more general *COMPLEX, which applies only to a subclass of consonant clusters, and only when these clusters appear in word-final position. As it stands, the constraint is too specific. For instance, deletion of [p] from words like *ramp*, *wisp*, etc., and deletion of [k] from words like *whisk*, *task*, etc. are also observed. To account for these deletions, the constraint should probably be generalized so that it penalizes all [...C+stop] sequences. However, the literature contains virtually no information on the deletion of [p] and [k], probably because there are so few [...Cp] and [...Ck] words in English. For this reason, we assume the more specific constraint here. See Coetzee (2004: Chap. 5) for an exploration of a more general constraint.

MAX-PRE-PAUSE Assign one violation mark for each segment that appears in pre-pausal context in the input, and that does not have a correspondent in the output (no deletion before a pause).

Steriade proposes that a segment is protected by special faithfulness constraints in contexts where its perceptual cues are robustly licensed. The consonant release burst can cue both place (Lahiri et al. 1984; Stevens and Blumstein 1978) and manner information (Stevens and Keyser 1989). The formant transitions into a following vowel also carry information about both place (Martínez-Celdrán and Villalba 1995; Eek and Meister 1995; Fowler 1994; Fruchter and Sussman 1997; Kewley-Port 1983; Kewley-Port et al. 1983; Nearey and Shammas 1987; Stevens and Blumstein 1978; Sussman et al. 1991; etc.) and manner (Diehl and Walsh 1989; Walsh and Diehl 1991). To motivate the existence of the positional versions of MAX, it is therefore necessary to show that release bursts and formant transitions are more robustly licensed in pre-vocalic and pre-pausal position than in pre-consonantal position.

In pre-consonantal position, the likelihood of a consonantal release being realized is relatively small. Zsiga (2000:78) reports a release rate of as low as 18 % in this context for English (see also Browman and Goldstein 1990). Except when the following consonant is a sonorant, there is also no opportunity for the realization of formant transitions, and even into a following sonorant, robust transitions are less likely than into a following vowel. Pre-consonantal position is hence the context in which *t/d* is least well cued, so that there is no special faithfulness constraint that protects against deletion specifically in this context.

In pre-pausal position, formant transitions into a following segment cannot be realized. However, it is possible to release stops in this position—Byrd found that 57 % of alveolar stops were released in the TIMIT corpus (Byrd 1992:37). There is also evidence that utterance-final released consonants are perceived more accurately than unreleased consonants (Malécot 1958). In pre-vocalic position, both formant transitions and releases can be realized. Only one of the cues can therefore be realized pre-pausally while both cues can be realized pre-vocalically. On the other hand, the pre-vocalic cues can only be realized across a word boundary. The crossing of the word boundary may result in a penalty for cue robustness in pre-vocalic position. The listener may, for instance, incorrectly perceive the *t/d* as the first segment of the following word rather than the last segment of the preceding word. As such, the additional acoustic cue available in this context would not necessarily result in easier perception and lexical access for the listener. A question is whether there is a universal difference in cue robustness between pre-pausal and pre-vocalic contexts. In Steriade's "licensing by cue" model of faithfulness, constraints protecting inherently more robust sponsoring contexts universally rank higher than constraints protecting less robust sponsoring contexts. If there is an inherent robustness difference between pre-vocalic and pre-pausal contexts, the two positional versions of MAX will therefore be in a universally fixed ranking.

Exactly how the ranking between cue-licensing constraints is established is still an unresolved topic. These rankings could be hard-wired into Universal Grammar or they could emerge during acquisition, influenced by misperception on the side of the

language learner (Boersma 2008). In English dialects where pre-pausal *t/d* is seldom released, the child acquiring the grammar will more often not perceive *t/d* in this position, even if his/her parents actually produced *t/d* in this context. Such a learning situation might lead to the lower ranking of the constraint MAX-PRE-PAUSE in the grammar of such a child. On the other hand, a child acquiring a dialect where pre-pausal stops are more often released may actually perceive *t/d* more often in this context, resulting in a higher ranking of MAX-PRE-PAUSE in the grammar of such a child. The rankings could therefore result from the concrete experience of the language learner as a listener. This is also in agreement with Kawahara's claims that rankings between cue-based faithfulness constraints are based on the actual perceptibility of contrasts in different contexts (Kawahara 2006). On the other hand, Moreton (2008, 2010) has shown that some typological tendencies may result from hard-coding of rankings into UG rather than from experience with actual perceptibility.

Given that this issue is still unresolved, we will not take a stance here on how exactly the ranking between cue-based faithfulness constraints comes about. We do note that, given the data reported in (12), it is necessary to allow MAX-PRE-PAUSE and MAX-PRE-V to rank differently in the grammars of different dialects/languages in order to account for the difference between dialects that show more deletion in pre-pausal position and those that show more deletion in pre-vocalic position.

3.2.2 The learning simulation and results

The constraint weights for Columbus English were determined by running a learning simulation with *Praat*'s noisy HG learning algorithm (Boersma and Weenink 2009). For details on this learning algorithm, see Boersma and Pater (2008) and Coetzee and Pater (2008). In creating an input file for the algorithm, we assumed that each of the contexts (pre-consonantal, pre-vocalic, pre-pausal) appears 100 times. Deletion was represented in the 100 tokens in each context proportional to the overall deletion rates from (12)—i.e., in pre-consonantal context, 80 tokens were coded as pronounced with deletion and 20 with a final *t/d*, in pre-pausal context 63 with deletion and 37 with retention, and in pre-vocalic context 76 with deletion and 24 with retention.[18] We based the learning input file on the overall deletion rate, following the tradition in the literature. The account that we develop here will therefore not take into account the contribution of the usage frequency of individual words. In the next section, we will augment our account by implementing weight scaling. In running the learning simulation, we set the "decision strategy" to "Linear OT" (*Praat*'s implementation of the noisy HG learning algorithm). All other settings were kept at *Praat*'s defaults.[19]

[18]The *Praat* input file is available as "supplementary material" on the Springer link for this article, or from the first author upon request.

[19]In particular, the following settings were used: (i) The initial weights of all constraints were set to 100. Changing the initial weights may influence the speed of learning, but as long as sufficient learning time is allowed, it will not influence the final grammar that is learned; (ii) An evaluation noise of 2.0 was used. Changing the evaluation noise may influence the absolute difference in weight between constraints, but will not influence the eventual performance of the grammar; (iii) The initial plasticity was set to 1.0, with 4 decrements of 0.1 in plasticity at every 100,000 replications. As explained by Boersma and Hayes (2001) with regard to their GLA for stochastic OT, starting out with a higher initial plasticity results in faster initial

Once the grammar had been learned, *Praat*'s "To output Distributions" function was used to test the predicted output of the grammar.[20]

The constraint weights that were learned are given in (14). Before this grammar is used to evaluate output candidates, noise is added to the constraint weights. In the noisy HG implementation in *Praat*, this noise is randomly selected from a normal distribution with a mean of zero. Under the default *Praat* setting, the standard deviation of the distribution is 2.[21]

(14)

*CT]$_{\text{Word}}$	101.16
Max	98.84
MAX-PRE-V	−1.51
MAX-PRE-PAUSE	0.96

In (15), we show the output patterns generated by a grammar with the weights in (14). As expected, there is a close match between the observed deletion rates (on which the learning input file was based), and the deletion rates predicted by the grammar. As has been shown before, noisy HG can replicate variation rates extremely well (Coetzee 2009a; Coetzee and Pater 2011; Jesney 2007). However, as we had shown earlier, words of different frequencies are subject to deletion at very different rates. Since high frequency words contribute more to the overall deletion rate, the deletion rate predicted by the grammar learned in this section, based on the overall deletion rate in the corpus, is relatively close to the deletion rates observed for high frequency words. Low frequency words, on the other hand, show deletion rates that are considerably lower than this overall deletion rate. In the next section, we augment this grammar to take into account the difference between words of different frequencies.

(15)

Context	Observed deletion	Expected deletion
Pre-C	80 %	79.6 %
Pre-V	76 %	76.2 %
Pre-Pause	63 %	62.8 %

3.3 Incorporating the frequency bias through weight scaling

In order to apply weight scaling, the scaling factors for words of different frequencies need to be determined, and to do that, the values of the parameters (α, β, and ρ) of

learning. Decreasing plasticity later in learning results in more accurate frequency matching of the learning input. An equally good grammar could be learned by starting out with a small plasticity, but more learning time might be required.

[20] For this production-oriented simulation, we also used *Praat*'s default settings: (i) An evaluation noise of 2.0 was used—the same value used during the learning simulation; (ii) Each input type (pre-consonantal, pre-vocalic and pre-pausal) was submitted to the grammar 100,000 times, and the frequency with which each output candidate (deletion or retention) was selected was tallied.

[21] If the sum of a constraint's weight and the noise added to this weight at a particular evaluation occasion is less than zero, *Praat* resets it to zero during evaluation. This adjustment prevents a candidate from being rewarded in its H-score for violating a constraint—a negative constraint weight multiplied by the negative integer used to mark constraint violation would have increased the H-score.

the *beta* distribution associated with words of different frequencies need to be determined. We start by showing how the values of α and β are determined. As explained in Sect. 2.3, the value of α is set to the logarithm of the reference frequency—i.e., that frequency that divides the words into the frequent and infrequent sets. Following the procedure illustrated in (9) in Sect. 2.3, the log reference frequency, α, for our Columbus English *t/d*-deletion corpus was determined to be 5.1. For all words, the value of α is hence set to 5.1. The value of β is set to the log frequency of the bin to which the word belongs. For the word *and*, for instance, α is set to 5.1, and β to the log bin to which *and* belongs, namely 5.8.

As shown in Sect. 2.3, ρ only influences the size of the scaling factors and not their signs. Its role is to determine how much influence usage frequency (via weight scaling) can have on the functioning of the grammar. We propose that the value of ρ is determined by fitting the model to the data. This value therefore has to be determined separately for each language (represented by some corpus). To determine the value of ρ that results in the best fit to our data, we ran multiple simulations, keeping the values of α and β constant while increasing the value of ρ by whole number steps from 1 upwards. We then compared the weight scaled models with the baseline model without weight scaling in terms of their mean square errors relative to the observed deletion rates. The improvement of the weight scaled grammars at different integer values of ρ could then be compared, and the value of ρ could be selected where the improvement reaches its maximum.[22]

(16) Scaling factors for words of different frequencies, at different values of ρ

		ρ						
		Baseline	3	4	5	6	7	
Frequency bins	2.0	0.0	1.82	2.43	3.04	3.65	4.26	Scaling factor
	2.6	0.0	1.32	1.76	2.20	2.63	3.07	
	3.0	0.0	1.03	1.38	1.72	2.06	2.41	
	3.5	0.0	0.73	0.97	1.21	1.45	1.69	
	4.0	0.0	0.47	0.62	0.78	0.93	1.09	
	4.4	0.0	0.28	0.37	0.47	0.56	0.65	
	5.8	0.0	–0.24	–0.32	–0.40	–0.47	–0.55	

In (16), we give the scaling factors associated with words belonging to different frequency bins in our corpus at different values of ρ. As the frequency increases (top to bottom), the scaling factors decrease, corresponding to the fact that faithfulness constraints play a less important role in the evaluation of more frequent words. For the most frequent frequency bin (5.8), the scaling factor is negative, since for words in this bin α (the reference value, 5.1) is smaller than β (the log frequency of the bin, 5.8), resulting in a right-skewed *beta* distribution with a negative mode. As the value of ρ increases (from left to right), the absolute values of all the scaling factors increase, even though their signs do not change. This correlation corresponds to the fact that frequency has a larger influence (via the scaling factors) at larger val-

[22] Using whole number increments for ρ is motivated by practical considerations. If smaller increments were used, it is possible that a slightly better fit could be achieved.

ues for ρ. The "baseline" column represents the basic grammar without frequency scaling.

Once the scaling factors for words of different frequencies at different values of ρ have been determined, weight scaling can be implemented formally. We use the scenario with $\rho = 5$ as an example. The same procedure is followed for all other values of ρ. The scaling factors listed in (16) represent the amount with which the weight of each faithfulness constraint has to be increased or decreased in the evaluation of words with a specific usage frequency. For instance, when evaluating a word with a usage frequency of 2.0 the weight of all faithfulness constraints has to be increased by 3.04. When evaluating a word with a frequency of 5.8, the weight of all faithfulness constraints has to be decreased by 0.40, etc. In (17), we show the weight scaled grammars for different frequency bins when $\rho = 5$. To get these grammars we added the scaling factors from (16) to the faithfulness constraint weights of the baseline model from (14). Once these weight scaled grammars were determined, we manually edited the *Praat* grammar file for the baseline model that was learned in Sect. 4.2 above. Specifically, we created separate grammar files for each of the different frequency bins by changing the weights of the faithfulness constraints according to the scaling factor for each of the frequency bins, as reflected in (17). Once different grammar files for each frequency bin have been created, we again used *Praat*'s "To output Distributions" function to determine the deletion frequency predicted by each of these frequency scaled grammars.

(17) Frequency scaled grammars at $\rho = 5$

		Scaling factor	*CT]$_{\text{Word}}$	MAX	MAX-PRE-V	MAX-PRE-PAUSE
Baseline		0	101.16	98.84	–1.51	0.96
Frequency bin	2.0	3.04	101.16	101.88	1.53	4
	2.6	2.20	101.16	101.04	0.69	3.16
	3.0	1.72	101.16	100.56	0.21	2.68
	3.5	1.21	101.16	100.05	–0.3	2.17
	4.0	0.78	101.16	99.62	–0.73	1.74
	4.4	0.47	101.16	99.31	–1.04	1.43
	5.8	–0.40	101.16	98.44	–1.91	0.56

In (18) we show the deletion rates in pre-consonantal position predicted for a selection of frequency bins, at the different values of ρ from (16). Since frequency has no influence in the baseline grammar, the same deletion rate is expected for all frequency bins. For all of the other values of ρ, deletion rates increase as frequency increases (top to bottom), given that the scaling factors decrease as frequency increases. Lower scaling factors imply lower effective weights for faithfulness constraints, and hence higher rates of unfaithfulness. For all but frequency bin 5.8, deletion rates decrease as the value of ρ increases (left to right). These frequency bins represent words that appear less often than the reference frequency, and as shown in (16), these bins are therefore associated with positive scaling factors. Also shown in (16) is that the values of the scaling factors increase with ρ. At higher values of ρ, the faithfulness constraints will hence have higher effective weights, and therefore exert more

influence on the selection of the output, with the resulting higher rates of faithfulness. Frequency bin 5.8 is the only bin with a frequency higher than the reference frequency of 5.1. As shown in (16), the scaling factors associated with this bin are hence negative, and decrease as ρ increases. As a result, for this frequency bin, deletion rates increase as ρ increases. The contribution of ρ to the model should now be clear. Higher values of ρ result in an increased contribution of frequency to the selection of the output. If a word is frequent and therefore has a higher than overall deletion rate, its deletion rate will be even higher at higher values of ρ. On the other hand, if a word is infrequent and therefore has a lower than overall deletion rate, its deletion rate will be even lower at higher values of ρ.

(18) Predicted deletion rates (%) in pre-consonantal context at different values of ρ

		ρ						
		Baseline	3	4	5	6	7	
Frequency bins	2.0	79.4	56.9	48.4	39.7	32.0	24.5	Predicted % deletion
	2.6	79.4	63.7	58.0	51.4	45.9	39.5	
	3.0	79.4	67.7	63.0	57.9	53.6	48.9	
	3.5	79.4	71.1	68.3	64.4	62.0	58.6	
	4.0	79.4	74.5	72.8	70.3	68.5	67.1	
	4.4	79.4	76.5	75.6	74.1	73.2	72.2	
	5.8	79.4	81.6	82.4	82.9	83.9	84.2	

The table in (19) compares the performance of the model at different values for ρ in terms of mean square errors.[23] For each value of ρ, we also give the percentage of improvement of the model relative to the baseline model. The performance of the model steadily increases up to a value of 5 for ρ, after which it starts declining again. Based on this, we set the value of ρ for the Columbus English *t/d*-deletion corpus at 5. Figure 4 shows the performance of the baseline model relative to a frequency scaled model with $\rho = 5$. The broken line represents the baseline model, and the solid line the frequency scaled model. The scaled model predicts a higher than overall deletion rate for words in frequency bin 5.8, and lower than overall deletion rates for other frequency bins. This figure also shows that the frequency scaled model fits the data better than the baseline model. In fact, as shown in (19), it improves on the baseline by nearly 80 %.

[23] Mean square error is calculated according to the formula $\sum_{i=1}^{n}(P_i - O_i)^2$, where P_i is the value predicted for observation i, and O_i the observed value for observation i. This value is an overall index of the deviation between the model prediction and the actually observed data. Improvement relative to the baseline model is calculated by first determining the difference in mean square error between the baseline and the model being evaluated—this difference represents the improvement of the new model relative to the baseline in terms of mean square error. This difference is then converted into an improvement percentage. For instance, to determine the improvement of a model with $\rho = 5$ relative to the baseline in (19), we first determine the difference in mean square error between the two models (i.e., $1009.7 - 208.2 = 801.5$). We then convert this to a percentage (i.e. $801.5/1009.7 \times 100 = 79.4$ %).

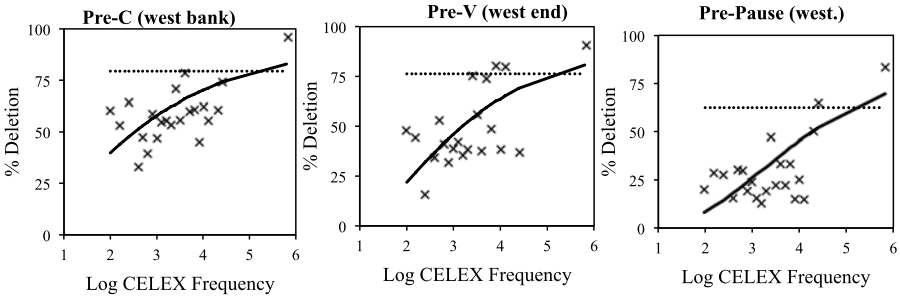

Fig. 4 Observed and predicted *t/d*-deletion rates in Columbus English. The *broken line* indicates the predictions based on the baseline, unscaled HG. The *solid line* shows the predictions based on the frequency weighted HG with a ρ-value of 5

(19) Mean square errors and percentage of improvement relative to the baseline, unscaled grammar at different values of ρ

		ρ				
	Baseline	3	4	5	6	7
Mean Square Error	1,009.7	354.0	280.4	208.2	218.8	425.9
Improvement Percentage Relative to Baseline	0 %	64.9 %	72.2 %	79.4 %	78.3 %	57.8 %

The fact that the scaled model fits the data better is not surprising—the scaled model incorporates one more parameter (frequency) than the baseline model, and given that frequency significantly impacts application of *t/d*-deletion, it is to be expected that a model with this additional parameter will fit the data better. To determine whether this improvement of 80 % is sufficient to warrant the additional complexity we used the Akaike Information Criterion (AIC; Akaike 1973, 1983). Roughly speaking, AIC is an estimate of the amount of information lost when using a specific model relative to the true model. A smaller AIC value associated with a model therefore indicates that the model more closely approximates the true model. To calculate the AIC for the baseline and scaled models, we use the partial AIC derivation (Burnham and Anderson 2004:268–269). The formula used is given in (20) where *MSE* is the mean square error associated with a model and k is the number of parameters used in the model. A model that fits the data better will have a smaller *MSE* and hence, all else being equal, a smaller (or better) AIC. On the other hand, the larger the number of parameters included in a model, the larger k will be. All else being equal, a model with more parameters will therefore have a higher (or less good) AIC than a model with fewer parameters. AIC therefore rewards a model for a better fit with the data (lower *MSE*), but penalizes a model for including more parameters (higher k), so that AIC gives a measure of the tradeoff between model complexity and model fit. The value of n is the number of observations in the dataset being modeled.

(20) $$\text{AIC} = n \log_e(MSE) + 2k$$

In calculating AIC for the baseline and frequency scaled models, we assume that each of the constraints in our HG grammar counts as one parameter. The baseline

model therefore has 4 parameters. The frequency scaled model has one additional parameter (i.e., 5) due to the addition of the frequency scaling factor to this model. Using the *MSE*s for the respective models reported in (19) above, the formula in (20), and setting $n = 65$ (since there are 65 total data points in the corpus), the AICs for the two models can be calculated: $AIC_{Baseline} = 457.6$, $AIC_{Scaled} = 357.0$. As Burnham and Anderson (2004:271) note, a model with an AIC that is more than 10 units larger than the best model has "essentially no support". Given that the frequency scaled model has an AIC that is 100 units smaller than the baseline model, we can hence conclude that the additional complexity of the scaled model is well warranted by the better fit that this model achieves relative to the baseline model.

4 Geminate devoicing in borrowings in Japanese

4.1 The data

In this section, we present another case study to show the generality of the model that we developed above. Although Japanese native phonology does not tolerate voiced geminates, these sounds have been introduced into Japanese via borrowings. Due to Japanese coda restrictions (Itô 1988), closed syllables are frequently borrowed with an epenthetic vowel. Additionally, when the coda consonant in a borrowed word is preceded by a lax vowel, the consonant is often geminated (Katayama 1998). When the coda consonant is also a voiced obstruent, the combination of these processes results in a voiced geminate. In words that contain another voiced obstruent, the geminate optionally devoices, as in the examples in (21) (all examples from Kawahara 2006:538).

(21)

guddo	~	gutto	'good'
beddo	~	betto	'bed'
deibiddo	~	deibitto	'David'
doggu	~	dokku	'dog'
baggu	~	bakku	'bag'
doraggu	~	dorakku	'drug'
biggu	~	bikku	'big'

This optional devoicing in loanwords has received a lot of attention in recent years so that the factors that condition its application are now well understood. We refer the reader to the literature for a discussion of these factors (Crawford 2009; Kaneko and Iverson 2009; Kawahara 2005, 2006, 2008, 2011a, 2011b; Nishimura 2003, 2006; Tanaka 2009 and references cited there). Our focus here will be on how this process is influenced by usage frequency. In two recent studies, Kawahara has found a strong positive correlation between geminate devoicing and word frequency (Kawahara 2011a, 2011b). We will develop an account of the results of Kawahara (2011a) here. We summarize the most important aspects of his results below, and refer the reader to the original paper for more details on the design of the experiment.

Kawahara presented 52 native Japanese speakers with 28 loanwords like those in (21) with the task of rating the naturalness of a pronunciation in which the voiced

geminate has been devoiced. Participants indicated their responses on a 5-point scale, with [5] corresponding to "very natural", and [1] to "very unnatural". The raw usage frequency of each loan word token was taken from the Amano and Kondo Japanese lexical corpus (Amano and Kondo 2000), and log-transformed. Figure 5 plots the average naturalness rating that each token received against its log-transformed frequency. Performing a linear regression on these data confirms that log frequency and naturalness are positively correlated ($r^2 = 0.43$, $p < 0.01$).

The best way to collect data on devoicing rates in actual speech production would be to investigate the prevalence of devoicing in a large, phonetically transcribed corpus of spoken Japanese—similar to how we investigated the prevalence of *t/d*-deletion in the *Buckeye Corpus* above. Unfortunately, no such corpus exists for Japanese that is large enough to contain enough examples of loanwords. A second option would be to conduct a production experiment, designed to collect data on loanwords. Participants in such experiments usually use a rather formal speech style in which optional processes, such as geminate devoicing, are often inhibited. We therefore work under the assumption that naturalness ratings such as those in Kawahara (2011a) originate in the same grammar that governs speech production, and that these naturalness ratings therefore also reflect the frequency with which devoicing will apply to the loanwords in actual speech. Even if this is accepted, it is still necessary to convert the 5-point naturalness scale to devoicing rate in some manner. Little is known about how naturalness ratings are related to production patterns (though see Kempen and Harbusch 2008 for some ideas involving syntactic data), and we therefore explored several different options for transforming the naturalness ratings of Kawahara (2011a) to devoicing rates. In all of the transformations that we explored, the positive correlation between frequency and rate of devoicing was preserved. We report here on only one of these transformations, a simple linear transformation.[24] This is the transformation on which our model had the best performance.

In order to transform the natural ratings to devoicing rates, we made the assumption that a rating of [5] corresponds to a token that is always produced with devoicing, a rating of [4] to a token that is produced with devoicing four-fifths of the time (i.e., with 80 % devoicing), etc. The formula used to transform the naturalness ratings is given in (22). Figure 6 plots the deletion rate under this transformation against the log frequency of the tokens. As this figure shows, the correlation between frequency and devoicing is preserved under this transformation ($r^2 = 0.43$, $p < 0.01$).

[24] Specifically, in addition to the linear transformation defined in (22), we also used an exponential and sigmoid transformation. The formulas used in these two transformations are given below. Under both of these transformations, the positive correlation between frequency of devoicing and usage frequency is preserved: exponential: $r^2 = 0.34$, $p < 0.01$; sigmoid: $r^2 = 0.41$, $p < 0.01$.

Let r be the average naturalness rating that some token t received, and $devoice(t)$ the rate of devoicing in token t. Let $norm^r$ be the standardized value of r. Then:

$$\text{Exponential transformation} \quad devoice(t) = \left(\frac{e^r}{e^5}\right)(100)$$

$$\text{Sigmoid transformation} \quad devoice(t) = \left(\frac{1}{1 + r^{-norm^r}}\right)(100)$$

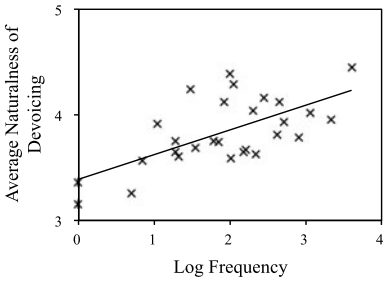

Fig. 5 The relation between frequency and devoicing in Kawahara (2011a). The *x-axis* represents log-transformed frequencies from Amano and Kondo (2000). The naturalness rating of devoicing is plotted on the *y-axis*. The *line* indicates the best-fit linear regression line

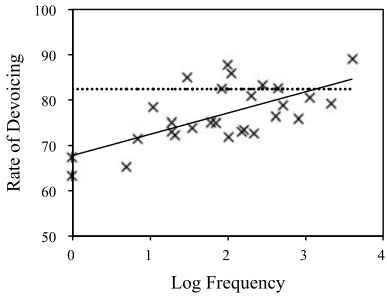

Fig. 6 The relationship between frequency and rate of geminate devoicing under a simple linear transformation of the natural ratings from Kawahara (2011a). The *solid line* represents the result of a linear regression. The *broken line* represents the overall devoicing rate

(22) Let r be the average naturalness rating that some token t received, and $devoice(t)$ the rate of devoicing in token t. Then:

$$devoice(t) = \left(\frac{r}{5}\right)(100)$$

To determine the overall devoicing rate under this transformation, we created a corpus assuming that each loanword appears in the corpus with its frequency in Amano and Kondo (2000). The loanword /budda/ 'Buddha', for instance, has a frequency of 99 in Amano and Kondo, and /budda/ was hence represented 99 times in our corpus. Each token was represented with devoicing according to the transformation given in (22). Devoicing in /budda/ received an average rating of 4.39. Performing the transformation on this score results in a devoicing rate of 87.8 %, and this percentage of the 99 occurrences of /budda/ in the corpus was hence represented with devoicing (i.e., 87 tokens with and 12 without devoicing). The same was done for all loanwords in the corpus. The overall devoicing rate in the corpus was then calculated to be 82.4 %. This overall rate is marked with a broken line in Fig. 6. As with the overall rate of *t/d*-deletion in the *Buckeye Corpus* (see Fig. 1), the overall rate of devoicing is closer to the rate observed for the more frequent words.

In the rest of this section, we develop an account for this transformed corpus. As with *t/d*-deletion, we first develop a purely grammatical model based on the overall devoicing rate in the corpus. We then augment this model with weight scaling according to the method described above in Sects. 2.2 and 2.3.

4.2 A grammatical account

We rely on the three constraints in (23)—see Nishimura (2003), Kawahara (2006) and Pater (2009) for analyses using slightly different constraints. As with *t/d*-deletion, we used the noisy HG learning algorithm in *Praat* to learn the weights associated with these constraints. The learning input file contained 100 tokens, with the proportion of tokens represented with devoicing determined by the overall rate of devoicing in the corpus (i.e., 82 out of the 100 tokens).[25] The learning file was submitted to *Praat*'s learning algorithm, using all of the default settings in *Praat*. The constraint weights that were learned are given in (24). Once the grammar had been learned, the "To output Distributions" function in *Praat* was used to determine the predicted rate of devoicing for the learned grammar. This returned an expected devoicing of 82.2 %, which very closely matches the observed deletion rate of 82.4 % in our corpus. However, as before, this grammar produces devoicing at the overall devoicing rate in the corpus, and treats all words of all frequencies the same. In the next section, we augment this account to incorporate the contribution of usage frequency to the rate of devoicing.

(23)

*GEMINATE	Assign one violation mark for every consonant linked to two timing slots.
*VOICEDOBS	Assign one violation mark for every voiced obstruent.
IDENT[voice]	Assign one violation mark for every output segment that has a different specification for the feature [voice] than its input correspondent.

(24)

*GEMINATE	100.0
*VOICEDOBS	101.3
IDENT[voice]	98.7

4.3 Incorporating the frequency bias through weight scaling

We incorporate the contribution of usage frequency into the model developed in the previous section in the same way as we did for *t/d*-deletion in Sect. 3.3. What is required is to scale the weight of the faithfulness constraint IDENT[voice] up for infrequent words so that they are more likely to be treated faithfully, and conversely to scale the weight of IDENT[voice] down for frequent words. First, we determined the reference point between frequent and infrequent words according to the method described in (9). In total, our corpus contains 11,000 tokens. The two most frequent words account for over half of the 11,000 tokens (/bagudaddo/ 'Baghdad', frequency: 3951;[26] /baggu/ 'bag', frequency: 2103). The reference point is hence halfway between the log frequency of /baggu/ (3.32) and the log frequency of the next most

[25] The learning input file is available as "supplementary material" on the Springer link for this article, or from the first author upon request.

[26] The high frequency of /bagudaddo/ in Amano and Kondo (2000) is a result of their frequency counts being taken from a corpus of newspapers including the time after the American invasion of Iraq. Although it is not clear that /bagudaddo/ will still have such a high frequency for the average Japanese speaker,

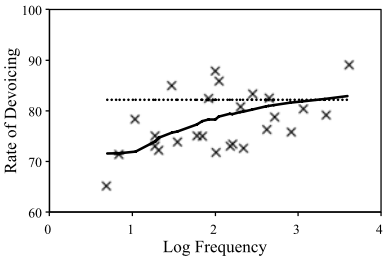

Fig. 7 Observed and predicted devoicing rates. The *broken line* indicates the prediction based on the basic, unscaled HG. The *solid line* shows the predictions based on the frequency weighted HG

frequent word, /bajji/ 'badge' (3.05),[27] or 3.19. With this reference value in hand, the *beta* distribution associated with each word can now be determined. For all words, the value of α is the reference log frequency of 3.19, and the value of β is the log frequency of the specific word. The value of the range parameter ρ is set to maximize the fit of the model's predictions with the data being modeled exactly as it was done for *t/d*-deletion above in Sect. 3.3. For the corpus with which we are working here, this value for ρ was found to be 1.

Once the value of ρ for a corpus has been determined, the scaling factor associated with each word can be determined. The weight of the faithfulness constraint can then be scaled according to this scaling factor for each word, and the predicted rate of devoicing can be determined for individual words using the "To output Distributions" function in *Praat*. Figure 7 shows how the baseline, unscaled HG model compares with the frequency scaled model. The broken line plots the prediction of the baseline model, and the solid line the prediction of the scaled model. This figure clearly shows that the scaled model fits the data better. This is confirmed by the mean square errors (*MSE*) of each of the models. The *MSE* of the baseline model is 52.7, and that of the scaled model 24.5, so that the scaled model represents a 53.5 % improvement over the baseline model.[28] As with the *t/d*-deletion account above, this improved fit is to be expected, given that the scaled model contains an extra parameter (frequency) that is known to be relevant. In order to determine whether the additional complexity of the scaled model is warranted by the increase in fit, we calculated AIC values for the baseline and scaled models, as we did above for *t/d*-deletion. The AIC value for the baseline model was found to be 109.1, and that for the scaled model was found to be 91.2. Since the scaled model has an AIC that is more than 10 units smaller than the baseline model, we conclude with confidence that there is sufficient support for the additional complexity of the scaled model.

we opted not to adjust its frequency for the purposes of this paper. The participants in Kawahara's experiment were mostly university students who were probably familiar with this event, so that /bagudaddo/ would have had a high frequency for them. The fact that /bagudaddo/ pronounced with devoicing, i.e., as [bagudatto], received a high naturalness rating in Kawahara (2011a) suggests that this might be correct.

[27] Following standard conventions in the literature on Japanese phonology, we use /j/ here for the affricate /dʒ/.

[28] As explained in footnote 24, we also explored an exponential and sigmoid transformation of the naturalness ratings. Frequency scaled models for corpora based on these transformations also performed better than baseline models, although the improvement was slightly less good than what we found for the linear transformation reported in the text. Improvement of the frequency scaled model over the baseline model was as follows: exponential transformation = 49.0 %; sigmoid transformation = 42.1 %.

5 Discussion

The model of phonological variation that we developed above incorporates the effects of usage frequency into a generative phonological grammar. Two case studies have shown that this model performs better than a model based on grammar alone. In this section, we discuss some general properties of our model, as well as some still unresolved and underexplored issues.

5.1 Grammar dominance

Although the model that we propose in this paper allows non-grammatical factors such as usage frequency to influence phonological variation, it is a grammar dominant model. Grammar sets the limits of what patterns of variation are possible, and all that the frequency can do is to determine how variation is realized within these limits. The dominance of grammar realizes itself in both universal terms and in the grammars of individual languages.

First consider the universal aspects of grammar dominance. In HG (as in OT), Universal Grammar is represented in the constraint set. Classic OT (Prince and Smolensky 1993, 2004) assumes that the constraint set is universal, so that the grammar of every language contains exactly the same constraints. From this assumption it follows that there are certain logically possible grammatical constraints that do not exist, and if some constraint does not exist then some logically possible grammatical patterns cannot be expressed. For example, in our analysis of *t/d*-deletion, we proposed positional MAX constraints for pre-vocalic and pre-pausal position, but argued that no such positional constraint exists for pre-consonantal position. If this is a true restriction on the constraint set, deletion in pre-consonantal context will always violate only a subset (MAX) of the faithfulness constraints violated by deletion in pre-vocalic (MAX, MAX-PRE-V) or pre-pausal (MAX, MAX-PRE-PAUSE) position. In (25), we show the consequences that this stringency relationship has for the H-score of deletion candidates in the different contexts. The H-score of deletion in pre-consonantal position will always be higher than that of deletion in the other two contexts. This effect cannot be overridden by weight scaling in our model, since we assume that all faithfulness constraints are scaled by the same factor (i.e., the scaling factor is not indexed to a particular faithfulness constraint). In any language, for a word of any frequency, deletion will always be most likely in pre-consonantal position. All that frequency can do is to increase or decrease the likelihood of deletion in all three contexts, but it will do so to the same extent in all three contexts.[29]

[29] Since a process cannot apply at a rate of higher than 100 %, this statement has to be qualified. Imagine a grammar where pre-consonantal context has a base deletion rate of 80 % and pre-pausal context of 50 %. Deletion in pre-consonantal position can be increased by at most 20 % by the contribution of scaling factors. The same holds for scaling factors that reduce the application of a simplification process and the floor of application, 0 %.

(25)

	MAX	MAX-PRE-V	MAX-PRE-PAUSE	H
/wɛst bæŋk/ → [wɛs_ bæŋk]	−1			$-w$(MAX)
/wɛst ɛnd/ → [wɛs_ ɛnd]	−1	−1		$-w$(MAX) − w(MAX-PRE-V)
/wɛst/ → [wɛs_]	−1		−1	$-w$(MAX) − w(MAX-PRE-PAUSE)

A similar point can be made with regard to geminate devoicing in Japanese. In our analysis, we assumed a markedness constraint that penalizes voiced obstruents, but no constraint that penalizes voiceless obstruents. If no constraint against voiceless obstruents exists, a language that has context-free voicing of obstruents (whether as a categorical or variable process) is impossible. It does not matter how frequent a word is: Since this process is ruled out by the grammar, it is predicted never to be observed.

Grammar also takes precedence over usage frequency at the level of individual languages. In the grammar developed for Columbus English above, the weight of MAX-PRE-V (−1.51) is lower than that of MAX-PRE-PAUSE (0.96), corresponding to the fact that this dialect of English shows more deletion in pre-vocalic than pre-pausal position. Since the weights of all faithfulness constraints are scaled by the same amount, the relative difference in the effective weights of MAX-PRE-V and MAX-PRE-PAUSE will be preserved under all scaling conditions. No matter how frequent a specific word is, on average a pre-vocalic deletion candidate will have a higher H-score than a pre-pausal deletion candidate. The grammar of Columbus English dictates that deletion is more likely in pre-vocalic context, and frequency cannot override this.

This dominance of grammar depends on the assumption that at a given instance of using the grammar (evaluation of a specific word, at a specific instance) the weights of *all* faithfulness constraints are scaled by the same amount. If weight scaling could variably affect different faithfulness constraints, the dominance of grammar could be lost. In this regard, our proposal diverges from the related proposal made by Boersma and Hayes (2001: Appendix C). Their model is developed in stochastic OT, and they therefore assume constraint ranking rather than weighting. They propose that the ranking values of some constraints can be changed in different speech situations. But crucially, they propose that some constraints can be ranked higher, others lower, and that constraint rankings do not have to be changed by the same amount. As a consequence, their model does not have the property of grammar dominance.

The dominance of grammar is also not a property of other models of phonological variation. In some implementations of usage-based models (Bybee 2001, 2006, 2007; etc.), or exemplar models (Gahl and Yu 2006 and papers therein; Pierrehumbert 2001; etc.), no formal distinction is made between grammatical and non-grammatical factors. In fact, in describing usage-based grammar, Bybee first defines the usage-based conceptualization of grammar as "the cognitive organization of one's experience with language" (Bybee 2006:711). Later on the same page she describes how this organization is done as follows: "… the general cognitive capabilities of the human brain, which allow it to categorize and sort for identity, similarity, and difference, go to work on the language events a person encounters, categorizing and entering in memory these experiences." Grammar is the result of cognitive organization achieved with general cognitive abilities, not with grammar or language specific abilities. Exactly

the same cognitive abilities that organize our experience with social interactions and with our physical environment organize our experience with language. No formal distinction is made between how language and other aspects of our experience are processed or stored in the mind. If a child acquiring a language were to be exposed to a set of experiences where deletion happens to be observed more often in pre-vocalic than pre-consonantal context, the general abilities of the mind to classify would notice this pattern, and codify this as the grammar. This view of grammar is fundamentally different from the type of approach that we advocate above. Under our approach, there are language specific cognitive capacities (Universal Grammar represented in the constraint set, as well as the principles for how constraints interact via their weights). Language is processed according to these principles and not with general cognitive capabilities. This places a limit on the types of grammars that can be learned. As we showed above, the assumptions about Universal Grammar under which we operate imply that no grammar that produces more deletion in pre-vocalic than pre-consonantal context is possible.

More research is necessary to determine to what extent certain types of grammars are truly impossible. A long tradition of typological research has established strong universal patterns across languages, a result that could be interpreted as favoring a system that includes a strong Universal Grammar. Recent research in artificial grammar learning has also shown that linguistic patterns that counter such universal trends are either unlearnable or at least not easily learnable (Carpenter 2006, 2010; Coetzee 2009b; Moreton 2008; Pater and Tessier 2006). On the other hand, there are also unambiguous examples of languages with grammars that counter universal trends (Coetzee and Pretorius 2010; Hyman 2001), showing that it should be possible for language learners to acquire grammars that do not fit neatly into the limits of Universal Grammar. Along similar lines, Bybee (2002:275) shows that in one dialect of English some words, under some circumstances, show more word-final *t/d*-deletion in pre-vocalic than pre-consonantal context. With conflicting data from the current literature it is impossible to choose definitively between a model with grammar dominance and a model in which grammar is afforded no special place. However, given that the evidence for strong universal tendencies is currently more copious than evidence for linguistic systems that counter these tendencies, we opt for the more restrictive model where Universal Grammar places limits on possible languages.

5.2 What processes are influenced by frequency?

In the model that we developed above, only the weights of faithfulness constraints are affected by frequency. From this restriction it follows that all and only those phonological processes that violate some faithfulness constraint will be affected by frequency scaling. In this paper, we have focused on two such processes—consonant cluster simplification and geminate devoicing. In both of these processes, it is the relative weight of some faithfulness constraint(s) (MAX/MAX-PRE-V/MAX-PRE-PAUSE or IDENT[voice]) and some markedness constraint(s) (*CT]$_{\text{Word}}$ or *VOICEDOBS/*GEMINATE) that determines whether the process applies. Since weight scaling affects the weights of the faithfulness constraints, it affects the relative weights of faithfulness and markedness constraints, and hence the likelihood that these processes will apply.

The processes on which we focused in this paper are both examples of simplification or reductive processes—i.e., the form that has undergone the process is in some sense articulatorily simpler or more reduced than the input. There is ample evidence from the literature that such reductive processes are indeed subject to the influence of frequency as predicted by the model that we developed above. In Sect. 1.3, we provided references for word-final obstruent deletion, unstressed vowel deletion, obstruent devoicing, and *l*-vocalization as examples.

However, the application of augmentation processes also depends on the relative weights of markedness and faithfulness constraints. In a language that avoids tautosyllabic consonant clusters via epenthesis, for instance, the application of epenthesis (arguably not a reductive process) depends on the relative weights of the anti-cluster markedness constraint *COMPLEX and the anti-epenthesis faithfulness constraint DEP. In a language in which such a process applies variably, the model developed above would predict that epenthesis will be observed more often in more frequent words than in less frequent words. Although there are examples in the literature that discuss such variable augmentation processes (see Auger 2001 on variable epenthesis in Vimeu Picard; Nevins 2007 on variable epenthesis in Brazilian Portuguese), we do not know of any example where the application of these processes is discussed in relation to usage frequency. If indeed variable augmentation processes are affected by frequency in the same way as variable reductive processes, it would be additional evidence for the model that we developed above. On the other hand, if augmentation processes are not affected by frequency in the same manner, the model would need to be revised in some way in order to differentiate between augmentation and reduction processes.

Given that only the weights of faithfulness constraints are affected by frequency scaling in the model developed above, variable phonological phenomena that do not depend on faithfulness constraints should not be affected by frequency in the same way. Under the assumption that there are no faithfulness constraints for prosodic structures (McCarthy 2003: Sect. 6), variable prosodification is not expected to be sensitive to frequency. As an example, consider Hammond's analysis of variable stress placement in Walmatjari (Hammond 1994; see also Anttila 2002b). In Walmatjari, tri-syllabic words are either stressed on the initial or the second syllable so that the underlying form /kaŋani/ 'carried' can be realized as [káŋani] or [kaŋáni]. Neither surface form violates any faithfulness constraints. The selection between the candidates is hence done by markedness constraints alone—in Anttila's account, by the constraints TROCHEE, FTBIN and *LAPSE (Anttila 2002b). Since only faithfulness constraints are sensitive to weight scaling, and since faithfulness constraints are irrelevant in the choice between these two variants, this choice cannot be influenced by frequency in the model developed above. We do not know of any literature that discusses such variable phenomena in relation to usage frequency, and we therefore cannot determine whether this prediction is borne out by actual data. If, in fact, processes such as these are also sensitive to frequency, the model developed above will need to be augmented in some way to account for this.

There is another set of variable phenomena that are known to be sensitive to usage frequency, but that are not accounted for in the model that we developed above. Morphological regularization (analogical leveling) is less likely to apply to more frequent

words. As an example, Bybee (1985:119–120; also Hooper 1976) shows that regularization of the English past tense is more likely to apply to infrequent words than to frequent words—a regular past tense form for infrequent *weep* (*weeped* instead of *wept*) is more likely than for frequent *keep* (*keeped* instead of *kept*). See also Phillips (1984, 2001) for more similar examples. Processes such as these are governed by the relations between morphologically related words, and hence by output-output correspondence constraints (Benua 2000) rather than by regular faithfulness constraints. Although our model cannot account for the role of frequency in these types of phenomena, the model could be extended in a straightforward manner to do so. An infrequent word (such as *weep*) is more likely to have a uniform paradigm. This implies that the OO-correspondence constraints that are responsible for enforcing paradigm uniformity should have higher weights in the evaluation of infrequent words than in the evaluation of frequent words. In the same way that we scale the weight of faithfulness constraints up for infrequent words, the weights of OO-correspondence constraints can be scaled up for infrequent words. However, we leave full development of this option for future research.

5.3 Modeling acquisition

In Sects. 3.2 and 4.2, we illustrated how a variable grammar can be learned using the noisy HG learning algorithm implemented in *Praat*. We also showed how this model can be augmented to account for the influence of usage frequency on variation. Two more questions need to be considered in this regard: (i) What predictions does this approach make with regard to the acquisition of variable phonological processes, and (ii) do these predictions correlate with how variable processes are acquired in reality? Although both of these questions are worth considering, we also want to make explicit that our goal in this paper is not to model the actual acquisition process of variable phenomena, but rather to show what a grammatical model would look like that can account for the variation observed in speech, and to show that such a grammatical model is in principle learnable. The goal of learnability theory is not to model how language is actually acquired, but to show whether a specific grammar is learnable from a given set of data (Pullum 2003:432.) Although we consider the possible implications of our model for acquisition, we do not believe that the value of our model crucially depends on how well it models actual acquisition processes.

We first want to set aside two simplifying assumptions that we made, and that do not constitute claims about actual acquisition. We assume that the learner has access to the correct underlying form of the words encountered. In the English *t/d*-deletion case, for instance, upon hearing an utterance like [wɛs bæŋk] for 'west bank', we assume that the learner knows that the underlying form of the first word in the utterance is really /wɛst/. This assumption is part of all of the main learning algorithms used in phonology (Boersma and Hayes 2001:51; Tesar and Smolensky 1998:237). The learning discussed here is hence learning at a later stage of acquisition, after underlying forms have already been acquired. For a development of the formal mechanisms involved in learning underlying forms in a constraint-based grammar, see Tesar and Smolensky (1996:40–44) and especially Merchant and Tesar (2005) and Tesar (2006).

The second simplifying assumption has to do with the role of usage frequency during grammar learning. We modeled the grammar learning stage above as if usage frequency of individual words plays no role during grammar learning, since we augmented the grammar with weight scaling only after the grammar has been learned (see also Hayes and Londe 2006 for a similar two-stage approach to learning). Another option that should be explored is one where usage frequency is incorporated into the grammar learning stage itself.

The noisy HG learning algorithm implemented in *Praat* is an error-driven learning algorithm. The basic steps in the learning process are: (i) The learner receives a learning input (a surface form produced by an adult); (ii) the learner determines the underlying form of the learning input, and submits this underlying form to his/her current grammar; and (iii) the learner compares the output generated by his/her current grammar to the learning input. If these two forms differ (i.e., if the learner's grammar generates an error), the learner adjusts his/her grammar to increase the likelihood that the grammar will generate an output identical to the learning input. In step (ii) of the learning cycle, the grammar is used to generate an output. In our modeling of learning above, this step did not include weight scaling. An alternative model of acquisition could incorporate weight scaling during this stage of grammar learning. The final state of the grammar that will be learned if weight scaling is incorporated during learning will be comparable to the final weight scaled grammars that we developed above. The most important difference between these two approaches is expected to be in the path of acquisition—i.e., how the grammar changes slowly during the learning period.

Although we did not incorporate frequency scaling during learning in our model, we can speculate about what would be expected from a model in which this is done. We use *t/d*-deletion as an example, but we expect the same basic pattern to be observed also in the acquisition of other variable processes. During the earlier stages of learning, when the learner has not yet built up a large corpus of learning inputs, chances are that the learner would have encountered mostly more frequent words. A child learning English, for instance, is more likely to hear a frequent word like 'want' (CELEX log frequency = 4.1) than an infrequent word like 'gourd' (CELEX log frequency = 0.9). Since more frequent words have higher deletion rates overall, and since the child is expected to hear mostly more frequent words, the corpus of learning inputs to which the child is exposed will have a higher *t/d*-deletion rate than the actual, complete adult production corpus. If the child aims to replicate the deletion rate in the learning corpus that he/she is exposed to, we would expect the child to show a higher overall deletion rate than what adults actually produce overall. This prediction agrees with the fact that child speech is often characterized by more reduction and simplification than adult speech.

Additionally, since during early acquisition the child will mostly be exposed to words from the higher end of the frequency spectrum, the range of the frequency distribution in the child's learning corpus is expected to be smaller than that in the actual adult speech corpus. (The range between the highest and lowest frequency words in the child's corpus is expected to be smaller than that in an adult's speech corpus.) In the model that we developed above, weight scaling is done based on how much the usage frequency of a specific word differs from the reference frequency in the corpus. In the child's early learning corpus, the usage frequencies are expected to differ

less than in the adult corpus. The expectation is hence that usage frequency will have less of an influence during the early stages of acquisition than in an adult grammar. During early acquisition, all words are expected to be treated more or less the same. Only during the later stage of acquisition will the difference between how frequent and infrequent words are treated emerge more clearly. To the best of our knowledge, there is no study that specifically investigates how usage frequency interacts with first language acquisition of variation. There is, however, suggestive evidence from second language acquisition that learners make less fine distinctions in terms of usage frequency than native speakers. Lacoste (2008) studies the acquisition of standard Jamaican English by Jamaican primary school children, with a focus on words that end in *-Ct/-Cd* clusters (i.e., exactly the words to which *t/d*-deletion could apply). She shows that the teachers make at least a three-level distinction in terms of usage frequency (2008:198), while children in the early stages of acquisition make only a two-level frequency distinction (2008:190).

Ultimately, more research is necessary to probe in detail how children acquire variation and to track specifically how the production of individual words changes during the course of acquisition. Similarly, the learning algorithm needs to be augmented to include weight scaling. Only once both of these things have been done will it be possible to go beyond speculation with regard to how variation is acquired, and with regard to how well the predictions of the model developed above matches the actual acquisition trajectory.

5.4 Final remark: integrating generative and usage-based grammars

In this paper, we developed a model of phonological variation that incorporates influences from both grammatical and non-grammatical factors. Our model retains some of the core characteristics of a classic generative grammar, while also embracing insights from usage-based and exemplar models of grammar. In the phonological literature, the generative approach and the usage-based/exemplar approaches have often been presented as opposites and as incompatible with each other. We believe this to be a false dichotomy. Not only is it possible to integrate these approaches, but such an integration also enables phonological theory to account better for many phenomena than what either of the two approaches could do in isolation. If such an integration is indeed the correct route to go, then future research will have to focus on two issues. First, the proper way to integrate the contributions from the two types of models needs to be determined. This paper contains one proposal, and the success of this proposal leads us to believe that it has merit. But other ways of integration are possible, and more research is necessary to determine all of the viable options, and to evaluate their success. Secondly, more targeted data collection would need to be performed. The data on phonological variation that are currently available are usually not suited to address the questions raised by an integrated model such as that proposed in this paper. We hope that the line of research reported in this paper will stimulate research into these issues.

Acknowledgements The ideas expressed in this paper were presented in various forms at NAPhC 5, NELS 38, NELS 41, the University of Michigan, the University of Massachusetts, Michigan State University, Stanford University, and SUNY Stony Brook. The feedback and reaction of the audiences at these

events contributed significantly to the development of our thoughts. This work has also been discussed in detail with many individuals, and we acknowledge our gratitude for their contribution. This list includes Joe Pater, John McCarthy, John Kingston, Anne-Michelle Tessier, Pam Beddor, San Duanmu, Ricardo Bermúdez-Otero, William Labov, Paul Smolensky, Matt Goldrick, Colin Wilson, Kevin McGowan, and Susan Lin. We also acknowledge the help of Amelia Compton in running many of the *Praat* simulations in this paper. The three reviewers and the associate editor similarly helped us to improve the paper and to express our ideas more clearly. As always, any remaining errors and shortcomings are our own.

References

Akaike, Hirotugu. 1973. Information theory as an extension of the maximum likelihood principle. In *Second international symposium on information theory*, eds. Boris N. Petrov and Frigyes Csaki, 267–281. Budapest: Akademiai Kiado.

Akaike, Hirotugu. 1983. Information measures and model selection. *International Statistical Institute* 44: 277–291.

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

Amano, Shigeaki, and Tadahisa Kondo. 2000. *NTT database series: Lexical properties of Japanese*, 2nd release. Tokyo: Sanseido.

Anttila, Arto. 1997. Deriving variation from grammar. In *Variation, change and phonological theory*, eds. Frans Hinskens, Roeland van Hout, and Leo Wetzels, 35–68. Amsterdam: John Benjamins.

Anttila, Arto. 2002a. Morphologically conditioned phonological alternations. *Natural Language & Linguistic Theory* 20: 1–42.

Anttila, Arto. 2002b. Variation and phonological theory. In *Handbook of language variation and change*, eds. Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 206–243. Oxford: Blackwell.

Anttila, Arto. 2006. Variation and opacity. *Natural Language & Linguistic Theory* 24: 893–944.

Anttila, Arto. 2007. Variation and optionality. In *The Cambridge handbook of phonology*, ed. Paul de Lacy, 519–536. Cambridge: Cambridge University Press.

Anttila, Arto, Vivienne Fong, Stefan Benus, and Jennifer Nycz. 2008. Variation and opacity in Singapore English consonant clusters. *Phonology* 25: 181–216.

Auger, Julie. 2001. Phonological variation and Optimality Theory: Evidence from word-initial vowel epenthesis in Vimeu Picard. *Language Variation and Change* 13: 253–303.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. 1995. *The CELEX Lexical Database (CD-ROM)*. Philadelphia: Linguistic Data Consortium.

Baese-Berk, Melissa, and Matt Goldrick. 2009. Mechanisms of interaction in speech production. *Language and Cognitive Processes* 24: 147–185.

Bane, Max. to appear. A combinatoric model of variation in the English dative alternation. In *Proceedings of the 36th annual meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society.

Bane, Max. 2011. Deriving the structure of variation from the structure of non-variation in the English dative. In *Proceedings of the 28th annual meeting of the West Coast conference on formal linguistics*, eds. Mary Byram Washburn, Katherine McKinney-Bock, Erika Varis, Ann Sawyer, and Barbara Tomaszewicz, 42–50. Somerville: Cascadilla Press.

Bayley, Robert. 1995. Consonant cluster reduction in Tejano English. *Language Variation and Change* 6: 303–326.

Bayley, Robert. 2002. The quantitative paradigm. In *The handbook of language variation and change*, eds. Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, 117–141. Oxford: Blackwell.

Bell, Alan, Jason Brenier, Michelle Gregory, Cynthia Girand, and Daniel Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60: 92–111.

Benua, Laura. 2000. *Phonological relations between words*. New York: Garland.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *University of Amsterdam Institute of Phonetic Sciences Proceedings* 21: 43–58.

Boersma, Paul. 2008. Emergent ranking of faithfulness explains markedness and licensing by cue. Ms., University of Amsterdam. www.fon.hum.uva.nl/paul/papers/EmergeFaith.pdf. Accessed 11 July 2012.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45–86.
Boersma, Paul, and Joe Pater. 2008. Convergence properties of a Gradual Learning Algorithm for Harmonic Grammar. Ms., University of Amsterdam and University of Massachusetts, Amherst. www.fon.hum.uva.nl/paul/papers/boersmaPaterHGGLA.pdf. Accessed 11 July 2012.
Boersma, Paul, and David Weenink. 2009. *Praat: Doing phonetics by computer version 5.1.20.* [Computer Program.] http://www.praat.org. 31 Accessed October 2009.
Browman, Catherine P., and Louis Goldstein. 1990. Tiers in articulatory phonology, with some implications for casual speech. In *Papers in laboratory phonology I: Between the grammar and physics of speech*, eds. John Kingston and Mary E. Beckman, 341–376. Cambridge: Cambridge University Press.
Burnham, Kenneth P., and David R. Anderson. 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research* 33: 261–304.
Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: Benjamins.
Bybee, Joan L. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In *Usage-based models of language*, eds. Michael Barlow and Suzanne Kemmer, 65–85. Stanford: CSLI.
Bybee, Joan L. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.
Bybee, Joan L. 2002. Word frequency and context of use in lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14: 261–290.
Bybee, Joan L. 2006. From usage to grammar: the mind's response to repetition. *Language* 82: 711–733.
Bybee, Joan L. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.
Byrd, Dani. 1992. A note on English sentence-final stops. In *UCLA Working Papers in Phonetics*, Vol. 81, eds. Pat Keating and Dani Byrd, 37–38. Los Angeles: Department of Linguistics, UCLA.
Carpenter, Angela. 2006. Acquisition of a natural versus an unnatural stress system. Ph.D. diss., University of Massachusetts.
Carpenter, Angela. 2010. A naturalness bias in learning stress. *Phonology* 27: 345–392.
Coetzee, Andries W. 2004. What it means to be a loser: Non-optimal candidates in Optimality Theory. Ph.D. diss., University of Massachusetts.
Coetzee, Andries W. 2005. The Obligatory Contour Principle in the perception of English. In *Prosodies*, eds. Sónia Frota, Marina Vigário, and Maria João Frietas, 223–245. Berlin: Mouton de Gruyter.
Coetzee, Andries W. 2006. Variation as accessing "non-optimal" candidates. *Phonology* 23: 337–385.
Coetzee, Andries W. 2008. Grammaticality and ungrammaticality in phonology. *Language* 84: 218–257.
Coetzee, Andries W. 2009a. An integrated grammatical/non-grammatical model of phonological variation. In *Current issues in linguistic interfaces: Volume 2*, eds. Young-Se Kang, Jong-Yurl Yoon, Hyunkyung Yo, Sze-Wing Tang, Yong-Soon Kang, Youngjun Jang, Chul Kim, Kyoung-Ae Kim, and Hye-Kyung Kang, 267–294. Seoul: Hankookmunhwasa.
Coetzee, Andries W. 2009b. Learning lexical indexation. *Phonology* 26: 109–145.
Coetzee, Andries W. 2009c. Phonological variation and lexical frequency. In *NELS 38*, Vol. 1, eds. Anisa Schardl, Martin Walkow, and Muhammad Abdurrahman, 189–202. Amherst: GLSA.
Coetzee, Andries W. 2012. Variation: Where laboratory and theoretical phonology meet. In *Oxford handbook of laboratory phonology*, eds. Abigail C. Cohn, Cécile Fougeron, and Marie K. Huffman, 62–75. Oxford: Oxford University Press.
Coetzee, Andries W., and Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory* 26: 289–337.
Coetzee, Andries W., and Joe Pater. 2011. The place of variation in phonological theory. In *Handbook of phonological theory: 2nd Edition*, eds. John Goldsmith, Jason Riggle, and Alan Yu, 401–434. Cambridge: Blackwell.
Coetzee, Andries W., and Rigardt Pretorius. 2010. Phonetically grounded phonology and sound change: The case of Tswana labial plosives. *Journal of Phonetics* 38: 404–421.
Côté, Marie-Hélène. 2004. Syntagmatic distinctness in consonant deletion. *Phonology* 21: 1–41.
Crawford, Clifford James. 2009. Adaptation and transmission in Japanese loanword phonology. Ph.D. diss., Cornell University.
Diehl, Randy L., and Margaret A. Walsh. 1989. An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America* 85: 2154–2164.
Eek, Arvo, and Einar Meister. 1995. The perception of stop consonants: locus equations and spectral integration. In *ICPhS XIII: Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, Vol. 1, 18–21.

Fasold, R. 1972. *Tense marking in Black English*. Arlington: Center for Applied Linguistics.
File-Muriel, Richard J. 2010. Lexical frequency as a scalar variable in explaining variation. *The Canadian Journal of Linguistics* 55: 1–25.
Fowler, Carol A. 1994. Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics* 55: 597–610.
Frank, Austin F., and T. Florian Jaeger. 2008. Speaking rationally: Uniform Information Density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the cognitive science society*, eds. Brad C. Love, Ken McRae, and Vladimir M. Sloutsky, 939–944. Austin: Cognitive Science Society.
Fruchter, David, and Harvey M. Sussman. 1997. The perceptual relevance of locus equations. *The Journal of the Acoustical Society of America* 102: 2997–3008.
Gahl, Susanne. 2008. *Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84: 474–496.
Gahl, Susanne, and Alan Yu, eds. 2006. *Special issue on exemplar-based models in linguistics*. Vol. 23(3) of *The Linguistic Review*. Berlin: De Gruyter Mouton.
Goeman, Ton. 1999. T-deletie in Nederlandse dialecten. Kwantitatieve analyse van structurele, ruimtelijke en temporele variate. Ph.D. diss., Vrije Universiteit, Amsterdam. The Hague: Holland Academic Graphics.
Goeman, Ton, and Pieter van Reenen. 1985. *Word-final t-deletion in Dutch dialects. The roles of conceptual prominence, articulatory complexity, paradigmatic properties, token frequency and geographical distribution*. Amsterdam: Vakgroep Algemene Taalwetenschap Vrije Universiteit.
Goldinger, Stephen D., Paul A. Luce, David B. Pisoni, and Joanne K. Marcario. 1992. Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 18: 1211–1238.
Goldsmith, John. 1993. Harmonic phonology. In *The last phonological rule: Reflections on constraints and derivations*, ed. John Goldsmith, 21–60. Chicago: Chicago University Press.
Goldsmith, John, ed. 1995. *The handbook of phonological theory*. Oxford: Blackwell.
Gupta, Arjun K., and Saralees Nadarajah, eds. 2004. *Handbook of the beta distribution and its applications*. New York: Marcel Dekker.
Guy, Gregory R. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3: 1–22.
Guy, Gregory R. 2011. Variability. In *The Blackwell companion to phonology: Volume 4*, eds. Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume, and Keren Rice, 2190–2213. Oxford: Blackwell.
Guy, Gregory R., and Charles Boberg. 1997. Inherent variability and the Obligatory Contour Principle. *Language Variation and Change* 9: 149–164.
Hammond, Michael. 1994. An OT account of variability in Walmatjari stress. Ms., University of Arizona.
Hayes, Bruce, and Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23: 59–104.
Hooper, Joan B. 1976. Word frequency in lexical diffusion and the source of morphological change. In *Current progress in historical linguistics*, ed. William M. Christie, 95–105. Amsterdam: North-Holland.
Hyman, Larry. 2001. On the limits of phonetic determinism in phonology: *NC revisited. In *The role of speech perception phenomena in phonology*, eds. Elizabeth Hume and Keith Johnson, 141–185. New York: Academic Press.
Itô, Junko. 1988. *Syllable theory in prosodic phonology*. New York: Garland.
Itô, Junko, and Armin Mester. 2001. Covert generalizations in Optimality Theory: The role of stratal faithfulness constraints. *Studies in Phonetics, Phonology, and Morphology* 7: 273–299.
Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61: 23–62.
Jesney, Karen. 2007. The locus of variation in weighted constraint grammars. Poster presented at the Workshop on Variation, Gradience and Frequency in Phonology. Stanford University, July 2007 [Downloaded on December 27, 2007 from http://people.umass.edu/kjesney/papers.html].
Jurafsky, Daniel, Alan Bell, Michelle Gregory, and William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency and the emergence of linguistic structure*, eds. Joan L. Bybee and Paul Hopper, 229–254. Amsterdam: Benjamins.
Kaisse, Ellen M., and Patricia A. Shaw. 1985. On the theory of lexical phonology. *Phonology Yearbook* 2: 1–30.
Kaneko, Emiko, and Gregory K. Iverson. 2009. Phonetic and other factors in Japanese on-line adaptation of English final consonants. In *Papers from the eighth annual conference of the Japanese Society for*

Language Science, eds. Shunji Inagaki and Makiko Hirakawa, Vol. 8 of *Studies in language sciences*, 179–185. Tokyo: Kuroshio Publications.

Kang, Hye-Kyung. 1994. Variation in past-marking and the question of the system in Trinidadian English. In *CLS 30: Papers from the 30th regional meeting of the Chicago Linguistic Society Volume 2: The parasession on variation in linguistic theory*, ed. Katherine Beals, 15–164. Chicago: Chicago Linguistic Society.

Katayama, Motoko. 1998. Optimality Theory and Japanese loanword phonology. Ph.D. diss., University of California, Santa Cruz.

Kawahara, Shigeto. 2005. Voicing and geminacy in Japanese: An acoustic and perceptual study. In *Univeristy of Massachusetts occasional papers in linguistics*, Vol. 31, eds. Katherine Flack and Shigeto Kawahara, 87–120. Amherst: GLSA.

Kawahara, Shigeto. 2006. A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in Japanese. *Language* 82: 536–574.

Kawahara, Shigeto. 2008. Phonetic naturalness and unnaturalness in Japanese loanword phonology. *Journal of East Asian Linguistics* 18: 317–330.

Kawahara, Shigeto. 2011a. Aspects of Japanese loanword devoicing. *Journal of East Asian Linguistics* 20: 169–194.

Kawahara, Shigeto. 2011b. Japanese loanword devoicing revisited: A wellformedness rating study. *Natural Language & Linguistic Theory* 29: 705–723.

Kempen, Gerard, and Karin Harbusch. 2008. Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in German clauses. In *The discourse potential of underspecified structures*, ed. Anita Steube, 179–192. Berlin: Walter de Gruyter.

Kewley-Port, Diane. 1983. Time-varying features as correlates of place of articulation in stop consonants. *The Journal of the Acoustical Society of America* 73: 322–335.

Kewley-Port, Diane, David Pisoni, and Michael Studdert-Kennedy. 1983. Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *The Journal of the Acoustical Society of America* 73: 1779–1793.

Kiparsky, Paul. 1985. Some consequences of Lexical Phonology. *Phonology Yearbook* 2: 85–138.

Kiparsky, Paul. 1993. An OT perspective on phonological variation. Ms. Stanford University. Paper presented at the Rutgers Optimality Workshop. October, 1993. http://www.stanford.edu/~kiparsky/Papers/nwave94.pdf. Accessed 11 July 2012.

Labov, William. 1966. *The Social stratification of English in New York City*. Washington: Center for Applied Linguistics.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45: 715–762.

Labov, William. 1989. The child as linguistic historian. *Language Variation and Change* 1: 85–97.

Labov, William. 2004. Quantitative analysis of linguistic variation. In *Sociolinguistics: An international handbook of the science of language and society, volume 1: 2nd edition*, eds. Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier, and Peter Trudgill, 6–21. Berlin: Mouton de Gruyter.

Lacoste, Véronique. 2008. Learning the sounds of Standard Jamaican English: Variationist, phonological and pedagogical perspectives on 7-year-old children's classroom speech. Ph.D. diss., University of Essex.

Lahiri, Aditi, Letitia Gewirth, and Sheila E. Blumstein. 1984. A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *The Journal of the Acoustical Society of America* 76: 391–404.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990. Can connectionism contribute to syntax? Harmonic Grammar, with an application. In *Proceedings of the 26th regional meeting of the Chicago Linguistic Society*, eds. Michael Ziolkowski, Manuela Noske, and Karen Deaton, 237–252. Chicago: Chicago Linguistic Society.

Lin, Susan, Patrice S. Beddor, and Andries W. Coetzee. 2011. Gestural reduction and sound change: An ultrasound study. In *ICPhS XVII: Proceedings of the 17th International Congress of Phonetic Sciences*, eds. Wai-Sum Lee and Eric Zee, 1250–1253.

Lindblom, Bjorn. 1990. Explaining phonetic variation: A sketch of the H and H theory. In *Speech production and speech modeling*, eds. William J. Hardcastle and Alain Marchal, 403–439. Dordrecht: Kluwer Academic.

Luce, Paul A., and David B. Pisoni. 1986. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19: 1–35.

Martínez-Celdrán, Eugenio, and Xavier Villalba. 1995. Locus equations as a metric for place of articulation in automatic speech recognition. In *ICPhS XIII: Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, Sweden, Vol. 1, eds. Kjell Elenius and Peter Branderud, 30–33. Stockholm: Stockholm University.

Malécot, André. 1958. The role of releases in the identification of released final stops: A series of tape-cutting experiments. *Language* 34: 274–284.

McCarthy, John J. 2003. Sympathy, cumulativity, and the Duke-of-York gambit. In *The syllable in Optimality Theory*, eds. Caroline Féry and Ruben van de Vijver, 23–76. Cambridge: Cambridge University Press.

McCarthy, John J., and Alan Prince. 1995. Faithfulness and reduplicative identity. In *Papers in Optimality Theory*, eds. Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey. Vol. 18 of *University of Massachusetts occasional papers in linguistics*, 249–384. Amherst: GLSA.

McNamara, Timothy P. 2005. *Semantic priming: Perspectives from memory and word recognition*. New York: Psychology Press, Taylor and Francis.

Merchant, Nazarré, and Bruce Tesar. 2005. Learning underlying forms by searching restricted lexical subspaces. In *CLS 41: Proceedings from the 41st annual meeting of the Chicago Linguistic Society*, Vol. 2, eds. Rodney L. Edwards, Patrick J. Midtlyng, Colin L. Sprague, and Kjersti G. Stensrud, 33–48. Chicago: CLS.

Mitterer, Holger, and Mirjam Ernestus. 2006. Listeners recover /t/'s that speakers reduce: Evidence from /t/-lenition in Dutch. *Journal of Phonetics* 34: 73–103.

Moreton, Elliott. 2008. Learning bias as a factor in phonological typology. *Phonology* 25: 83–127.

Moreton, Elliott. 2010. Underphonologization and modularity bias. In *Phonological argumentation: Essays on evidence and motivation*, ed. Stephen Parker, 79–101. London: Equinox.

Nearey, Terrance M., and Sherrie E. Shammas. 1987. Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics* 15: 17–24.

Nevins, Andrew. 2007. Review of Scheer 2004. *Lingua* 118: 425–434.

Nishimura, Kohei. 2003. Lyman's law in loanwords. M.A. thesis, Nagoya University.

Nishimura, Kohei. 2006. Lyman's Law in loanwords. *Phonological Studies [Onin Kenkyuu]* 9: 83–90.

Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33: 999–1035.

Pater, Joe, and Anne-Michelle Tessier. 2006. L1 phonotactic knowledge and the L2 acquisition of alternations. In *Inquiries in linguistic development: Studies in honor of Lydia White*, eds. Roumyana Slabakova, Silvina A. Montrul, and Philippe Prévost, 115–131. Amsterdam: Benjamins.

Patrick, Peter L. 1992. Creoles at the intersection of variable processes: *t*, *d* deletion and past-marking in the Jamaican mesolect. *Language Variation and Change* 3: 171–189.

Patterson, David, and Cynthia M. Connine. 2001. Variant frequency in flap production: A corpus analysis of variant frequency in American English flap production. *Phonetica* 58: 254–275.

Phillips, Betty S. 1984. Word frequency and the actuation of sound change. *Language* 60: 320–342.

Phillips, Betty S. 2001. Lexical diffusion, lexical frequency, and lexical analysis. In *Frequency and emergence of linguistic structure*, eds. Joan L. Bybee and Paul Hopper, 123–136. Amsterdam: John Benjamins.

Phillips, Betty S. 2006. *Word frequency and lexical diffusion*. New York: Palgrave Macmillan.

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In *Frequency effects and the emergence of lexical structure*, eds. Joan L. Bybee and Paul Hopper, 137–157. Amsterdam: Benjamins.

Pitt, Mark A., Laura Dilley, Keith Johnson, Scott Kiesling, William D. Raymond, Elizabeth Hume, and E. Fosler-Lussier. 2007. *Buckeye corpus of conversational speech*, 2nd Release. Columbus: Department of Psychology, Ohio State University. www.buckeyecorpus.osu.edu.

Postal, Paul. 1966. Review of "Elements of general linguistics" by André Martinet. *Foundations of Language* 2: 151–186.

Postal, Paul. 1968. *Aspects of phonological theory*. New York: Harper and Row.

Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in Generative Grammar. Ms., New Brunswick, Rutgers University.

Prince, Alan, and Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in Generative Grammar*. Oxford: Blackwell.

Pullum, Geoffrey K. 2003. Learnability. In *The international encyclopedia of linguistics*, ed. William J. Frawley, 431–434. Oxford: Oxford University Press.

Raymond, William D., Robin Dautricourt, and Elizabeth Hume. 2006. Word-medial /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors. *Language Variation and Change* 18: 55–97.

Reynolds, Bill. 1994. Variation and phonological theory. Ph.D. diss., University of Pennsylvania.
Santa Ana, Otto. 1991. Phonetic simplification processes in the English of the barrio: A cross-generational sociolinguistic study of the Chicanos of Los Angeles. Ph.D. diss., University of Pennsylvania.
Scarborough, Rebecca. 2004. Coarticulation and the structure of the lexicon. Ph.D. diss., UCLA.
Scarborough, Rebecca. 2010. Lexical and contextual predictability: Confluent effects on the production of vowels. In *Papers in laboratory phonology X: Variation, phonetic detail and phonological modeling*, eds. Cécile Fougeron, Barbara Kühnert, Mariapaola D'Imperio, and Nathalie Vallée, 557–586. Berlin: Mouton de Gruyter.
Schouten, Marten E.H. 1982. T-deletie in de stad Utrecht: Schoolkinderen en grootouders. *Forum der Letteren* 23: 282–291.
Schouten, Marten E.H. 1984. T-deletie in Het Zuiden van die provincie Utrecht. *Taal en Tongval* 36: 162–173.
Smolensky, Paul, and Géraldine Legendre, eds. 2006. *The harmonic mind: From neural computation to Optimality-Theoretic grammar, Volume 1: Cognitive architecture, Volume 2: Linguistic and philosophical implications*. Cambridge: MIT Press.
Steriade, Donca. 1999. Phonetics in phonology: The case of laryngeal neutralization. In *UCLA working papers in linguistics 2 (Papers in phonology 3)*, ed. Matthew K. Gordon, 25–146. Los Angeles: Department of Linguistics, UCLA.
Steriade, Donca. 2001. Directional asymmetries in place assimilation. In *The role of speech perception in phonology*, eds. Elizabeth Hume and Keith Johnson, 219–250. San Diego: Academic Press.
Stevens, Kenneth N., and Sheila E. Blumstein. 1978. Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America* 64: 1358–1368.
Stevens, Kenneth N., and Samuel J. Keyser. 1989. Primary features and their enhancement in consonants. *Language* 65: 81–106.
Sussman, Harvey M., Helen A. McCaffrey, and Sandar A. Matthews. 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America* 90: 936–946.
Tanaka, Shin-Ichi. 2009. The eurhythmics of segmental melody. *Journal of the Phonetic Society of Japan* 13: 44–52.
Tesar, Bruce. 2006. Learning from paradigmatic information. In *NELS 36: Proceedings of the thirty-sixth annual meeting of the North East Linguistic Society*, Vol. 2, eds. Christopher Davis, Amy-Rose Deal, and Youri Zabbal, 619–638. Amherst: GLSA.
Tesar, Bruce, and Paul Smolensky. 1996. *Learnability in Optimality Theory (long version).* Technical report JHU-CogSci-96-4, Department of Cognitive Science, The Johns Hopkins University, Baltimore, Md.
Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29: 229–268.
van Oostendorp, Marc. 1997. Style levels in conflict resolution. In *Variation, change and phonological theory*, eds. Frans Hinskens, Roeland van Hout, and Leo Wetzels, 207–229. Amsterdam: John Benjamins.
Versace, Rémy, and Brigitte Nevers. 2003. Word frequency effect on repetition priming as a function of prime duration and delay between the prime and the target. *British Journal of Psychology* 94: 389–408.
Vitevitch, Michael S., and Paul A. Luce. 1998. When words compete: Levels of processing in spoken word recognition. *Psychological Science* 9: 325–329.
Vitevitch, Michael S., and Paul A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40: 374–408.
Walsh, Margaret A., and Randy L. Diehl. 1991. Formant transition duration and amplitude rise time as cues to the stop/glide distinction. *The Quarterly Journal of Experimental Psychology A* 43: 603–620.
Zsiga, Elisabeth. 2000. Phonetic alignment constraints: Consonant overlap and palatalization in English and Russian. *Journal of Phonetics* 28: 69–102.