The place of variation in phonological theory

1. Introduction

Over the past 15 years, the study of variation has become increasingly important in phonology. 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 of theoretical phonology 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 the some of the issues that arise in making a place for it in phonological theory.

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, a difficulty that, as we will see, is exacerbated when the details of variation are taken into account.

Variation has been studied from a number of perspectives in phonology, ranging from sociolinguistic approaches in the Labovian tradition to recent studies in Optimality Theory (OT; Prince and Smolensky 1993/2004), usage-based models of phonology (Bybee 2001, 2006) and exemplar-based models of phonology (Bybee as well as Pierrehumbert 2001, 2002). We will trace developments in the study of variation by focusing on the locus of variation in a model of phonological grammar. Variation is sometimes considered to be limited to the late stages of phonological derivation, i.e. towards the end where phonetic implementation takes over from phonology. We will review evidence showing that this is, in fact, not accurate – variation is also found in phonological alternations that are incontrovertibly part of the “early” phonology, since they are morphologically conditioned. We will focus particularly on OT and related constraint-based theories, which have made considerable headway not only on the analysis of variation, but also in providing explicit accounts of its learning. We will discuss the successes of these approaches, but also point out some of challenges that they face. In our view, successfully meeting these challenges may require revisions of fundamental assumptions about the nature of phonological grammar and the phonological lexicon. In taking this view, we are following the precedents of Bybee (2001, 2006) and Pierrehumbert (2001, 2002), but unlike that work, our suggestions for theoretical revisions focus more on the grammar than on the lexicon.

Since our discussion will be structured around the question of the locus of variation in the phonological component, we give a brief overview here of standard assumptions about the architecture of this component of grammar. Within generative linguistics, the phonology is standardly assumed to have roughly 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. Within this 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 of changes assigned to each of them are given in (2).

(2)  

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

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 et seq.) 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 the weighted constraints model of variation to Labov’s (1969) Variable Rules theory, especially to its elaboration as a log-linear model of probabilistic 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 2007). 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 4.5 on how probability of rule application is calculated).

(4) Labov (1969:738)

$$X \rightarrow (Y) / \left[ \begin{array}{c} \alpha \text{fe}_i \\ \vdots \\ \beta \text{fe}_k \\ \vdots \\ \delta \text{fe}_l \end{array} \right]$$
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 4.5.

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 Donegan 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 3.1.

(5)  Nasal place assimilation in English
a.  Lexical = obligatory
   e[nt]er, *e[mt]er, *e[nt]er
   a[mb]er, *a[nb]er, *a[nt]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[n k]ard

The variable phrasal 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 (1990) 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. 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 gradient variation 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 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 words 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 place the process outside of phonology proper: “… 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 3.1). In section 3.2, 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. We begin this section by showing that t/d-deletion is conditioned by factors similar to those that condition categorical phonological processes. 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 whipped 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*.\(^1\)

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 of the robustness of this generalization.

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

<table>
<thead>
<tr>
<th></th>
<th>Regular past (missed)</th>
<th>Semi-weak past (kept)</th>
<th>Monomorpheme (mist)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philadelphia English (Guy 1991b)</td>
<td>17%</td>
<td>34%</td>
<td>38%</td>
</tr>
<tr>
<td>Chicano English (Santa Ana 1992)</td>
<td>26%</td>
<td>41%</td>
<td>58%</td>
</tr>
<tr>
<td>Tejano English (Bayley 1997)</td>
<td>24%</td>
<td>34%</td>
<td>56%</td>
</tr>
</tbody>
</table>

The fact that the application of the variable, gradient process of \(t/d\)-deletion is conditioned by morphology is problematic for 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, 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.

---

\(^1\) Labov (1989) actually gives the following hierarchy from consonants that are most likely to induce deletion 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) for evidence in agreement with Guy and Boberg.
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.\(^2\) 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). Kiparsky (1993) proposes that variation between optimal outputs
for can be derived by placing the constraints in a partial order. Each time the grammar is used to
evaluate a candidate set, one of the possible rankings consistent with the partial order is randomly
chosen. When some of these total orders pick different candidates as optimal, variation results.
We will refer to this theory, which is developed in much more detail by Anttila (1997 *et seq.*) as
the partially ordered constraints (POC) model of variation.

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

---

\(^2\) Anttila’s (1997) data have in fact been analyzed in three of the theories we discuss below: see Boersma and Hayes (2001) for a stochastic OT analysis, Goldwater and Johnson (2003) for one in MaxEnt-HG, and Paolillo (2002) for one in the variable rules theory.
ranking, and \textit{cand}_2 under the other ranking. This is therefore a language where /input_/ \ will be variably realized as either \textit{cand}_1 or \textit{cand}_2, but will never surface as \textit{cand}_3.

(8) Grammar: \textit{C}_1 >> \textit{C}_2, \textit{C}_1 >> \textit{C}_3

\begin{itemize}
  \item \textbf{a. First possible ranking:} \textit{C}_1 >> \textit{C}_2 >> \textit{C}_3
  \begin{tabular}{|c|c|c|c|}
    \hline
    /input_/ & \textit{C}_1 & \textit{C}_2 & \textit{C}_3 \\
    \hline
    \textit{cand}_1 & \* & \* & \* \* \\
    \textit{cand}_2 & \* & \* & \* \* \\
    \textit{cand}_3 & \* & \* & \* \* \\
    \hline
  \end{tabular}

  \item \textbf{b. Second possible ranking:} \textit{C}_1 >> \textit{C}_3 >> \textit{C}_2
  \begin{tabular}{|c|c|c|c|}
    \hline
    /input_/ & \textit{C}_1 & \textit{C}_3 & \textit{C}_2 \\
    \hline
    \textit{cand}_1 & \* & \* & \* \* \\
    \textit{cand}_2 & \* & \* & \* \* \\
    \textit{cand}_3 & \* & \* & \* \* \\
    \hline
  \end{tabular}
\end{itemize}

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 \( n \) out 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, this is perhaps the first adoption of Labov’s (1969) proposal that a grammar can encode a probability distribution over outcomes.\(^3\)

Kiparsky (1993) proposes the POC model in the context of an analysis of the morphological and phonological factors influencing the rate of \textit{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 likelihood that \textit{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 \textit{t/d}. As Labov (1989) points out, a clear generalization that emerges from the variationist literature on \textit{t/d}-deletion is that the rate of deletion is always highest in pre-consonantal position (\textit{west bank}). Both pre-vocalic (\textit{west end})

\(^3\) This is our interpretation, not Kiparsky’s. He frames his proposal in terms of competition between grammars, rather than in terms of a probabilistic grammar. Anttila and Andrus (2006) distinguish between the Multiple Grammars theory of Kiparsky (1993), and the Partially Ordered Grammars of Anttila and Cho (1998), though Kiparsky (1993) does discuss both possibilities.
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 away 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 [r] (e.g. more deletion in best luck than in best rock) may be due to the fact that [t-] is a possible onset cluster but [tl-] is not.

(10) Percent deletion in different contexts

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

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, to appear). These constraints

---

4 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).
protect 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-vocally 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 4 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 *Cr. 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.
Crucial rankings, number of corresponding total orders, outcomes

<table>
<thead>
<tr>
<th>Crucial rankings</th>
<th>Total # of rankings</th>
<th>Pre-V</th>
<th>Phrase-final</th>
<th>Pre-C</th>
</tr>
</thead>
</table>
a. \( \text{MAX} >\gg *\text{CT} \) | 12 | No | No | No |
b. \( \text{MAX-PRE-V} >\gg *\text{CT} >\gg \{\text{MAX, MAX-FINAL}\} \) | 2 | No | Yes | Yes |
c. \( \text{MAX-FINAL} >\gg *\text{CT} >\gg \{\text{MAX, MAX-PRE-V}\} \) | 2 | Yes | No | Yes |
d. \( \{\text{MAX-PRE-V, MAX-FINAL}\} >\gg *\text{CT} >\gg \text{MAX} \) | 2 | No | No | Yes |
e. \( *\text{CT} >\gg \{\text{MAX, MAX-PRE-V, MAX-FINAL}\} \) | 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-dialectical 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). 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 out 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 3.3 below.
### Probabilities of deletion from quantitative interpretation of partial orders

<table>
<thead>
<tr>
<th>Partial order</th>
<th># rankings</th>
<th>Pre-V</th>
<th>Phrase-final</th>
<th>Pre-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. None</td>
<td>8/24</td>
<td>8/24</td>
<td>12/24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.33</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>b. MAX-PRE-V &gt;&gt; *CT</td>
<td>0/12</td>
<td>4/12</td>
<td>6/12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.33</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>c. *CT &gt;&gt; MAX-PRE-V</td>
<td>6/12</td>
<td>4/12</td>
<td>6/12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.33</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>d. MAX-FIN &gt;&gt; *CT</td>
<td>4/12</td>
<td>0/12</td>
<td>6/12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>e. *CT &gt;&gt; MAX-FIN</td>
<td>4/12</td>
<td>6/12</td>
<td>6/12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

### 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 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 possible phonological systems. For example, a version of the t/d-deletion system in which pre-consonantal deletion has a lower probability 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 4.5 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 3.2, 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 obviously 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 Boersma and Londe 2006 for relevant data). Insofar as the phonological grammar is at least 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 1995, 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 et seq.) 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 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. 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. 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).

\begin{table}
\begin{tabular}{|c|c|c|}
\hline
 & 101.6 & 98.2 \\
\hline
/CtC/ & *Ct & Max \\
\hline
.10 & CtC & * \\
.90 & CC & * \\
\hline
\end{tabular}
\end{table}

(14) Skewed probability distributions in stochastic OT

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. The Constraint Demotion Algorithm (CDA; Tesar 1995, Tesar and Smolensky 1998, 2000) has 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 some value \( x \) from the ranking values of each constraint that is violated more in the correct form

---

5 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.
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 $/CtC/ \rightarrow [CtC]$, 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: $/CtC/ \rightarrow [CtC]$

Learner’s parse:

<table>
<thead>
<tr>
<th>/CtC/</th>
<th>*Ct</th>
<th>MAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>CtC</td>
<td>*!</td>
<td></td>
</tr>
<tr>
<td>$\not{\cdot}$ CC</td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>

Grammar $H_1$: *Ct 101, MAX 99

When provided with sufficient examples of $/CtC/ \rightarrow [CtC]$, 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 $/CtC/ \rightarrow [CtC]$ and $/CtC/ \rightarrow [CC]$ 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. If numerical constraint values are required for learning and variation, it is seems natural to also use them to choose the optimal candidate. Because this theory retains stochastic OT’s noisy evaluation, we refer to it as Noisy HG. Noisy HG first appeared in a Praat implementation (Boersma and Weenink 2007).
In HG, the optimal candidate is the one with the highest *Harmony*, which is the sum of the weighted constraint scores. We adopt Legendre, Sorace and Smolensky’s (2006) convention of converting OT violation marks to negative integers. In (16), we provide a tableau in which 

\[ \text{*CT has a greater weight than MAX; the constraint weights are indicated in the top row. Each candidate’s Harmony is indicated in the rightmost column; this figure is obtained by multiplying each violation score by the weight, and then summing these. The candidate with deletion has the highest (closest to zero) Harmony, and is thus optimal. For more detailed discussion of HG and its relation to OT, see Smolensky and Legendre (2006) and Pater, Potts and Bhatt (2007).} \]

\[
\begin{array}{ccc}
\text{/CtC/} & \text{*CT} & \text{MAX} \\
\varphi \text{ CC} & -1 & -1 \\
\text{CtC} & -1 & -2 \\
\end{array}
\]

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 in that it is capable of producing cumulative constraint interaction. An example of cumulativity that involves variation, which is also discussed in HG terms in Pater et al. (2007), comes from the phonology of Japanese loanwords (Nishimura 2003, Kawahara 2006). In native Japanese words, 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 (*VCE-GEM*). 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 (17).

\[
\begin{array}{ccc}
1.5 & 1 & H \\
\text{/bobu/} & \text{ID-VCE} & \text{OCP-VCE} \\
\varphi \text{ bobu} & -1 & -1 \\
\text{bopu} & -1 & -1.5 \\
\end{array}
\]

\[
\begin{array}{ccc}
1.5 & 1 & H \\
\text{/webbu/} & \text{ID-VCE} & \text{*VCE-GEM} \\
\varphi \text{ webbu} & -1 & -1 \\
\text{weppu} & -1 & -1.5 \\
\end{array}
\]

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 (18). 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.
When we introduce noise into the evaluation process, we will get variation in (18) if the sum of the weights of OCP-VCE and *VCE-GEM is sufficiently close to that of ID-VCE, producing different choices between the two candidates across instances of evaluation. If ID-VCE has at the same time a sufficiently greater weight than either single one of the constraints, no variation will occur for the cases in (17). As with stochastic OT, a set of constraint values producing this result can be obtained by submitting a data distribution to a learning algorithm.

Jäger (to appear) points out that there is a learning algorithm for weighted constraint grammars that closely resembles the GLA for stochastic OT. This learning procedure is broadly used in statistical and connectionist learning frameworks, where it is referred to as stochastic gradient ascent (Jäger to appear) or the Perceptron update rule (Pater 2008). The sole difference from the 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 (to appear) 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 the GLA. We will refer to this weighted constraint learning algorithm as the HG-GLA.

A difference between the HG-GLA and the OT-GLA is that the 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 much simpler Perceptron proof, as well as an extension to the case of learning with Noisy HG. 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. The first learner operated with a stochastic OT grammar, and the OT-GLA learning algorithm. The results are shown in the row labeled “St-OT” in (19). 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

<table>
<thead>
<tr>
<th></th>
<th>OCP-VCE</th>
<th>ID-VCE</th>
<th>*VCE-GEM</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>/guddo/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>guddo</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>gutto</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1.5</td>
</tr>
</tbody>
</table>

17
of a second voiced obstruent. The second learner operated with a noisy HG grammar, and the 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 (19).

(19)  | Grammars learned by stochastic OT and noisy Harmonic Grammar |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency of devoicing in learning data</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>OCP-VOICE</td>
<td>*VCE-GEM</td>
<td>IDENT-VOICE</td>
<td>bobu webbu guddo</td>
</tr>
<tr>
<td>St-OT</td>
<td>3113.9</td>
<td>3113.9</td>
<td>3113.7</td>
</tr>
<tr>
<td>N-HG</td>
<td>66.8</td>
<td>67.6</td>
<td>134.4</td>
</tr>
</tbody>
</table>

As well 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.  **MaxEnt HG**

Johnson (2002) shows how an OT grammar can be transformed into a log-linear probabilistic model. Goldwater and Johnson (2003), Jäger (to appear) and Wilson (2006) apply the resulting model of variation to phonology. In this section, we show how this theory relates to the noisy HG model we have just described.

The log-linear model calculates a probability distribution over the candidate set. The probability of a candidate is proportional to the exponential of its Harmony score. The tableau in (20) illustrates this for the variation in the Japanese loanword case.

(20)  | 2 | 1 | 1 | H | e^H | p |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>/guddo/</td>
<td>ID-VCE</td>
<td>OCP-VCE</td>
<td>*VCE-GEM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>guddo</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>0.14</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>gutto</td>
<td>-1</td>
<td></td>
<td>-2</td>
<td>0.14</td>
<td>.50</td>
<td></td>
</tr>
</tbody>
</table>

The log-linear model of probabilistic grammar is often referred to as a *Maximum Entropy* model. Maximum Entropy defines an optimal probability distribution given a set of data: the one that makes the fewest assumptions about the shape of the probability distribution beyond the available evidence. We will not need to refer to this principle here, but we will follow previous work in referring to this as a MaxEnt model. We will call it MaxEnt-HG to distinguish it from other applications of Maximum Entropy, and also to emphasize its relatedness to the connectionist model of grammar proposed by Legendre and Smolensky (2006).

---

6 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 esp. Kawahara 2006 on the data).
Because they both use weighted constraints, Noisy HG and MaxEnt-HG can both represent cumulative constraint interaction, which distinguishes them from stochastic OT, and standard OT. Here we have discussed a case in which this can be seen as advantageous (see also Guy 1997 on cumulativity in phonological variation and Jäger and Rosenbach 2006 on cumulativity in syntactic variation). However, Prince and Smolensky (1993/2004) and Legendre, Sorace and Smolensky (2006) claim that weighted constraints produce implausible linguistic systems. The typological consequences of the greater power of weighted constraints are currently the subject of ongoing research: see Pater, Potts and Bhatt (2007) for a critical review of extant arguments for strict domination, as well as discussion of outstanding difficulties for weighted constraints.

Noisy HG and MaxEnt-HG differ in that Noisy HG produces a single optimal output each time the grammar is used, 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 can give a portion of the probability mass to a candidate that is harmonically bounded; this cannot happen in Noisy HG (Jesney 2007). The empirical consequences of this and other differences between the two theories remain to be investigated.

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 outlined 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 (21). 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. The 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.\(^7\)

\(^7\) 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.
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 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

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

(21) Observed and learned t/d-deletion rates for different dialects of English
the Noisy HG result, we note first that a categoric version of HG that restricts weights to positive values generates the same five languages as OT. 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 N-HG grammar for Tejano', the negatively weighted constraints have no effect on evaluation. The effect of negative weights is illustrated by the ME-HG result in (21); the versions of ME-HG implemented in Boersma and Weenink (2007) and discussed in the literature cited above has no non-negativity restriction. It was thus able to find a weighting that disfavors 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.

Before turning to a comparison of these constraint-based models with the variable rules theory, we should emphasize that these results depended crucially on the constraints selected for the analysis. First, the restriction against greater frequency of pre-consonantal deletion is of course dependent on the absence of a MAX-PRE-C constraint penalizing deletion only in this context. Second, we also obtained much worse fits with observed frequencies when we used other constraint sets that did generate the same categorical typology as our constraint set. When we implemented an analysis that explained the higher rate of pre-vocalic retention as due to the possibility of syllabification of the t/d as an onset (Kiparsky 1993, Reynolds 1994, Guy 1997; cf. Labov 1997), we ran into a hidden structure problem (Tesar 1995 et seq.) that hindered accurate probability matching. While this hidden structure problem is probably not in-superable (see e.g. Tesar and Smolensky 2000, Boersma 2003), it is important to note that, like the Constraint Demotion Algorithm, the convergence properties of the weighted constraint learners are dependent on having full access to structure. This less positive result illustrates the fact that there is considerable room for future development of these models of variation and their associated learning algorithms.

4.5. Variable Rules

As we pointed out earlier, Labov’s (1969) variable rules 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 likelihood of rule application is 1.0, and whenever any of these components is absent, the likelihood of application is zero. Labov introduced the notion of weighting the components in the structural description such that each component contributes to the likelihood of rule application. Under 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).

9 Readers interested in examining the less positive result can submit the file "aave-syllable.collection" from "coetzee-pater-variation.zip" to Praat.
interpretation of rewrite rules, it becomes possible to say, for instance, that the presence of C __ in the context increases the likelihood of rule application by some specific factor, and similarly for the presence of __ D. The rule can now apply even if not both C __ and __ D are present, and the likelihood of application can take on any value between zero and 1.0.

As pointed out by Cedergren and Sankoff (1974:338), Labov also discovered that the different components of the structural description usually make an independent contribution to the likelihood of rule application. To continue with the same example, the contributions of C __ and __ D to the likelihood of rule application are independent from each other. This implies that the presence of only one the components can still result in rule application. But more importantly, since many standard statistical methods depend on the assumption of independence between factors, this independence between the components of the structural description means that grammatical variation can be analyzed using these standard statistical methodologies. Specifically, assume that we have a dataset that contains information about some variable process, for instance /t/d-deletion. Each token in the dataset can be coded for whether or not the rule of /t/d-deletion has applied to it, and also for all of the factors that are suspected to influence the likelihood that the rule will apply – following the discussion in the earlier sections, each token can be coded for whether the /t/d appears in pre-consonantal, pre-vocalic or phrase-final position. Once the dataset has been coded like this, it can be subjected to a regression analysis. As Paollilo (2002:177) points out, when the dependent variable in a regression consists of count data (how many times did the rule apply vs. how many times did it not apply), the log-linear regression model is the most appropriate. In this model, the observed values are log-transformed before the regression is performed. Such a regression will find the factor values for each of the factors (pre-consonantal, pre-vocalic, and phrase-final in the current example) that results in the best fit between the observed data and the values predicted by regression model.

Along with Labov’s original proposal for the formalization of a probabilistic generative grammar, equally responsible for the ensuing paradigm change in (socio-)linguistic research was Cedergren and Sankoff’s (1974) further mathematical formalization of the theory, and their implementation of the model as the VARBRUL software package. The most recent version of the software is Goldvarb X (Sankoff, Tagliamonte and Smith 2005). The specific regression model that Goldvarb X employs is given in (22). In this equation, $p$ is the probability of the rule applying. $p_0$ is what is called the “input factor weight”, and represents the probability of the rule applying independently from any of the conditioning factors. $p_1 \ldots p_n$ represent the weight of the different factors.

(22) **Regression model used in Goldvarb X**

$$\frac{p}{(1-p)} = \frac{p_0}{(1-p_0)} \times \frac{p_1}{(1-p_1)} \times \frac{p_2}{(1-p_2)} \ldots \frac{p_n}{(1-p_n)}$$

To illustrate the application of this model, we created a corpus for the Tejano data (Bayley 1995). Since we did not have access to Bayley’s original corpus, we made the assumption that each of the three contexts (pre-consonantal, pre-vocalic, and phrase-final) appears 100 times. In our corpus, pre-consonantal position was marked with deletion (application of /t/d-deletion) 62 times, phrase-final position 46 times, and pre-vocalic position 25 times, in accordance with the deletion rates that Bayley reports for these three contexts. We then submitted this corpus to Goldvarb X. The result generated by Goldvarb X is given in the first line of (23). The expected values are calculated using the formula from (22). For instance, to
calculate the expected deletion rate in pre-vocalic position, we substitute .44 for $p_0$ and .30 for $p_1$. Since the three contexts are mutually exclusive (one occurrence of $t/d$ can be only pre-vocalic, pre-consonantal or phrase-final), only the pre-vocalic factor term is used in addition to the input term. The resulting formula is shown in (24). Solving for $p$ in this formula gives 0.25, which thus specifies the probability of the $t/d$-deletion rule applying in pre-vocalic position.

(23) Goldvarb $X$ outputs

<table>
<thead>
<tr>
<th></th>
<th>Factor weights</th>
<th>Observed and expected deletion rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Input CtV Ct CtC</td>
<td>O E</td>
</tr>
<tr>
<td>Tejano</td>
<td>.44 .30 .52 .68</td>
<td>25 25.03</td>
</tr>
<tr>
<td>Tejano'</td>
<td>.44 .68 .52 .30</td>
<td>62 61.97</td>
</tr>
</tbody>
</table>

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

\[
\frac{p}{1-p} = \frac{.44}{1-.44} \times \frac{.30}{1-.30}
\]

We also submitted the Tejano' distribution to Goldvarb $X$, just as we did for the constraint-based models discussed in the last section. Unsurprisingly, factor weights were found that allowed the model to equally match this unattested distribution. Cedergren and Sankoff (1974) see the development of universal restrictions on the variables affecting probability of rule application as an important direction for future research, but to the best of our knowledge this research program was never carried out, at least as far as phonological universals are concerned. Thus, there is no restriction in variable rules theory against having a rule that yields greater probability of deletion in pre-consonantal position. Paolillo (2002: ch. 10) provides an explicit comparison of POC with the variable rules model by reanalyzing Anttila’s (1997) Finnish data, and defends the lack of substantive restrictions on variable rules. It is worth noting, though, that the distinction between plausible and implausible phonological rules has often guided practice in formulating variable rules analyses; see especially Labov (2004) for discussion.

5. Lexically conditioned variation

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 show 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 will serve as a springboard for our discussion of directions for further development of constraint-based theories of phonological variation. Much of the data that we discuss here has formed the basis of recent arguments that a standard

---

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

23
assumption about lexical representation in generative phonology is inadequate, that instead of each morpheme being phonologically represented in terms of a single abstract underlying form, each one is associated with a set of phonetically detailed exemplars (Bybee 2001, Pierrehumbert 2001, 2002). 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 discussing the possibilities that numerically valued constraints offer for the formalization of the interaction between the grammar and the lexicon, including a lexicon that differs from the standard generative one.

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 (25), we provide two examples from Pater (2000); see that paper for further discussion and references to earlier work.

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

- **Stressed:** augmentation, condensation, importation, chimpanzee, incarnation, ostentation
- **Stressless:** information, segmentation, transportation, Mozambique
- **Variable:** advantageous, authenticity, condemnation

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

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

The data in (25) abstract from important subregularities. For example, it seems that 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).

Bybee (2001) draws particular attention to the role that lexical frequency plays in the propensity for 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 (25): 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 this correlation should be captured in a model of phonological grammar, though we do briefly return to this question below.

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 (25) 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.

If the lexicon played no role in the frequency of application of variable processes, feast and yeast should not differ from each other in terms of deletion rate in the three contexts we discussed above: pre-vocalic, phrase-final, and pre-consonantal. In order to test whether this prediction is borne out we conducted a small experiment. We selected 15 monomorphemic, monosyllabic English words that end in /-st/: crust, feast, yeast, mast, moist, nest host, dust, fast, test, list, rest, best, last, most. We then embedded each of these 15 words in frame sentences so that each word occurred in one sentence where it is followed by a consonant, one where it is followed by a vowel, and one where it is followed by a pause. There were hence a total of 45 sentences. We recruited six undergraduate students from the University of Michigan to participate in the experiment. All six grew up in south-eastern Michigan, so that they share the same basic dialect. After illustrating to participants that word-final /t/ can sometimes be dropped in pronunciation, participants were presented with the 45 sentences in randomized order. They were asked to rate for each of the token words how likely they are to delete the /t/ in a casual speech situation. Rating was done on a 10 point scale where [1] meant that the /t/ is nearly always pronounced, and [10] that it is nearly never pronounced. The task that the participants performed is similar to a well-formedness rating: they rated each token for how well-formed a pronunciation without a [t] would sound.

Since the participants did not all use the same portion of the [10]-point scale, we normalized the ratings for each participant. A positive score for an item then means that a subject showed a higher than average preference for deletion on a specific token, while a negative score indicates a lower than average preference for deletion. The graph in (26) plots the mean standardized scores for all six participants for a selection of the tokens in the three contexts.
The first thing to note about this graph is that it confirms the influence of the following context on the rate of \textit{t/d}-deletion – overall, the deletion preference scores for the pre-consonantal context (the white bars) are higher than that for the other two contexts. To test whether this difference is significant, we conducted two-tailed paired-samples \textit{t}-tests to compare the deletion preference score in pre-consonantal context with pre-vocalic and pre-pausal contexts. Both comparisons yielded a highly significant result (pre-C vs. pre-#: \(t(14) = 9.49, p < .001\); pre-C vs. pre-V: \(t(14) = 10.12, p < .001\)). In general, the deletion preference scores are also higher for pre-vocalic than pre-pausal contexts (black bars vs. spotted bars). This is also borne out by a \textit{t}-test (\(t(14) = 2.71, p = 0.017\)). The dialect of the subjects in our experiment therefore shows most deletion in pre-consonantal context, and least in pre-pausal context.

However, more important for our purpose here is to note that there are large differences between individual words. For instance, \textit{feast} and \textit{yeast} that are practically identical differ markedly from each other. Similarly for \textit{host} and \textit{most}. Since there are no relevant phonological differences between these words, these differences in deletion preference cannot follow from the grammar. It must be lexically encoded in some way.

As we noted above, we do not aim to provide an account of the correlation between frequency of use of a word and frequency of deletion. However, we will briefly discuss the role of frequency here, since if frequency of use, along with phonological factors, fully determined the probability of the application of the process, then there would be no need to invoke lexical idiosyncrasy. Like English vowel reduction and other lenition processes, \textit{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 \textit{t/d}-deletion often leave out words like \textit{just}, \textit{went}, \textit{and}, and \textit{n’t}, since these words typically show anomalously high deletion rates. As pointed out by both Bybee (2000:70) andPatrick (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 that 35 times per million. As shown in (27), she found significantly higher deletion rates in the high frequency than in the low frequency words.

(27) Rate of /d/-deletion in Chicano English

<table>
<thead>
<tr>
<th></th>
<th>Deletion</th>
<th>Retention</th>
<th>% Deletion</th>
</tr>
</thead>
<tbody>
<tr>
<td>High frequency</td>
<td>898</td>
<td>752</td>
<td>54.4%</td>
</tr>
<tr>
<td>Low frequency</td>
<td>137</td>
<td>262</td>
<td>34.4%</td>
</tr>
</tbody>
</table>

Our data also give some support for this. We calculated the log Francis–Kučera frequency for each word in the IPhOD corpus, and divided the words into two frequency groups: (i) Words with a lower frequency than the average word frequency in Francis and Kučera (crust, feast, yeast, mast, moist, nest) and (ii) words with a frequency higher than the average (host, dust, fast, test, list, rest, best, last, most). The mean log frequency of the low frequency list is 0.68, and of the high frequency list is 2.25. A one-tailed two-sample t-test comparing the mean normalized scores for the fifteen tokens (averages across all three contexts) tends toward significance (t(13) = 1.76, p = .051). However, it is also clear from our data that usage frequency does not explain all of the between word differences. Feast and yeast, for instance, both have a Francis–Kučera frequency of 3 per million, yet the figure in (26) shows that their deletion preference scores differ greatly. It therefore seems unavoidable that the likelihood of participation in a variable process is conditioned to some extent by lexical idiosyncrasy.

Dutch has the exact same process of /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 ligt ‘light’ and bracht ‘brought’ both have CELEX log frequencies of 3.8, yet they differ widely in their deletion rates: ligt undergoes deletion 18% of the time and bracht 32% (Goeman 1999:182; Phillips 2006:65).12

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. One that provides a nice parallel to English /d/-deletion is French schwa deletion (see Eychenne 2006 for a recent overview of the generative literature). It similarly has some of the hallmarks of

11 IPhOD stands for the Irvine Phonotactic Online Dictionary. This is a publicly available corpus of 33,432 words with their CMU pronunciation transcriptions and their Francis–Kučera frequencies. The corpus is available at www.iphod.com.

12 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.
a late phonological process. Deletion of schwa from the initial syllable of a polysyllabic word (e.g. [sœmɛn] vs. [sœmen] 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. Dell also notes that the exceptional words tend to be rarely used or literary (along with proper names, which seem to be regular exceptions to deletion). 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. 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.

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 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 a pre-publication version of Pater (2000) as a means of dealing with influences of the lexicon on English secondary stress like those illustrated in (25).

As a simple illustration of how this would work for t/d-deletion, we provide the result 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. For both words, frequency of deletion was greatest in pre-consonantal position, intermediate in pre-vocalic position, and least in pre-pausal position. In each position, the frequency of deletion was greater for most than for feast. 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 (28), 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.
Results of a learning simulation with lexically specific constraints

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100.71</td>
<td>1.18</td>
<td>3.58</td>
<td>100.07</td>
<td>1.16</td>
<td>3.27</td>
<td>99.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>feast_V</th>
<th>feast</th>
<th>feast_C</th>
<th>most_V</th>
<th>most</th>
<th>most_C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of deletion in learning data</td>
<td>0.40</td>
<td>0.20</td>
<td>0.60</td>
<td>0.50</td>
<td>0.30</td>
</tr>
<tr>
<td>Probability of deletion in learned grammar</td>
<td>0.40</td>
<td>0.20</td>
<td>0.59</td>
<td>0.50</td>
<td>0.30</td>
</tr>
</tbody>
</table>

While this is an apparently effective solution to the problem of lexically conditioned variation and its learning, we think that it can be improved upon. Our main worry is that we have given the system too much freedom to match word-specific frequency distributions. For example, if each word-specific faithfulness constraint can occupy its own position in the hierarchy, then it should be possible for one word to show higher rates of deletion in pre-pausal than pre-vocalic position (e.g. MAX-PRE-V-/FEAST/ weighted higher than MAX-FIN-/FEAST/), while another word displays the reverse pattern (e.g. MAX-FIN-/MOST/ weighted higher than MAX-PRE-V-/MOST/). Although it is certainly conceivable that such systems might exist, in the absence of evidence of their existence, it is worth developing a more conservative theory. It is worth noting that the variable rules model would not be able to yield this pattern, if individual lexical items were included as factors in its regression model. The next section sketches another way that a weighted constraints theory can deal with lexically restricted variation, which does not seem to produce this sort of result, and which has some added benefits for the modeling of variation.

5.3. Scaled constraint weights

The values of constraint weights can be straightforwardly rescaled, which opens up ways of modifying a weighted constraint model of grammar that are unavailable to a theory that eschews numerical values for constraints. By scaling weights of faithfulness constraints according to the lexical item under evaluation, we can reproduce some of the effects of lexically specific faithfulness constraints. Perhaps the most restrictive way of doing this is to have a single constant for each lexical item that is added to, or multiplied by, the basic weight of every faithfulness constraint. For example, if we set the constant encoding the faithfulness strength of feast at 3, and that of most at 1, and multiply the weights of faithfulness constraints by these constants, then with appropriate weights for the constraints, we get deletion of the final consonant of most, but not of feast, as shown with the simple two-constraint examples in (29). By adding a stochastic component to this system, a variable system with greater probability of deletion for most could be produced.
A simple illustration of a lexically scaled faithfulness constraint

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3*1</th>
</tr>
</thead>
<tbody>
<tr>
<td>/fijst/</td>
<td>*CT</td>
<td>MAX</td>
</tr>
<tr>
<td>[fijs]</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td><em>fijst</em></td>
<td>-1</td>
<td>-2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>1*1</th>
</tr>
</thead>
<tbody>
<tr>
<td>/mowst/</td>
<td>*CT</td>
<td>MAX</td>
</tr>
<tr>
<td><em>mows</em></td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>[mowst]</td>
<td>-1</td>
<td>-2</td>
</tr>
</tbody>
</table>

While a system in which each lexical item is associated with a single constant that affects the basic weights of all faithfulness constraints will reproduce some of the effects of lexically specific constraints, it cannot reproduce the putative pattern of t/d-deletion mentioned in the previous section, in which different lexical items show different relative rates of deletion across contexts. A somewhat richer system would imbue specific features of each lexical item with a numerical strength. Even then, if a consonant had a given strength of representation, that strength would be constant across contexts.

Not only does scaling of faithfulness constraints have the potential of dealing with lexically conditioned variation, but it also offers an approach to a second challenge to OT models of variation: the effect of register or style. While variation is often conditioned by register, as far as we are aware, no OT account has successfully integrated this factor. 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/grammar. This seems unsatisfactory in the face of alternative models that succeed in encoding probability distributions that correspond to observed relative frequencies of variants. A theory of variation with numerically valued constraints allows an alternative formalization of van Oostendorp’s account of register differences: that increasing formality is an increase in the weights of the faithfulness constraints, which shifts probabilities in the direction of decreasing application of variable processes. 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.

The fact that the model of the interaction of the grammar with the lexicon and with register/style that we have sketched has some resemblance to Lindblom’s theory of the phonetic details of speech production brings us back to where we started this chapter. The phonetically gradient outcomes of many variable processes, including our central case study of English t/d-deletion, is probably the main reason that a phonologist might assume that these processes should be handled by the phonetics. We have argued that their morphological and lexical
sensitivity places them squarely within the domain of phonological theory. How then is the phonology to handle between category outcomes? Here we can do no more than point to Flemming’s (2001) demonstration that weighted constraints permit a unified account of categorical and gradient phenomena.

A theory that provides lexical items with numerical indices of their strength of representation departs considerably from standard generative assumptions about the nature of lexical representation, and its interaction with the grammar. Like lexically specific constraints, this theory would blur the distinction between the lexicon and the grammar, allowing them to interact in novel ways. 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) for a review.

A properly elaborated 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 develop a highly interactive theory of lexical representation and phonological grammar would be to build on research on statistical 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.

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. However, 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 probabilistic component of the theory is largely independent of the rewrite rule formalism. The situation is quite different for the constraint-based theories of phonological variation that we have surveyed in this chapter. The factors that determine the probability of a process are the phonological constraints themselves, and their relative ranking or weight, the very same factors that determine whether a categorical process applies in standard OT. That constraint-based models of probabilistic phonology are firmly grounded in the core formal mechanisms of the theory bodes well for the continued placement of variation as a central topic of research in generative phonology.
References


Vaux, Bert. 2007. Why the phonological component must be serial and rule-based. Ms, University of Cambridge.

