Bibliographic Details

The Blackwell Companion to Phonology

Edited by: Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume and Keren Rice
eISBN: 9781405184236
Print publication date: 2011

92. Variability

Gregory R. Guy

Sections

1 Introduction

Variability is a term used in phonology with several meanings. One common meaning is sociolinguistic diversity: speakers of different social backgrounds speak differently, and all speakers vary in speech style and register. Thus in New York City, as **Labov (1966)** famously demonstrates, a consonantal articulation of coda /r/ is preferred by higher status speakers, and by all speakers in their more careful styles, while a vocalized or deleted realization is preferred by lower status speakers and in casual styles.

Another sense refers to acoustic and articulatory diversity: English voiced stops can be articulated with voice onset times ranging from negative values to +10 msecs, with a mode around +5 msecs, while their voiceless counterparts typically show VOTs from +20 to +90 msecs. Successive articulations by the same speaker under the same conditions are not identical, but wander around these ranges. Vowels are similarly variable in realization. As **Peterson and Barney (1952)** showed, a vowel cannot be defined as an articulatory point, or as a particular acoustic realization, but rather as a region in articulatory and acoustic space; a series of measurements of a speaker saying a given vowel will show considerable scatter in this range, and sometimes items that fall outside it.

A third sense of variability addresses simple optionality: phonological characteristics or processes that may or may not occur in certain circumstances. Thus English voiceless stops in final position may be aspirated or not, although those in initial position show aspiration systematically. Hence we can describe aspiration as obligatory in initial position, but optional or variable in final position.

But, despite this range of meanings, variability is simple to define: in its broadest sense it is the inverse of generality. A phonological generalization is a statement about a sound system that is true everywhere, in every relevant occurrence; when some statement is not true everywhere, we encounter a case of variability. A phonological phenomenon that occurs in some, but not all, of its possible instances is not fully general; it is in some respect variable in occurrence (sometimes it happens, and sometimes it doesn't) or in realization (sometimes it happens one way, and sometimes another).

Variability as lack-of-generality is therefore a chronic problem for linguistic analysis. Generalizations are, of course, privileged in linguistics. Linguists are trained to seek generalizations; indeed, we see regularly recurring structures as a defining property of language, and the absence of all regularity as the defining property of noise (i.e. non-language). Consequently, phonology has repeatedly addressed issues associated with defining the frontiers of generality and the treatment of partial generality.

Limits on generality. The search for ever broader and deeper generalizations has been a prominent theme in the history of linguistics. The broadest generalizations about sound systems are phonological universals: statements that are true of all utterances in all languages. A truly universal property cannot be absent or contradicted in some language or some speakers, cannot have lexical exceptions, and must be apparent in 100 percent of cases. Hence a context-free universal can be characterized as in (1):

(1) *Maximum generality: Phonological universals*

For all human speakers (of all languages),
in all linguistic contexts,
in all lexical items,
x is always true.

Any phonological generalization that cannot satisfy all the quantifiers in (1) is less than universal. But this is true of most of phonology; although universals have important status, most work in phonology deals with generalizations that are limited in some respect, for example to a particular language or particular context. Hence, in a sense, most of phonology deals with variability, with partial generalizations that leave a region of variation where non-conforming realizations occur. The phonologist who pursues generality and regularity is therefore always confronted with the task of identifying the limits on the generalization, and the alternatives that occur beyond those limits. This is the problem of variability.

The problem can be approached in terms of the various quantifiers in (1). These are of two types: the “all” quantifiers, which deal with issues of scope, and the “always” quantifier, which addresses prevalence. When these are less than universal, they delineate the several problems of variability. First is social scope: generalizations that are true of only some human speakers (some language, speech community, or ethnic or other social group) constitute cases of social variation. Second is contextual scope: many generalizations are context-sensitive, i.e. valid only for items in a particular structural position. But the definitions of contexts, and indeed of what may count as a relevant context, are substantive theoretical and empirical questions. Various forms for the interaction between context and item have been proposed: a context may be seen as having a categorical effect, a gradient effect, or some other non-categorical outcome. And third is the problem of lexical scope: do phonological processes apply to all relevant words (all that have a given phonological structure in the right context)? Or do some words of the appropriate phonological shape nevertheless fail to conform to an applicable generalization by virtue of their lexical identity? If phonological statements are limited in their applicability to subsets of the lexicon, leaving words or sets of words in which different conditions prevail, we confront problems of lexical variability.

Orthogonal to the problems of scope is the problem of variable prevalence – does a given state of affairs always prevail, or is it encountered only some of the time? Although some theoretical approaches treat this as a scope problem, for example by seeking to define a narrow social or contextual domain in which prevalence is categorical, the logical problem of prevalence exists nonetheless: if all relevant dimensions of scope are held constant, is a given phonological generalization valid for all successive occurrences of relevant forms? If we listen to the same speakers producing the same words in the same contexts, do we always hear the same productions, or do they vary? Does a given process apply 100 percent of the time in the relevant domains, or less than 100 percent? And how does phonology cope with the two scenarios – how does it model categorical prevalence, and how does it account for variable prevalence?

In what follows we consider in turn each of these potential limitations on generality – each of these types of variability. We begin with the question of prevalence.

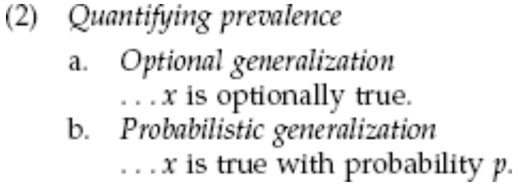
2 The quantitative limits of generality: Variable prevalence

A linguistic universal has universal prevalence: it always occurs wherever possible. We can describe a phonological generalization that is always true as categorical or obligatory. But how does phonology treat a state of affairs with less than universal prevalence – a generalization that is not categorically true, or a process that is not obligatory? This question lies at the core of what many linguists consider variability.

To illustrate this issue, consider common alternations, found in a number of languages, between presence and absence of coda consonants (CHAPTER 68: DELETION). In natural speech in English, words containing final coronal stops, such as *best, old*, are often articulated without those stops (cf. *bes' friend, ol' man*). In colloquial Caribbean Spanish and Brazilian Portuguese, syllable- and word-final /s/ is similarly variable in realization: *estamos, menos* are often articulated as *etamo, meno*. Such facts can be accounted for by phonological deletion processes, which are plausibly motivated by markedness considerations that are likely universal: simpler codas are universally less marked than more complex ones. But in the cases cited, the prevalence of these deletion processes is less than categorical. Not all utterances of eligible words undergo deletion.

These are not questions of scope. The alternations are not restricted to particular words. Although social groups vary in deletion rates, these societies are not composed of some groups or individuals who always delete and others who never do (**Guy 1980, 1981)**. And although these processes show contextual conditioning – for example, all three languages delete more before consonants than before vowels – the contexts do not define domains in which deletion always applies or never applies. No matter how we slice up the data in terms of scope, we always encounter both deleted and non-deleted forms. Hence, the prevalence of these phenomena is less than "always." How does phonology deal with such cases? Or indeed, is phonology responsible for such facts at all?

A phonological state of affairs that does not always prevail needs some statement of limitations on its prevalence. This can be done with quantitative vagueness, by replacing the "always" quantifier in (1) with non-universal quantifiers like "sometimes" or "optionally," or with existential quantifiers like "may occur." Or, a more precise quantification can be used, specifying some frequency or probability of occurrence (CHAPTER 90: FREQUENCY EFFECTS). Thus the prevalence clause on generality can be restated as (2a) or (2b):



(2) *Quantifying prevalence*

a. *Optional generalization*
. . . *x* is optionally true.

b. *Probabilistic generalization*
. . . *x* is true with probability *p*.

As we shall see, these choices are the subject of considerable theoretical debate. Some schools of thought strenuously argue that grammar has nothing to say about frequency or quantification, and consequently deny that statements like (2b) are permissible in formal phonology. Other frameworks embrace to a greater or lesser extent the quantification implied by (2b), seeking to account for variable prevalence by grammatical means. Let us consider the range of arguments bearing on variable prevalence.

The preference for categorical prevalence. Since so much linguistic analysis is inductive, it is unsurprising to observe a long, articulate tradition in linguistic theory of preferring generalizations that are categorical, i.e. true of 100 percent of relevant cases. Generalizations that are not fully general are often treated as valueless. The typical heuristic in linguistic analysis is to hypothesize a generalization, and, if counterexamples are discovered, to seek a reformulation of the generalization that either properly excludes the counterexamples, or accounts for them in another way, perhaps as a consequence of some other generalization (see also **CHAPTER 106**: EXCEPTIONALITY).

The historical prototype for this heuristic is **Verner's (1877)** refinement of Grimm's Law, which describes the sound changes characteristic of the Germanic languages. In the early nineteenth century Grimm and others discovered a set of common correspondences between the obstruents of the Germanic branch and those of other Indo-European languages (**Rask 1818**; **Grimm 1819**). Among these was the correspondence between Proto-Indo-European voiceless stops and Germanic voiceless fricatives (thus Latin *ped, tres, cornu vs.* English *foot, three, horn*). But there were also many recognized exceptions to these correspondences, so they appeared to fall well short of being categorical "laws." Thus in many words PIE voiceless stops end up as voiced stops or fricatives in Germanic (cf. Old English *fader, hundred vs.* Latin *pater, centum*). Based on the data known to early nineteenth-century linguists, the Germanic sound-shift might more accurately have been described as a variable process with several outcomes, among which one particular set (those known as Grimm's Law) were quantitatively prominent.

In 1877, this picture was dramatically altered when Karl Verner published a paper showing that one large class of exceptions was regularly conditioned by the position of the word-stress in PIE: the regular fricative outcome for voiceless stops occurs only in initial and post-tonic positions, while the exceptional voiced outcome occurs elsewhere. This discovery removed many counterexamples to Grimm's Law, reducing the domain of apparent variability, and greatly increasing the prevalence of the Grimm's Law generalizations. This allowed the conjecture that a correct definition of contexts would yield two categorical generalizations – two separate conditioned sound changes, each of which *always* occurred in its appropriate context. This conjecture received explicit formulation as the Neo-grammarian hypothesis: sound change is "exceptionless," admitting no variation (**Osthoff and Brugmann 1878**). By this hypothesis, any variability in the data was spurious, deriving from alternate sources such as borrowing, neologism, or dialect mixture, or from an inaccurate statement of the context of the change, calling for a Verner-like amendment of the generalization.

With the emergence of modern linguistics since **Saussure (1916)**, the idea of exceptionlessness became dominant in synchronic phonology: generalizations should be categorical. If a generalization is nearly categorical, the analyst should seek to make it categorical by redefining the context or by explaining the exceptions. Variability, in the sense of a phonological state of affairs for which no redefinition of context yields categorical prevalence, is considered an unhappy outcome by many schools of phonological theory, including structuralism, generative phonology and its various developments, and mainstream Optimality Theory, and it receives little theoretical attention in these frameworks, being tolerated only where empirically necessary.

Optionality. Despite this widespread theoretical preference for categorical statements, the task of accounting for a body of data has always led linguists to the necessity of non-categorical descriptive statements. In American structuralism, this necessity was formally treated in phonemic theory in terms of "free variation" (**Swadesh 1934**; **Hockett 1942**). Thus, to account for variable realizations of final voiceless stops in English, a structuralist analysis would present a list of possible allophones in this position which included the aspirated and unaspirated variants. Declaring that these allophones occurred in free variation was equivalent to a statement that their occurrence was random, not subject to further principled (i.e. structural or contextual) analysis. Such an account therefore adopted the strategy of (2a), declaring only that these alternatives all occurred with limited prevalence, but making no attempt to quantify their respective frequencies.

In rule-based generative frameworks, the equivalent mechanism is the "optional" rule: a rule which may apply, generating its output, or fail to apply, which leaves its input to surface unaltered (**Chomsky and Halle 1968**). Thus the English voiceless stop alternations could be modeled in generative phonology with an optional rule that rewrites the feature matrix of a voiceless stop to include aspiration in final position. When it applies, aspirated forms are generated, but when it optionally fails to apply, an unaspirated realization is generated. Again, this reflects the (2a) approach to variable prevalence, stating only that it exists, without further quantification.

Variable prevalence in OT. The treatment of variability-as-optionality has survived the transition from rule-based to constraint-based models. In constraint-based phonology, optionality can be captured with variable or incompletely specified constraint rankings (**Anttila 1997**). In Optimality Theory, a fixed constraint ranking is expected to yield a unique outcome for every evaluated form. This is the ranking equivalent of "obligatory" rules: if constraint A outranks constraint B, and the requirements of A and B conflict in some situation, candidate forms must obligatorily satisfy A over B in order to be selected. In most circumstances, the conjunction of the ranked demands of a set of constraints will rule out all but a unique candidate, excluding the possibility of variation.

But Optimality Theory also allows for the existence of a great variety of rank orders, and makes extensive explanatory use of ranking differences. Although OT postulates a universal constraint set, it captures phonological differences between languages by differing constraint rankings. Similarly, dialect differences, and, by extension, all sorts of sociolinguistic differences between individuals, class and ethnic groups, even speech styles and registers, can be modeled as differences in constraint rankings. The theory also models change across time as changes in the rank order of the constraint set. Since different orders generate different outcomes, variable orders within a given grammar will generate variable outcomes, and variable prevalence of any given outcome.

To illustrate, consider the English case mentioned above, final coronal stop deletion (CSD), where alternations like *best ˜ bes', old ˜ ol'* occur. To simplify for illustrative purposes, we might model this alternation in OT with two constraints, one that disfavors complex codas (*Complexcoda), and a generalized faithfulness constraint stating that underlying segments should be realized on the surface (Faith). An OT grammar with these two constraints will generate full forms (*best, old*) if Faith is ranked higher, but will generate deleted forms (*bes', ol'*) if *ComplexCoda has the higher ranking. Consequently, if these constraints are variably ranked, the grammar generates both forms in variation. Variable (or underspecified) constraint ranking is thus the OT equivalent of the optional rule.

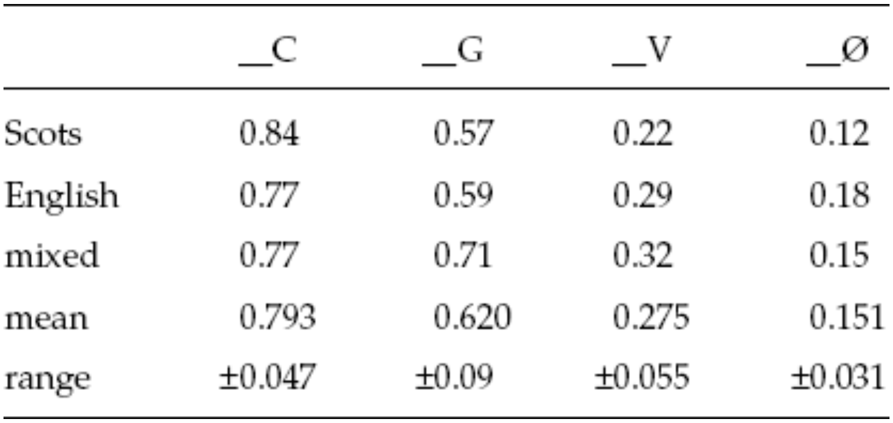
Competence and performance. All the above accounts – structuralist, generative, and optimality-theoretic – treat variable prevalence in terms of optionality, as in (2a), while eschewing the quantification in (2b). In these schools of thought, the "optionality" approach is considered empirically adequate. For the structuralists, it was typically adequate to account for all the structural patterns in a corpus; hence, a list of options was sufficient. An adequate generative grammar, according to **Chomsky (1965)**, must be able to generate all and only the possible grammatical utterances of a language (or, more properly, of an idealized homogeneous idiolect), and this criterion is largely maintained in OT. In all these frameworks the theory and grammar are responsible only for accounting for the *existence* of a possible form, but not for a more precise account of its likelihood or frequency of occurrence.

The formal models we have considered – the sequence of frameworks that runs from Saussure through the structuralists and generativists to OT – thus say nothing about whether a form is common or rare, preferred or exceptional. In fact, they mostly define such facts as lying outside the purview of grammar and formal linguistics. The limited empirical responsibility of the grammar is seen as a theoretical necessity in these models, as a consequence of a set of assumptions about the organization of language. They claim a fundamental distinction between the system of language (which defines grammaticality, possible structures, etc.), and the usage of that system and the utterances it generates. Linguistic theory, and the grammars that encapsulate linguistic knowledge, are concerned with the former, termed *langue*, competence, or i-language, according to the terminology of the day. The usage speakers make of the system, their productions and utterances, is treated as a separate phenomenon, termed *parole*, performance, or e-language. Performance and production are argued to bear an uncertain relationship to the system, subject to non-linguistic constraints such as errors, interruptions, or memory lapses. And the system says nothing about prevalence. Grammars are postulated to be essentially non-quantitative; in one formulation, "grammars can't count." Properly speaking, such models are weakly quantified: the grammar permits the specification of at most three levels of quantification: "always" (anything obligatory), "never" (anything the grammar doesn't generate), and "sometimes" (anything optional or occurring in free variation). But any more quantitative detail than this is, by definition, a consequence of usage, and hence need not and cannot be modeled by the grammar. Statement (2a) is a possible element of competence, but (2b) is, by this definition, a statement about performance.

The position just described lies on one side of a major fault-line in phonological thought. On the other side lie several theoretical frameworks that take a more expanded view of the empirical responsibilities of grammar, and a different view of the capabilities of grammar. As we see in the next section, these frameworks undertake to quantify prevalence, and admit statements like (2b) as elements of grammar.

Quantified prevalence. Opposed to the models just described are theories in which phonological analysis explicitly engages with the quantitative facts of variable prevalence. Some examples of these are the "quantitative paradigm" arising from sociolinguistic research (exemplified by the variable rule model; **Labov 1969**; **Cedergren and Sankoff 1974**), the variable OT models associated with scholars such as **Anttila (1997, 2009)** and **Nagy and Reynolds (1997)**, the Stochastic OT model (**Boersma and Hayes 2001**), and "usage-based" models such as Exemplar Theory (**Bybee 2001**; **Pierrehumbert 2001, 2006)**. The point of departure for these frameworks lies in the empirical evidence demonstrating that many variants have distinctive quantitative tendencies: some forms are recognizably rare, while others are common, frequent, even highly preferred. For example, in American English, final voiceless stops are rarely aspirated, while in other dialects, such as Irish English, aspiration is common (**Kirke 2005**). Speakers appear to be aware of these quantitative facts in the sense that they faithfully reproduce the aspiration rates typical of their speech communities, and are capable of recognizing speech that shows a different rate as distinctive.

Empirical studies of variation massively document the systematic nature of such quantitative patterns (**Labov 1966, 1969**; **Cedergren 1973**; **Poplack 1979**; **Guy 1981**). Every language has phonological variables that systematically occur at certain frequencies. This is the quantitative form of structure, and the quantified models we will consider seek to account for this structure. The typical kinds of patterning and systematicity are illustrated in **Table 92.1**. These data are drawn from a study of CSD in early New Zealand English (conducted by the author with Jen Hay and Abby Walker using the ONZE corpus; cf. **Guy *et al.* 2008**; **Hay *et al.* 2008**). As many studies have demonstrated, coronal stop deletion rates are sensitive to following context, here classified as consonants, glides, vowels, or zero (i.e. utterance-final position). The study sample included speakers from several national backgrounds; the table separates those with Scottish backgrounds, English backgrounds, and mixed backgrounds (including Scottish, English, and/or Australian parents and settlement histories).



	__C	__G	__V	__Ø
Scots	0.84	0.57	0.22	0.12
English	0.77	0.59	0.29	0.18
mixed	0.77	0.71	0.32	0.15
mean	0.793	0.620	0.275	0.151
range	±0.047	±0.09	±0.055	±0.031

Table 92.1 Probabilities of coronal stop deletion in Early New Zealand English, by following context and national background of speakers

These data show several highly regular patterns. First, they are not random. Randomness does occur in linguistic production: some phonetic variability in articulation, such as the scatter observed in vowel articulations, is a consequence of random variability in physical gestures; random ordering of constraints is postulated to be a basic feature of the selection mechanism in variable and Stochastic OT models. But random is not the same as non-categorical (**CHAPTER 89**: GRADIENCE AND CATEGORICALITY IN PHONOLOGICAL THEORY). The statistical meaning of randomness is that all possible outcomes are equally likely: when there are two alternants, each should occur 50 percent of the time, like coin flips. Since the possibilities in this case are deletion or non-deletion, a grammar that randomly generated forms would yield deletion probabilities of 0.5. Second, the results show a systematic effect of following context: consonantal contexts promote deletion (glides somewhat less than obstruents), while vowels disfavor deletion and null contexts are associated with the least deletion of all. Third, these contextual effects are regularly observed in all the speaker groups. The rank order of favorability to deletion is C > G > V > Ø for all three groups, and the actual numerical values for each context fall into non-overlapping ranges, tightly clustered around a mean. Each of these speaker groups, of course, constitutes an independent experiment, being made up of separate individuals who had little or no lifetime contact with the other speakers. If they were behaving randomly, it would be essentially impossible for them to converge on common values.

Such findings, echoed repeatedly in studies of variation, are a principal motive for the quantified models considered here. Given quantitative properties of language that are non-random, systematic, and linguistically conditioned, these theories all seek to account for them by grammatical means, which implies statements of prevalence along the lines of (2b). However, the various quantified theories differ substantially in their assumptions and formal models; in the rest of this section we consider the principal models in turn.

Variable rule model. The earliest formally quantified approach was the "variable rule" model (VR), developed by **Labov (1969)** and **Cedergren and Sankoff (1974)**. This model was conceived as a straightforward extension of generative grammar, in which the rules are all quantified by probability of occurrence. "Obligatory" rules receive the same quantitative treatment as in conventional models: they have a probability of 1. But the distinctiveness of VR lies in the treatment of "optional" processes, here termed variable processes: for these, the probability can be any real number between 0 and 1. This permits a VR model to make specific quantitative predictions for any phonological variable. The values in **Table 92.1** are taken from such an analysis, in which the likelihood of deletion was related to following context; VR incorporates a treatment of contextual constraints on a process. Context-sensitivity is of course an essential feature of an adequate account of variability, just as it is essential for adequate accounts of invariant outcomes, as we saw in the Germanic sound-shift case. Contextual limits on generality are discussed further in §4, below.

The probabilistic quantification of VR illustrated in **Table 92.1** lends the model all the interpretive significance of real numbers. First, different numbers mean differing effects. A model that only defines options could only say that deletion is optional in all four contexts, but a quantitative model shows relationships of more and less: in this case, that consonants and glides favor deletion most, followed by vowels and then by zero. Second, difference is scalar, distinguishing proximal values from distant values; here following glides are closer in effect to consonants than they are to vowels: comparing means for C (0.793), G (0.620), and V (0.275), the distance from G to C (0.173) is just half of the distance from G to V (0.345). Third, the model is quantitatively comparable and falsifiable: we compared across the speaker groups and found that they all had similar values for these contexts. We could look at other speakers to see if they have similar values, or look at lexical or morphological subsets of the data to see if the context effects remain constant. If they do not, we might explore other predictors of these differences. Finally, precise quantification lets us use the standard apparatus of statistics: tests of significance, central tendencies, dispersion, etc. For example, the values reported in **Table 92.1** are significant predictors at the 0.05 level or better.

An important characteristic of VR is that it uses a multivariate analysis to partial out the effects of various predictors or contexts. This has an important practical consequence: it controls for differing distributions of the data across the independent variables. In CSD, for example, the morphology of a word has a substantial effect on deletion rates: past tense forms like *missed, packed* are deleted less than monomorphemes like *mist, pact*. This factor was controlled for in the analysis shown in **Table 92.1**. But an analysis that failed to do this could yield numbers like those in **Table 92.1** purely as an epiphenomenon. Thus if most of the words with following vowels happened to be verbs (in phrases like *messed up, baked it*), while most words with following consonants were monomorphemes (in phrases like *best friend, old man*), a univariate analysis would be open to doubt: do monomorphemes have high deletion rates because they usually occur before consonants, or do following consonants show high deletion because they are most often preceded by monomorphemes? Multivariate analysis, used in VR (and in Exemplar Theory studies; **Jannedy and Hay 2006**), routinely controls for such problems.

As noted above, the "variable rule" model was originally designed for a rule-based grammar, but it is not wedded to such a formalism. The conceptual framework of VR treats linguistic production as set of choices in which each alternative is associated with a probability. The model is agnostic with respect to whether those choices are modeled as rules, constraint orders, branching graphs (as in systemic functional grammar, cf. **Halliday 1985**), selections among allophonic inventories (**CHAPTER 11**: THE PHONEME), or other formal devices.

Stochastic OT. A prominent quantified model in a constraint-based framework is Stochastic Optimality Theory

(**Boersma and Hayes 2001**). As we have noted, conventional OT can generate optional outcomes by means of variable constraint ordering, but says nothing about relative frequencies of particular constraint orders, nor about relative frequencies of occurrence of alternative outputs. This arises from the fact that constraint ranking in conventional OT is purely ordinal: one can say that A outranks B, but not by how much. There is no concept of A being slightly higher than B, or a lot higher. Stochastic OT adds the concept of proximity. Ranking values in Stochastic OT are real numbers, not sequential positions – values like 92.7, 94, and 101.3, not A >> B >> C. An analogy can be drawn with the results of a horse race: conventional OT tells us which horse finished first, second, third, etc., but doesn't say if the first place finisher won by a nose or a mile. Stochastic OT, however, metaphorically gives the finishing times, from which we can deduce whether A beat B by a tenth of a second, or by several minutes. This concept of proximity is then translated into a likelihood of ranking orders being reversed: if the race is re-run with a certain amount of randomness affecting the outcomes, horses or constraints that are separated by very little might well end up in a reverse order, but a horse that trails by a mile is unlikely to ever make up the distance and come out ahead.

Formally, Stochastic OT postulates that each constraint varies randomly within a probability envelope around its central ranking value whenever it is called upon to evaluate a form; at some points in this range it may outrank other nearby constraints, but at other points it will fall below them in ranking value. By suitably adjusting the ranking values of constraints, this model can match frequency distributions observed in natural data. In the English CSD example, a speaker who deletes 50 percent of the time gets equal ranking values for the markedness constraint (*Complexcoda) and the faithfulness constraint (Faith), so their order of dominance is completely random. But a speaker who deletes at a low rate like 10 percent would have a ranking value for Faith that is appreciably higher than the value for *ComplexCoda, and the distance between them is such that their probability envelopes overlap in just 10 percent of their total area, representing the region in which the markedness constraint prevails, selecting for deletion.

Variable OT. An interesting variant of Optimality Theory that generates quantitative predictions without an explicit quantitative apparatus within the grammar is found in the works of scholars such as **Anttila (1997, 2007)** and **Nagy (1996**; see also **Nagy and Reynolds 1997**). These approaches treat quantitative patterning in the data as a structural consequence of variable constraint ordering. As noted above, two randomly ordered constraints would, in the long run, occur in the order A >> B at 50 percent of the time, and B >> A the remaining 50 percent. If these orders select different candidate outputs, a random model predicts the two variant realizations should occur in a 50: 50 ratio. Anttila, Nagy, and their associates have extended this simple observation to larger sets of randomly or partially ordered constraints, sometimes with remarkable results.

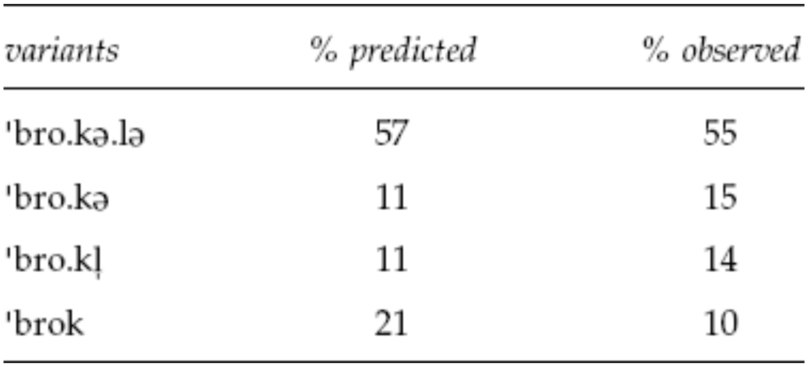
The basic device of these approaches depends on the identification of a number of interacting constraints which, if randomly ordered, would generate the appropriate proportions of observed variants. These proportions depend crucially on the number of relevant constraints. Two constraints have just two possible orders, but the number of possible orders (O) increases as the factorial of the number (*n*) of constraints (O = *n*!). Thus three constraints have six possible orders (ABC, ACB, BAC, BCA, CAB, CBA), four constraints have 24 possible orders, and so on. If these orders occur at random, the variants they select should occur in the ratios of the number of orders that select them.

This can be illustrated with an extension of our simplified example for CSD. If, in addition to the Faith and *complexCoda constraints, the more general markedness constraint NoCoda (disfavoring all coda consonants) were included in the variably ordered set, there would be six possible orders, of which only two – those with Faith highest ranked – would select an undeleted form like *best* or *old*. Hence such a system would generate full forms one-third of the time, and deleted forms two-thirds.

The works mentioned above differ in the details of the model. Anttila mostly relies on fully random subsets of constraints, as in the example just presented, while Nagy and Reynolds work with orders that are fixed for some constraints while others “float” within a certain range. This is illustrated in their study of variable word-truncation in Faetar (a Franco-Provençal language in Southern Italy). A word like *brokele* ‘fork’ has alternate realizations ['brokələ] ˜ ['brokḷ] ˜ ['brokə] ˜ [brok] (plus a few other rarer forms not considered here). Nagy and Reynolds treat this variation as a consequence of variable ordering of the floating constraint Align-prWd with respect to the members of a specified set of constraints; the ones relevant to this word are the ranked set *Coda >> Parse >>

{HNuc, *Schwa} (HNuc = HarmonicNucleus, disfavoring syllabic consonants). When Align-PrWd (which aligns the right edges of the stressed syllable and the prosodic word) is highest ranked, the form [brok] is selected, because the post-tonic syllables in the other alternatives violate Align-PrWd. When Align-PrWd ranks below *Coda, either ['brokl̩] or ['brokə] is selected; the choice between these two depends on the relative order of HNuc and *Schwa (disfavoring reduced vowels), which are variably ordered with respect to each other. Finally, when Align-PrWd falls below Parse (which requires segments in the input to be maintained on the surface), the full form ['brokələ] surfaces.

Counting up all possible tableaux resulting from the variable orders of the constraints they consider, Nagy and Reynolds find 28 possible orders, of which 16 (57 percent) select ['brokələ], 6 (21 percent) select ['brok], while ['brokə] and ['brokl̩] are selected by 3 tableaux (11 percent) apiece. These are the percentages that Nagy and Reynolds's model predicts should be observed in actual data, assuming they have correctly identified the right constraints and orderings. **Table 92.2** shows that the data they observed match these predictions fairly closely.



variants	*% predicted*	*% observed*
'bro.kə.lə	57	55
'bro.kə	11	15
'bro.kl̩	11	14
'brok	21	10

Table 92.2 Variable OT model of word-truncation in Faetar (from **Nagy and Reynolds 1997**)

As in this example, variable OT studies have successfully modeled the empirical distribution of phonological variants in several cases (e.g. **Anttila 1997, 2009** on case inflections and vocalic phonotactics in Finnish). But the strict linkage they require between number of varying constraints and predicted frequencies of prevalence raises questions about their general applicability. As we shall see, social variability commonly involves differences in prevalence: class differences, gender differences, even stylistic differences in the same individual are typically realized as higher or lower overall rates of use of some form. For example, for the English (-*ing*) variable, seen in alternations like *running/runnin*', higher rates of the velar variant (-*ing*) are found in more careful speech styles and in higher status speakers and females. But studies rarely report any differences in constraints among these social groups. To generate quantitative differences, variable OT models would require different sets of variably ranked constraints for women and men, for different social classes, and even for a speaker's casual and careful speech styles. Nagy and Reynolds suggest that this is true for some of their speakers:

> women, particularly younger women [favor] full forms of the words. For those speakers, Align-prWd floats at the lower end of its domain, below Parse, so that the optimal candidate more frequently has all its segments surface … In the grammar of the males and oldest women … Align-PrWd has a greater tendency to float at the higher end of its domain, above Parse. (1997: 47)

But such results are atypical in the sociolinguistic literature on variation; indeed, the empirical evidence suggests that speakers in a community tend to share important grammatical properties like constraint rankings (cf. **Labov 1969**; **Guy 1980**).

Exemplar Theory. The usage-based models that have achieved prominence in recent years begin with a very different set of assumptions about mental representations and grammatical processes (**Johnson 1997**; **Bybee**

2001, 2002; **Pierrehumbert 2001, 2006**; **Hay and Sudbury 2005**; see also **CHAPTER 1**: UNDERLYING REPRESENTATIONS). Instead of the conventional abstract mental representations of words and speech sounds, these approaches postulate that lexical and phonological units are stored in the mind as concrete memories ("exemplars") of the tokens that a speaker has previously encountered. This set of memories is potentially vast – in principle, the exemplar set for a given word may include all the utterances of that word that one has heard in one's lifetime.

Consequently, variability is directly represented in memory. If a speaker has heard a sound or a word pronounced in variant forms, those exemplars are available. Hence the exemplar set provides each speaker with direct knowledge of variation, of the quantitative ratios at which each variant occurs, and of the contextual facts about which contexts favor or disfavor a variant. Each speaker therefore "knows" precisely the values of parameters like those in (2b), quantifying prevalence. (They also know all the details of scope – social, contextual, and lexical, but these are treated in subsequent sections.)

This knowledge is used in production: when speakers compose utterances, they select production targets from the relevant exemplar clouds. Since these targets vary in the same proportions that the speaker has encountered in the input, the speaker faithfully reproduces this variability in production (subject to limitations by certain other factors discussed in §5). Given so rich a set of mental representations, Exemplar Theory relies very little on the abstract processes that are so prominent in other schools of phonology. Indeed, some versions of the theory deny that such processes exist. Consequently, this framework is little concerned with many of the theoretical debates we have mentioned, such as the domain of linguistic description, competence *vs.* performance, etc.

3 The social limits of generality: Sociolinguistic variation

The most obvious limitation on generality in phonology is that so many phonological phenomena are language-specific. Most of the phonology of Chinese has little resemblance to a phonological account of English, Arabic, or Seneca. The features that differ among these languages are thus, in effect, variable elements of universal grammar: different languages vary in terms of whether they have lexical tone, triliteral roots, nasal vowels, and voicing contrasts in the stops. Linguists often take it for granted that the domain of phonological analysis and generalization is a language, but a moment's reflection shows that there are substantive issues at stake here. A "language" is not necessarily a well-defined domain of description. Some things commonly considered languages, like English and Arabic, clearly encompass a wide range of dialects, social varieties, registers, and speech styles, and the differences between any two language varieties can occupy any point on a continuum from near-identity to complete disparity. Some varieties commonly treated as different languages, like Serbian and Croatian, occupy very similar points on such a continuum. Hence phonology faces a systematic issue of how to define the social limits of a generalization. The universal statement must be modified as follows:

(3) *Quantifying social scope*
For speakers in some social domain *i* . . .

(At this level of abstraction we leave open the question of whether the domain *i* defines a language, dialect, speech community, idiolect, speech style, social class, ethnic group, etc., but we return to this matter below.)

This issue has often gone unaddressed by phonologists. When a phonological analysis refers to the social limits of its range of applicability, these are often vaguely defined or deliberately restricted. Some approaches that are commonly encountered are: (i) informal definition of the social domain, using popular labels for languages ("The sound pattern of English"), localities or dialects ("Juchitán Zapotec") or social identities ("upper middle class white speech"); (ii) description of a defined social set ("my informants," "my idiolect"); (iii) definition by linguistic means, such as shared intuitions (speakers who judge a given form to be grammatical). This approach is of course tautological – the grammar accounts for the speakers who use the grammar.

Such approaches leave open questions. Vague definitions make it difficult to identify which speakers are included or excluded; narrowly restricted ones leave it unclear what relevance the analysis has for speakers outside the limits (How might your idiolect differ from mine? How representative of a broader social universe are the speakers who were studied, or who shared a given intuition?) Of course, in some theoretical schools these are not considered linguistic questions. Vague or narrow social scopes may serve a theoretical end: they externalize diversity, facilitating a more homogeneous description.

These issues reflect an unresolved theoretical debate about the domain of linguistic description. In the generative tradition, this domain is narrowly focused on the concept of a mental grammar containing the knowledge required to produce grammatical utterances. Each speaker possesses such a mental grammar, acquired through the interaction between innate capacities (the language faculty or universal grammar) and individual experience. Since the experiences of each individual are unique, it follows that each individual can potentially possess a unique grammar. In this tradition, the scope of a grammar is the speech of the individual who possesses it, the “idiolect.” Social variability is therefore extragrammatical, reflecting only the variability between individuals in linguistic experience. The theory and methods of linguistics per se have little to say about sociolinguistic variation.

The principal alternative to the idiolect treats language as a social construct, with an existence independent of any given individual; indeed, the knowledge of the language that exists in a community of speakers is more comprehensive and arguably more systematic than the partial subset available to a single individual (**Labov 1966**). Linguistic description therefore takes some larger social entity as its object: a language, dialect, or speech community. Some level of social variability therefore lies within the domain of the grammar of the speech community; indeed, some knowledge of social variation in the community is possessed by each individual, and utilized in commonplace linguistic activities such as accommodation to context and interlocutors.

This social focus has deep roots in linguistic theory. Saussure sees linguistics as principally the science of *langue* – an object that characterizes the systematic and general properties of a language as a whole, rather than of *parole* – in which individual and idiosyncratic properties are located. Similarly, the methodological strictures of American structuralism typically define the appropriate domain of linguistic description as a corpus of observations, which could as easily be drawn from a speech community as from a single informant. Since the 1960s this position has been especially associated with sociolinguistic research (see for example, the discussion of “communicative competence” in **Hymes 1972**). It is from this tradition that most of what is known about social variability has emerged. Two important theoretical issues arise from this work: the nature of linguistic similarity and difference, and the content of grammar.

A central finding of the research on social variability in phonology is what **Weinreich *et al.* (1968)** term “orderly heterogeneity”: social diversity in language use is neither random nor highly idiosyncratic; rather, it shows great consistency and order. The simplest summary of this orderliness is that social proximity correlates with linguistic similarity: you talk like the people you talk to.

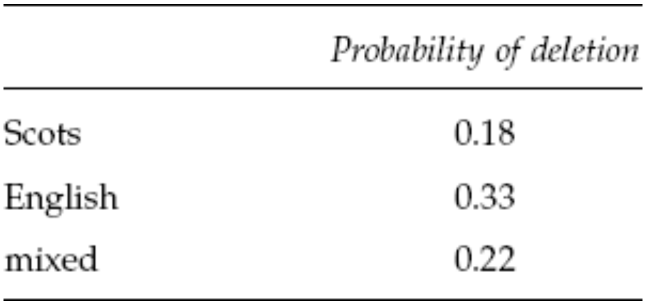
The theoretical problem this presents for phonology is to define what it means for the usage of one speaker to be “like” another. Formal phonology has a good account of identity and non-identity; having the “same” grammar means being identical in all respects, while any differences in what speakers accept as grammatical means that they are “not the same” in grammar. But the theory lacks an account of linguistic similarity. My phonology is nearly identical to my brother's, very similar to those of the friends I grew up with, broadly like other speakers of American English whose ethnic and social backgrounds are comparable to my own, quite different from but recognizably related to that of speakers of Australian English, and drastically different from that of a speaker of Vietnamese. How does phonology capture this scale?

Finally, studies of social variability in language shed important light on the content of grammar – what level of diversity is one grammar capable of modeling?

It seems uncontroversial that Kimbundu and Portuguese should be modeled by two different grammars, and that bilingual Angolans who speak both are therefore possessed of two mental grammars. But what about stylistic variation in one language in an individual: does the switch from a casual style, chatting with one's friends in a bar, to careful style, being interviewed for a job, involve different grammars, or modest adjustments to some stylistic parameters within a single grammar?

Research on social variability in language has yielded some answers to these questions. Consider for example the above-mentioned study of final coronal stop deletion in early New Zealand English. As noted, this corpus includes speakers from several national backgrounds. **Table 92.1** showed the three nationality groups had very similar effects of following context on deletion. But other aspects of their usage were not so similar.

The first is difference in prevalence: the groups deleted at different rates. **Table 92.3** shows overall rates of deletion by nationality. The differences among the groups are significant. The English delete the most, Scots the least, and the mixed group falls in between.

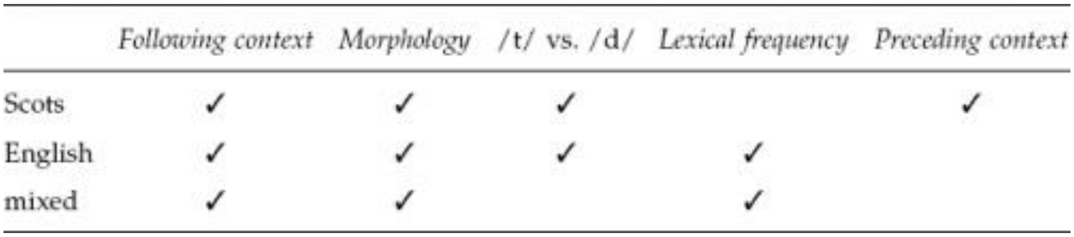


	Probability of deletion
Scots	0.18
English	0.33
mixed	0.22

Table 92.3 Overall rates of coronal stop deletion in early New Zealand English, by national background of speakers

These results lead to an immediate conclusion: prevalence and constraint effects are independent and orthogonal. The nationality groups are significantly different in prevalence of deletion, but nearly identical in following context effects. This is a result that shows up repeatedly in studies of social variation. For example, speakers of different social classes in a speech community may have markedly different levels of prevalence of a socially significant variant, but show the same constraints: thus **Labov's (1966)** New Yorkers varied wildly in rates of rhotic articulations, but all produced more /r/ in final than in internal position. And speakers vary their styles by adjusting prevalence, using more or less of prestige variants, but without changing constraint effects.

This leaves the question of whether and when contextual effects can differ. In the New Zealand study, the nationality groups showed some differences with respect to several other linguistic contexts, as can be seen in **Table 92.4**.



	Following context	*Morphology*	*/t/ vs. /d/*	*Lexical frequency*	*Preceding context*
Scots	✓	✓	✓		✓
English	✓	✓	✓	✓	
mixed	✓	✓		✓	

Table 92.4 Significant constraints on coronal stop deletion in early New Zealand English, by national background of speakers

Beside following context, all three groups showed a significant effect of morphology, but only two showed an effect of whether the target stop was /t/ or /d/, a different two showed a lexical frequency effect, and only one showed a preceding context effect.

These results demonstrate that different speech communities often differ in constraint effects. A variety of research results show the same finding: the following zero effect on CSD, although constant for the three national groups in **Table 92.1**, differs significantly in many speech communities (**Guy 1980**); the contexts for tensing of /æ/ vary substantially in American English dialects (**Labov *et al.* 2006**; **Labov 2007**); the contexts affecting subject pronoun expression vary among Spanish dialects (**Cameron 1993**; **Otheguy *et al.* 2007**); metrical constraints on Portuguese /-s/ deletion vary among speech communities in Brazil (**Guy 2002**). However, these and other studies show that, within speech communities, speakers are mostly similar or identical in contextual effects.

Such findings provide an answer to the question posed above about the content of grammar. Grammatical similarity is measured by shared constraints, not by prevalence; within a grammar, prevalence may vary, but not contextual constraints. Speakers within a speech community share a grammar; therefore, stylistic and social class differences in a community consist of differences in prevalence, but not in constraints. Differences in constraint effects imply different grammars, and different speech communities.

What are the implications for phonological theory? The unquantified models considered above, such as generative phonology and conventional OT, do not engage with such facts. Among the quantified models, these results are naturally accommodated in the VR model, which makes a basic distinction between contextual constraints and prevalence, expressed by an overall probability associated with each rule or grammatical choice-point. They are also easily modeled in Exemplar Theory, since all relevant data on both prevalence and context are stored in each speaker's memory to guide their production; indeed, **Foulkes and Docherty (2006)** argue that an exemplar account "offers the most productive means of modeling sociophonetic variation." But such results are awkward for OT-based models, because they lack an independent representation of prevalence.

This difficulty can be illustrated with the data on overall prevalence in **Table 92.3**. In a variable OT model like **Anttila's (1997)**, the differences in deletion rates between the three NZ speaker groups can be modeled only by postulating different sets of variably ordered constraints. The 0.33 deletion probability for the English group could be generated by three constraints (P, Q, R) that select for deletion only when P is ranked highest. But to approximate the deletion rate of 0.22 for the mixed nationality group would require adding an additional constraint S to the variably ordered mix, which also prevents deletion if it outranks P. Then the model predicts deletion in the 25 percent of cases in which P was highest ranked of these four. And modeling the 0.18 rate for the Scots group would require still another mix of five or six constraints. This approach thus implies substantively different grammars for any differences in surface ratios, making no distinction between constant grammar and varying prevalence. A Stochastic OT model has a somewhat easier time, since it can predict different prevalence levels in a constant set of constraints, by varying the distances among them on the ranking scale. But how does Stochastic OT model different prevalence for speech styles or social classes in a community where everybody has the same constraint effects? It is a complex, perhaps impossible, mathematical task to generate constant constraint effects like those illustrated in **Table 92.1**, while simultaneously generating differences in overall prevalence, like those in **Table 92.3**, simply by sliding the same constraints up or down the ranking value scale.

4 The contextual limits of generality: Context-sensitivity and variability

Phonological theory commonly makes a distinction between statements, processes, properties, etc. that are context-free ("all syllables have a nucleus") and context-sensitive ("English vowels are ordinarily oral, but may be nasalized before a nasal consonant"). In the latter case, context-sensitive operations involve variability: English vowels vary in their realization on the nasal-oral dimension. This implies that the scope of many generalizations must be defined in some statement such as (4):

(4) *Quantifying contextual scope*

. . . in some linguistic context *j* . . .

It is an interesting fact about the discipline that such limitations on contextual scope are a routine part of phonological analysis, while limitations on social or lexical scope are less often examined and are typically seen as problematic. In fact, phonological theory doesn't even consider contextual limitations as variability, provided prevalence is categorical within the context. Following the Grimm/Verner example, a clearly defined contextual scope in which a unique outcome occurs is not a violation of exceptionlessness. The analytical heuristic that seeks to partial out variable outcomes to categorical contexts holds a privileged status in the conceptual armory of phonology. Nevertheless, context-sensitivity raises several issues associated with variability.

As we have seen, contexts are often associated not with categorical prevalence, but with probabilistic effects. Often a linguistic context favors one outcome without precluding others. This is clear in **Table 92.1**: each following context has a distinct quantitative effect on CSD, but none categorically demands or prohibits deletion.

Importantly, the following segment effects in **Table 92.1** have a phonological explanation consistent with a broad spectrum of work on phonotactics and syllable structure. The words that undergo this deletion process end in consonant clusters: … CC#. When they occur in running speech with a following word, sequences like … CCC … (e.g. *west side*), … ccg … (*west wing*) and … CCV … (*west end*) result. Universal principles of markedness, sonority sequencing, etc. all agree that CCC is more marked and less acceptable than CCV (**chapter 49**: sonority; chapter 46: positional effects in consonant clusters). ot constraints like noCoda, *ComplexCoda, and Onset are designed to capture such generalizations. Sometimes these principles have categorical effect: many languages prohibit CCC sequences completely. The quantitative findings in English reflect the same principles, but with probabilistic effect.

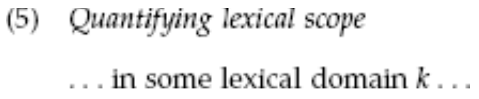
Such results, echoed in numerous studies of variation, have been termed "stochastic generalizations" (**Bresnan *et al.* 2001**; **Clark 2005**): generalizations that are categorically true in one language but are probabilistic constraints in another language or social variety. This constitutes crucial evidence against the more extreme theoretical positions favoring categorical prevalence discussed in §2, such as denying the validity of non-categorical generalizations, and exiling all quantitative facts to the grammatically irrelevant terrain of usage and performance. If phono-tactic markedness is a continuum (**chapter 4**: markedness), generating categorical effects in some social or contextual domains and probabilistic effects in others, phonological theory should treat it as an integrated phenomenon.

These results also shed light on another area of variability mentioned in the introduction: phonetic gradience. The intermediate effect of following glides in **Table 92.1** (promoting deletion more than vowels but less than consonants) reflects an intermediate status for glides on phonetic/phonological scales. Many phono-logists treat this as a sonority scale: glides are more sonorous than obstruents but less sonorous than vowels (**chapter 15**: glides; chapter 49: sonority). Syllable structure and the markedness of consonantal sequences also depend on sonority sequencing: for example, many languages prohibit obstruent-obstruent sequences (CC) but permit obstruent-glide sequences (GC or CG). Viewed in this light, the results of **Table 92.1** reflect gradient sonority effects: deletion is inversely correlated with sonority of following context. Phonetic properties thus parallel the quantitative behavior of generalizations: in some social and linguistic domains they are involved in discrete (categorical) phenomena, but in other domains they have probabilistic effects. (For an insightful discussion of gradient variability in an OT framework, see **Anttila 2008**.)

Finally, as seen in §3, context is also a defining element of linguistic similarity: differences in contextual effects define a difference in grammar, while differences in overall prevalence do not. Members of a speech community converge to a remarkable degree on the contextual limits of phonological generalizations, whether categorical or probabilistic, while they vary considerably in prevalence. Collectively, these results suggest that the typical heuristic of seeking linguistic structure by pursuing Neogrammarianesque categorical contexts is misguided, if the analyst ignores the possibility that such types of contextual conditioning are just one point on a continuum. Context is indeed a fundamental element of linguistic structure, but it includes probabilistic as well as categorical conditioning.

5 The lexical limits on generality: Lexical exceptions

The lexical scope of phonological generalizations has been a recurring topic of debate for over a century. The central question is lexical variability in phonology: does a statement or process apply to all relevant lexical items, or are there words that exceptionally fail to show some generalization (**chapter 106**: exceptionality)? If so, how does the theory account for lexical limits on generality? A generalization that varies across the lexicon, applying to some words, but not all, requires a specification of lexical scope, as in (5):



(5) *Quantifying lexical scope*

. . . in some lexical domain *k* . . .

The debate over the lexical limits of phonology first achieves prominence in the nineteenth century, in connection with the emergence of the Neogrammarians. The Neogrammarian hypothesis specifically asserts lexical universality of phonological change: "exceptionlessness" means no lexical exceptions to a generalization (**Osthoff and Brugmann 1878**). In fact, the Neogrammarians propose a phonemic model that rules out lexical variability (**Paul 1880**): words are constituted as sequences of phoneme-like units; phonologically, they have no independent existence apart from the string of phones of which they are composed. Sound change operates on these phones, so when one of them changes, all words containing them necessarily change as well.

At the same time an anti-Neogrammarian critique emerged, arguing that the historical record actually does show lexical variability (**Schuchardt 1885**). One slogan for this position was "each word has its own history" – a position that affirms lexical variability as forcefully as the Neogrammarians denied it. In the twentieth century, this position re-emerged under the label "lexical diffusion," in work by Wang and his associates (**Wang 1969, 1997**; **Chen and Wang 1975**). These scholars argued against the "lexically abrupt" application of phonological change (i.e. categorical processes applying to all relevant words simultaneously), and in favor of a "lexically gradual" model of phonological process, which spreads across the lexicon word by word, in a manner reminiscent of analogical change (see also **CHAPTER 93**: SOUND CHANGE).

The focus on synchronic phonology in linguistics after Saussure continued to confront lexical generality. As we have noted, the dominant formal theories emphasized the pursuit of invariant generalizations, and largely assumed, with the Neogrammarians, that words are merely assemblages of the phonological units (such as phonemes, feature arrays, autosegments) on which phonological processes operate. Hence these theories give short shrift to lexical limits on generality other than those that can be given a segmental, prosodic, or morphological formulation. They assume that phonological statements normally apply without lexical limits.

But synchronic phonology has witnessed the reappearance of theories that give primacy to the word, and envisage significant lexical variability. Thus Exemplar Theory argues that the word is the primary unit that speakers recognize, remember, and manipulate, and, indeed, that speakers rely on their massive inventories of remembered auditory images (exemplars) of words they have heard as their primary mental database for most of phonology. Phonemes, features, and the like are emergent abstractions or generalizations across those exemplars, rather than primary units of perception and production. For some versions of the theory, abstract phonological operations do not exist; all of phonology is reduced to "phonetic" processes (neuromechanical events such as gestural overlap, gestural weakening, etc.) and "generalizations" that function more like analogies than like phonological operations.

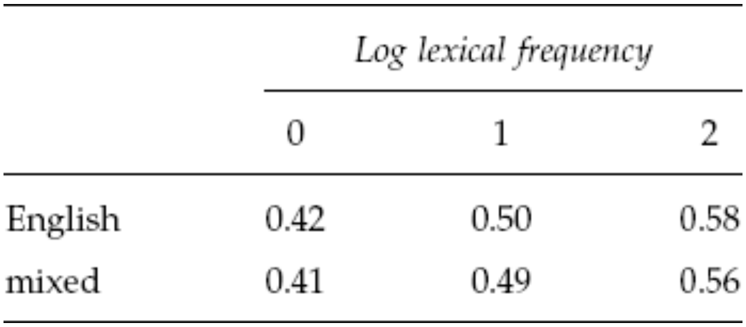
Such a model consequently assumes that lexical variability is a normal state of affairs. Each word has its own phonological identity, and the mental representation of a word, since it incorporates a broad array of remembered exemplars, necessarily includes all the variability that a speaker has encountered hearing people say that word. The theory permits, even predicts, that any statement that might be made about the phonology of a language might have to be lexically qualified word-by-word across the vocabulary.

Empirical evidence. The issues in this debate are partly empirical. Do we observe lexical limitations on phonology? In the historical record, there are quite a few cases that suggest lexical exceptionality. For example, Latin onset clusters containing /l/ changed in several directions in Old Portuguese (**Williams 1938**; **CHAPTER 30**: THE REPRESENTATION OF RHOTICS; CHAPTER 31: LATERAL CONSONANTS). Sometimes /l/ becomes /r/, thus *branco, praia, praça, fraco, cravo, regra* (compare cognates in Spanish, which lacks this change: *blanco, playa, plaza, flaco, clavo, regla*). Other words show the entire cluster changing to the palatal fricative /ʃ/: *clavem* > *chave, plenum* > *cheio, flammam* > *chama.* And in numerous cases Portuguese retains the historical /l/: *flor, flamengo, claro, classe, planta*. There are no obvious contexts that predict one or the other outcome, nor do the usual Neogrammarian tactics for resolving lexical effects appear to offer solutions: the exceptional cases do not obviously arise from dialect borrowing, paradigm leveling, neologisms or classicisms, etc. But it should be emphasized that history records many cases where Neogrammarian regularity does prevail. The Grimm's Law changes in Germanic included a shift from IE voiced aspirates ([bh dh gh]) to Germanic simple voiced stops ([b d g]). This change left no lexical residue whatsoever in the Germanic languages, no words that retain a voiced aspirate. Similarly the loss of the voiced velar fricative in Middle English left behind not a single lexical item in which it was preserved, despite the fact that this sound continues to be spelled in English orthography as <gh>, in words like *cough, though, night*. The empirical evidence thus indicates that both patterns occur, although the literature suggests there are far more cases of regular sound change than of lexical diffusion.

In synchronic phonology the situation is similar: the literature attests many processes that apply without lexical limitations, but a substantial number of cases have been reported showing lexical limits; some of these involve just a few words that are exceptions to a general pattern, while others define substantial subsets of the lexicon that show distinctive phonology, such as the Chinese-origin loanwords in Japanese (**CHAPTER 95**: LOANWORD PHONOLOGY). These cases of lexical variability have attracted considerable attention in phonological theory.

Theoretical solutions. For a survey of theoretical thought on lexical exceptionality, readers are referred to **CHAPTER 106**: EXCEPTIONALITY. The principal approaches fall into two camps: lexical strategies, which represent exceptionality in the lexical entries, and phonological strategies, which use the phonological apparatus (features, rules, constraints, etc.) to generate distinctive outcomes for exceptional words. Although these are sometimes treated as technical questions about the workings of a theory, they actually raise substantive questions about the workings of the human mind. The lexical approach – encoding exceptionality in underlying representations – effectively assumes that speakers use minimalist, local, memory-based strategies: they simply remember that a word is anomalous. But the phonological approach implies that speakers strive to extract maximum generalities, and are willing to rejig the whole system if there are any efficiencies to be gained. These are matters worthy of direct investigation.

Lexical frequency. The emergence of Exemplar Theory has brought forward additional questions about lexical variability, and stimulated a range of research that has revealed important evidence bearing on lexical variability. Since Exemplar Theory postulates that speakers retain a rich set of memories of utterances of words, they have available a great deal of evidence on which to base their linguistic behavior that other theories ignore. One prominent prediction of the theory is that phonological processes should be sensitive to lexical frequency – how often a speaker encounters a word (**CHAPTER 90**: FREQUENCY EFFECTS). The exemplar set obviously contains frequency information: frequent words have more exemplars. Scholars in this framework have argued that lenition processes should correlate directly with lexical frequency, because lenition is an articulatory process favored by repetition. This prediction has been confirmed in a range of research in the ET framework. For example, in the New Zealand study previously cited, significant frequency effects were found for CSD (**Guy *et al.* 2008**). Some relevant findings appear in **Table 92.5**.



	Log lexical frequency		
	0	1	2
English	0.42	0.50	0.58
mixed	0.41	0.49	0.56

Table 92.5 Lexical frequency effect on coronal stop deletion in early New Zealand English

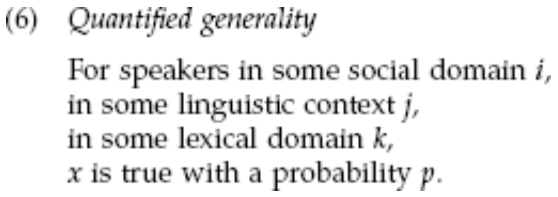
For both groups shown, increasing lexical frequency (measured here by the logarithm of the number of occurrences of the item in the ONZE database) is associated with increasing rates of deletion, consistent with the predictions of **Bybee (2001, 2002)**. Although this effect is treated discretely here, by dividing the frequency range into blocks, **Guy *et al.* (2008)** demonstrate a continuous correlation.

Lexical frequency effects have been investigated in a variety of studies in usage-based models and elsewhere (e.g. **Phillips 1984**). Many find significant effects, but the evidence is mixed: some studies show no frequency effects (e.g. the Scots nationality group in the NZ study showed no frequency effect). But the theoretical implications of frequency are extensive. Such facts are difficult to accommodate in any model that assumes economical (i.e. impoverished) lexical representation – in other words, in most theoretical currents from the Neo-grammarians to Optimality Theory. Traditional abstract lexical representations provide no place to record how often and in what form the word has been heard, activated, or articulated. If the lexical frequency of items turns out to regularly affect their phonological treatment, richer representations that can incorporate frequency information will be required.

6 Conclusions

One historical task of phonology reflects the perceptual task that hearers face when listening to a speech signal: the signal contains many components, from which the meaningful elements need to be extracted, and the noise and other non-linguistic elements removed. Consequently, phonologists have sought to maximize generality and predictability, to find constancies and regularities, to tell the signal from the noise. Variability in its various forms presents a challenge to this undertaking – irregularities and limitations on generality. Not unreasonably, one response to variability has therefore been to exclude or minimize it, to treat it as part of the noise.

But in taking this approach, phonology runs the risk of reifying the task as the object. Since our method involves seeking generalities, we risk assuming that generalities are all there is to be sought, that phonology is exceptionless, always regular, invariant. This is a perilous course: if we predefine what we seek, we may not be able to find anything else. If we look only for forests, we may not see the trees. A careful scrutiny of the empirical evidence shows that language does contain “irregularities” that are not noise, variability that is generated by and part of the linguistic system. All the schools of thought we have considered recognize this at some level, at least by acknowledging optionality and contextual variability, and some level of lexical variability. Some schools additionally seek to model social variability and quantitative aspects of variable prevalence. A comprehensive account, which departs from the ideal generality – the phonological universal – in each relevant dimension, will have the following form:



(6) *Quantified generality*

For speakers in some social domain *i*,
in some linguistic context *j*,
in some lexical domain *k*,
x is true with a probability *p*.

The evidence so far available further suggests that the value of *p* will be some function of *i, j*, and *k*. Phonology has made great progress on the task of seeking maximal generality, even universals. It is now beginning to confront the task of exploring the limits of generality, and the linguistic uses of variability.

REFERENCES

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

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

Anttila, Arto. 2008. Gradient phonotactics and the Complexity Hypothesis. *Natural Language and Linguistic Theory* (26) . 695–729.

Anttila, Arto. 2009. Derived environment effects in colloquial Helsinki Finnish. In Kristin Hanson & Sharon Inkelas (eds.) *The nature of the word: Essays in honor of Paul Kiparsky*, 433–460. Cambridge, MA: MIT Press.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* (32) . 45–86.

Bresnan, Joan, Shipra Dingare & Christopher Manning. 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In Miriam Butt & Tracy Holloway King (eds.) *Proceedings of the LFG 01 Conference, University of Hong Kong, online proceedings*. Stanford: CSLI. Available (July 2010) at **http://csli-publications.stanford.edu/LFG/6/lfg01.pdf**.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* (14) . 261–290.

Cameron, Richard. 1993. Ambiguous agreement, functional compensation, and nonspecific *tú* in the Spanish of San Juan, Puerto Rico, and Madrid, Spain. *Language Variation and Change* (5) . 305–334.

Cedergren, Henrietta J. 1973. *The interplay of social and linguistic factors in Panama*. Ph.D. dissertation, Cornell University.

Cedergren, Henrietta J. & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* (50) . 333–355.

Chen, Matthew & William S.-Y. Wang. 1975. Sound change: Actuation and implementation. *Language* (51) . 255–281.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.

Clark, Brady. 2005. On stochastic grammar. *Language* (81) . 207–217.

Foulkes, Paul & Gerard J. Docherty. 2006. The social life of phonetics and phonology. *Journal of Phonetics* (34) . 409–438.

Grimm, Jacob. 1819. *Deutsche Grammatik*. Göttingen: Dieterich.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In William, Labov (ed.) *Locating language in time and space*, 1–36. New York: Academic Press.

Guy, Gregory R. 1981. *Linguistic variation in Brazilian Portuguese: Aspects of the phonology, syntax, and language history*. Ph.D. dissertation, University of Pennsylvania.

Guy, Gregory R. 2002. A identidade lingüística da comunidade de fala: Paralelismo interdialetal nos padrões de variação lingüística. *Organon* (14) . 17–32.

Guy, Gregory R., Jennifer Hay & Abby Walker. 2008. *Phonological, lexical and frequency factors in coronal stop deletion in early New Zealand English*. Poster presented at Laboratory Phonology 11, Wellington.

Halliday, M. A. K. 1985. *An introduction to functional grammar*. London: Edward Arnold.

Hay, Jennifer & Andrea Sudbury. 2005. How rhoticity became /r/-sandhi. *Language* (81) . 799–823.

Hay, Jennifer, Margaret A. Maclagan & Elizabeth Gordon. 2008. *New Zealand English*. Edinburgh: Edinburgh University Press.

Hockett, Charles F. 1942. A system of descriptive phonology. *Language* (18) . 3–21.

Hymes, Dell H. 1972. On communicative competence. In J. B. Pride & J. Holmes (eds.) *Sociolinguistics*, 269–293.

Harmondsworth: Penguin.

Jannedy, Stefanie & Jennifer Hay (eds.) 2006. Modelling sociophonetic variation. *Special issue, Journal of Phonetics* (34) . 405–530.

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson & John W. Mullenix (eds.) *Talker variability in speech processing*, 145–165. San Diego: Academic Press.

Kirke, Karen D. 2005. When there's more than one norm–enforcement mechanism: Accommodation and shift among Irish immigrants to New York City. *Penn Working Papers in Linguistics* (11) .

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

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

Labov, William. 2007. Transmission and diffusion. *Language* (83) . 344–387.

Labov, William, Sharon Ash & Charles Boberg. 2006. *The atlas of North American English: Phonetics, phonology and sound change*. Berlin & New York: Mouton de Gruyter.

Nagy, Naomi. 1996. *Language contact and language change in the Faetar speech community*. Ph.D. dissertation, University of Pennsylvania.

Nagy, Naomi & Bill Reynolds. 1997. Optimality Theory and variable word–final deletion in Faetar. *Language Variation and Change* (9) . 37–55.

Osthoff, Hermann & Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Leipzig: Hirzel.

Otheguy, Ricardo, Ana Celia Zentella & David Livert. 2007. Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language* (83) . 770–802.

Paul, Hermann. 1880. *Prinzipien der Sprachgeschichte*. Halle: Niemeyer. English translation published 1889, *Principles of the history of language*. New York: McMillan.

Peterson, Gordon E. & Harold L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* (24) . 175–184.

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

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan Bybee & Paul Hopper (eds.) *Frequency and the emergence of linguistic structure*, 137–157. Amsterdam & Philadelphia: John Benjamins.

Pierrehumbert, Janet B. 2006. The next toolkit. *Journal of Phonetics* (34) . 516–530.

Poplack, Shana. 1979. *Function and process in a variable phonology*. Ph.D. dissertation, University of Pennsylvania.

Rask, Rasmus. 1818. *Undersögelse om det gamle Norske eller Islandske Sprogs Oprindelse*. Copenhagen: Gyldendal.

Saussure, Ferdinand de. 1916. *Cours de linguistique générale*. Lausanne and Paris: Payot.

Schuchardt, Hugo. 1885. *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Oppenheim.

Swadesh, Morris. 1934. The phonemic principle. *Language* (10) . 117–129.

Verner, Karl. 1877. An exception to Grimm's Law. Published 1978 in Philip Baldi & Ronald N. Werth (eds.) *Readings in historical phonology: Chapters in the theory of sound change*, 3–22. University Park: The Pennsylvania State University Press. Translation by R. Stanley from *Eine Aushnahme der ersten Lautverschiebung*.

Wang, William S.–Y. 1969. Competing changes as a cause of residue. *Language* (45) . 9–25.

Wang, William S.–Y. (ed.) 1997. *The lexicon in phonological change*. The Hague: Mouton.

Weinreich, Uriel, William Labov & Marvin Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann & Yakov Malkiel (eds.) *Directions for historical linguistics*, 95–195. Austin: University of Texas Press.

Williams, Edwin B. 1938. *From Latin to Portuguese*. Philadelphia: University of Pennsylvania Press.

Cite this article

Guy, Gregory R. "Variability." *The Blackwell Companion to Phonology*. van Oostendorp, Marc, Colin J. Ewen, Elizabeth Hume and Keren Rice (eds). Blackwell Publishing, 2011. Blackwell Reference Online. 14 March 2013 <http://www.companiontophonology.com/subscriber/tocnode.html?id=g9781405184236_chunk_g978140518423694>