1 Variation and phonological theory

GREGORY R. GUY

Introduction

The study of linguistic variation is often perceived to be quintessentially engaged with phonological phenomena. This is a manifest misperception: variationist work on morphosyntactic issues began with the original foundational articles that launched the "variable rule" framework (Labov [1969] on the English copula, and Labov [1972d] on negative concord), and continues to be among the most active areas in the field. But it is instructive to consider *why* such a misperception persists. There are two factors that drive this view. First, there exists an almost prescriptive attitude that phonology is the only domain in which linguists *should* speak of variation, arising from an uneasy suspicion that any alternations found at other levels of linguistic structure might involve intentional differences in meaning. In Labov's informal definition, variation involves "different ways of saying the same thing," and for most linguists it is easy to conclude that *runnin'* and *running* are different versions of the "same thing," but rather worrisome to make the same claim about *Kyle got arrested* and *Kyle was arrested*. Hence the view that variationists tidily confine their labors to the vineyard of phonology alleviates this existential angst about the status of morphosyntactic variation.

But a second, more interesting, reason for this view is that it is indeed quite true that work on phonological variation has been deeply intertwined with phonological theory. Phonological variation in all languages is massively structured and orderly; there is a random component, such that the surface realization of given utterance cannot be predicted categorically, but the patterns of realizations in particular contexts are probabilistically structured with great regularity – particular realizations are strongly favored by particular phonological contexts. Most of these patterns of contextual constraints on phonological variables find clear explanation in principles of phonological organization; in other words, phonological theory can (and should) explain the variable aspects of phonology along with the categorical facts. And this relationship, as with all scientific theories, is reciprocal and reinforcing: the evidence from phonological variation has been brought to bear on a variety of theoretical questions in phonology. This includes quantitative evidence and quantitative argumentation, approaches which were historically uncommon and unfamiliar in phonological theory, but which are becoming increasingly evident in recent years (cf. for example, the work of Anttila [1997]

and Kiparsky [in press] on partial constraint rankings, and of Boersma and Hayes on Stochastic Optimality Theory [Boersma 2003, Boersma and Hayes 2001]. In this respect, work on phonological variation is comparable to the development of laboratory phonology, in that it provides new kinds of data to inform and illuminate the development of phonological theory.

This chapter explores the reciprocal, mutually illuminating relation between phonological variation and phonological theory. First, we will consider some examples of how theory contributes to explaining the data; in particular, we will see how the linguistic constraints evident in phonological variation are consistently interpretable in terms of the principles and mechanisms proposed in phonological theory. Second, we will examine some of the ways that variation data has contributed to clarifying or even resolving theoretical issues in phonology. Finally, we will discuss the general theoretical question of how to best construct a theory that models both the variable and invariant facts about the sound systems of human language, and hence explains how language can be both discrete and continuous in its organization.

Explaining the patterns: what phonological theory does for the study of variation

The fundamental observation of research on linguistic variation is that it displays, in the words of Weinreich, Labov, and Herzog (1968), "orderly heterogeneity"; in other words, the alternating variants occur in probabilistically regular patterns, not in a random distribution. These orderly patterns exhibit social regularities (e.g. higher status speakers always use more of the socially highly valued variants), which are discussed elsewhere in this volume. Our focus here is on the linguistic regularities that are also apparent. These take the form of contextual conditioning: certain linguistic contexts favor the occurrence of particular variants. Thus phonological reduction processes, if sensitive to stress, typically occur more often in unstressed syllables, assimilation processes typically occur more often word-internally than across word boundaries, and vocalization of sonorants occurs more often in coda positions than onsets. Such results are unsurprising. To a phonologist, none of the examples just cited contravenes any theoretical principle, while all of them resemble numerous cases involving categorical alternations. The central observation here is that variable processes display the same patterns of occurrence and non-occurrence that are found for categorical alternations, and hence are likely governed by the same principles and generated by the same processes of grammar. Since alternations are what phonological theories have classically been designed to account for, we can reasonably expect that extant theories incorporate explanatory principles and generalizations about linguistic structure that are relevant to variable alternations.

As an example, consider the patterns of alternation between occurrence and non-occurrence of word-final consonants. We find many cases of categorical alternation, such as French liaison, where a consonant is articulated at the ends of given words when the following word in the utterance begins with a vowel, but is absent when the next word begins with a consonant. Such cases are typically described in phonological theory as involving an underlying consonant that is suppressed under certain conditions that would be phonologically infelicitous – in this case, when it is in the coda – but retained in more favorable conditions, e.g. when it can be syllabified as an onset. In the terminology of Optimality Theory, a markedness constraint like *Coda outweighs considerations of faithfulness to the underlying form.

Parallel patterns showing the same kind of constraint but involving variable rather than categorical conditioning are also easy to find. English (also Dutch) has alternating presence and absence of final coronal stops, and this alternation is affected by whether a following word begins with a vowel or consonant, but the alternation is not categorical. That is, a word like *east* can occur as *eas'* in any following context, but the form with deletion is much more common when there is a following consonant. Thus the pattern is:

frequent, preferred:	*east end*	*eas' side*
possible, but rarer:	*eas' end*	*east side*

The generalization is that the language prefers retention before vowels and deletion before consonants. This is the same generalization that could be made about French liaison. The difference between the two cases is that in French, the dispreferred cases are absent, while in English coronal stop deletion they are not entirely absent, but simply occur less often.

This is an example of what has been described as the "stochastic generalization" relating variable and categorical observations in linguistics (Clark 2005:209, Bresnan, Dingare, and Manning 2001). Many of the principles and processes proposed in phonological theory to account for categorical facts are also evident in variable operations, in a probabilistic form. Some principle enunciated on the basis of the observation that in language A, structure X never occurs, turns out in language B to explain why structure X is very rare, although not categorically absent.

In the balance of this section, we consider some examples of how general phonological principles are reflected in the probabilistic distributions found for phonological variation. The exposition focuses on one variable which is typical of the kinds of patterns evident in variable phonology: the alternation in Brazilian Portuguese between presence and absence of word-final sibilants.

Final sibilant deletion in Brazilian Portuguese

In vernacular speech, Brazilian Portuguese shows great variation in the realization of word-final sibilants: words such as *menos* "less, minus" and *ônibus* "bus" are,

Table 1.1 *Constraints on final sibilant deletion in Brazilian Portuguese (Data from Guy 1981)*

Factor	*N*	% deleted	Factor weight
Word stress			
Stressed monosyllable	7504	6	.24
Stressed polysyllable	1375	10	.34
Unstressed	1392	53	.86
Following segment			
Vowel	3625	8	.40
Consonant	4876	16	.60
Voicing of following consonant			
−voice	2270	9	.42
+voice	2606	21	.58
Place of following consonant			
Labial	1600	14	.53
Alveolar	2240	21	.66
Velar	1036	6	.31
TOTAL	10271	13	–

for many speakers, more often realized as *meno*, *ônibu* without the final consonant. This variation is subject to a number of constraints, which are illustrated in Table 1.1.

These data raise basic linguistic questions: why do we find these patterns, and not others? Why do these contexts have the observed effects? These are the kinds of issues that phonological theory is intended to answer. Let us consider each constraint in turn.

Stress

Word stress is found to condition phonological operations and distribution in virtually every language that has a stress contrast. The direction of effect observed here is that stressed syllables have greater retention (i.e. are more faithful to underlying form), while unstressed syllables are more congenial to deletion. This is consistent with theories of prosody, positional prominence, etc., and with categorical alternations in many languages. It is also consistent with diachronic principles: in language change, stressed positions are more resistant to lenition and deletion processes.

Following segment

Increased rates of deletion in preconsonantal contexts are widely observed in variation studies. The theoretical explanation for this lies in principles of syllable structure. A word-final consonant resides underlyingly in coda position, which is universally marked and disfavored. Theories of syllable structure state this in

various ways; thus, CV phonology (Clements and Keyser 1983) treats CV as the universally unmarked syllable type, while Optimality Theory postulates NoCoda as part of the universal inventory of phonological constraints. So coda deletion is an expected repair, and a common diachronic change. However, a following vowel licenses the consonant as an onset, which is an optimal position for retention. Word-internally in Portuguese, as in many other languages, prevocalic consonants are obligatorily syllabified rightwards, as onsets. Across word boundaries, this is optional, and the outcomes are variable.

Voicing of following consonant

The data show appreciably more deletion before voiced than voiceless consonants. A theoretical explanation of this result requires one additional observation about Brazilian Portuguese. Voicing of sibilants is not phonemically distinctive in coda position; hence final sibilants assimilate obligatorily to the voicing of a following consonant. The pattern shown here therefore reduces to the generalization that voiced fricatives are deleted more than voiceless ones, which has a ready explanation in markedness. Voiced fricatives are universally more marked than their voiceless counterparts; they are also typologically rarer, and raise aerodynamic problems in articulation, since the glottal impedance associated with voicing reduces the airflow required to generate the turbulence of frication.

Place of following consonant

The figures in the table indicate a robust effect of the place of a following consonant, with highest deletion rates before an alveolar, second highest before a labial, and least deletion before a velar. This is a clear example of the Obligatory Contour Principle (OCP), which states that adjacent identical elements are dispreferred. It was first proposed in phonological theory to account for the avoidance in tonal languages of sequences of adjacent identical tones, but it has been generalized to phonological processes that avoid adjacent identical segments and features (cf. Yip 1988).

As the name implies, the OCP was originally postulated to account for obligatory, categorical phenomena, but numerous gradient or variable phenomena also confirm a general preference for "contoured" sequences (where adjacent elements are dissimilar) over "level" sequences where adjacent elements are identical or similar. For example, Guy and Boberg (1997) found that English coronal stop deletion shows an OCP effect of the preceding consonant: there is more deletion after segments that are phonologically similar to the targeted /t,d/, i.e. those that share more features. Thus deletion is favored by preceding stops (e.g. *act, apt* – same in continuancy and obstruency) and alveolar fricatives (*last* – same in place and obstruency), but disfavored by preceding liquids (*cold, hard*) and labial fricatives (*left*), which share fewer features with the target.

The place data in 1.1 show essentially the same pattern. A conventional distinctive feature treatment of place contrasts velar, alveolar, and labial in terms of several features, as in the following matrix:

	[coronal]	[back]
labial	−	−
alveolar	+	−
velar	−	+

In this treatment, alveolar place shares one feature with labial place, but none with velar. Hence the deletion target, a coronal sibilant, is most similar in place to a following coronal consonant (like t,d,n); partially similar to labials (p,b,m), and most different from velars (k,g). The deletion facts in Table 1.1 follow this cline of similarity, implying that they are governed by a Contour Principle that is not obligatory, but probabilistic.

Constructing the theory: what variation does for phonological theory

The above examples illustrate the explanatory value of phonological theory for the analysis of variation. Now we turn to the utility of variation data for the evaluation and construction of phonological theory. As with any data, evidence of variation can be used in several ways: it can provide empirical tests of theoretical issues, it can confirm or deny the predictions of theoretical models, or it can provide facts that theory must account for. But the greatest theoretical significance of the study of phonological variation is that it has the potential to resolve theoretical issues that cannot be addressed by other means. Categorical alternations lack nuance: given a defining set of conditions, they abruptly select a single outcome. But the continuous frequency ranges of phonological variables, displaying sensitivity to a number of features of the context, offer a subtler analytical tool that can probe more finely into phonological structure. In this section I will offer an extended example of how variation data provide a unique empirical test of a theoretical issue in phonology: the treatment of lexical exceptions to phonological processes.

Phonological theory is centrally concerned with identifying generalizations about sound systems and hypothesizing mental grammatical structures that explain why and how those generalizations come about. Generative and post-generative models of phonology typically assume a bipartite architecture consisting of a phonological component, in which the generalizations are captured, and a lexicon, which lists the ungeneral, specific characteristics of individual words. For example, in the word *act*, the fact that the coda cluster /kt/ shows a constant value for the feature [−voice] throughout, and has the /k/ preceding, rather than following, the /t/, are general features of English phonology, but the fact that the vowel is /æ/ rather than /ey/ or /iy/ is one of the distinctive properties of this lexical item that distinguish it from *ached, eked*, and other words of English. The basic organizing principle is: general properties = phonology, specific properties = lexicon.

The problem that arises, however, is that there are many phonological generalizations that do not apply to the entire lexicon; rather, some lexical items are exceptional in certain respects when compared to most other words in the language. Thus English shows a vowel laxing alternation in *serene–serenity, obscene–obscenity*, but not in *obese–obesity*. Also, in Philadelphia English, the vowel /æ/ is tense before tautosyllabic anterior nasals and fricatives (hence tense *man, mansion, ham, hamster, half, after*, vs. lax *hang, hammer, planet, scaffold, have, that, sad, sack*, etc.); however, *mad, bad, glad* are tense despite the following /d/ (cf. lax *sad, Dad, had, fad*, etc.). How are such cases to be accounted for?

Although the theoretical literature on lexical exceptions has focused on categorical alternations, the same issue also arises in phonological variation. There it takes the form of lexical items that undergo certain processes at an exceptional rate, compared to other words of comparable structure. For example, the word *and* is produced without a final stop far more often than phonologically similar words like *hand* or *band*. So an adequate phonology of variation faces the same problems confronted by a categorical phonology.

Given the phonology-with-lexicon architecture, there are just two ways that lexical exceptions have been handled without dropping the generalization from the phonology. First, exception features can be attached to lexical items to co-index them with phonological processes; this is the mechanism suggested by Chomsky and Halle (1968). A lexical item that fails to undergo rule *n* can be annotated in the lexicon with a feature [−rule *n*]; similarly, a set of lexical items that undergo some rule *m* that other words do not can be annotated with a feature [+rule *m*]. Second, the exceptional outcomes can be directly represented in the underlying representation of the exceptional words, preempting the phonological processes that would otherwise apply or fail to apply.

These two approaches to lexical exceptionality have survived the theoretical shift in phonology from rules to constraint-based formalisms. Optimality-theoretic treatments of exceptionality use the same two strategies, relying either on preemptive structural marking of underlying representations or on lexically-specific constraints that apply only to co-indexed lexical items (cf. Pater and Coetzee 2005). It therefore appears that the roots of these approaches lie in the dichotomous architecture of phonology vs. lexicon – one repository for general facts, one for particular facts. The existence of exceptions implies that there are "generalizations" that are only partially true, i.e. partly general and partly specific. The dichotomy between phonology and lexicon therefore gives us two choices. We can focus on the supralexical generality of the pattern, thus retaining the phonological mechanisms that would capture it (whether they are rules, representations, or constraints), but delimiting their lexical scope by means of exception features; this is the "phonological" approach. Or, we can focus on the particularity of the exceptions by writing them directly into the lexical representations, thereby preempting the phonological mechanisms from accounting for them; this is the "lexical" approach.

The question for phonological theory is: which of these approaches is preferable, or in some sense "superior"? Chomsky and Halle had a formal algorithm for answering this question based on economy: write a rule whenever it saves more features than it costs. But it's not clear that language and mind work on so strict a parsimony principle. Empirical evidence is often unhelpful in deciding this issue, because the two analyses end up making the same predictions. Thus the Philadelphia /æ/ example could be treated either way: an exception-feature treatment would assign *mad, bad, glad* a diacritic to indicate that they undergo /æ/-tensing, and a lexical treatment would simply mark these words as tense in the lexicon. So the issue has remained undecided through four decades of theoretical development in phonology.

Happily, quantitative evidence from phonological variation offers the prospect of an empirical test of the two approaches. The examples discussed above are undecidable partly because of their lexically categorical nature: a given word either is or is not an exception to some phonological generalization, so there is no possibility of interaction with other conditions that might clarify the question. But variable processes, as we have seen, typically are strongly conditioned by features of the context. If these contextual conditions are equally present in exceptional and unexceptional words, this would suggest that all words are operated on by the phonology, hence an exception-feature analysis. If, however, the exceptional words show different conditioning from the non-exceptional, it would suggest that the exceptions are not undergoing the phonological processes in the same way as other words, hence favoring a lexical analysis in which the exceptions have distinct underlying representations.

Consequently, for certain cases of lexical exceptionality in variable processes, the two approaches make different quantitative predictions, which can be empirically tested in a suitable mathematical model. I will illustrate using the familiar variable rule (VR) framework (cf. Chapter 10 of this volume for an extended discussion), but other mathematical models should yield comparable results.

To see how this works, consider the case of English *and,* which is observed to have an exceptionally high rate of absence of the final coronal stop. In a phonological (exception-feature approach), *and* is indexed with a feature that tells the deletion process to raise the probability of affecting this word. This is easily represented in the VR model by associating the word with a factor weight that captures the effect of that particular word on the probability of occurrence of the variable. This lexically specific factor weight for *and* would have a value greater than .5, which in the variable rule model would boost the probability of deletion in this word above the rate experienced by words lacking such an exception feature.

In the lexical approach, however, the exceptionality of *and* is captured by an alternative underlying representation lacking the final *-d*: i.e. *an'*. I submit that this is what is implied by the common orthographic device of spelling *'n'* in phrases like *rock 'n' roll*, an orthographic recognition of this mental representation. In this approach, speakers have two mental representations for this word: when they

select underlying *and*, it undergoes deletion at the same rate as other words, like *band* and *land*, yielding the expected proportions of full and deleted forms. But sometimes speakers will select the underlying form *an'*; in this case the deletion rule is irrelevant and the word will always surface without the final /d/. What we observe on the surface is thus the sum of two different pathways to /d/ absence, with the mathematical effect of boosting the observed cases of missing /-d/s in this word. If, for example, a speaker with a 30% coronal stop deletion rate also selected the exceptional *an'* representation half the time, they would have a surface rate of absent *-d* of 65% in this word, composed of the 50% of tokens derived from underlying *an'*, plus the usual 30% deletion of the other 50% of tokens derived from underlying *and* (30% of 50% equals an additional 15% of the corpus).

This is crucially relevant to deciding the theoretical issue at hand because the two analyses are *not* mathematically symmetrical! Rather, they make different predictions with respect to how they interact with other constraints on the variable process. The phonological or exception-feature approach predicts that other constraints on the process should be entirely independent of the status of lexical items, while the lexical approach predicts the other constraints should be attenuated or nullified in a set of observations. (Note that a variationist study of this problem must rely on a corpus rather than single cases. For any single utterance we cannot say whether it was selected from the lexicon or generated by the phonology. But in a quantitative analysis of optional processes, we can often find statistical regularities in a *corpus* of utterances that allow us to draw inferences about what is going on. This is the method that is relevant to the present example.)

The independence of linguistic constraints has been confirmed by most of the research in the variationist framework, so it is the default expectation for lexical exceptions. Independence in this sense means that the effect of one constraint is evident and proportionally constant regardless of what other constraints are operative in a given case. Thus the constraints on English coronal stop deletion mentioned above – the OCP effect (more deletion after a preceding coronal segment, and less after a non-coronal, hence, more deletion in *west* than *left*) – and the following segment effect (more deletion before a following word beginning with a consonant than a vowel) show this relationship: phrases with following consonants but differing OCP values like *west side* and *left side* both display more deletion, by proportionately the same amount, than phrases with following vowels like *west end* and *left end*, while at the same time, words like *west* (that reflect OCP avoidance of double-coronal sequences) have proportionately more deletion in all phrasal contexts, in comparison with words like *left* (with no OCP effect).

In a VR analysis, therefore, an exception feature should work just like any other contextual constraint. A word that came with such a feature, like *and*, would experience an independent adjustment in its probability of deletion in comparison with other words, *but the effect of other contextual features will continue to operate at the same magnitude*! Thus we would predict that *and* followed by a vowel should show the same reduction in deletion as *west, left*, or *land* followed

P1: FXS/FGC P2: FXS
9780521871273c01.xml CUUK1061B-Bayley & Lucas July 16, 2007 13:40

by a vowel, vis-a-vis the same words followed by consonants. Hence, *ham and eggs* should have less deletion than *cheese and crackers*, in the same proportion as *second effort* compared with *second son*. In this approach, all other contextual effects should be constant across exceptional and non-exceptional words.

However, the lexical approach makes a quite different quantitative prediction. Under this hypothesis, whenever a speaker selects underlying *an'* without the final *-d*, such tokens never undergo deletion, and so show *no* effect of contextual constraints. Hence one should be just as likely to say *ham 'n' eggs* as *cheese 'n' crackers*. Intuitively, I feel that this is correct. Retaining the -d in *ham and eggs* sounds overly precise.

However, since speakers do occasionally pronounce a /d/ in *and*, even if rarely, they must still have an underlying representation that retains a final -d; when this form is selected, coronal stop deletion can still apply, in which case the contextual constraints on the process will still operate. Therefore, the surface corpus is actually a composite of two data sets produced by different derivational pathways; one in which contextual effects are evident and one in which they are not. The statistical effect of this conjunction will be to weaken the apparent contextual effects in the observed corpus, because they apply to only some of the words in the corpus, not all. Selection of -d-less allomorphs acts as statistical noise in the corpus, attenuating the statistical evidence for the external constraint effect.

Therefore, phonological variables with lexical exceptions offer straightforward empirical tests of the two approaches. If speakers' grammars use exception features, they should have the same magnitude of constraint effects for exceptional and unexceptional words, but if their mental grammars rely on alternative lexical entries for exceptional words, these should exhibit surface attenuation of the effects of contextual constraints.

Quantitative testing of phonological theory

As it happens, there are several phonological variables described in the literature that are known to have lexical exceptions and can provide testing sites for this theoretical problem. As a first example, we return to the case discussed above of final sibilant deletion in popular Brazilian Portuguese. The general alternation between presence and absence of -s is systematic across the lexicon: all words with final -s enter into it, and virtually all show the same contextual effects seen in Table 1.1. Hence, the straightforward analysis is to postulate the alternation as a general pattern belonging to the phonology. The relevant words appear in the lexicon with underlying final -s, while surface presence or absence of -s is governed by a phonological process. In a generative phonology this is a variable -s deletion rule; in an Optimality Theoretic approach, this is captured by a version of the NoCoda constraint, like *–S##, which is variably ordered with respect to an appropriate faithfulness constraint and the constraints that capture the contextual effects.

Table 1.2 *Final* -s *deletion in Brazilian Portuguese: Following context effects and lexical exceptionality (factor weights; data from five cities in VARSUL corpus)*

Features of following consonant		Non-exceptional words	Lexical exceptions (*-mos* forms)
Voice/Manner	sonorant	.69	.49
	voiced obstruent	.44	.58
	voiceless obstruent	.36	.44
Range		.33	.14
Place	labial	.32	.58
	coronal	.61	.53
	velar	.44	.39
Range		.29	.19
N		5880	1225
Log likelihood		– 704.8	– 791.5

For present purposes, the most relevant contextual constraints are those that are external to the lexicon, the following context. We focus on two of them: the *place* and the *voicing/manner* of following consonants. The place effect is OCP-like, with *more* deletion of -s before another coronal consonant, less deletion when the following consonant has a different place. The voicing effect is presented here in more detail than in Table 1.1: the deletion-promoting voiced segments are here broken down into sonorants and obstruents, with the result that sonorants favor deletion more than voiced obstruents, with the least deletion occurring before voiceless obstruents.

Previous studies have demonstrated that these constraints affect the lexicon as a whole. But recent work on this variable, by myself and colleagues in the VARSUL consortium in southern Brazil, has revealed a significant lexical exception that was not adequately treated in earlier studies. This is the verbal morpheme *-mos*, which marks first person plural: *nós falamos, nós comemos* "we speak, we eat." In running speech in the VARSUL corpus (Zilles 2005), these typically occur without the final -s at a higher rate than other unexceptional final -s words, like *menos* and *ônibus*. So people say *falamo, comemo* more often, relatively speaking, than they say *meno*.

The overall figures on this point are quite striking. The deletion frequency in this corpus for *-mos* forms is 41%, versus only 10% for other unstressed non-inflectional -s (in words like *menos*) and just 2% for stressed -s (e.g. *demais, rapaz*). Accordingly, we ran separate analyses of the *-mos* forms vs. other words, and the results are shown in Table 1.2, organized according to place and manner of following consonant.

For non-exceptional words, the results confirm the findings seen in Table 1.1: for voicing/manner, more deletion before voiced segments, with peak deletion

before sonorants, and for place, maximum deletion before other coronals. The magnitude of these effects can be measured by means of the range of values in the factor group from highest to lowest: strong effects should have large ranges, while weak effects have values clustered close to .5, with smaller ranges. In this case, both factor groups have substantial ranges: .33 for voicing/manner and .29 for place.

When we look at the same constraints on the *-mos* words, however, the picture is quite different. First, the generalizations about which contexts are most favorable are both lost. Sonorant is no longer the most favorable voicing category, and remarkably, coronal is not the most favorable place! This is striking, given the systematic evidence for OCP preferences in many phonological processes, both variable and invariant. This is strong evidence of a non-phonological process affecting these data. Now consider the ranges; in both factor groups the range of values has shrunk in the exceptional cases, by a factor of one-third for the place effect, from .29 to .19, and by more than half for the voicing effect, from .33 to .14. The phonological effects appear weakened in these data, suggesting a pre-phonological variable accounts for increased -s absence in *-mos* forms.

Another comparison of the exceptional and non-exceptional cases can be made with the log likelihood (l.l.) statistics – the goodness of fit measure incorporated in the VARBRUL procedure. This is a negative number whose absolute value increases with respect to two parameters: the number of tokens in the corpus, and the goodness of fit between model and data – a worse fit gives a bigger l.l. In these data, the non-exceptional corpus of 5900 words has a l.l. of −705. The lexical exceptions, with a corpus only one-fifth the size (1225 words), show a *larger* l.l., of −792! The fact that non-exceptional words have a smaller l.l., even with many more tokens in the analysis, means they fit the model much better than the exceptional items. The appropriate conclusion is that the exceptional items are not well predicted by purely phonological factors; something else is going on. That "something else," I suggest, is lexical: many verbal plurals lack a final -s in underlying representation, in the input to the phonology; therefore, the phonological context does *not* explain their absence very well.

A second empirical test of the treatment of lexical exceptions comes from Salvadoran Spanish, which also has a final -s deletion process. Hoffman (2004) finds exceptional behavior in several discourse markers that show exceptionally high rates of -s absence, namely *entonces, pues,* and *digamos*. When she analyzed these tokens separately, she found different results for the phonological constraints on the process. The results for two constraints, stress and following segment, are presented in Table 1.3.

The following segment effect parallels the Portuguese case: vowels and voiceless consonants disfavor deletion, but voiced consonants and sonorants favor. But the magnitude of the effect is smaller in lexical exceptions: the range of values is reduced from .42 to .25. The stress effect, in which unstressed tokens favor deletion, also shows a reduction in magnitude in exceptional words, by a factor of one-third. Remember, the exceptional cases show a higher rate of absence overall,

Table 1.3 -s *deletion in Salvadoran Spanish: Stress and following context effects and lexical exceptionality (factor weights; data from Hoffman 2004)*

Factor group	Factor	Non-exceptional words	Lexical exceptions (*entonces, digamos, pues*)
Following context	sononant	.60	.63
	voiced obstruent	.75	.55
	voiceless obstruent	.33	.38
	vowel	.36	.38
	pause	.44	.56
	range	.42	.25
Syllable stress	stressed	.38	.42
	unstressed	.62	.58
	range	.24	.16

so it is not simply the case that effects are attenuated by a lack of evidence. Rather, these results parallel the Portuguese case, suggesting that the increased absence of final segments in the exceptional cases is due to the inclusion of items that are not conditioned by context, because they do not have the final -s present in their underlying representation.

Finally, for a third example, let us return to the case of English *and*. Precisely because of the exceptionality of *and*, there are few published studies that deal with it. Since it was recognized in the earliest work on coronal stop deletion that *and* doesn't behave like other lexical items, the practice was adopted of excluding *and* from studies investigating the general process. But one published study that did look at *and* is Neu (1980). The data in Table 1.4 are drawn from Neu's work.

Since Neu presented her data in univariate frequency tables, no VARBRUL analysis is possible, and the figures in Table 1.4 represent percentages, not factor weights. Overall, Neu finds that *and* surfaces without a -d some 90% of the time – an extraordinarily high figure compared with an overall deletion rate of about 30% for other words. As noted above, English coronal stop deletion is strongly conditioned by following segment effects, and this shows up in Neu's results: non-exceptional words have 39% deletion when followed by consonants, vs. under 16% when followed by vowels, for a range of 23%. But what happens in the exceptional word, *and*? In the percentage data, the figures are 95.7% deletion before consonants and 82.1% before vowels, for a range of 13.6%, which is only about half the range found for the non-exceptional words. So on these facts, the data support the lexical selection model: there is an additional lexical entry for *and*, without a final -d, which is selected some 70–80% of the time.

All of the examples we have considered support the same conclusion: exceptional lexical items in cases of variable phonological processes are best treated lexically, by means of alternative underlying representations, rather than by means

Table 1.4 *Coronal stop deletion in English: Following context effect and lexical exceptionality (% deleted; data from Neu 1980)*

	Non-exceptional words		Lexical exception (*and*)	
Following context	N	% deleted	N	% deleted
Obstruent consonant	572	39.3	441	95.7
Vowel	495	15.8	312	82.1
Range		23.5%		13.6%

of an exception-feature approach. If the exception- feature treatment were valid, at least some cases ought to show a constant effect of other phonological constraints across both exceptional and non-exceptional words, which would be the empirical manifestation of a model in which both kinds of words are subject to the same processes and constraints, albeit at different overall rates. But this is not the case. There is no evidence that exception features are operative in the mental grammars governing these cases of phonological variation.

These facts suggest a further prediction. In principle, the exception- feature approach permits both positive and negative exceptions – that is, there should be words that undergo phonological processes at both exceptionally high and exceptionally low rates. But the lexical approach, which encodes outcomes directly in the lexicon, does not permit exceptionally low rates of occurrence, at least in cases of deletion. There is no reasonable way to construct alternative entries for *and*, *-mos*, or *entonces* that will resist the deletion processes more than other words. Hence the lexical approach predicts that only words with exceptionally *high* rates of occurrence should be found. It is my impression that this prediction is consistent with the cases discussed in the literature.

Assuming this prediction is also confirmed, we will have strong quantitative evidence bearing on the theoretical issue at hand: all the data are consistent with the predictions of the lexical approach to exceptionality, and none are consistent with an exception-feature treatment. This is a potentially decisive resolution from phonological variation that has not been achieved in four decades of work on categorical processes.

Towards an adequate theory of phonology

Formal theories of phonology have, for the most part, been constructed to account for invariant facts. This is due to the dominant bias in structuralist and post-structuralist linguistics favoring categorical, invariant models and generalizations over probabilistic ones. Informally, this bias reflects an assumption of invariance; that only categorical generalizations are valid products of grammar. Hence a generalization that is not always true is of little value; encountering these, the

linguist is trained to refine the statement of the generalization so as to exclude the variable bits, seeking an absolute definition that is always true within a certain domain. Outside of linguistics, of course, there is a general recognition of the validity and utility of probabilistic generalizations: men are taller than women, dark clouds bring rain, it's bad to drink and drive. None of these is categorically true, but all are useful principles for organizing one's observations of reality; they constitute valid generalizations that a "theory" of life would do well to incorporate.

In the domain of phonology, any careful observation of the way people talk reveals that there are both categorical and probabilistic patterns in the data. An adequate theory of phonology should be able to account for both. As we have seen, the same principles govern both categorical and variable patterns: for example, the OCP, which governs categorical alternations such as the realization of the -s and -ed suffixes in English (which always have an epenthetic vowel inserted whenever they are attached to a root ending in a like consonant, but not otherwise – *passes* vs. *puffs, raided* vs. *raked*), also governs the variable alternation between presence and absence of final coronal stops in English and of final sibilants in Portuguese. In the case of English -ed and -s, adjacent *identical* segments are categorically prohibited, while in the variable deletion examples in English and Portuguese, adjacent *similar* segments are probabilistically disfavored. But both observations reflect a common harmonic principle that has the status of a phonological universal: phonologies prefer contours – sequences of differing articulations – over non-contoured sequences involving repetition of the same articulatory gestures. A phonology that captures only the categorical generalization, while ignoring the probabilistic one or consigning it to be treated by a separate non-categorical principle, is self-defeating. Such an approach is either inadequate, because it leaves some of the facts unexplained, or logically unsound, because it violates Occam's razor by unnecessarily multiplying explanatory principles. A sound and adequate treatment, however, would see such cases as manifestations of a single principle governing outcomes with a range of probabilities: sometimes the predicted probability is 1 (the categorical cases), sometimes it is less than 1 (the variable cases). But in all cases, the common prediction is made that non-contoured outputs are less favored than contoured ones.

Such stochastic generalizations of phonological principles are a central element of an adequate theory of phonology. Abandoning the assumption of invariance enables a broader range of facts to be brought under the explanatory scope of the grammar, including the quantitative patterns evident in variation.

It is important to note that accounting for variation does not, for the most part, demand or preclude any particular formal framework, once the assumption of invariance is suspended. Any framework that incorporates a representation of optionality can, in principle, be adapted to account for phonological variation. Theoretical treatments of variation have been proposed within most of the influential phonological frameworks of the last half-century. Labov's early accounts of phonological variables such as (r) and (th) are couched in a structuralist terminology, in which units at one level of description (in this case the variable)

subsume several realizations at another level, much as the structuralist phoneme comprises several contextually-selected allophones. The quantified, probabilistic model known as the "variable rule" (VR) model, developed by Labov (1969) and Cedergren and Sankoff (1974), was formulated in the terminology of generative phonology. In this model, each rule of grammar is assumed to be associated with a probability of application, which is 1 for categorical rules, and less than 1 for optional – i.e. variable – rules. Additionally, quantified constraint effects are represented in the VR model by attaching probabilities to particular features of the context of a rule.

Subsequent developments in formal phonology have also been adapted to treat variation. In non-linear phonologies, representational structures such as phonological tiers and feature geometries have been utilized to model variation by means of devices such as variable placement of association lines; for example, Guy (1991) treats English coronal stop deletion as a variable attachment of the segmental tier to its licensing position in the CV tier, with deletion arising from a generalized process of stray erasure. This analysis explained the higher rates of retention in prevocalic position as a result of resyllabification – the coronal stops are relicensed as onsets of the following syllable – and it made the novel and unexpected prediction that following /l/-initial words would favor deletion more than /r/-initial words, because of the English prohibition on *tl-, *dl- onsets. This prediction has subsequently been empirically confirmed.

Morphophonological aspects of variability have also received a formal account within the framework of Lexical Phonology (LP). A well-known example of morphological conditioning of phonological variation is the contrasting rates of coronal stop deletion in various morphological classes of English words: deletion is highest in underived words (e.g. *past*, *pact*, *bold*), lowest in regular past tense forms (*passed, packed, tolled)*, and intermediate in irregular past tense forms (*lost, kept, told*). Guy (1991a, 1991b) as well as Santa Ana (1992) and Bayley (1994a), explain this in terms of differing derivational histories in LP. The higher deletion rate in monomorphemic words is a consequence of multiple exposures of these words to a deletion process that iterates at each derivational level. By contrast, regular past tense verbs have lower rates of deletion because they only become candidates for deletion at the postlexical level.

This theoretical treatment also led to an unforeseen prediction which empirical research confirms: the relationship between retention rates in word classes with different derivational histories should be an exponential function. If x is the rate at which final coronal stops words are retained in words exposed to deletion only once (in this case, the regular verbs), then words exposed twice (irregular verbs, available for deletion both at one lexical level and again postlexically) have a retention rate of x^2, and words exposed three times (monomorphemic words, subject to deletion at one postlexical and two lexical levels) will have a retention rate of x^3. Several studies confirm this prediction: speakers who retain, say, 90% of coronal stops in regular past tense forms also retain about 81% (the square of .9) in irregular verbs, and about 72% (the cube of .9) in monomorphemic or

underived words. This result offers an unparalleled confirmation of the vision of the overall architecture of phonology that LP incorporates.

Variation in Optimality Theory

The most far-reaching development in recent phonological theory has been the emergence of constraint-based approaches, in which general, possibly universal, principles expressing desirable phonological states do most of the work of accounting for phonological generalizations. In Optimality Theory, these principles are summarized in a ranked list of constraints, each of which will prevail unless in a given case it would cause a violation of a higher-ranked constraint. Alternative realizations of a word or utterance ("candidate forms") are evaluated by the grammar according to the number and severity of constraint violations that they incur; the evaluation metric selects the candidate that incurs the least severe (lowest ranked) violations as the optimal output.

This model was originally conceived as deterministic and categorical: only one optimal candidate should exist for any set of circumstances, and that form should occur categorically in the output. This is accomplished by means of a fixed and comprehensive rank-ordering of the constraints: if constraint A always outranks B, then for any candidate set where they conflict, a form that satisfies A will always be preferred over one that violates A but satisfies B. However, the model is straightforwardly adaptable to account for variation by means of variable or indeterminate rankings for some of the constraints. Where A and B conflict, and are variably ordered, then sometimes the candidate that satifies A will be selected, and sometimes the candidate that satisfies B.

A number of scholars have taken this step and postulated OT models that can account for both variable and invariant facts in the same grammar, by means of variable or partial constraint ranking. Particularly notable are the works of Kiparsky (in press) and Anttila (1997, 2002), and in a somewhat different vein, the Stochastic OT model of Boersma and Hayes (2001). The approach taken by Kiparsky and Anttila (cf. also similar work by others, e.g. Nagy and Reynolds [1997]), relies on the different selections made by different rankings to predict the frequencies of occurrence of competing forms. In the coronal stop deletion case, for example, if the only constraints implicated were a faithfulness constraint and a markedness constraint against complex codas, and these were freely ranked, then whenever the faithfulness constraint ranked higher, final -t,d would be retained, and whenever the markedness constraint prevailed, -t,d would be deleted. If these orderings were random, each should occur half the time, predicting a surface coronal stop deletion rate of 50%. However, more complicated cases generate different quantitative predictions. If a following vowel favors retention because of a constraint that prefers onsets, then we might postulate three variably-ranked constraints affecting cases like *east end*. Deletion would occur only when *ComplexCoda outranked both Onset and Faith. Among the six possible orders of these

three constraints, this would occur in only two of them, or one-third of the time, predicting a deletion rate in such phrasal contexts of just 33%.

The jury is still out on the empirical adequacy of this model at predicting the actual frequencies of occurrence of phonological variables in differing contexts. Anttila (1997, Anttila and Cho 1998) has achieved remarkable quantitative accuracy with this approach, but Nagy and Reynolds (1997) were only partially successful. Guy (1997) has criticized this procedure because the frequencies end up being mere epiphenomena, a pure function of how many constraints are involved in the variable ordering – as we have seen, with just two constraints involved, the only possible frequency prediction is a 50–50 split between two outcomes, while with three, the only possible frequency predictions are 1/6, 1/3, 1/2, 2/3, or 5/6. A linguist who encountered some variable phenomenon with a robust 25% frequency rate would therefore have to conclude that at least four constraints were implicated, whether or not there was any theoretical or empirical evidence to support this conclusion.

Stochastic OT takes the variable ordering insight a step further, by distributing constraints as probability functions along a continuous linear scale rather than assigning them discrete ordinal rankings. Thus a constraint A centered at .9 on the scale would normally outrank a constraint B centered at .85, but in the production of actual utterances, both fluctuate over a range. In the evaluation of a particular candidate set, constraint A might on some occasion locate at the value .87, while B located at .88, with the result that a different candidate is considered optimal.

The crucial quantitative difference between this procedure and the variable ranking approach is that the distance between any two constraints in Stochastic OT can assume a range of values: two constraints can be very close together, or quite far apart. When they start out close together, their probability of overlapping will be high, but if their central distribution is far apart, they will rarely or never occur in an inverted order. Consequently, the likelihood of selecting particular candidates can be expressed as a function of the proximity of two constraints, rather than as a function of the number of constraints affecting an evaluation. In principle this should offer substantial improvement in the quantitative adequacy of the model for explaining the observed patterns of variation.

Conclusion

One of the promising trends in phonological theory in recent decades has been a widening of the data horizons, with more and more evidence from non-traditional sources being adduced in the construction and evaluation of theoretical models. Phonetic evidence, including experimental work in articulation and perception, child language data, data from language contact such as the treatment of loanwords, neurolinguistic and psycholinguistic evidence, statistical analyses of the phonotactics of lexical items – all of these have been brought to bear on theoretical questions in phonology. The patterns and probabilities of phonological

variation are part of this expanding landscape. An adequate theory of phonology will offer explanations of the broadest range of sound patterns, including non-categorical, probabilistic patterns, and the study of variation will inform the construction of such a theory. The assumption of invariance, which has dominated linguistic theory since the neogrammarians, has been useful in the history of linguistics as a debating strategem in certain theoretical arguments, and as a heuristic device for driving the research agenda, but it is not a design principle of human language. Phonological theory now has the tools in hand to replace it with more realistic models that can hope to achieve elementary observational and descriptive adequacy, in addition to pursuing the capacity to explain.