Andries Coetzee and Joe Pater. Draft 10/6/09 for 2nd Edition of the Handbook of Phonological Theory, Goldsmith, Riggle and Yu (eds.).

The place of variation in phonological theory*

1. Introduction

Over the past 15 years, the study of variation has become increasingly important in phonological theory. As recently as 1995, the previous edition of this *Handbook* did not have a chapter on variation. In fact, the term "variation" does not even appear in its subject index. Today, any volume that attempts to give an overview of the current status of the field cannot go without a chapter dedicated to variation. In this chapter, we review how the fortunes of variation have changed over the past fifteen years, and discuss some of the issues that arise in making a place for variation in a theory of phonology.

For the purposes of this chapter, we understand the term "phonological variation" to describe a situation in which a single morpheme can be realized in more than one phonetic form in a single environment. This definition is intentionally broad. We do not take an *a priori* position on whether phonological variation includes only differences in terms of categories like [+/-voice], or whether it also includes sub-categorical distinctions in terms of phonetic features like voice onset time. We also do not wish to exclude from our definition of phonological variation an alternation between two forms that are too far apart to be related by a phonological derivation (i.e. variation involving a suppletive form). Our reluctance to draw thick lines around *phonological* variation is due to the well-known difficulty in identifying a principled way of separating phonology from morphology and phonetics (see Inkelas, this volume and Ladd, this volume respectively). As we will see, this difficulty is exacerbated when the details of variation are taken into account.

Since our discussion will be structured around the question of the locus of variation in phonology, we give a brief overview here of standard assumptions about the architecture of this part of the grammar. Within generative linguistics, the phonological component is usually at least implicitly assumed to have the following shape:

(1) Lexicon → Early Phonology → Late Phonology → Phonetic Implementation

We use the theory neutral terms of early and late phonology rather than more theory-specific terms such as lexical and post-lexical phonology (Kiparsky 1982) or the traditional division between morphophonology and phonology, which also entails a relatively specific characterization of the distinction between the levels. Within this general model, syntax supplies the morphemes through the operation of lexical insertion. Each morpheme has a single lexical form (except in cases of suppletion), which may be changed during the course of phonological derivation. The derivation begins with the application of a set of changes that we refer to as early phonology, which are then followed by a second set of changes that we refer to as late phonology. The exact content given to early and late phonology varies between different phonological theories, but typical characteristics assigned to each of them are given in (2).

* Acknowledgments: Thanks to Matthew Goldrick, Bernard Laks, Jeroen van de Wejer, and the editors of this volume for comments on an earlier draft. The preparation of this paper was partially supported by grant BCS-0813829 from the National Science Foundation to the University of Massachusetts Amherst.

(2)		*Early Phonology*	*Late Phonology*
	i.	Sensitive to morphology (Because of direct interaction with lexicon, in which words are morphologically decomposed)	Insensitive to morphology (Because this level has no contact with the lexicon)
	ii.	May have exceptions (Since these are encoded in the lexicon)	Exceptionless (Because of lack of contact with lexicon)
	iii.	Makes only categorical changes (Because only categories are represented in the lexicon)	Can introduce non-categorical changes (Because of its contact with phonetics, which requires richer representations)
	iv.	Word bounded (Because only single words can be input to this level)	Sensitive to cross-word contexts (Because whole utterances are input to this level)
	v.	Insensitive to factors like speech rate (Because this level has no contact with phonetics)	Sensitive to factors like speech rate (Because of direct contact with phonetics)

Though this modular structure with its associated typology of phonological changes is most closely associated with Lexical Phonology, nearly all work in generative phonology, including that in relatively non-derivational frameworks like OT, at least implicitly assumes much of it. We also note that theories differ in what they consider to be the domain of phonology: parts of early phonology are sometimes argued to be purely morphological or lexical (e.g. Hooper 1976, Dressler 1985, Ford and Singh 1985), and some of late phonology is often left to the phonetic component (on variation see recently Hale, Kissock and Reiss 2007). Even though the details of this architecture, and how it is applied, vary tremendously, we adopt this broad outline as a means of structuring our discussion.

In the second section of the chapter, we discuss the view that variation is limited to late phonology. This position is made explicit in Lexical Phonology, and is in line with the phonetically gradient nature of many variable phenomena. In the third section we discuss evidence showing that variable processes have characteristics of early phonology, focusing on examples where they are conditioned by morphology. In this section, we introduce approaches to variation in OT, paying special attention to the Partially Ordered Constraints theory of Kiparsky (1993) and Anttila (1997 *et seq*.). In the fourth section, we discuss models of variation in OT-like theories in which constraints are placed on a numerical scale. These include Boersma's (1997, 1998) stochastic OT, as well as models of grammar in which ranking is entirely replaced by numerical weighting, as in OT's predecessor Harmonic Grammar (HG; Legendre, Miyata and

Smolensky 1990; Smolensky and Legendre 2006). We discuss the strengths of these theories of grammar in terms of their ability to model quantitative aspects of phonological variation, and in terms of the existence and robustness of associated learning algorithms. We also discuss their relationship to the original generative theory of probabilistic grammar: the Variable Rules model of Labov (1969).

In the fifth section, we discuss evidence that the lexicon influences variation. Variable processes can apply differently to two lexical forms that are identical in terms of all relevant morphological and phonological properties. These data pose difficulties for any theory that advocates a strict separation between the lexicon and the level at which variation takes place, as well as for some OT approaches to variation and its learning. In this final section, we provide a brief account of these phenomena in terms of weighted lexically specific constraints. We also briefly discuss alternative formalizations of the influence of the lexicon on variation in HG, and ways in which this model can be extended to deal with differences between registers or styles that are often associated with variation. Because of the formal resemblance of weighted constraints models of variation to Labov's (1969) Variable Rules theory, especially to its elaboration as an explicitly probabilistic model of grammar (Cedergren and Sankoff 1974), there is reason to be optimistic about future strengthening of connections between research on phonological variation in sociolinguistics, and in generative phonological theory.

2. Variation limited to late phonology

In rule-based phonology, variation is standardly handled by simply marking a rule as [+optional] (see recently Vaux 2008). Labov (1969) suggests that this could be formalized by writing parentheses around the structural change of a rule, as in (3).

(3) Labov (1969:737)

$X \rightarrow (B) / Y \_ Z$

Labov goes on to propose an augmentation of this rule writing convention, so that it is possible to include contextual factors that influence the probability that the rule will apply. In his *variable rule* notation, Greek letter variables are used to indicate features that play this role (see section 0 on how probability of rule application is calculated).

(4) Labov (1969:738)

$$X \rightarrow (Y) \;/\; \begin{bmatrix}\alpha \text{fea}_i \\ \vdots\end{bmatrix}\begin{bmatrix}\gamma \text{fea}_j \\ \vdots\end{bmatrix}\begin{bmatrix}\beta \text{fea}_k \\ \vdots \\ \delta \text{fea}_l\end{bmatrix}$$

Although the "paradigm change" (Cedergren and Sankoff 1974) entailed by Labov's introduction of a probabilistic component to generative grammar did have a profound effect on subsequent research in sociolinguistics, it had little impact elsewhere in theoretical phonology until relatively recently. We will return to Labov's proposal, and some aspects of its subsequent development, in section 0.

In Lexical Phonology (Kiparsky 1982), it is sometimes proposed that only post-lexical rules can be subject to optional application (Kaisse and Shaw 1985: 6; Kiparsky 1985: 86; see also Donnegan and Stampe 1979: 145 for the related claim that Natural Phonology's processes, but not its rules, can be optional). Post-lexical rules show most of the characteristics that we ascribed above to late phonology. They are insensitive to morphology, they are exceptionless, and they can be conditioned by cross-word contexts. Moreover, Kaisse and Shaw (1985:6) connect the variability of post-lexical rules specifically to speech rate: "We also suspect that only postlexical rules can be optional and subject to variation due to rate of speech, though this requires further investigation." As an example of the difference between lexical and post-lexical rules, Kiparsky (1985:86) cites nasal place assimilation in English, a rule that applies both lexically and post-lexically. Kiparsky remarks that the lexical application of this rule is obligatory, while the post-lexical application is variable. Kiparsky does not mention the facts concerning assimilation at the prefix-root boundary; we briefly discuss them in section 0.

(5) Nasal place assimilation in English

a. Lexical = obligatory

e[nt]*er*, **e*[mt]*er*, **e*[ŋt]*er*

a[mb]*er*, **a*[nb]*er*, **a*[ŋb]*er*

pra[ŋk], **pra*[nk], **pra*[mk]

b. Post-lexical = optional

gree[n b]*ox* ~ *gree*[m b]*ox*

i[n b]*ed* ~ *i*[m b]*ed*

gree[n k]*ard* ~ *gree*[ŋ k]*ard*

The variable nasal place assimilation in (5b) is also an example of a process that is sometimes claimed not to be phonological at all, but instead the result of phonetic implementation. Barry (1985), Nolan (1992) and Ellis and Hardcastle (2002) provide articulatory evidence that for at least some speakers, the nasal in a phrase like *green box* retains to a variable degree its alveolar closure, even when it is perceived as fully labial (see Baayen and Ernestus, this volume, on other realizations of these sequences). This can be taken as evidence that the process is at least sometimes phonetic, rather than phonological, insofar as this intermediate articulation cannot be produced by a categorical phonological rule (though cf. Hayes 1995 on this example), and to the extent that this sort of gradience is taken to be diagnostic of rules of phonetic implementation, outside of the domain of categorical phonology (though cf. Ohala 1990, Flemming 2001).

Probably the most intensively studied variable phonological process is another English example: the variable deletion of an alveolar stop from word-final consonant clusters, which results in variation between [wɛst] and [wɛs] for a word like *west*. We return to this process again later in this chapter, and will review some of the relevant literature there. Here we want to point out that this is another process that has been claimed to be the result of phonetic implementation rather than of the variable application of a phonological rule. At first glance, *t/d*-deletion seems like the prototypical late phonological process. It is more likely to occur in casual or fast than in formal or slow speech. It is also reported to result in gradient, rather than categorical outcomes. Browman and Goldstein (1990) investigated the production of phrases like "perfect memory" – i.e. where the first word ends in a [-Ct] cluster and the second word begins in a consonant. They recorded subjects reading these phrases in careful speech style and then in a more casual speech style. Simultaneously with the acoustic recording, they also collected articulatory data by X-ray. In the acoustic data they found evidence of an alveolar stop [t] in the careful but not in the casual speech condition, showing that the process is sensitive to speech style. However, the articulatory data indicated that the tongue blade moves towards the alveolar ridge to form the [t] closure in both the careful and the casual speech condition; the difference between the speech styles was in the timing of the gestures, rather than in the presence of alveolar closure. Bybee (2000:73) uses this sort of evidence to argue that there is no need for a phonological account of the process: "… there is no variable rule of *t/d*-deletion. Rather there is a gradual process of shortening or reducing the lingual gesture." (See also Bybee 2001:75-76.)

3. Variation in early phonology

We now turn to some examples of morphologically conditioned phonological variation that provide evidence that variation has characteristics of early phonology (section 0). In section 0, we present the OT theory of variation proposed by Kiparsky (1993), and developed by Anttila (1997 et seq.), and show how it handles some of the English *t/d*-deletion data. In section 3.3, we discuss some issues with this model, and briefly review other OT approaches to variation.

3.1 Morphologically conditioned variation

We introduced the variable process of *t/d*-deletion in English just above, pointing out that it is sometimes considered to be the result of phonetic implementation rather than of variable phonological rule application, which is consistent with its categorization as a "late" process in the typology in (2). We begin this section by showing that *t/d*-deletion is conditioned by factors similar to those that condition the (apparently) categorical processes typically studied by phonologists. We focus particularly on its morphological conditioning, which is sometimes claimed to be a characteristic of only early phonological processes.

Variationist sociolinguists have extensively studied *t/d*-deletion over the past three or four decades. We therefore have data on this process for many different dialects of English. In a 1989 paper, Labov synthesizes the current state of knowledge about this process. He identifies several factors that seem to influence the rate of *t/d*-deletion in every dialect that had been described up until then. Some of the factors that he identifies are given in (6). Guy (1994) and Coetzee (2004) provide updated literature reviews, which confirm Labov's original synthesis.

(6) Cross-dialectical generalizations about *t/d*-deletion in American English

a. *Stress*: *t/d* is more likely to delete from an unstressed than a stressed syllable – i.e. more deletion from *safest* than from *resist*.

b. *Third consonant*: More deletion from three consonant than two consonant clusters – i.e. more deletion from *whisked* than from *picked*.

c. *Preceding consonant*: The more similar the preceding consonant is to *t/d*, the more likely *t/d* is to delete – i.e. more deletion from *west* than from *left*.[1]

---

[1] Labov (1989) actually gives the following hierarchy from consonants that are most likely to induce deletion of a following stop to those that are least likely: /s/ > stops > nasals > other fricatives > liquids. However, Guy and Boberg (1997) show convincingly that what is really relevant is the number of features that the preceding consonant shares with *t/d* – the more shared features, the higher the deletion rate. See also Coetzee (2004:chapter 5) for evidence in agreement with Guy and Boberg.

d. *Morphological status of t/d*: More deletion if *t/d* is part of a monomorpheme than if it functions as the past tense suffix – i.e. more deletion from *mist* than *missed*

e. *Following segment*: The more sonorous the following segment, the less likely *t/d* is to delete – i.e. less deletion from *best work* than from *best book*.

These factors are all typical of ones that condition the application of non-variable phonological processes. A theory that provides a unified account of variable and categorical processes is thus likely to avoid considerable duplication of formal machinery (see especially Guy 1997 for discussion of this point with respect to the *t/d*-deletion data).

From the perspective of a theory that distinguishes between early and late phonology in the manner outlined in (2) above, the problematic observation is that *t/d*-deletion is conditioned by morphology. Guy (1991b) has shown that the morphological conditioning is even more fine-grained than reported by Labov (1989). Not only does *t/d* as the final consonant of a monomorpheme (*mist*, *land*) delete more frequently than *t/d* that functions as a past tense morpheme (*missed*, *banned*), but semi-weak past tense forms (*kept*, *told*) have an intermediate degree of *t/d* deletion. This pattern has been documented for many different dialects of English, and we report only a smattering of the available data in (7). Guy (1994) and Labov (2004:15-16) provide further evidence for the robustness of this generalization.

(7) Deletion rate of *t/d* in different dialects of English

	Regular past (*missed*)	*Semi-weak past* (*kept*)	*Monomorpheme* (*mist*)
Philadelphia English (Guy 1991b)	17%	34%	38%
Chicano English (Santa Ana 1992)	26%	41%	58%
Tejano English (Bayley 1997)	24%	34%	56%

The fact that the application of the variable, gradient process of *t/d*-deletion is conditioned by morphology is counter to the strawman typology of phonological processes we laid out in (2). Guy's 1991 analysis of this pattern is cast within a modified version of Lexical Phonology (Guy 1991a), one in which a variable rule can apply within the lexicon (even in Level 1; see Guy 1994 for discussion of other ways in which his analysis departs from classic Lexical Phonology), while Kiparsky (1993) proposes a single-level OT analysis.

The English nasal place assimilation facts discussed by Kiparsky (1985) and in the previous section also show evidence of morphological conditioning. Assimilation is categorical within words derived with the "Level 1" prefix *in*-, as well as within underived words. Assimilation is gradient between the "Level 2" prefix *un*- and the following stem, as well as across word boundaries. Thus, the gradient process does apply within some words. One line of explanation for this and other apparent exceptions to morphological invisibility in late phonology is to invoke a prosodic difference; *un*- could be placed outside the prosodic word that contains the root, and thus be made to behave as an independent word for the purposes of assimilation.

However, it is hard to see how this sort of account would generalize to the fine-grained morphological sensitivity of *t/d*-deletion.

Many more examples of variable processes that interact with morphology can be found in the literature, and some of these have been prominent in OT analyses of variation. Anttila's (1997) much-discussed case of Finnish genitive plural allomorphy is notable in that the variation between the allomorphs shows every indication of being an early process. Not only are the alternations limited to the genitive, but as Anttila (1997: fn. 2) notes, it is unclear whether the changes are produced by phonological processes, or whether they are choices between stored allomorphs. Nonetheless, Anttila argues that the factors that produce preferences between variants are clearly phonological. Other well-known examples include variation between forms of reduplication in Ilokano, which was first discussed in Hayes and Abad (1989), and later formally analyzed by Boersma and Hayes (2001) and Coetzee (2006), and the variable application of vowel harmony observed in Hungarian, in which individual stems vary in the extent to which they select harmonic and disharmonic suffixes (Hayes and Londe 2006). Given examples like these, it is clear that phonological theory is responsible for providing an account of variation; it cannot be entirely left to phonetic implementation. In the next section, we begin to discuss the constraint-based analyses of variation that have recently emerged in OT and related theories.

3.2 The partially ordered constraints theory of phonological variation

In the version of OT proposed by Prince and Smolensky (1993/2004), which we will refer to as standard OT, the grammar of a language is a total ordering of a ranked set of constraints. Standard OT yields a single optimal output (Surface Representation in phonology) for each input (Underlying Representation). In independent developments, Kiparsky (1993) and Reynolds (1994) proposed elaborations of OT that yield variation between optimal outputs over instances of evaluation. We will focus on one of Kiparsky's proposals, developed in much more detail by Anttila (1997 *et seq*.), which we will refer to as the partially ordered constraints (POC) model of variation. In this theory, the grammar states a partial, rather than a total order on the constraint set. Each time the grammar is used to evaluate a candidate set, one of the total orders consistent with the partial order is randomly chosen. When some of these total orders pick different candidates as optimal, variation results.

In (8), we give a schematic example. In this example, the grammar is not a total ordering of the constraints – although both C_2 and C_3 are ranked beneath C_1, their relative order is unspecified. Every time that an input is submitted to the grammar, one of the possible rankings between C_2 and C_3 is chosen. As the example shows, *cand*$_1$ is selected as optimal under one ranking, and *cand*$_2$ under the other ranking. This is therefore a language where /input$_1$/ will be variably realized as either *cand*$_1$ or *cand*$_2$, but will never surface as *cand*$_3$.

(8) Grammar: $C_1 >> C_2$, $C_1 >> C_3$

a. First possible ranking: $C_1 >> C_2 >> C_3$

/input$_1$/	C_1	C_2	C_3
☞ *cand*$_1$			*
cand$_2$		*!	
cand$_3$	*!		

b. Second possible ranking: $C_1 >> C_3 >> C_2$

/input$_1$/	C_1	C_3	C_2
cand$_1$		*!	
☞ *cand*$_2$			*
cand$_3$	*!		

The model also predicts the frequency with which different variants will be observed, according to the principle in (9), from Anttila (1997).

(9) Quantitative interpretation of multiple rankings

Let *t* be the total number of total orders corresponding to a partially ordered constraint set. If a candidate is selected as optimal in *n* of these rankings, then this candidate's probability of occurrence is *n*/*t*.

In the grammar in (8), there are two possible total orders, and each of the candidates is selected as optimal under one of these rankings. Each of these candidates therefore has equal probability. In "mainstream" generative phonology, the POC model and Reynold's (1994) alternative are perhaps the first adoptions of Labov's (1969) proposal that a grammar can encode a probability distribution over outcomes (see Bod, Hay and Jannedy 2003 for an overview of other probabilistic models of grammar, in phonology and elsewhere).

Kiparsky (1993) proposes the POC model in the context of an analysis of the morphological and phonological factors influencing the rate of *t/d*-deletion in English (see also Reynolds 1994). Here we provide an illustration of the POC theory by providing a slightly amended version of Kiparsky's analysis of the effect of phonological context. As we showed in (6), there are many different factors that influence the probability that *t/d*-deletion will apply in a specific instance. Like Kiparsky, we focus here only on the influence of what Guy (1991a) calls the "external" context – that is, what follows on the word-final *t/d*. As Labov (1989) points out, a clear generalization that emerges from the variationist literature on *t/d*-deletion is that the rate of deletion is always highest in pre-consonantal position (*west bank*). Both pre-vocalic (*west end*) and phrase-final (*west*) position often have a lower rate of deletion, with dialects varying in which of these positions most resists deletion. In (10), we give a sample of the data from the literature on the influence of the external context on *t/d*-deletion. Chicano and Philadelphia

English are examples of dialects with more deletion in pre-vocalic than phrase-final contexts, and the other dialects all have more deletion in phrase-final than pre-vocalic position. We abstract from some aspects of the data by lumping all consonants together. Labov (1989) and Guy (1991a, 1994), amongst others, show that some consonants are more likely than others to induce deletion. Specifically, less sonorous consonants typically are more likely to result in deletion (i.e. more deletion in *best book* than in *best week*). Syllable structure constraints may also play a role (though cf. Labov 1997). For instance, as Guy (1991, 1997) points out, the fact that more deletion is observed before [l] than before [ɹ] (e.g. more deletion in *best luck* than in *best rock*) may be due to the fact that [tɹ-] is a possible onset cluster but [tl-] is not.[2]

(10) Percent deletion in different contexts[3]

	Pre-V *west end*	Pre-Pause *west*	Pre-C *west side*
AAVE (Washington, DC)	29	73	76
Chicano English	45	37	62
Jamaican English	63	71	85
New York City English	66	83	100
Tejano English	25	46	62
Trinidadian English	21	31	81
Philadelphia English	38	12	100

Our analysis draws in particular on that of Coetzee (2004: chapter 5). The constraints assign violation marks as follows.

(11) *CT — Assign a violation mark to a consonant cluster ending in a coronal stop.

MAX — Assign a violation mark to an input consonant that is not present in the output.

MAX-PRE-V — Assign a violation mark to an input consonant in pre-vocalic position that is not present in the output.

MAX-FINAL — Assign a violation mark to an input consonant in phrase final position that is not present in the output.

The markedness constraint (*CT) penalizes specifically a consonant cluster that ends in a [t] or [d] (see Coetzee 2004:252-255 for ways in which the analysis might be extended to labial and dorsal stops). MAX-PRE-V and MAX-FINAL are contextual faithfulness constraints in the spirit of Steriade's licensing by cue constraints (Steriade 2001, 2008). These constraints protect

[2] An alternative account of this asymmetry, consistent with Labov's (1997) observation that the allophonic details do not support the resyllabification account, is that homorganicity blocks release in [tl] but not [tr], and that unreleased consonants have higher rates of deletion (thanks to Lisa Selkirk for discussion).

[3] The data reported in this table come from the following sources: AAVE (Fasold 1972), Chicano (Santa Ana 1991), Jamaican (Patrick 1992), New York City (Guy 1980), (Tejano Bayley 1995), Trinidad (Kang 1994), Philadelphia (Guy 1980).

consonants from deletion where they are more perceptible, i.e. in contexts where their perceptual cues are more robustly licensed. To correctly identify a consonant, it is necessary to perceive both its place and manner of articulation. Information about the place and manner of articulation of consonants is realized in the consonantal release and in the formant transitions into and out of the consonant (Lahiri et al. 1984; Malécot 1958; Nearey and Shammas 1987; Stevens and Keyser 1989; Sussman et al. 1991; Walsh and Diehl 1991; etc.). In pre-consonantal position, it is unlikely that either releases or transitions will be realized, so that there is no special faithfulness constraint that protects *t/d* from deletion in pre-consonantal position. In phrase-final position, consonant releases can be realized so that consonants are more likely to be perceived accurately in this position and hence less likely to delete. This motivates the existence of MAX-FINAL. Before a vowel, both releases and formant transitions can be realized, motivating the existence of MAX-PRE-V.

Although both cues can be realized pre-vocalically and only one phrase-finally it does not mean that pre-vocalic position is always a better sponsor for the consonant. In pre-vocalic position, the cues can only be realized across a word-boundary, while no word-boundary needs to be crossed in phrase-final position. This may explain why some dialects of English have more deletion in phrase-final position and others in pre-vocalic position: dialects with more deletion in pre-vocalic position are dialects that are less tolerant of realizing consonantal cues across word-boundaries. An account that encodes perceptual factors more directly might state this as a constraint, and eliminate MAX-FINAL and MAX-PRE-V in favor of constraints on the preservation of perceptual cues themselves, but we use the MAX-FINAL and MAX-PRE-V constraints for simplicity.

In adopting a perceptually oriented analysis of *t/d*-deletion, we are following not only Coetzee's (2004) markedness-based OT account, but also Labov's (1997, 2004) hypothesis that the differences in rates of deletion across contexts are due to perceptual factors. It is likely that this account could be extended to other aspects of the process, such as the distinctions between consonants that follow *t/d* discussed by Labov (1997), but we leave that for future research.

The four constraints in (11) can give rise to five different categorical systems, given in the table in (12). The first column in this table presents the rankings that must obtain to yield a particular system. The second column gives the total number out of the 24 possible rankings that contain the crucial rankings given in the first column. The final three columns indicate whether or not deletion is observed in each of the three contexts under the crucial rankings in the first column. For example, the first line shows the situation that would hold if MAX outranks *CT. This ranking is observed in half of the 24 possible rankings, indicated by the 12 in the second column. Under this ranking, deletion is blocked in all three contexts, as indicated in the last three columns.

(12) Crucial rankings, number of corresponding total orders, outcomes

	Crucial rankings	Total # rankings	Deletion produced? Pre-V	Phrase-final	Pre-C
a.	MAX >> *CT	12	No	No	No
b.	MAX-PRE-V >> *CT >> {MAX, MAX-FINAL}	2	No	Yes	Yes
c.	MAX-FINAL >> *CT >> {MAX, MAX-PRE-V}	2	Yes	No	Yes
d.	{MAX-PRE-V, MAX-FINAL} >> *CT >> MAX	2	No	No	Yes
e.	*CT >> {MAX, MAX-PRE-V, MAX-FINAL}	6	Yes	Yes	Yes

The generalization to be captured is that pre-consonantal position always has the highest rate of deletion, while there are cross-dialectal differences in the relative rates of pre-vocalic and phrase-final deletion. This is reflected in the categorical patterns in (12). Every possible ranking that yields deletion in pre-vocalic and phrase-final position also yields deletion in pre-consonantal position – rows (b), (c) and (e) in the table. However, there are some rankings that yield deletion only in pre-consonantal position – row (d) (see Anttila 2006 and Prince 2006 for alternative ways of presenting the implications). Kiparsky notes that if variation is the result of speakers varying in the grammars that they use, it is impossible to have a pattern of variation in which pre-consonantal position has the lowest rate of deletion.

As an example of a POC grammar that encodes a probability distribution over outcomes according to the quantitative interpretation in (9), we can take a grammar that imposes no ranking at all on this constraint set. There is only one faithfulness constraint that protects *t/d* in pre-consonantal position (MAX). Whenever MAX ranks below *CT, deletion will be observed in pre-consonantal position, so that 12/24 rankings will result in *t/d*-deletion in pre-consonantal position. Two faithfulness constraints protect *t/d* from deletion in pre-vocalic position, MAX and MAX-PRE-V. In this context, deletion will only be observed if both MAX and MAX-PRE-V rank below *CT, so that 8/24 rankings will result in deletion in pre-vocalic position. Since there are also two faithfulness constraints that are active phrase-finally, deletion will also be observed in 8/24 rankings phrase finally. The predicted deletion rates for POC grammars where the ranking between some of the four constraints is fixed can be determined in a similar manner. The table in (13) gives the predictions for a sample of the possible POC grammars. The first column gives the partial ordering between the constraints that holds for each specific POC grammar. The next three columns show the number of rankings that will result in deletion in each of the three contexts, as well as the predicted probability of deletion in each context. The first POC grammar in the table is the one we discussed in the previous paragraph. The other four are examples of POC grammars with a single fixed ranking, which provide differential rates of deletion between pre-vocalic and phrase-final position. We evaluate the success of this POC analysis in accounting for the actually observed variation in different English dialects in section 0 below.

(13) *Probabilities of deletion from quantitative interpretation of partial orders*

Partial order			Pre-V	Phrase-final	Pre-C
a.	None	# rankings	8/24	8/24	12/24
		p of deletion	0.33	0.33	0.50
b.	MAX-PRE-V >> *CT	# rankings	0/12	4/12	6/12
		p of deletion	0	0.33	0.50
c.	*CT >> MAX-PRE-V	# rankings	6/12	4/12	6/12
		p of deletion	0.50	0.33	0.50
d.	MAX-FIN >> *CT	# rankings	4/12	0/12	6/12
		p of deletion	0.33	0	0.50
e.	*CT >> MAX-FIN	# rankings	4/12	6/12	6/12
		p of deletion	0.33	0.50	0.50

3.3 Predictions of the POC theory

Above we pointed out that the POC theory shares with Labov's (1969) variable rules the property of defining a probability distribution over variants. Anttila (1997), however, draws a distinction between the POC theory and variable rules models: that the POC theory makes stronger predictions about the range of possible variable phonological systems. These predictions come from two sources: OT's universal constraint set, and the POC theory of probability distribution.

Because the POC model is cast within OT, it assumes a constraint set that imposes substantive limits on the range of possible phonological systems. For example, a version of the *t/d*-deletion system in which deletion has a lower probability in pre-consonantal context than in the other environments is ruled out by the absence of a MAX-PRE-C constraint that protects consonants in exactly that environment. This attribute is shared with all OT theories of variation that assume a universal constraint set. We further discuss this difference between OT models of variation and the Variable Rules model in section 0 below.

The POC theory makes even stronger predictions than other OT models of variation about quantitative patterns, ones that appear too strong, as Boersma and Hayes (2001:72) have pointed out. Because the quantitative interpretation of partial orders derives probabilities from the number of rankings that yield a particular pattern, the constraint set imposes restrictions not only on the relative probability of different processes, but also on the absolute probability of each of the processes themselves. For example, in the analysis of English *t/d*-deletion presented in section 0, whether or not t/d-deletion occurs in pre-consonantal position is determined by the ranking of two constraints: MAX and *CT. There are therefore only three probabilities of deletion in this context that the POC theory can derive: 0, .50, and 1. One could always increase the size of the constraint set to yield other probability distributions, such as those observed in the dialects in (10), but this strategy becomes implausible very quickly. Boersma and Hayes (2001:72) point out how it becomes particularly difficult to maintain in cases where the probability distribution between two variants is strongly skewed in favor of one of them. To model a situation where the probability distribution of the two variants is .99 vs. .01 in the POC theory, at least 100 different rankings are required, which necessitates at least 5 constraints (5 constraints can be ranked in 5! = 120 different ways). But five unranked constraints alone would not suffice. To get the correct probability distribution, only one or two (1% of the 120) possible rankings must favor one variant, while the other variant must be favored by 118 or 119 of the possible rankings. This is a very unlikely scenario.

One possible reaction to this shortcoming of the POC approach is to remove the responsibility for producing probability distributions from the grammar. Coetzee (2004, 2006), for instance, claims that grammar only imposes relative probabilities on variants – i.e. grammar dictates that one variant is more probable than another without specifying the absolute probability of the different variants. However, under such an approach, there is no grammatical difference between two systems with probability distributions between two variants of .10 and .90, and .40 and .60. In both, the first variant is less probable than the second. We suspect that native speakers would express a much stronger dispreference for the less frequent variant in the first case (see Boersma and Hayes 2001 and Hayes and Londe 2006 for relevant data). Insofar as the phonological grammar is at least partially responsible for such judgments, if not also for the distributions themselves, it would be preferable to adopt a theory that can distinguish between such systems.

Another possible reaction to this weakness of the POC approach is to change the model in some way so that it can encode probability distributions beyond those allowed by POC. Section 4 is dedicated to the discussion of OT-like theories of grammar that do this, as well as to a brief discussion of their relationship to Labov's (1969) variable rules theory. The theories discussed in section 4 place constraints on a numerical scale. We note that another approach to quantitative aspects of variation in OT is to designate a non-numerical range over which the ranking of a constraint can vary (e.g. Reynolds 1994, Hayes and MacEachern 1998, Davidson, Jusczyk and Smolensky 2004). See Boersma and Hayes (2001) for discussion of the relationship of such a theory to one that incorporates a numerical scale.

4. Probabilistic models of phonology with numerically valued constraints

4.1 Stochastic OT

Boersma (1997, 1998) proposes an elaboration of OT that he refers to as stochastic OT. Boersma and Hayes (2001) provide an introduction to the theory and applications to several cases of phonological variation. Though we follow this tradition in calling this theory stochastic OT, we emphasize that there are other versions of OT, including those discussed in the previous section, that include a stochastic component. In stochastic OT, constraints are given values along a real-numbered scale. However, each time the grammar is used to evaluate a candidate set, the values are converted to a corresponding ranking. The size of the numerical differences between the constraints is irrelevant after this conversion: if C_1 has a value greater than C_2, then the corresponding ranking is $C_1 >> C_2$, irrespective of the size of the $C_1 – C_2$ difference.

The distance between constraints on the numerical scale does play a role in the conversion process itself. Before transforming the numerical values into a ranking, each one is perturbed by adding a different positive or negative number, taken from a normal distribution with a mean of zero. In successive evaluations, constraints that have numerical values sufficiently close to one another will vary in their ranking. This stochastic element of the theory is called “noisy evaluation”.

Stochastic OT can yield probability distributions beyond those produced by the POC theory.[4] As a simple example, we can consider the interaction of *CT and MAX. The tableaux in (14) provide numerical values for the constraints that yield probability distributions that are highly skewed in favor of deletion and retention respectively. The probabilities of the candidates were estimated by submitting the candidate set to evaluation 100,000 times for each of the two sets of constraint values, with an evaluation noise of 2.0 (using Praat’s “get output distributions” function; Boersma and Weenink 2007).

(14) *Skewed probability distributions in stochastic OT*

	101.6	98.2
/CtC/	*CT	MAX
.10 CtC	*	
.90 CC		*

	101.6	98.2
/CtC/	MAX	*CT
.90 CtC		*
.10 CC	*	

A particularly attractive aspect of stochastic OT is the fact that it is accompanied by a learning theory, called the Gradual Learning Algorithm (GLA). The POC theory lacks a learning algorithm, which raises both theoretical and practical difficulties. On the theoretical side, it does not inherit from standard OT the attribute of possessing a provably convergent learning algorithm (see Albright and Hayes, this volume, for further discussion of these and other theories of phonological learning). The Constraint Demotion Algorithms (CDA; Tesar 1995, Tesar and Smolensky 1998, 2000) have never been extended from standard OT to the POC theory. On the practical side, an implemented learning algorithm can aid analysts in working with the theory, as does OT-Soft’s implementation of the GLA and the CDA (Hayes, Tesar and Zuraw 2003) and Praat’s implementation of the GLA (Boersma and Weenink 2007). Without such help, it can be difficult to determine whether a given set of constraints can yield an observed pattern of variation (though see Anttila and Andrus 2006 for a partial solution).

The GLA is an on-line error driven learner, like some versions of the CDA. The learner is presented with one correct input-output pair at a time, and it determines the optimal output for that input, given the current state of its grammar. If that generated output differs from the learning datum, learning is triggered. The GLA updates the constraint values by subtracting

---

[4] The theories are not in a subset relation: see Anttila (2007) for abstract examples of patterns of variation that POC can generate but that cannot be produced by the standard version of stochastic OT.

some value *x* from the ranking values of each constraint that is violated more in the correct form than in the learner's own "error", and adding that same value *x* to all constraints that are violated more in the error.

The constraint values in (14) were obtained by using the implementation of the GLA in Praat. In this simulation, and all of the others we report below, the constraints start out with a value of 100, and the rate of change (*x* in the last paragraph) starts out at 1. The rate of change decreases over the course of learning (by 0.1 after each of four sets of 100,000 learning trials). When provided with the learning datum /CtC/ → [CtC], a learner with this initial state, and with noisy evaluation, might parse it incorrectly, as shown in (15). This would lead to the updated set of values shown as Grammar H_1.

(15) *A learning step in the GLA with stochastic OT*

Grammar H_0: *CT 100, MAX 100

Values with noise: *CT 100.4, MAX 99.8

Learning Datum: /CtC/ → [CtC]

Learner's parse:

/CtC/	*CT	MAX
CtC	*!	
☞ CC		*

Grammar H_1: *CT 99, MAX 101

When provided with sufficient examples of /CtC/ → [CtC], the learner's grammar will eventually reach a state in which *CT is far enough above MAX that errors become vanishingly improbable, and the constraint values cease to change. If the learner is provided with data in which the two mappings /CtC/ → [CtC] and /CtC/ → [CC] both occur, then the learner tends to converge on values that result in probability matching. That is, the learned grammar, with the same evaluation noise, will select each of the inputs with a probability matching their relative frequency in the learning data.

4.2 Noisy Harmonic Grammar

By abandoning numerical constraint weights in favor of ranking, OT distinguishes itself from its predecessor Harmonic Grammar (HG: Legendre, Miyata and Smolensky 1990, Smolensky and Legendre 2006; see Goldsmith 1990, 1993 for early phonological applications). Stochastic OT is essentially a hybrid of the two, in that it reintroduces numerical constraint values for the purposes of modeling variation and learning, while still maintaining ranked constraint evaluation. In this section, we discuss a version of this theory that uses numerical weights in evaluation as well. Because this theory retains stochastic OT's noisy evaluation, we refer to it as Noisy HG. Noisy HG in the form we are using it first appeared in a Praat implementation (Boersma and Weenink 2007); see Goldrick and Daland (2009) on the similar use of noise in connectionist models of speech processing.

In HG, the optimal candidate is the one with the highest numerical *Harmony*, which is the sum of the weighted constraint scores (16). For each constraint *k* in a constraint set of size *K*, the candidate's violation score (s_k) is multiplied by that constraint's weight (w_k). To yield Harmony (*H*), the results are then summed, indicated by the large epsilon in equation (16).

(16) *A candidate's Harmony in HG*

$$H = \sum_{k=1}^{K} w_k \cdot s_k$$

We adopt Legendre, Sorace and Smolensky's (2006) convention of converting OT violation marks to negative integers. In (17), we provide a tableau in which *CT has a greater weight than MAX; the constraint weights are indicated immediately beneath the constraint names. Each candidate's Harmony is indicated in the rightmost column. The candidate with deletion has the highest (closest to zero) Harmony, and is thus optimal. For more detailed discussion of HG and its relation to OT, including the question of whether HG overgenerates typologically relative to OT, see especially Smolensky and Legendre (2006), Pater (2009) and Potts, Pater, Jesney, Bhatt and Becker (to appear).

(17) *A weighted constraint tableau*

/CtC/	*CT 2	MAX 1	*H*
☞ CC		-1	-1
CtC	-1		-2

Variation can be obtained just as in stochastic OT by perturbing the constraint values by noise each time the grammar is used. One way in which this theory differs from stochastic OT is that it is capable of producing cumulative constraint interaction. An example of cumulativity that involves variation, analyzed in HG terms in Pater (2009), comes from the phonology of Japanese loanwords (Nishimura 2003, Kawahara 2006). In native Japanese words, multiple voiced obstruents are categorically banned (by 'Lyman's Law'), which Itô and Mester (1986) account for in terms of an OCP-VOICE constraint (cf. Itô and Mester 2003). Voiced obstruent geminates are also absent in native words, motivating a *VOICED-GEMINATE constraint. In loanwords however, voiced geminates occur (e.g. [webbu] 'web') as do multiple voiced obstruents (e.g. [bobu] 'Bob'). In HG, the loanword pattern requires a weighting of the faithfulness constraint IDENT-VOICE above both of the markedness constraints, as shown in the pair of tableaux in (18).

(18)

/bobu/	ID-VCE 1.5	OCP-VCE 1	
☞ bobu		-1	-1
bopu	-1		-1.5

/webbu/	ID-VCE 1.5	*VCE-GEM 1	*H*
☞ webbu		-1	-1
weppu	-1		-1.5

Cumulativity becomes evident in words that contain both a voiced geminate and another voiced obstruent. As Nishimura and Kawahara show, such words are subject to a process of optional geminate devoicing (e,g, [gutto] ~ [guddo] 'good') that does not affect geminates outside of the Lyman's Law context. The geminate devoicing outcome is shown in (19). In this tableau, the sum of the violations of the constraints with lower weight, OCP-VCE and *VCE-GEM, is greater than that of the constraint with the higher weight, ID-VCE. No OT ranking of these constraints will produce this result.

(19)

/guddo/	ID-VCE 1.5	OCP-VCE 1	*VCE-GEM 1	
guddo		-1	-1	-2
☞ gutto	-1			-1.5

By introducing noise into the evaluation process, we can model the observed variation between the outcomes in (19). The equation to calculate Harmony for a given candidate in Noisy HG is given in (20). It differs from the equation in (16) in that for each constraint, we first sample a random number from a normal distribution with mean zero (N_k), and add it to that constraint's weight before multiplying the result by the violation score.

(20) *A candidate's Harmony in Noisy HG*

$$H = \sum_{k=1}^{k} (w_k + N_k) \cdot s_k$$

As with stochastic OT, a set of constraint values can be obtained by submitting a data distribution to a learning algorithm, and as with stochastic OT, the currently available algorithm for Noisy HG is an on-line error-driven one, termed HG-GLA by Boersma and Pater (2008). The update rule used in HG-GLA is broadly applied in statistical and connectionist learning frameworks, and was perhaps first used in generative linguistics by Jäger (2007) for learning of Maximum Entropy grammar (see the next section on this alternative stochastic version of HG; see Boersma and Pater 2008 for further references on HG-GLA). The sole difference from OT-GLA as described above is that the degree of change for a constraint's value depends on the degree of difference between the correct form and the error. For each constraint, the difference between the number of violations in the error and in the correct form is calculated, and that difference is multiplied by a constant, and added to the constraint's weight to get the updated value. Jäger (2007) and Pater (2008) note that when the error and the correct form differ by a maximum of one violation, the update rule is identical to that of OT-GLA.

A difference between HG-GLA and OT-GLA is that HG-GLA has proofs of convergence. Fischer (2005) provides an adaptation of a stochastic gradient ascent proof for learning Maximum Entropy grammars (see the next section), while Boersma and Pater (2008) provide an adaptation of the Perceptron convergence proof, though this proof is limited to the learning of cases without variation. The stochastic OT/GLA combination has been shown to fail on a relatively simple abstract categorical pattern (Pater 2008).

We supplied a distribution of 50% devoicing for /guddo/, and consistent faithful realization for each of /bobu/ and /webbu/ to two learners implemented in Praat (we assume an even distribution between the outcomes for illustration; we have no information on their relative frequency). The first learner operated with a stochastic OT grammar, and OT-GLA. The results are shown in the row labeled "St-OT" in (21). Because stochastic OT cannot represent this pattern of variation, the OT-GLA failed to converge on a set of values for the constraints. The high values shown in the columns headed by the constraint names are indicative of this non-convergence. The last three columns show the frequency of devoicing that this grammar produces for each input. This set of values does display a limited "cumulative" effect, shown in the higher frequency of devoicing for /guddo/, which will devoice if either OCP-VOICE or *VCE-GEM outranks IDENT-VOICE. However, stochastic OT cannot produce the categorical cumulativity observed in the Japanese data, in which full devoicing of a geminate only occurs in the presence of a second voiced obstruent.[5] The second learner operated with a Noisy HG grammar, and HG-GLA. The weighting values in the final state grammar are as described above: the sum of OCP-VOICE and *VCE-GEM equals that of IDENT-VOICE. This grammar produces a distribution that matches the frequency distribution in the learning data, as shown in (21).

(21) *Grammars learned by stochastic OT and Noisy HG*

	Frequency of devoicing in learning data			*0.0*	*0.0*	*0.50*
	OCP-VOICE	*VCE-GEM	IDENT-VOICE	bobu	webbu	guddo
St-OT	3113.9	3113.9	3113.7	0.15	0.15	0.25
N-HG	66.8	67.6	134.4	0.0	0.0	0.50

As well as demonstrating cumulative constraint interaction, the Japanese loanword example also provides a striking further demonstration that variation is sensitive to "early" phonology. The Lyman's Law restriction against multiple voiced obstruents has all the characteristics of early phonology: it is morphologically restricted (Itô and Mester 1986), and as the loanwords show, has exceptions. Nonetheless, a Lyman's Law violation contributes to the possibility of variable devoicing.

4.3 Maximum Entropy grammar

Johnson (2002) shows how an OT grammar can be transformed into a log-linear probabilistic model, usually referred to as Maximum Entropy grammar; we will use the abbreviation Max-Ent-HG to emphasize that it is formally a stochastic version of HG. Goldwater and Johnson (2003), Wilson (2006), Jäger (2007) and Hayes, Zuraw, Síptár and Londe (to appear) and others apply the resulting model of variation to phonology.[6] In this section, we discuss how this theory relates to Noisy HG.

Max-Ent-HG directly defines a probability distribution over the candidate set: the probability of a candidate is proportional to the exponential of its Harmony. The tableau in (22) illustrates Maximum Entropy grammar with the simple case of variation in Japanese loanwords

[5] We do note that with more phonetically detailed representations and constraints, it might be possible to create a stochastic OT system that yields only gradient devoicing for geminates in isolation, and categorical devoicing for geminates in the context of another voiced obstruent (see especially Kawahara 2006 on the data).

[6] Hayes and Wilson (2008) present another application of the Maximum Entropy framework to phonology. They use a log-linear model to define a probability distribution over the space of possible words, that is, as a model of phonotactics.

(see the above-cited papers for more complete formal presentations). Columns are added showing the result of raising *e* to the power of H (the exponential function), and the probability *p* resulting from dividing a candidate's e^H score by the sum of these scores over the candidate set.

(22)

/guddo/	ID-VCE 2	OCP-VCE 1	*VCE-GEM 1	*H*	e^H	*p*
guddo		-1	-1	-2	0.14	.50
gutto	-1			-2	0.14	.50

Because they both use weighted constraints, Noisy HG and MaxEnt-HG can both represent cumulative constraint interaction, which distinguishes them from stochastic versions of OT. Noisy HG and MaxEnt-HG differ in that Noisy HG produces a single optimal output each time the grammar is used (modulo ties), while MaxEnt-HG defines a probability distribution over the candidate set, which is then sampled to yield an outcome for a given utterance. One result of this difference is that MaxEnt-HG gives a portion of the probability mass to a harmonically bounded candidate (Jäger and Rosenbach 2006); this cannot happen in Noisy HG if the sum of the weight and noise terms is positive (Jesney 2007). The empirical consequences of this and other differences between the two models of stochastic grammar remain to be investigated.

In terms of learning, both grammatical models can be learned with HG-GLA; when applied with Max-Ent-HG, this is stochastic gradient ascent (Jäger 2007). Given an input-output pair as learning data, the learner samples from the candidate set according to the probability distribution defined by the current weights; updating proceeds as outlined in 4.3. An attractive property of Max-Ent-HG is that there are existing proofs of convergence for associated learning algorithms that can be applied to the learning of variation (e.g. Fischer 2005); this has yet to be shown for Noisy HG. It is important to note that there is much room for further development of the learning algorithms for all of these models, since none of the convergence proofs apply when some of the structure of the learning data is hidden (see Boersma and Pater 2008 on Noisy HG, and Eisenstat 2009 and Pater, Smith, Staubs, Jesney and Mettu 2010 on Max-Ent-HG).

4.4 Applications to dialectical differences in t-d deletion

To test the ability of these models of grammar to encode a range of probability distributions, we submitted distributions of *t/d*-deletion matching those from each of the dialects in (10) to learners implemented in Praat. The learners operated with stochastic OT (St-OT), Noisy HG (N-HG) and MaxEnt HG (ME-HG) grammars, using the OT-GLA and HG-GLA learning algorithms described above. For the Noisy HG learner, a non-negativity condition on weights was imposed in evaluation: if the ranking value (post-noise) was less than zero, it was replaced by zero (this is termed Linear-OT in Praat; see Keller 2000, 2006). We will discuss the motivation for the non-negativity condition shortly. The results are presented in (23). For each dialect, the top row indicates the observed proportion of deleted instances of *t/d* in each environment (pre-vocalic = CtV, pre-pausal = Ct, and pre-consonantal = CtC). The following rows show the final state constraint values for each model of grammar, and the encoded probability distributions (estimated using Praat's "get output distributions" method). In all cases but one, all the grammars encode probabilities closely matching the observed frequencies. In fact, the different models match the probabilities so closely that a statistical comparison of closeness of fit is unnecessary.

The deletion patterns are reflected in the values of the constraints: when /CtV/ has the lowest rate of deletion, MAX-P-V has a higher value than MAX-FIN, and when /Ct/ has the lowest rate of deletion, the relationship is reversed.[7]

(23) Observed and learned *t/d*-deletion rates for different dialects of English

		*CT	MAX-P-V	MAX-FIN	MAX	CtV	Ct	CtC
AAVE (Washington, DC)						*0.29*	*0.73*	*0.76*
	St-OT	101.0	102.3	96.8	99.0	0.29	0.73	0.76
	N-HG	101.0	5.8	-1.5	97.2	0.29	0.73	0.77
	ME-HG	100.6	2.1	0.2	99.4	0.30	0.74	0.77
Chicano						*0.45*	*0.37*	*0.62*
	St-OT	100.4	99.7	100.6	99.6	0.45	0.37	0.62
	N-HG	100.4	1.0	1.8	99.6	0.43	0.36	0.60
	ME-HG	100.2	0.7	1.0	99.8	0.44	0.36	0.61
Jamaican						*0.63*	*0.71*	*0.85*
	St-OT	101.4	100.0	99.2	98.6	0.63	0.70	0.84
	N-HG	101.5	1.7	0.8	98.5	0.63	0.70	0.85
	ME-HG	100.9	1.2	0.8	99.1	0.64	0.73	0.85
New York City						*0.66*	*0.83*	*1.00*
	St-OT	107.6	106.5	104.9	92.4	0.66	0.84	1.00
	N-HG	141.1	80.9	79.0	58.9	0.65	0.83	1.00
	ME-HG	140.4	80.3	79.3	59.6	0.65	0.83	1.00
Tejano						*0.25*	*0.46*	*0.62*
	St-OT	100.4	101.9	99.6	99.6	0.25	0.46	0.62
	N-HG	100.3	1.5	0.7	99.7	0.25	0.47	0.62
	ME-HG	100.4	3.2	0.7	99.6	0.27	0.46	0.63
Trinidad						*0.21*	*0.31*	*0.81*
	St-OT	101.2	103.4	102.5	98.8	0.21	0.31	0.80
	N-HG	101.2	5.2	4.1	98.8	0.21	0.31	0.80
	ME-HG	100.7	2.8	2.2	99.3	0.21	0.32	0.81
Philadelphia						*0.38*	*0.12*	*1.00*
	St-OT	107.2	108.2	110.6	92.8	0.37	0.12	1.00
	N-HG	139.2	79.4	82.4	60.8	0.38	0.12	1.00
	ME-HG	139.5	79.5	81.0	60.5	0.38	0.12	1.00
Tejano'						*0.62*	*0.46*	*0.25*
	St-OT	99.8	-6511.3	-523.2	100.2	0.45	0.45	0.45
	N-HG	99.8	-6382.1	-735.2	100.2	0.44	0.44	0.44
	ME-HG	99.4	-1.6	-0.8	100.6	0.61	0.42	0.24

The one case in which the learners did not all succeed in probability matching is labeled "Tejano-prime". This distribution was created by trading the proportions of deletion between

[7] All of the input files used in the learning simulations reported here are included in "coetzee-pater-variation.zip", which is available from the authors, or from http://people.umass.edu/pater/coetzee-pater-variation.zip.

pre-consonantal position and pre-vocalic position from real Tejano. The result is a pattern that exists in no known dialect: lowest frequency of deletion in pre-consonantal position. Stochastic OT and Noisy HG were unable to capture this pattern. For stochastic OT, as in the POC theory, this is because no ranking of the constraints yields deletion in only pre-consonantal position. Since stochastic OT produces a probability distribution over rankings, its restrictions on relative rates of variable processes have the same basic character as those of the POC theory. Turning to the Noisy HG result, we note first that a categorical version of HG that restricts weights to positive values generates the same five languages as OT.[8] Since Noisy HG produces a probability distribution over weightings, in cases like these where the OT and HG typologies converge, it also yields the same basic restrictions on relative rates of variable processes as POC and stochastic OT. This result is dependent on banning negative weights for the constraints, since a constraint that is negatively weighted will *prefer* the structure that violates the constraint. For the Noisy HG grammar for Tejano', the negatively weighted constraints have no effect on evaluation, since the values are converted to zero at evaluation time. The effect of negative weights is illustrated by the Max-Ent-HG result in (23); the version that we used had no non-negativity restriction (this is not a necessary property of Max-Ent-HG models *per se*, since they can also be limited to positive weights). It was thus able to find a weighting that disprefers pre-consonantal deletion, by rewarding deletion in the pre-vocalic and phrase-final positions. Since constraint violations are marked with negative numbers, if the constraint weight is negative, then the product of the weight and the number of violations results in a positive increase of the Harmony of the candidate.

4.5 Variable Rules

As we pointed out earlier, Labov's (1969) variable rule notation specifies elements of the context of a rule as affecting the probability of application of the rule. We now show how this probability is calculated, and compare this probabilistic theory of grammar to the constraint-based ones just discussed. In a standard categorical rewrite rule of the form A → B / C __ D, A is changed to B every time that it occurs in the context C __ D, and only then. Whenever all the components in the structural description of the rule (C __ D) are present, the probability of rule application is 1.0, and whenever any of these components is absent, the probability is zero. Labov introduced the notion of weighting the components in the structural description such that each component contributes to the probability of rule application. Under this interpretation of rewrite rules, it becomes possible to say, for instance, that the presence of C __ in the context increases or decreases the probability of rule application by some specific factor, and similarly for the presence of __ D. The rule can now apply even if neither, or only one of, C __ and __ D is present, and the probability of application can take on any value between zero and 1.0.

Several mathematical models have been proposed over the years for relating the observed application rate of some rewrite rule to the presence/absence of different components of the rule's context (Cedergren and Sankoff 1974; Rousseau and Sankoff 1978; Guy 1993; etc.). The one that has become the standard in the field, and that is implemented in the widely used software packages of VarbRul and Goldvarb, performs a multivariate stepwise logistic regression over observed token counts (Paollilo 2002:177; Sankoff, Tagliamonte and Smith 2005). In this

[8] Readers interested in verifying this result can submit the file "typology.txt" from "coetzee-pater-variation.zip" to OT-Help (Becker, Pater and Potts 2007).

analysis, application/non-application of the rule is taken as the dependent variable, and different factors hypothesized to influence the probability of application are taken as independent variables. Given a corpus of observed tokens to which the rule could apply, and in which each token is coded for application/non-application of the rule, as well as for the value for each of the independent variables that hold of the specific token, VarbRul/Goldvarb estimates the contribution that each independent variable makes to the probability of rule application, using a maximum likelihood algorithm. In this model, the probability that some rule will apply is given by the expression in (24). In this expression, p_0 is the "input probability", or the probability that the rule will apply independent of any of the contextual factor variables. The different independent variables are represented by *1* to *n*, and p_1 to p_n are then the contribution that each of these variables makes to the probability of rule application, as determined by the maximum likelihood algorithm.[9] Stated in more concrete terms, an underlying form like /w□st/ is subject to a variable deletion rule and therefore has two possible surface forms [w□st] and [w□s]. The expression in (24) defines a probability distribution over the two possible surface forms with the probability of the surface form to which the rule has applied ([w□s]) being *p*, and the probability of the form to which the rule has not applied being (1 – *p*). These probabilities depend on the values of p_0 to p_n.

(24) *Probability of rule application in the variable rule framework*

$$p = \frac{p_0 \times \ldots \times p_n}{[p_0 \times \ldots \times p_n] + [(1 - p_0) \times \ldots (1 - p_n)]}$$

To provide an illustration of this model that can be easily compared with those of the grammar models discussed above, we created a corpus for the Tejano data (Bayley 1995) that we also discussed in sections 3.2 and 4.4 above. Since we did not have access to Bayley's original corpus, we made a toy corpus with the three contexts (pre-consonantal, pre-vocalic, and phrase-final) appearing 100 times each, giving 300 tokens total. Each token was coded for whether or not *t/d*-deletion applied, using the frequency of deletion that Bayley reports for each context (62 times for pre-consonantal position, 46 times for phrase-final position, and 25 times for pre-vocalic position). We also coded each token for one independent variable, namely the following phonological context (i.e. either pre-consonantal, pre-vocalic, or pre-pausal). We then submitted this corpus to Goldvarb X – the most recent version of the software package developed by David Sankoff to implement the statistical analysis described just above (Sankoff, Tagliamonte and Smith 2005). The output generated by Goldvarb X is given in the first row of (25); variables with weights greater than .5 increase the probability of rule application, and lower weights decrease it. As explained above, p_0 is the input probability. Since we coded our data for only one independent variable (the following context), there are values only for p_1 in addition to the input probability. The expected deletion rates in the three contexts can now be calculated by substituting the value for p_0 and the appropriate value for p_1 into the equation in (24). For instance, to calculate the expected deletion rate in pre-vocalic position, we substitute .44 for p_0 and .30 for p_1. The resulting formula is shown in (26). Solving for *p* in this formula gives .25, which thus specifies the expected application rate of the *t/d*-deletion rule applying in pre-vocalic position. Substituting .52 for p_1 and solving for *p* gives .46 as the expected deletion rate in pre-

[9] For a detailed discussion of the mathematical model underpinning variable rule analyses, see Paollilo's instructive and accessible study (Paollilo 2002).

pausal context, and substituting .68 for p_1 gives .62 as the expected deletion rate in pre-consonantal context.

(25) *Goldvarb X outputs*[10]

	Factor weights				Observed and expected deletion rates					
	p_0	p_1			CtV		Ct		CtC	
		CtV	Ct	CtC	O	E	O	E	O	E
Tejano	.44	.30	.52	.68	25	25.03	46	46.00	62	61.97
Tejano'	.44	.68	.52	.30	62	61.97	46	46.00	25	25.03

(26) *Expected probability of t/d-deletion in pre-vocalic position in Tejano English*

$$p = \frac{.44 \times .30}{[.44 \times .30] + [(1 - .44) \times (1 - .30)]} = .25$$

Variable rules, like stochastic versions of OT and HG, define a probability distribution over the possible surface forms for a given underlying representation. One difference between these two approaches is that OT and HG models are usually assumed to be restricted in the types of pattern that they can express by their use of a universal constraint set (though cf. Boesma 1998, Hayes, Zuraw, Síptár and Londe to appear). We illustrated this aspect of these models in section 4.4 with the case of Tejano' – a dialect similar to actual Tejano except that the deletion rate in pre-vocalic and pre-consonantal position were inverted. Because the constraint set did not include a constraint blocking deletion in pre-consonantal position, the greater frequency of deletion in pre-vocalic position in Tejano' could not be modeled (with strictly positive weights in HG). Perhaps unsurprisingly, as shown in (25), our application of the variable rule theory was able to match the Tejano' distribution just as well as it could match Tejano. Paolillo (2002: ch. 10) provides an explicit comparison of partially ordered constraint theory (POC) with the variable rules model by reanalyzing Anttila’s (1997) Finnish data, and defends the lack of substantive restrictions on variable rules. The question of whether and how substantive restrictions on phonological processes should be encoded is of course a generally controversial issue in phonological theory; see Hale and Reiss (2000) and Hayes, Kirchner and Steriade (2004) for two poles of the debate (see further Hansson, this volume, see also Odden, this volume on rules and constraints). Though the variable rule framework does not encode the distinction between plausible and implausible phonological rules, it is worth noting that this distinction has often guided practice in formulating variable rule analyses; see especially Labov (2004) for discussion.

Another difference between research using variable rules and stochastic OT/HG is that social factors are usually included in variable rule analyses, but rarely, if ever, figure in analyses using the constraint-based frameworks. In variable rule theory, a variable corresponding to a social factor like register can appear alongside linguistic variables like phonological context. For example, if *t/d*-deletion is more frequent in an informal register in Tejano English, this can be expressed with a second independent variable p_2. To illustrate, we show in (27) the result if this variable has the value .70 for the informal register, which raises the probability of rule

[10] The files that were used as input to Goldvarb X are included in the aforementioned “coetzee-pater-variation.zip”.

application to .44. If the formal register had value 0 for this variable, it would have the .25 deletion rate calculated in (26).

(27) *Expected probability of pre-vocalic t/d-deletion in a hypothetical informal register of Tejano English*

$$p = \frac{.44 \times .30 \times .70}{[.44 \times .30 \times .70] + [(1-.44) \times (1-.30) \times (1-.70)]} = .44$$

Some research in OT has proposed accounts of style/register differences. For example, van Oostendorp (1997) proposes that increasingly formal registers have increasingly high rankings of faithfulness constraints (see also Itô and Mester 2001). In this proposal, however, each register is associated with its own categorical grammar, and variation results only from the selection of a different register and grammar. It is hard to see how such a model could capture the observed differences in frequency of deletion observed across languages/dialects, as in the English *t/d*-deletion case.

Boersma and Hayes (2001: Appendix C) suggest an approach to stylistic differences in stochastic OT in which the ranking value of a constraint is modified by a term expressing the degree to which that constraint's value changes in a given style. Concretely, they propose that *Style* is a real-numbered variable ranging from 0 to 1 (with 1 the formality maximum), and that *style-Sensitivity* is a constraint-specific variable that can take on positive and negative values. These variables are multiplied, and the result is added to the constraint's ranking value. In HG, this proposal would result in the Harmony equation in (28).[11]

(28) *A candidate's Harmony in a style-sensitive HG*

$$H = \sum_{k=1}^{K} (w_k + style \cdot styleSens_k) \cdot s_k$$

Boersma and Hayes do not provide a learning algorithm for the stochastic OT version of this model. In HG at least, such algorithms do seem to be available. Given a version of HG that defines probability distributions over candidate sets, along with learning data annotated for style, one learning objective would be to find the values of the variables in (28) that minimize the difference between the observed distributions and the expected ones, that is, that minimize error. This sort of numerical optimization problem can be solved by a range of algorithms. The practical question of how to find weights for an analysis thus seems easily resolvable; a model that makes predictions about learning paths would be more challenging to construct.

We see the development of this and other variants of HG that include social variables as a particularly promising direction for further research. The development of these models could build on sociolinguistic research that examines the manner in which social variables impact the application of variable rules.[12] In the probabilistic model proposed by Cedergren and Sankoff (1974), as well in Labov's (1969) earlier additive model and the later Goldvarb implementations,

---

[11] As Edward Flemming (p. c.) points out, this approach has a precedent in the 'carefulness' weight of Lindblom's (1990) H & H theory of speech production, which controls the extent to which target undershoot is minimized.

[12] Thanks to John McCarthy for discussion of the issues raised in this paragraph, and especially for pointing us to Lim and Guy (2005).

factors are independent. In terms of social and phonological factors, this predicts that the relative contribution of phonological factors cannot vary across registers. For example, the standard variable rules model cannot accommodate a variety of English that had the lowest rate of *t/d*-deletion pre-vocalically in one register, and pre-pausally in another. Although we do not know if such a dialect in fact exists, recent research suggests that the assumption of independence may be too strong (Bayley 2002:130-134, Lin and Guy 2005). The model in (27) would allow for such non-independence, in that any subset of the constraints can be made sensitive to style (e.g. MAX-FINAL, but not MAX-PRE-V). A more restrictive model would make style sensitivity uniform across constraints, and/or limit the ways that constraint weights can be altered across registers, as in van Oostendorp's (1997) proposal; see relatedly Coetzee (2009a, 2009b). Further empirical research and development of these and other models is needed to choose between them.

5. Lexically conditioned variation

We now return to the main rhetorical thread of this chapter: the argument that variation cannot be left to "late phonology". In the last section, we discussed cases of phonological variation from English and other languages that display a characteristic of an early phonological process: sensitivity to morphological category. In this section, we discuss evidence that variable processes also test positive on another purported diagnostic of early processes: sensitivity to lexical idiosyncrasy. This demonstration serves three purposes. First, it cements the case that it is insufficient to relegate variation to late phonology (or phonetic implementation), insofar as late phonology operates strictly on the output of early phonology, and is disallowed access to lexical representations. Second, it serves to raise some issues for the accounts of variation in OT and OT-like models discussed in the previous section. Third, it serves to introduce some of the data that have formed the basis of recent arguments that a standard assumption about lexical representation in generative phonology is inadequate, that instead of a single abstract underlying form, each morpheme is associated with a distribution over a set of phonetically detailed exemplars (Bybee 2001, Pierrehumbert 2001, 2002; see Ladd, this volume and Baayen and Ernestus, this volume). We start by examining the consequences of these data for theories of phonological grammar, in particular the constraint-based ones discussed in section 4, before moving on to briefly discuss the implications for theories of the phonological lexicon.

5.1 Lexically conditioned variation in English

One of the cases of lexically conditioned variation discussed by Bybee (2001) has figured prominently in generative phonology ever since its conception: English secondary stress and vowel reduction. As Chomsky and Halle (1968) and many subsequent investigators have noted, words of the same phonological shape often have different secondary stress patterns. If we follow Chomsky and Halle and take Kenyon and Knott's (1953) pronunciation dictionary as our data source, words fall into three classes: a syllable of a particular type in a particular position can be consistently stressed (have a full vowel), consistently stressless (have a reduced vowel), or vary between stressed and stressless. In (29), we provide two examples from Pater (2000); see that paper for further discussion and references to earlier work.

(29) a. *Sonorant-final syllables that follow a heavy syllable and precede a stressed syllable*

Stressed: aug<u>men</u>tation, con<u>den</u>sation, im<u>por</u>tation, chim<u>pan</u>zee, in<u>car</u>nation, os<u>ten</u>tation
Stressless: in<u>for</u>mation, seg<u>men</u>tation, trans<u>port</u>ation, Mo<u>zam</u>bique

Variable: advantageous, authenticity, condemnation

b. *Sonorant- and obstruent-final syllables in initial pretonic position*

Stressed: bandana, pontoon, bacteria, cognition, emporium, excursus
Stressless: Atlantic, admire, companion, confection, embrace, excursion
Variable: ambassador, Atlanta, Kentucky, September, sincere, obscene, accelerate

The data in (29) abstract from important subregularities. For example, it seems that the stressed category in the (a) cases is less well populated than in the (b) cases, and is usually dependent on the presence of a stress in the base form of a derived word. In addition, stresslessness in the (b) cases seems more productive in words with (historic) Latinate prefixes. Dealing with these subregularities would take us too far afield. For present purposes, we note just that the lexical idiosyncrasy cannot be explained away by the subregularities: derived words in (a) vary in whether they preserve the stress of their bases, and not all the cases of stress in this position are in derived words. For the cases in (b), it is not just words with Latinate prefixes that show reduction. Thus, any descriptively adequate account of these facts will have to accord a role for the lexicon in determining whether or not reduction takes place. Crucially, the lexicon does not fully determine whether reduction occurs: see for example the discussion of the categorical absence of stress on non-initial pretonic light syllables in Pater (2000).

It is of course possible that what Kenyon and Knott (1953) transcribed as variation was confined to inter-speaker variation, and did not include any genuine cases of within-speaker variation. However, we see it as highly likely that for at least some speakers, there are words like those in (29) for which there are two acceptable pronunciations, which are both produced in utterances that are in all relevant respects identical. We also suspect that variation is underreported in Kenyon and Knott (1953).

English vowel reduction thus demonstrates that a variable process can display lexical idiosyncrasy in whether it applies or not. We now return to the case of English *t/d*-deletion to discuss evidence that the role of the lexicon in variation can be more fine-grained: that it can affect the frequency of application of a variable process. Thus, the determination of the probability of application of a phonological process must take into account not only the morphological and phonological factors discussed in the last section, but also the lexical item in question.

Like English vowel reduction and other lenition processes, *t/d*-deletion is more likely to apply to words with a higher usage frequency than words with a lower usage frequency (see especially Phillips 2006). Studies on *t/d*-deletion often leave out words like *just*, *went*, *and*, and *n't*, since these words typically show anomalously high deletion rates. As pointed out by both Bybee (2000:70) and Patrick (1992:172), these are words that are used with very high frequency. Motivated by this observation, Bybee reanalyzed the data collected by Santa Ana (1991) for Chicano English. She selected 2049 tokens of words that end on /-Ct/ or /-Cd/ from Santa Ana's corpus. All these tokens were then divided into two groups based on their Francis-Kučera (1982) frequency. The "high frequency" group all appeared 35 or more times per million, and the "low frequency" group less than 35 times per million. As shown in (30), she found significantly higher deletion rates in the high frequency than in the low frequency words. See also Coetzee (2009a, 2009b) for similar evidence from different dialects of English.

(30) Rate of *t/d*-deletion in Chicano English

	Deletion	*Retention*	*% Deletion*
High frequency	898	752	54.4%
Low frequency	137	262	34.4%

Bybee (2001) also draws attention to the role that lexical frequency plays in the propensity for vowel reduction. As first noted by Fidelholtz (1975), frequent words are more likely to show reduction. Indications of this correlation can be glimpsed in the words in (29): *transportation* and *information* are more frequently used than *importation*, and *embrace* and *excursion* are more common than *emporium* and *excursus*. We do not take a position here on how or even whether the correlation between usage frequency and the frequency of the application of a variable process should be captured in a model of phonological grammar (see Coetzee 2009a, 2009b for two different approaches). At a minimum, however, these examples show that the probability of participation in a variable process is conditioned to some extent by lexical idiosyncrasy. Furthermore, even in a model that does take word frequency into account, there is almost certainly a residue of lexical influence that is independent of frequency (that is, of "exceptionality"). For example, Coetzee (2009a) calculated the deletion *t/d*-deletion rate for several words in the *Buckeye Corpus* (Pitt *et al*. 2007).[13] Although he found a positive correlation between lexical frequency and deletion rate, there are many individual words that do not follow this general trend. The words *sound* and *friend*, for instance, have very similar frequencies in CELEX (Baayen *et al*. 1995) and are also phonologically and morphologically similar (both are monolsyllabic monomorphemes that end on /-nd/). However, they have very different deletion rates in pre-consonantal position, as shown in (31). The words *list* and *east*, which are also morphologically and phonologically similar (monosyllabic monomorphemes that end on /-st/), have very different CELEX frequencies, but more comparable deletion rates in pre-consonantal position. Lexical frequency, morphological status and phonological properties do not fully determine the probability of *t/d*-deletion, and accepting some lexical idiosyncrasy seems unavoidable.

(31) *Deletion rate of selected words from the Buckeye Corpus*

Word	*CELEX frequency*	*Log CELEX frequency*	*Deletion rate in Pre-Consonantal context*
sound	2,989	3.48	62
friend	3,087	3.49	81
list	1,360	3.13	56
east	2,409	3.38	59

[13] This corpus consists of phonetically transcribed audio recordings of sociolinguistic interviews with 40 speakers of the Columbus, Ohio, dialect of English. For more detail on the corpus, see Pitt *et al.* (2007). For more on how the deletion rates were determined, see Coetzee (2009a).

We have focused here on the lexical conditioning of two English variable processes because they have been the subject of such careful scrutiny, and because it has been possible for us to relatively easily directly observe their application. However, we note that there are other well-studied cases of phonological variation that clearly show the role of lexical idiosyncrasy. Dutch has the exact same process of *t/d*-deletion, and this process in Dutch also has the same properties as in English – the process applies at different rates to lexical items that are identical in all relevant phonological properties and that have the same usage frequency (Goeman 1999; Goeman and van Reenen 1985; Hinskens 1992, 1996; Schouten 1982, 1984). For instance, the words *blaast* 'blow-3SG PRESENT' and *danst* 'dance-3SG PRESENT' have virtually identical CELEX frequencies (104 and 105, respectively), yet they differ quite substantially in their deletion rates: *blaast* undergoes deletion 16% of the time and *danst* 9% (Goeman 1999:182; Phillips 2006:65).[14]

Another process that provides a nice parallel to English *t/d*-deletion is French "schwa" deletion.[15] It similarly has some of the hallmarks of a late phonological process. Deletion of schwa from the initial syllable of a polysyllabic word (e.g. [sœmɛn] vs. [smɛn] *semaine* 'week') is variable and sensitive to rate of speech and style, is sensitive to phrasal context, and it seems to sometimes produce between-category outcomes (Fougeron and Steriade 1997, Barnes and Kavitskaya 2003; cf. Côté and Morrison 2007). However, Dell (1973/1980:206) notes that deletion of schwas from word-initial syllables does have exceptions, and claims that careful study of the phonological properties of the exceptional and non-exceptional words reveals no "simple regularity" that predicts whether a word will be exceptional. Thus, we have another process that combines aspects of early and late phonology (see Walker 1996 for discussion of the challenges that schwa deletion poses for Lexical Phonology in particular). Dell also notes that the exceptional words tend to be rarely used or literary. Racine and Grosjean (2002) conducted a production study with speakers of Swiss French, and found a continuum of frequency of deletion across different lexical items. While they did find a correlation between frequency of use of the lexical items in a corpus and frequency of deletion in their experiment, they note that the correlation is not perfect. This supports the view that at least some between word differences are purely idiosyncratic. Thus, as with English *t/d*-deletion, the available evidence indicates that words can be idiosyncratically resistant, to various degrees, to the variable process of French schwa deletion. And as with English *t/d*-deletion, the wealth of research on French schwa deletion makes this an ideal empirical base for the further development of theories of variation.

5.2 Lexically indexed faithfulness constraints and variation

As well as being problematic for theories that deny access of variable processes to the lexicon, the data we have just discussed are problematic for the constraint-based models of variation

[14] Goeman calculates the deletion frequencies from the "Phonological and Morphological Properties of Dutch Dialects" database. See Goeman and Taeldeman (1996) for more on this database.

[15] The scare quotes are to indicate that the vowel is not typically a phonetic schwa. We cannot here do justice to the complexities of this phenomenon, nor to the voluminous literature on the topic. Our brief discussion is based mostly on Dell (1973/1980); see Durand and Laks (2000) for discussion of Grammont's classic treatment, Eychenne (2006) for a recent overview of the generative literature and OT analysis, Tranel (2000) for an earlier OT approach, and Kimper (2008) on schwa deletion in a serial version of OT.

overviewed in sections 3 and 4, insofar as they also provide no way for the lexicon to affect the application of a variable process. Here we discuss one solution that draws on an existing proposed elaboration of OT: the indexation of faithfulness constraints to individual lexical items. Lexically indexed faithfulness constraints were introduced into OT in Pater (2000) as a means of dealing with influences of the lexicon on English secondary stress like those illustrated in (29).

As a simple illustration of how this would work for *t/d*-deletion, we provide the results of another learning simulation, this time giving the learner a hypothetical set of frequencies of deletion for both *feast* and *most* in the three "external" contexts, with deletion being more frequent for *feast* in each case. For both words, frequency of deletion was greatest in pre-consonantal position, intermediate in pre-vocalic position, and least in pre-pausal position. The learner had a Noisy HG grammar, and we elaborated our constraint set by creating versions of the faithfulness constraints specific to each of these lexical items. The results, provided in (32), show that the learner succeeded in probability matching for the distributions of deletion for both words. The weights of the faithfulness constraints specific to *feast* are higher than those specific to *most*, so that the grammar produces lower rates of deletion for *feast* than *most*.

(32) *Results of a learning simulation with lexically specific constraints*

*Ct	MAX-PRE-V-/FEAST/	MAX-FIN-/FEAST/	MAX-/FEAST/	MAX-PRE-V-/MOST/	MAX-FIN-/MOST/	MAX-/MOST/
100.71	1.18	3.58	100.07	1.16	3.27	99.22

	feast_V	feast	feast_C	most_V	most	most_C
Probability of deletion in learning data	*0.40*	*0.20*	*0.60*	*0.50*	*0.30*	*0.70*
Probability of deletion in learned grammar	0.40	0.20	0.59	0.50	0.30	0.70

Lexically specific constraints are a departure from standard generative assumptions about the nature of lexical representation, and its interaction with the grammar, in that they blur the distinction between the lexicon and the grammar, allowing them to interact in novel ways.[16] Such an interactive theory of the phonological grammar and the lexicon seems to be required to allow the lexicon to impact the application of variable processes. There is also increasing evidence from psycholinguistic research for richer interactions between the lexicon and perception and production than standard models permit; see Baayen (2007) and Baayen and Ernestus, this volume for reviews.

There are a number of alternatives to lexically specific constraints that deserve to be explored, especially in conjunction with weighted constraint and other stochastic theories of grammar. A theory of the phonological lexicon with numerically enriched representations would likely resemble existing connectionist models; this resemblance could be exploited in the development of accounts of processing and learning (see e.g. Goldrick 2008). Another way to

[16] Lexically specific rules used in Chomsky and Halle (1968) and much subsequent work (see Zonneveld 1973), which give a similar power to the lexicon, became unfashionable with the rise of autosegmental phonology in the 1980s.

develop a highly interactive theory of lexical representation and phonological grammar would be to build on research on probabilistic approaches to language. If, as Pierrehumbert (2001) has proposed, the lexicon defines a probability distribution over phonetically detailed forms of each morpheme, then this probability distribution could be made to interact with the probabilities for different contexts given by the phonological grammar.

Finally, in the context of a discussion of the nature of the representations manipulated by a phonological grammar, we must return to the issue with which we began this chapter. The view that variation belongs in the late phonology is consistent with the observation that variable processes often yield between-category outcomes. As we mentioned above, this observation holds for the case that we have paid the most attention to, English *t/d*-deletion (Browman and Goldstein 1990). The analyses that we have discussed do not generate these outcomes, being limited to categorical deletion or preservation of the word-final consonant. Our discussion emphasized the advantages of numerically weighted constraints for modeling the probabilities of deletion across dialects. It is likely that weighted constraints are also better suited than ranked constraints for generating phonetically gradient outcomes (though *cf.* Boersma 1998), and may well help in this way to flesh out the accounts of *t/d*-deletion we have sketched (see Flemming 2001 and Kirchner and Moore 2008 on unified HG accounts of phonetics and phonology).

6. Conclusions

A theory of generative phonology that produces a probability distribution over outputs for an underlying representation has existed for nearly 40 years. Research using Labov's variable rules model has generated a wealth of information about the nature of phonological variation, of which we have discussed only a small sample. We have drawn on some of these data to show that variable processes are necessarily phonological, in that they are conditioned by morphological and lexical factors. We have also highlighted the overlap between the phonological factors conditioning the application of categorical and variable processes. Phonologists outside of the sociolinguistic tradition have been reluctant to embrace variable rules. Some (but clearly not all) of this reluctance may be attributed to the fact that the relative weights given to the variables in these rules are completely foreign to the formalisms used in standard rule-based phonology. The situation is quite different for the constraint-based theories of phonological variation that we have surveyed in this chapter. In these, the core formal mechanism of constraint prioritization (by ranking or weighting), determines cross-linguistic differences in probability of process application, just as constraint prioritization determines categorical differences between languages in the original version of OT. That constraint-based models of probabilistic phonology are firmly grounded in the core formalisms of the frameworks bodes well for the continued placement of variation as a central topic of research in generative phonology.

References

Albright, Adam and Bruce Hayes. This volume. Learnability.

Anttila, Arto. 1997. Deriving variation from grammar. In Frans Hinskens, Roeland van Hout and Leo Wetzels, eds. *Variation, Change and Phonological Theory.* Amsterdam: John Benjamins. p. 35-68.

Anttila, Arto. 2002. Variation and phonological theory. In J.K. Chambers, Peter Trudgill and Natalie Schilling-Estes, eds. *The Handbook of Language Variation and Change*. Oxford: Blackwell. p. 206-243.

Anttila, Arto. 2006. Variation in Optimality Theory. Handouts from the Indiana Phonology Fest. [Downloaded on September 25, 2009 from http://www.stanford.edu/~anttila/teaching/indiana/variation.html]

Anttila, Arto. 2007. *Variation, Typology, Universals*. Class handout for LSA.365: Variation in Optimality Theory. Class taught at the LSA Linguistic Institute, Stanford University. [Downloaded on December 27, 2007 from http://www.stanford.edu/~anttila/teaching/lsa-2007/variation.html.]

Anttila, Arto and Curtin Andrus. 2006. *T-Orders*. Ms. Stanford University.

Baayen, R. Harald. 2007. Storage and computation in the mental lexicon. In Gonia Jarema and Gary Libben, eds. *The Mental Lexicon: Core Perspectives*. Oxford: Elsevier. p. 81-104.

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

Barnes, Jonathan and Darya Kavitskaya. 2003. Phonetic analogy and schwa deletion in French. In Julie Larsen and Mary Paster, eds. *Proceedings of the Twenty-Eighth Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society. p. 39-50.

Barry, Martin. 1985. A palatographic study of connected speech processes. *Cambridge papers in phonetics and experimental linguistics* 4.

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

Bayley, Robert. 1997. Variation in Tejano English: evidence for variable Lexical Phonology. In C. Bernstein, T. Nunnaly, and R. Sabino, eds. *Language Variety in the South Revisited.* Tuscaloosa: University of Alabama Press.

Bayley, Robert. 2002. The quantitative paradigm. In J.K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, eds. *The Handbook of Language Variation and Change.* Oxford: Blackwell. p. 117-141.

Becker, Michael, Joe Pater and Christopher Potts. 2007. OT-Help 1.1. Software package. University of Massachusetts, Amherst. [http://web.linguist.umass.edu/~OTHelp/].

Bod, Rens, Jennifer Hay, and Stefanie Jannedy. 2003. Probabilistic Linguistics. MIT Press.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **21**. 43-58. [Available on Rutgers Optimality Archive, ROA-221.]

Boersma, Paul. 1998. *Functional Phonology: Formalizing the Interaction Between Articulatory and Perceptual Drives*. The Hague: Holland Academic Graphics. [Doctoral dissertation, University of Amsterdam.]

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 Learner in Harmonic Grammar*. Manuscript. Amsterdam and Amherst, MA: Meertens Institute and University of Massachusetts Amherst. [Available on Rutgers Optimality Archive, ROA-970.]

Boersma, Paul and David Weenink. 2007. *Praat: Doing Phonetics by Computer. (Version 5.0.01).* [Computer program]. Retrieved December 18, 2007, from http://www.praat.org/

Browman, Catherine P. and Louis Goldstein. 1990. Tiers in articulatory phonology, with some implications for casual speech. In John Kingston and Mary E. Beckman, eds. *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge: Cambridge University Press. p. 341-376.

Bybee, Joan. 2000. The phonology of the lexicon: evidence from lexical diffusion. In Michael Barlow and Suzanne Kemmer, eds. *Usage-Based Models of Language.* Stanford: CSLI Publications. p. 65-85.

Bybee, Joan. 2001. *Phonology and Language Use.* Cambridge: Cambridge University Press.

Cedergren, Henrietta J. and David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language*, 50:333-355.

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Coetzee, Andries W. 2004. *What It Means to Be a Loser: Non-Optimal Candidates in Optimality Theory.* Ph.D. dissertation, University of Massachusetts Amherst.

Coetzee, Andries W. 2006. Variation as accessing "non-optimal" candidate. *Phonology*, 23:337-385.

Coetzee, Andries W. 2009a. An integrated grammatical/non-grammatical model of phonological variation. In Young-Se Kang, Jong-Yurl Yoon, Hyunkyung Yoo, Sze-Wing Tang, Yong-Soon Kang, Youngjun Jang, Chul Kim, Kyoung-Ae Kim, & Hye-Kyung Kang, eds. *Current Issues in Linguistic Interfaces. Volume 2*. Seoul: Hankookmunhwasa. p. 267-294.

Coetzee, Andries W. 2009b. Phonological variation and lexical frequency. In Anisa Schardl, Martin Walkow, & Muhammad Abdurrahman, eds. *NELS 38: Proceedings of the 38th Meeting of the North East Linguistics Society*. *Volume 1*. Amherst: GLSA. p. 189-202. [Available on Rutgers Optimality Archive, ROA-952.]

Côté, Marie-Hélène and Geoffrey Stewart Morrison. 2007. The nature of the schwa-zero alternation in French clitics: experimental and non-experimental evidence. *Journal of French language studies* 17: 159-186.

Davidson, Lisa, Peter Jusczyk and Paul Smolensky. 2004. The initial and final sates: theoretical implications and experimental explorations of Richness of the Base. In René Kager, Joe Pater and Wim Zonneveld, eds. *Constraints in Phonological Acquisition*. Cambridge: Cambridge University Press. p. 321-368.

Dell, François.1973. Les règles et les sons. Orléans: Imprimerie Nouvelle. [trans. 1980 by Catherine Cullen as *Generative phonology and French phonology*. Cambridge University Press.]

Donnegan, Patricia Jane and David Stampe. 1979. The study of Natural Phonology. In Daniel A. Dinnsen, ed. *Current Approaches to Phonological Theory.* Bloomington: Indiana University Press. p. 126-173.

Dressler, Wolfgang U. 1985: *Morphonology*. Ann Arbor: Karoma Press.

Durand, Jacques and Bernard Laks. 2000. Relire les phonologues du français: Maurice Grammont et la loi des trois consonnes. Langue Française, 126: 29-39.

Eisenstat, Sarah. 2009. Learning Underlying Forms with MaxEnt. Master's Thesis, Brown University.

Ellis, Lucy and William J. Hardcastle. 2002. Categorical and gradient properties of assimilation in alveolar to velar sequences: Evidence from EPG and EMA data. *Journal of Phonetics*, 30: 373-396.

Ernestus, Mirjam and R. Harald Baayen. This volume. Corpora and exemplars in phonology.

Eychenne, Julien. 2006. Aspects de la phonologie du schwa dans le français contemporain optimalité, visibilité prosodique, gradience. PhD. dissertation, Université de Toulouse-Le Mirail.

Fasold, R. 1972. *Tense Marking in Black English*. Arlington: Center for Applied Linguistics.

Fidelholtz, J. 1975. Word frequency and vowel reduction in English. In *CLS*, 11:200-213.

Fischer, Markus. 2005. *A Robbins-Monro type learning algorithm for an entropy maximizing version of stochastic Optimality Theory*. MA thesis, Humboldt University, Berlin. [Available as *Rutgers Optimality Archive* 767.]

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

Fougeron, Cécile and Steriade, Donca. (1997). Does the deletion of French schwa lead to the neutralization of lexical distinctions? *Proceedings of the Eurospeech conference, Rhodes*. 943–946.

Francis, W. Nelson and Henry Kučera. 1982. *Frequency Analysis of English Usage*. Boston: Houghton Mifflin.

Goeman, A.C.M. 1999. *T-deletie in Nederlandse Dialecten. Kwantitatieve Analyse van Structurele, Ruimtelijke en Temporele Variatie*. Ph.D. dissertation, Vrije Universiteit, Amsterdam. LOT Series 26. The Hague: Holland Academic Graphics.

Goeman, A.C.M. and J. Taeldeman. 1996. Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaal-verzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, 48:38-59.

Goeman, A.C.M. and P.Th. van Reenen. 1985. *Word-Final t-Deletion in Dutch Dialects. The Roles of Conceptual Prominence, Articulatory Complexity, Paradigmatic Properties, Token Frequency and Geographical Distribution*. Vrije Universiteit Working Papers in Linguistics 16. Amsterdam: Vakgroep Algemene Taalwetenschap Vrije Universiteit.

Goldrick, Matt. 2008. A harmonic grammar account of lexically-conditioned phonetic variation. Paper presented at Linguistic Society of America Annual Meeting. Chicago.

Goldrick, Matt and Robert Daland. 2009. Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. Phonology, 26, 147-185.

Goldsmith, John. 1990. *Autosegmental and Metrical Phonology*. Oxford and Cambridge, MA: Blackwell.

Goldsmith, John. 1993. Harmonic phonology. In John Goldsmith (ed.) *The Last Phonological Rule: Reflections on Constraints and Derivations*. Chicago: University of Chicago Press. p. 21-60.

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl. eds. *Proceedings of the Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University. p. 111-120.

Guy, Gregory R. 1980. Variation in the group and the individual. In William Labov, ed. *Locating Language in Time and Space*. New York: Academic. p. 1-36.

Guy, Gregory R. 1991a. Contextual conditioning in variable lexical phonology. *Language Variation and Change*, 3:223-239.

Guy, Gregory R. 1991b. Explanation in variable phonology: an exponential model of morphological constraints. *Language Variation and Change*, 3:1-22.

Guy, Gregory R. 1993. The quantitative analysis of linguistic variation. In Dennis R. Preston, ed. *American Dialect Research. Celebrating the 100th Anniversary of the American Dialect Society, 1889-1989*. Amsterdam: John Benjamins. p. 223-250.

Guy, Gregory R. 1994. The phonology of variation. In *CLS 30: Papers from the 30th Regional Meeting of the Chicago Linguistic Society. Volume 2: The Parasession on Variation in Linguistic Theory*, eds. Katharine Beals *et al.*, 133-149. Chicago: Chicago Linguistic Society.

Guy, Gregory R. 1997. Competence, performance, and the generative grammar of variation. In Frans Hinskens, Roeland van Hout and Leo Wetzels, eds. *Variation, Change and Phonological Theory*. Amsterdam: John Benjamins. p. 125-143.

Guy, Gregory R., and Charles Boberg. 1997. Inherent variability and the obligatory contour principle. *Language Variation and Change*, 9:149-164.

Hale, Mark and Charles Reiss. 2000. "Substance Abuse" and "Dysfunctionalism": Current Trends in Phonology. Linguistic Inquiry, 31/1, 157-169.

Hale, Mark, Madelyn Kissock and Charles Reiss. 2007. Microvariation, variation, and the features of universal grammar. *Lingua*, 117:645-665.

Hansson, Gunnar. This volume. Diachronic explanations of sound patterns.

Hayes, Bruce. 1995. On what to teach the undergraduates: some changing orthodoxies in phonological theory. In Ik-Hwan Lee, ed. *Linguistics in the Morning Calm 3*. Seoul: Hanshin. p. 59-77.

Hayes, Bruce and May Abad. 1989. Reduplication and syllabification in Ilokano. *Lingua*, 77:331-374.

Hayes, Bruce, Robert Kirchner and Donca Steriade. 2004. *Phonetically-Based Phonology.* Cambridge: Cambridge University Press.

Hayes, Bruce and Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology*, 23:59-104.

Hayes, Bruce and Margaret MacEachern. 1998. Quatrain form in English folk verse. *Language*, 64:473-507.

Hayes, Bruce, Bruce Tesar and Kie Zuraw. 2003. *OTSoft 2.1.* Los Angeles and New Brunswick,NJ: UCLA and Rutgers University. [Software package available at http://www.linguistics.ucla.edu/people/hayes/otsoft/.]

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

Hayes, Bruce, Kie Zuraw, Péter Siptár, and Zsuzsa Londe. To appear. Natural and unnatural constraints in Hungarian vowel harmony. *Language*.

Hinskens, Frans. 1992. *Dialect Leveling in Limburg. Structural and Sociolinguistic Aspects*. Ph.D. dissertation, Katholieke Universiteit te Nijmegen.

Hinskens, Frans. 1996. *Dialect Leveling in Limburg: Structural and Sociolinguistic Aspects*. Tübingen: Niemeyer.Kaisse, Ellen M. and Patricia A. Shaw. 1985. On the theory of Lexical Phonology. *Phonology Yearbook*, 2:1-30.

Hooper, Joan Bybee. 1976. *An Introduction to Natural Generative Phonology*. New York: Academic Press.

Inkelas, Sharon. This volume. Phonology and morphology.

Itô, Junko and Armin Mester. 1986. The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry*, 17:49-73.

Itô, Junko and Armin Mester. 2001. Covert generalizations in Optimality Theory: the role of stratal faithfulness constraints. *Proceedings of the 2001 International Conference on Phonology and Morphology.* Seoul: The Phonology-Morphology Circle of Korea. p. 3-33.

Itô, Junko and Armin Mester. 2003. *Japanese Morphophonemics: Markedness and Word Structure*. Cambridge, MA: MIT Press.

Jäger, Gerhard. 2007. Maximum entropy models and stochastic Optimality Theory. In Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson and Annie Zaenen, eds. *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. Stanford, CA: CSLI Publications. 467-479.

Jäger, Gerhard, and Anette Rosenbach. 2006. The winner takes it all — almost: cumulativity in grammatical variation. Linguistics 44:937–971.

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.]

Johnson, Mark. 2002. Optimality-theoretic Lexical Functional Grammar. In Suzanne Stevenson and Paolo Merlo, eds. *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*. Amsterdam: John Benjamins. p. 59-73.

Kaisse, Ellen M., and Patricia A. Shaw. 1985. On the theory of Lexical phonology. *Phonology Yearbook*, 2:10–30.

Kang, Hyeon-Seok. 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*. Chicago: Chicago Linguistic Society. p. 150-164.

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

Kenyon, J. and T.A. Knott. 1953. *A Pronunciation Dictionary of American English. 4th edition*. Springfield, MA: Merriam.

Kimper, Wendell. 2008. Local Optionality and Harmonic Serialism. Ms, University of Amherst, Massachusetts. [Available on Rutgers Optimality Archive, ROA-988.]

Kiparsky, Paul. 1982. Lexical phonology and porphology. In In-Seok Yang, ed. *Linguistics in the Morning Calm*. Seoul: Hanshin. p. 1-91.

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

Kiparsky, Paul. 1993. An OT perspective on phonological variation. Handout from *Rutgers Optimality Workshop 1993*, also presented at *NWAV* 1994, Stanford University. Available at http://www.stanford.edu/~kiparsky/Papers/nwave94.pdf.

Kirchner, Robert and Roger Moore. 2008. Computing phonological generalization over real speech exemplars. Ms, University of Alberta and University of Sheffield.

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. 1997. Resyllabification. In F. Hinskens *et al*. (eds.) 145-179.

Labov, William. 2004. Quantitative analysis of linguistic variation. In Ulrich Ammon, Norbert Dittmer, Klaus J. Mattheier, and Peter Trudgill, eds. *Sociolinguistics: An International Handbook of the Science of Language and Society. 2nd edition*. Berlin: De Gruyter. p. 6-21.

Ladd, D. R. This volume. Phonetics in phonology.

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. *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 M. Ziolkowski, M. Noske and K. Deaton, eds. *Proceedings of the 26th Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society. p. 237-252.

Legendre, Géraldine, Antonella Sorace and Paul Smolensky. 2006. The Optimality Theory-Harmonic Grammar connection. In Paul Smolensky and Géraldine Legendre, eds. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar, Volume 2: Linguistic and Philosophical Implications*. Cambridge, MA: MIT Press/Bradford Books. p 339-402.

Lim, Laureen T. and Gregory R. Guy. 2005. The limits of linguistic community: speech styles and variable constraint effects. In Maciej Baranowski, Uri Horesh, Keelan Evans and Giang Nguyen, eds. Penn Working Papers in Linguistics. Volume 10.2. Selected Papers from NWAVE 32. Philadelphia: Penn Linguistics Club. p. 157-170.

Lindblom, Bjorn. 1990. Explaining phonetic variation: a sketch of the H & H theory. In W. J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling*, 403-439. Dordrecht: Kluwer Academic Publishers

Malécot, André. 1958. The role of releases in identification of released final stops. *Language*, 34:370-380.

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

Nishimura, Kohei. 2003. *Lyman's Law in Loanwords*. Manuscript. Tokyo: Tokyo University.

Nolan, Francis. 1992. The descriptive role of segments: evidence from assimilation. *Papers in laboratory phonology II*, ed. by Gerard J. Docherty and D. Robert Ladd, 261-80. Cambridge: Cambridge University Press.

Odden, David. This volume. Rules vs. constraints.

Ohala, John. 1990. There is no interface between phonetics and phonology. A personal view. *Journal of Phonetics*, 18:153-171.

Oostendorp, Marc van. 1997. Style Levels in Conflict Resolution. In Frans Hinskens, Roeland van Hout and Leo Wetzels, eds. *Variation, Change and Phonological Theory*. Amsterdam: John Benjamins. p. 207-229.

Paollilo, John C. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford: CLSI.

Pater, Joe. 2000. Non-uniformity in English secondary stress: the role of ranked and lexically specific constraints. *Phonology*, 17:237-274. [Published version of 1995 ms. available on Rutgers Optimality Archive, ROA-107.]

Pater, Joe. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39/2, 334-345 [Available on Rutgers Optimality Archive, ROA-917.]

Pater, Joe. 2009. Weighted Constraints in Generative Linguistics. Cognitive Science 33: 999-1035. [Available on Rutgers Optimality Archive, ROA-982.]

Pater, Joe, David Smith, Robert Staubs, Karen Jesney and Ramgopal Mettu. 2010. Learning Hidden Structure with a Log-Linear Model of Grammar. Paper presented at Linguistic Society of America Annual Meeting. Baltimore.

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.

Phillips, Betty S. 2006. *Word Frequency and Lexical Diffusion*. New York: Palgrave Macmillan.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In Joan Bybee and Paul Hopper, eds. Frequency Effects and the Emergence of Lexical Structure. Amsterdam: John Benjamins. p. 137-157.

Pierrehumbert, Janet B. 2002. Word-specific phonetics. In Carlos Gussenhoven and Natasha Warner, eds. *Laboratory Phonology VII*. Berlin: Mouton de Gruyter. p. 101-139.

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt and Michael Becker. To appear. Harmonic Grammar with Linear Programming: From linear systems to linguistic typology. *Phonology*. [Available on Rutgers Optimality Archive, ROA-984.]

Prince, Alan. 2006. Implication and Impossibility in Grammatical Systems: What it is & How to find it. Ms, Rutgers University. [Available on Rutgers Optimality Archive, ROA-880.]

Prince, Alan and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, MA, and Oxford, UK: Blackwell. [Revision of 1993 technical report, Rutgers University Center for Cognitive Science. Available on Rutgers Optimality Archive, ROA-537.]

Racine, Isabelle and François Grosjean. 2002. La production du E caduc facultatif est-elle prévisible? Un début de réponse. *Journal of French Language Studies*, 12: 307-326.

Reynolds, Bill. 1994. *Variation and Phonological Theory*. Ph.D. dissertation, University of Pennsylvania.

Rousseau, Pascal and David Sankoff. 1978. Advances in variable rule methodology. In David Sankoff, ed. *Linguistic Variation: Models and Methods*. New York: Academic. p. 57-69.

Sankoff, David, Sali Tagliamonte, Eric Smith. 2005. *Goldvarb X: A variable rule application for Macintosh and Windows*. Department of Linguistics, University of Toronto. [Software package available at http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm.]

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. Dissertation, University of Pennsylvania.

Santa Ana, Otto. 1992. Chicano English evidence for the exponential hypothesis: a variable rule pervades lexical phonology. *Language Change & Variation*, 4: 275-288.

Schouten, M.E.H. 1982. T-deletie in de stad Utrecht: schoolkinderen en grootouders. *Forum der Letteren*, 23:282-291.

Schouten, M.E.H. 1984. T-deletie in het zuiden van die provincie Utrecht. *Taal en Tongval*, 36:162-173.

Smolensky, Paul and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, MA: MIT Press/Bradford Books.

Steriade, Donca. 2001. Directional asymmetries in place assimilation. In Elizabeth Hume and Keith Johnson, eds. *The Role of Speech Perception in Phonology*. San Diego: Academic Press. p. 219-250.

Steriade, Donca. 2008. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. In Kristin Hanson & Sharon Inkelas (eds.), *The Nature of the Word: Studies in Honor of Paul Kiparsky*. Cambridge, MA: MIT Press. p. 151-180.

Stevens, Kenneth N., and Samuel Jay Keyser. 1989. Primary features and their enhancement in consonants. *Language*, 65:81-106.

Sussman, Harvey M., Helen A. McCaffrey, and Sandra A. Matthews. 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90:1309-1325.

Tesar, Bruce. 1995. *Computational Optimality Theory*. Doctoral dissertation, University of Colorado. [Available on Rutgers Optimality Archive, ROA-90.]

Tesar, Bruce and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry*, 29:229-268. [Available on Rutgers Optimality Archive, ROA-155. Reprinted in *Optimality Theory in Phonology: A Reader*, ed. by John J. McCarthy, Malden, MA and Oxford, Blackwell (2004).]

Tesar, Bruce and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.

Vaux, Bert. 2008. Why the phonological component must be serial and rule-based. In B. Vaux and A. Nevins, eds. *Rules, Constraints and Phonological Phenomena*. Oxford University Press.

Walker, Douglas C. 1996. The new stability of unstable -e in French. *French Language Studies*, 6. 211-229.

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*, 43A:603-620.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30:945-982.