21. The Phonological Basis of Sound Change

PAUL KIPARSKY

Subject	Theoretical Linguistics » Pholonogy
DOI:	10.1111/b.9780631201267.1996.00023.x

Tout est psychologique dans la linguistique, y compris ce qui est mécanique et matériel.

–F. de Saussaure 1910/1911

0 Sound Change

The neogrammarians portrayed sound change as an exceptionless, phonetically conditioned process rooted in the mechanism of speech production.[1] This doctrine has been criticized in two mutually incompatible ways. From one side, it has been branded a mere terminological stipulation without empirical consequences, on the grounds that apparent exceptions can always be arbitrarily assigned to the categories of analogy or borrowing.[2] More often though, the neogrammarian doctrine has been considered false on empirical grounds. The former criticism is not hard to answer (Kiparsky 1988), but the second is backed by a formidable body of evidence. Here I will try to formulate an account of sound change making use of ideas from lexical phonology, which accounts for this evidence in a way that is consistent with the neogrammarian position, if not exactly in its original formulation, then at least in its spirit.

The existence of an important class of exceptionless sound changes grounded in natural articulatory processes is not in doubt, of course. It is the claim that it is the *only* kind of sound change that is under question, and the evidence that tells against is primarily of two types. The first is that phonological processes sometimes spread through the lexicon of a language from a core environment by generalization along one or more phonological parameters, often lexical item by lexical item. Although the final outcome of such *lexical diffusion* is in principle indistinguishable from that of neogrammarian sound change, in midcourse it presents a very different picture. Moreover, when interrupted, reversed, or competing with other changes, even its outcome can be different.

Against the implicit assumptions of much of the recent literature, but in harmony with older works such as Schuchardt (1885) and Parodi (1923, p. 56), I will argue that lexical diffusion is not an exceptional type of sound change, nor a new, fourth type of linguistic change, but a well-behaved type of analogical change. Specifically, *lexical diffusion is the analogical generalization of lexical phonological rules*. In the early articles by Wang and his collaborators, it was seen as a process of phonemic redistribution spreading randomly through the vocabulary (Chen and Wang 1975; Cheng and Wang 1977). Subsequent studies of lexical diffusion have supported a more constrained view of the process. They have typically shown a systematic pattern of generalization from a categorical or near-categorical core through extension to new phonological contexts, which are then implemented in the vocabulary on a word-by-word basis. In section 1 I argue that lexical diffusion is driven by the rules of the lexical phonology, and that the mechanism is analogical in just the sense in which, for example, the regularization of *kine* to *cows* is analogical. In fact, the instances of "lexical diffusion" which Wang and his collaborators originally cited in support of their theory include at least one uncontroversial instance of analogical change, namely, the spread of retracted accent in deverbal nouns of the type *tórmènt* (from *tormént*). In most cases, of course, the analogical character of the change is less obvious because the analogy is nonproportional and implements distributional phonological regularities rather than morphological alternations. For example, the item-by-item and dialectally varying accent retraction in nonderived nouns like *mustache, garage, massage, cocaine* is an instance of nonproportional analogy, in the sense that it extends a regular stress pattern of English to new lexical items. What I contend is that genuine instances of "lexical diffusion" (those which are not due to other mechanisms such as dialect mixture) are *all* the result of analogical change. To work out this idea I will invoke some tools from recent phonological theory. In particular, radical underspecification and structure-building rules as postulated in lexical phonology will turn out to be an essential part of the story.

The second major challenge to the neogrammarian hypothesis is subtler, less often addressed, but more far-reaching in its consequences. It is the question how the putatively autonomous, mechanical nature of sound change can be reconciled with the systematicity of synchronic phonological structure. At the very origins of structural phonology lies the following puzzle: if sound changes originate through gradual articulatory shifts which operate blindly without regard for the linguistic system, as the neogrammarians claimed, why don't their combined effects over millennia yield enormous phonological inventories which resist any coherent analysis? Moreover, why does no sound change ever operate in such a way as to subvert phonological principles, such as implicational universals and constraints on phonological systems? For example, every known language has obstruent stops in its phonological inventory, at least some unmarked ones such as *p, t, k*. If sound change were truly blind, then the operation of context-free spirantization processes such as Grimm's Law to languages with minimal stop inventories should result in phonological systems which lack those stops, but such systems are unattested.

With every elaboration of phonological theory, these difficulties with the neogrammarian doctrine become more acute. Structural investigations of historical phonology have compounded the problems. At least since Jakobson (1929), evidence has been accumulating that *sound change itself*, even the exceptionless kind, is structure-dependent in an essential way. Sequences of changes can conspire over long periods, for example to establish and maintain patterns of syllable structure, and to regulate the distribution of features over certain domains. In addition to such top-down effects, recent studies of the typology of natural processes have revealed pervasive structural conditioning of a type hitherto overlooked. In particular, notions like underspecification, and the abstract status of feature specifications as distinctive, redundant, or default, are as important in historical phonology as they are synchronically. The neogrammarian reduction of sound change to articulatory shifts in speech production conflicts with the apparent structure-dependence of the very processes whose exceptionlessness it is designed to explain.

A solution to this contradiction can be found within a two-stage theory of sound change according to which the phonetic variation inherent in speech, which is blind in the neogrammarian sense, is selectively integrated into the linguistic system and passed on to successive generations of speakers through language acquisition (Kiparsky 1988). This model makes sound change simultaneously mechanical on one level (vindicating a version of the neogrammarian position), yet structure-dependent on another (vindicating Jakobson). The seemingly incompatible properties of sound change follow from its dual nature.

My paper is organized as follows. In the next section I present my argument that lexical diffusion is analogical and that its properties can be explained on the basis of underspecification in the framework of lexical phonology. I then spell out an account of sound change which reconciles exceptionlessness with structure-dependence (sec. 2). Finally in section 3 I examine assimilatory sound changes and vowel shifts from this point of view, arguing that they too combine structure-dependence with exceptionlessness in ways which support the proposed model of sound change, as well as constituting additional diachronic evidence for radical underspecification in phonological representations.

1 Lexical Diffusion

1.1 “It walks like analogy, it talks like analogy…”

If lexical diffusion is not sound change, could it be treated as a subtype of one of the other two basic categories of change? Clearly it is quite unlike lexical *borrowing*: it requires no contact with another language or dialect (i.e., it is not reducible to “dialect mixture”), it follows a systemic direction set by the language's own phonological system (it is a species of “drift”), and it involves a change in the pronunciation of existing words rather than the introduction of new ones.

On the other hand, it does behave like lexical *analogy* in every respect, as summarized in the following table.[3]

Table 21.1

	Sound change	Borrowing	Lexical analogy	Lexical diffusion
Generality	Across-the-board	Item by item	Context by Context, item by item	Context by context, item by item
Gradience	Gradient	Quantal	Quantal	Quantal
Origin	Endogenous	Contact	Endogenous	Endogenous
Rate	Rapid	Rapid	Slow	Slow
Effect on:				
Rule system	New rules	No change	Rules generalized	Rules generalized
Sound/phoneme inventory	New inventory	Peripheral	No change	No change
Vocabulary	No change	New words	No change	No change

It seems to be the case that lexical diffusion always involves neutralization rules, or equivalently that lexical diffusion is structure preserving (Kiparsky 1980, p. 412). This has been taken as evidence for locating lexical diffusion in the lexical component of the phonology (Kiparsky 1988). Being a redistribution of phonemes among lexical items, it cannot produce any new sounds or alter the system of phonological contrasts. Its nongradient character follows from this assumption as well, since lexical rules must operate with discrete categorical specifications of features.

An important clue to the identity of the process is its driftlike spread through the lexicon, by which it extends a phonological process context by context, and within each new context item by item. This is of course exactly the behavior we find in many analogical changes. An example of such lexical diffusion is the shortening of English /ū/, which was extended from its core environment (1a), where it was categorical, by relaxing its context both on the left and on the right (Dickerson 1975). In its extended environments it applies in a lexically idiosyncratic manner. The essential pattern is as follows:

(1)

(a) [−anterior] ___ $\begin{bmatrix}-\text{anterior}\\ -\text{coronal}\end{bmatrix}$

cook, hook, shook, rook, brook, crook, hookah (short)

(b) ___ $\begin{bmatrix}-\text{anterior}\\ -\text{coronal}\end{bmatrix}$

took, book, nook, look, forsook, Wookie (short)
snook, snooker, stook, boogie, Sook, gadzooks, spook (variable)
bazooka (long)

(c) [−anterior] ___

good, could, should, hood "covering", hoodwink (short)
roof, rooster, hoodlum, cooper, hoof, room, root, hoodlum, hood "ruffian", coop, proof (variable)
brood, shoot, hoot, behoove, scoop, coon, coot, roost, groove . . . (long)

We can provide a theoretical home for such a mechanism of change if we adopt lexical phonology and combine it with a conception of analogical change as an optimization process which eliminates idiosyncratic complexity from the system – in effect, as grammar simplification.[4] The mechanism that drives such redistribution of phonemes in the lexicon is the system of structure-building rules in the lexical phonology. The direction of the phonemic replacement is determined by the rule, and its actuation is triggered jointly by the generalization of the rule to new contexts, and by the item-by-item simplification of lexical representations in each context. When idiosyncratic feature specifications are eliminated from lexical entries, the features automatically default to the values assigned by the rule system, just as when the special from *kine* is lost from the lexicon the plural of *cow* automatically defaults to *cows*. The fact that in the lexical diffusion case there is no morphological proportion for the analogy need not cause concern, for we must recognize many other kinds of nonproportional analogy anyway.

To spell this out, we will need to look at how unspecified lexical representations combine with structure-building rules to account for distributional regularities in the lexicon. This is the topic of the next section.

1.2 The Idea behind Underspecification

The idea of underspecification is a corollary of the Jakobsonian view of distinctive features as the real ultimate components of speech. All versions of autosegmental phonology adopt it in the form of an assumption that a feature can only be associated with a specific class of segments designated as permissible bearers of it (P-bearing elements), and that such segments may be lexically unassociated with P and acquire an association to P in the course of the phonological derivation. But in phonological discussions the term "underspecification" has come to be associated with two further claims, mostly associated with lexical phonology, namely that the class of P-bearing segments may be extended in the course of derivation, and that lexical (underlying) representations are minimally specified.

How minimal is minimal? There are several alternative versions of under-specification on the market which differ in their answers to this question.[5] The most conservative position, *restricted underspecification*, is simply that redundant features are lexically unspecified. On this view, the feature of voicing in English would be specified for obstruents, where it is contrastive, but not for sonorants, which are redundantly voiced. An entirely nondistinctive feature, such as aspiration in English, would not be specified in lexical representation at all.

Radical underspecification (the version which I will assume later on) carries the asymmetry of feature specifications one step further, by allowing only one value to be specified underlyingly in any given context in lexical representations, namely, the negation of the value assigned in that context by the system of lexical rules. A feature is only specified in a lexical entry if that is necessary to defeat a rule which would assign the "wrong" value to it. The default values of a feature are assigned to segments not specified for it at a stage in the derivation which may vary language-specifically within certain bounds.

A third position, departing even further from *SPE*, and currently under exploration in several quarters, holds that the unmarked value is never introduced, so that features are in effect one-valued (privative).

Contrastive and radical underspecification both posit *redundancy rules* such as:

(2) [+ sonorant] → [+ voiced]

Radical underspecifications in addition posits *default rules*, minimally a context-free rule for each feature which assigns the unmarked value to it:

(3) [ ] → [−voiced]

The following chart summarizes the theoretical options, and exemplifies them with the values of the feature [voiced] which they respectively stipulate for voiceless obstruents, voiced obstruents, and sonorants, at the initial and final levels of representation:

(4)

		/p/	/b/	/r/
None (full specification)	Lexical: fully specified	−	+	+
	Phonetic: fully specified	−	+	+
Contrastive	Lexical: contrastive values	−	+	+
	Phonetic: fully specified	−	+	+
Radical	Lexical: minimal specifications		+	
	Phonetic: fully specified	−	+	+
Privative	Lexical: only marked values		+	
	Phonetic: only marked values		+	

As (4) shows, fully specified representations and privative representations are homogenous throughout the phonology. Contrastive underspecification and radical underspecification both make available *two representations*, by allowing an underlying minimal structure to be augmented in the course of the derivation.

Radical underspecification moreover assumes that default values are assigned by the entire system of structure-building lexical rules. For example, in a language with a lexical rule of intervocalic voicing such as (5),[6] the lexical marking of obstruents in intervocalic position would be the reverse of what it is in other positions, with voiced consonants unmarked and voiceless ones carrying the feature specification [−voiced] to block the rule.

(5) [ ] → [+voiced] / V___V

At what point are default values and redundant values to be assigned? I will here assume that default feature values are filled in before the first rule that mentions a specific value of that feature.[7] Many assimilation rules do not mention a specific feature value, but simply spread the feature itself, or a class node under which that feature is lodged. Such rules can apply before the assignment of default values, yielding the characteristic pattern "assimilate, else default."

To summarize:

(6) (a) For each feature F, a universal default rule of the form [ ] → [αF] applies in every language.
(b) In each environment E in underlying representations, a feature must be either specified as [αF] or unspecified, where E is defined by the most specific applicable rule R, and R assigns [−αF].
(c) Default feature values are filled in before the first rule that mentions a specific value of that feature.

(6a) guarantees that the basic choice of unmarked value of a feature is fixed language-independently, but leaves open the possibility that particular rules (universal as well as language-specific) may supersede it in special contexts. (6b) says essentially that the lexicon is minimally redundant: feature specifications are only allowed where needed to defeat rules. Subject to (6c), default feature values can be assigned either cyclically, at the word level, or postlexically. Redundant values are normally assigned post-lexically.

An early argument for radical underspecification was that it makes it possible to extend the first level of phonological rules to account for the structure of morphemes (Kiparsky 1982), eliminating from the theory the extremely problematic "Morpheme Structure Constraints (MSC)," never satisfactorily formalized, and heir to a multitude of embarrassing problems and paradoxes. The structure of morphemes in a language can now be treated simply as derivative of the rules and conditions on its earliest level of phonological representations.[8]

The distinction between structure-changing and structure-building (feature-filling) operations is important here. Feature-changing assimilations (i.e., those which override existing feature specifications) have been shown to consist of two independent processes, delinking of the features of the target, followed by spread of a feature to it (Poser 1982; Cho 1990). The introduction of structure-building rules, which make essential use of radical underspecification, has several striking consequences. It has provided the basis for new accounts of "strict cycle" effects (Kiparsky 1993) and of inalterability (Inkelas and Cho 1993). If these prove to be correct, they will provide the strongest kind of support for underspecification. My contention here is that it is also implicated in the explanation of lexical diffusion. In the next section, we will see how this works.

1.3 Lexical Diffusion as Analogy

Equipped with this theory of lexical rules and representations, let us go back to the ū-shortening process (1) to illustrate the general idea. [ū] and [ŭ] are in the kind of semi-regular distribution that typically sets off lexical diffusion processes. The core context (1a) has almost only [ŭ] to this day. Exceptions seem to occur only in affective or facetious words of recent vintage: *googol* (*-plex*), *googly, kook*. And the context most distant from the core, not included in any of the extensions of (1a), has overwhelmingly long [ū]: *doom, stoop, boom, poop, boob, snood, loose, Moomin, loom, baboon, spoof, snooze, snoot, snoop*, etc. Even here some subregularities can be detected. There are a few shortened [ŭ]'s before coronalas even if the onset is coronal or labial (*foot, stood, toots(ie), soot* versus *booth, moon, pool, tool, loose, spoon, food, mood, moose*... with long [ū]). Before labials, however, the exclusion of short [ŭ] is near-categorical.[9]

Let us suppose that the core regularity is reflected in the lexical phonology of English by a rule which assigns a single mora or vocalic slot to stressed /u/ between certain consonants, and two moras or vocalic slots elsewhere, provided that syllable structure allow. Suppose the original context of this rule was [-anterior]___[-anterior, -coronal]. As a structure-building rule it can, however, be extended to apply in the contexts (1b) and (1c). This part of the change is a natural generalization (simplification) of the rule's environment, in principle no different from the extension of a morphological element to some new context. But because structure-building rules are defeasible by lexical information, such an extension of the shortening rule need not effect any overt change at first: the extended rule simply applies (in the synchronic grammar) to the words which always had short [ŭ] in that context, now reanalyzed as quantitatively unmarked, while words with long [ū] in those contexts are now presepecified with two moras in the lexicon to escape the effect of the generalized shortening rule. But once the rule's context is so extended, words can fall under its scope, slowly and one at a time, simply by being "regularized" through loss of the prespecified length in their underlying representations. This is the lexical diffusion part of the process.

The model for this phase of the analogical regularization is the existence of a systematic context (the core shortening environment) where length is systematically predictable, which is extended on a case-by-case basis. The normal scenario of lexical diffusion, then, is contextual rule generalization with attendant markedness reversal and subsequent item-by-item simplification of the lexicon. In principal, it could proceed until the rule is extended to all contexts and all quantitive marking is lost in the lexicon. In this example, however, the robust exclusion of short [ŭ] in the context between labials sets a barrier to further extension of the rule to those contexts. The result is the pattern of partial complementation that we find in the modern English distribution of [ŭ] and [ū].

Let us now turn to the rule which thanks to Labov's work has become the most famous case of lexical diffusion: the "æ-Tensing" of Philadelphia and several other Eastern U.S. dialects, applying in the core environment before tautosyllabic -*f*, -*s*, -θ, -*n*, -*m*.

First, I would like to raise a terminological point, relating to a larger issue of fact which is tricky but luckily does not have to be settled here. Although usually referred to as æ- Tensing, æ- Lengthening would be a more appropriate term because the vowel is not always tense. Phonetically, it is typically a lax long [ɛ] in the dialects I am concerned here with (see, e.g., Bailey 1985, p. 174). Phonologically, that may be a better analysis as well, because it is the same vowel as the word-finally lengthened lax [ɛ] in the truncated form of *yes* (*"yeah"*). At least in the feature system that I will be using in section 3.2 below, this is a genuine [-Tense] vowel. But since it won't make much of a difference for present purposes, I'll just follow tradition and continue to talk of "Tensing," while writing the "tensed" vowel noncommittally as *A*.

What is the status of [æ] and [A] in these dialects? Are the two phonemically distinct? Is their distribution governed by rule? It is clear that they are two distinct phonemes, in the sense that there is an irreducible lexical contrast between them in certain environments. From the viewpoint of many phonological theories, that settles the second question as well: they contrast and they do not alternate with each other, so their distribution cannot be rule-governed.

The distribution of [æ] and [A] is however far from random. In the framework proposed in Kiparsky (1982c), the regularities that govern it have a place in the lexical module of the grammar as structure-building lexical rules which assign the appropriate default specifications of tenseness to the underlying unspecified low front vowel, which we can write /*a*/. The lexicon need specify only those comparatively few instances of lax /*æ*/ which fall out of line. This analysis follows from the requirement (6b) that the redundancy of the lexicon must be reduced to a minimum.

The Philadelphia version of *æ*-Tensing (Ferguson 1975; Kiparsky 1988; Labov 1981, 1993) includes all the core environments -*f*, -*s*, -θ, -*n*, -*m* as well as the extension -*d*, -*l*, as discussed further below.

(7) Philadelphia lexical æ Tensing rule:
æ A before tautosyllabic *f*, *s*, θ, *m*, *n*, (*d*, *l*)

In New York, the rule applies also more generally before voiced stops and before-š.

(8) New York lexical æ-Tensing rule:
æ A before tautosyllabic *f*, *s*, θ, š, *m*, *n*, *b*, *d*, ǰ, *g*

In accord with our previous discussion, (7) and (8) are structure-building rules which assign [+Tense] to *a* in regular words like (9a). The value [-Tense] is then assigned by default to *a* in regular words like (9b). The only cases of lexically specified Tenseness are exceptional words with [-Tense] in Tensing environments, such as [9c]:

(9) (a) pAss, pAth, hAm, mAn
(b) mat, cap, passive, panic
(c) alas, wrath

In fact, the unpredictable cases for which lexical specification of [±Tense] is required are probably even fewer than is apparent at first blush. Consider the contrast before consonant clusters in polysyllables illustrated by the words in (10):

(10) (a) astronaut, African, plastic, master (lax æ OK)
(b) After, Afterwards, Ambush, Athlete[10] (Tense A)

These data follow directly from rule (7) on standard assumptions about English syllable structure. English syllabification tends to maximize onsets, and *str–,fr–* are possible onsets, but *ft–,mb–,θ–* are not, so the relevant VC sequence has to be tautosyllabic in (10b) but tends to be heterosyllabic in (10a). Independent evidence for this syllabification is the fact that vowel reduction, restricted to unstressed open syllables, is possible before permitted onsets, as in *astronomy*, but not before other clusters, as in *athletic* (Kahn 1976).[11]

Rule (7) must apply at level 1 in the lexical phonology of English. Five arguments for this position were given in Kiparsky (1988). We can now add two more. First, the observations in the preceding paragraph show that (7) must precede the "left capture" rule that attaches onset consonants to a preceding stressed syllable (perhaps making them ambisyllabic). But left capture can be shown to apply at level 1 (as well as at later levels), so æ Tensing must apply at level 1 as well. The evidence that left capture applies at level 1 is the pattern of shortening seen in derived words such as (11):

(11) (a) cȳcle cy̌clic cy̌clicity
(b) trībe trībal trībality

Myers (1987) has shown that the various English shortening processes, including "Trisyllabic Shortening" and the shortening before *–ic* as in *cȳcle* □ *cy̌clic*, are special cases of a general lexical rule which shortens nuclei in closed syllables, including those which become closed through the application of "left capture" resyllabification. But the short initial syllable of *cy̌clicity* is clearly inherited from *cy̌clic*, since the conditions for shortening no longer hold in the derivative *cy̌clicity* (cf. *trībality*). It follows that the shortening must be cyclic. Therefore, the left capture rule that feeds shortening, as well as the æ Tensing rule (7) that itself precedes left capture, must also be cyclic. But cyclic phonology is located at level 1.

My second new argument for the level 1 status of æ–Tensing is that it explains the variation in the past tenses of strong verbs such as *ran*, *swam*, *began*. These /æ/–vowels are regularly lax in Philadelphia, a fact accounted for by ordering æ–Tensing before the æ → A ablaut rule which introduces /æ/in the past tense. Since ablaut is a level 1 rule, æ–Tensing, which precedes it, must also apply at level 1. The possibility of applying the rules in reverse order, still within level 1, predicts a dialect in which the vowels of these verb forms do undergo æ–Tensing. Such a dialect is in fact attested in New York, as Labov notes. In contrast, nonmajor category words such as *am, had, can* and the interjections *wham!,bam!* have lax æ in all dialects where æ–Tensing is lexical. The lack of variation in these cases is likewise predicted because nonlexical categories are not subject to the rules of lexical phonology.

With these synchronic preliminaries out of the way, let us turn to the rule's lexical diffusion. Labov shows that [+Tense] vowels have replaced (or are in the process of replacing) [–Tense] vowels in a class of words in Philadelphia, especially in the speech of children and adolescents. The innovating class of words includes: (1) words in which æ is in the proper consonantal environment of the tensing rule (7) but, contrary to what the core rule requires, in an open syllable, such as (12a), and (2) words in which æ is before *l* and *d*, voiced consonants not included among the rule's original triggers.[12] In cases like (12c), both extensions of the rule are combined.

(12) (a) plAnet, dAmage, mAnage, flAnnel
(b) mAd, bAd, glAd, pAl
(c) personAlity, Alley, Allegheny

There are several facts that need explaining about these developments. First, the environments into which tense *A* is being extended are not arbitrary phonologically. There is no "lexical diffusion" of *A* before voiceless stops, the class of consonants that is systematically excluded from the core tensing environment as well as from the Philadelphia and New York versions of the rule. Second, there are no reported cases of lax æ being extended into words which have regular tense *A* in accord with (7), e.g., in words like *man, ham, pass*. Third, [æ] changes not to any old vowel, but precisely to [A], the very vowel with which it is in partial complementation by(7).

If we assume that lexical diffusion is nothing more than the substitution of one phoneme for another in the lexical representations of words, we have no explanation either for the direction of the change, nor for the envelope of phonological conditions that continues to control it. Such a theory cannot distinguish the Philadelphia development from a wholly random redistribution of tense and lax *a*, nor even explain why it should involve this particular pair of vowels at all.

If we recognize that the distribution of tense and lax *a* in Philadelphia is an analogical extension of rule (7), then we are in a position to explain these facts. The phonological conditions under which tense *A* spreads through the lexicon are an extension of the rule's original context in two respects: (1) the condition requiring the triggering consonant to be tautosyllabic is dropped (here one might also explore the possibility that the tensing rule gets reordered after left capture), and (2) *l, d* are included among the conditioning consonants. This development conforms to the pattern of contextual generalization with item-by-item implementation of the extended environment that is typical of lexical diffusion. The scenario is similar to the one sketched out above for the shortening of /ū/. The old tensing rule, applicable before a class of tautosyllabic consonants, is generalized by some speakers to apply before certain additional consonants and the tautosyllabicity condition is dropped. Speakers who have internalized the rule in this generalized from can pronounce tense *A* in words of the type (12). But being structure-building (feature-filling), the rule applies only to vowels underspecified for the feature of tenseness, and speakers with the generalized rule can still get lax æ in the new contexts by specifying the vowels in question as law in their lexical representation. In the resulting variation in the speech community, the generalized rule, and the forms reflecting the unmarked lexical representations, will enjoy a selective advantage which causes them gradually to gain ground.

I conclude that æ-Tensing supports the claim that lexical diffusion is the analogical extension of structure-building lexical rules. We see that, on the right assumptions about the organization of phonology and about analogical change, lexical diffusion fits snugly into the neogrammarian triad, and all its by now familiar properties are accounted for. A wider moral that might be drawn from this result is that even "static" distributional regularities in the lexicon, often neglected in favor of productive alternations, can play a role both in synchronic phonology and in analogical change.

1.4 What Features are Subject to Diffusion?

According to the present proposal, the prerequisite for lexical diffusion is a context-sensitive structure-building lexical rule and its starting point is an existing site of neutralization or partial neutralization of the relevant feature in lexical representations. The original environment of the æ-Tensing rule (originally the broad *a* rule) was before tautosyllabic *f*, *s*, θ, -*nt*, -*ns*, as in *pass, path*, *laugh, aunt, dance*. It became generalized to apply before the nasals *n, m* in all the Mid-Atlantic dialects, and later before voiced stops as well (see (7) and (8)). the cause of this generalization of the lexical æ-Tensing rule is probably the merger with a *post-lexical* raising/tensing rule in those dialects where their outputs coincided (Kiparsky 1971, 1988). In those dialects which either lacked the lexical rule entirely (as in the Northern Cities), or retained it as a different rule (as in Boston, where broad *a* was pronounced as [a]), the post-lexical æ-Tensing rule can today be observed as a separate process in several variant forms. In the Northern Cities, it yields a continuum of tensing and raising, with most tensing before nasals and least tensing before voiceless stops.

(13)

Tensing environments in Northern Cities dialects:

nasals (tensest) ← voiced stops — fricatives → voiceless stops (laxest)

In Boston, only the environment at the top of the scale, the nasals, triggers tensing and raising; before other consonants, the dialect retains lax æ (Labov 1993).

The merger of the inherited lexical æ-Tensing rule with these two types of post-lexical æ-Tensing gives the Philadelphia and New York versions of lexical æ-Tensing, respectively. Specifically, by adding the environments of the original lexical æ-Tensing rule (*-f*, *-s*, -θ, *-ns*, *-nt*) and the environemnts of the postlexical æ-Tensing/Raising of the Boston type (nasals), we get exactly the environments of the post-lexical æ-Tensing/Raising of the Northern Cities type (13) (nasals, voiced stops, and fricatives), we get very nearly the New York rule (8). Only the failure of -ŋ to trigger æ-Tensing in New York remains unexplained.[13]

Having acquired lexical status in this way, Tensing then spreads to new lexical items, that is, it undergoes lexical diffusion. Thus, the lexical diffusion of æ-Tensing in the Mid-Atlantic dialects is due to its lexical status in those dialects, inherited from the lexical broad α rule of British English.

Labov (1981, 1993) makes the interesting suggestion that lexical diffusion is an intrinsic characteristic of some kinds of phonological features and neogrammarian sound change is characteristic of others. Lexical diffusion affects "higher order classes", phonological features such as tenseness and length, which are defined in terms of several unrelated phonetic properties, such as duration, height, peripherality, and diphthongization. Features like front/back and high/low, on the other hand, will not undergo lexical diffusion because their physical realization is more direct. If lexical diffusion really does depend on whether a feature is realized on a single physical dimension or on several, my account of lexical diffusion as teh analogical extension of structure-building lexical rules would have to be given up at least in its present form.

One problem with Labov's idea is that æ Lengthening, though it involves the same feature in all dialects, undergoes lexical diffusion in the Mid-Atlantic dialects and not in the Northern Cities. In response to that objection, Labov suggests that the rule operates at a "high level of abstraction" in the Mid-Atlantic dialects and at a "low level of abstraction" in the Northern Cities. But this amounts to using the term "abstraction" in two different senses. On the one hand, it is a phonetic property having to do with the degree of diversity and complexity of the feature's phonetic correlates. With respect to æ-Tensing, however, it has to be understood in a functional/structural sense, as something like the distinction between phonemic and allophonic status, or lexical and postlexical status - for that seems to be the one relevant distinction between the Mid-Atlantic and the Northern Cities versions of æ-Tensing. But there is no reason to believe that these two kinds of "abstraction" can be identified with each other. Certainly features differ in the intrinsic complexity and diversity of their phonetic realizations: stress and tenseness probably tend to have relatively complex and diverse phonetic effects, whereas fronting, lip rounding, height, and voicing probably tend to have more uniform phonetic effects. But this would appear to be true whether they are distinctive or redundant. I know of no evidence to show that the intrinsic complexity and diversity of the phonetic reflexes of a feature is correlated with its lexical/phonemic status, let alone that these two kinds of "abstractness" are the same thing.

The interpretation of lexical diffusion that I have advocated here would entail that the structural notion of abstractness is all we need, and the phonetic character of the feature should be immaterial. The generalization that only lexically distinctive features can undergo lexical diffusion, itself a rigorous consequence of LPM principles, predicts excatly the observed difference between the Mid-Atlantic dialects and the other U.S. dialects. The contrast between them shows that the same feature, assigned by one and the same rule in fact, can be subject to lexical diffusion in one dialect and not in another, depending only on whether it is lexically distinctive or redundant. In addition, it also correctly predicts the existence of lexical diffusion in such features as height and voicing, which on

Labov's proposal should not be subject to it.[14].

2 The Structure-dependence of Sound Change

2.1 Sound Change is Not Blind

The majority of structuralists, European as well as American, thought they could account for phonological structure even while conceding to the neogrammarians that sound changes are "blind" phonetic processes. In their view, the reason languages have orderly phonological systems is that leaners impose them on the phonetic data, by grouping sounds into classes and arranging them into a system of relational oppositions, and by formulating distributional regularities and patterns of alternation between them. The reason languages have phonological systems of only certain kinds would then have to be that learners are able to impose just such systems on bodies of phonetic data. But, on their scheme of things, fairly simple all-purpose acquisition procedures were assumed to underlie the organization and typology of phonological inventories, and the combinationrial regularities apprehended by learners.

It seems clear, however, that a battery of blind sound changes operating on a language should eventually produce systems whose phonemicization by the standard procedures would violate every phonological universal in the book. The linguist who most clearly saw that there is a problem here was Jakobson (1929). Emphasizing that phonological structure cannot simply be an organization imposed ex post facto on the results of blind sound change, he categorically rejected the neogrammarian doctrine in favor of a structure-governed conception modeled on the theory of orthogenesis (or nomogenesis) in evolutionary biology (a theory now thoroughly discrdited, but for which Jakobson always maintained a sneaking fondness). His basic thesis is that sound changes have an inherent direction ("elles vont selon des directions déterminées") towards certain structural targets.[15]

Jakobson was in fact able to cite fairly convincing long-term tendencies in the phonological evolution of Slavic, involving the establishment of proto-Slavic CV syllable structure by a variety of processes (degemination, cluster simplification, metathesis, prothesis of consonants, coalescence of C+*y*, coalescence of V + nasal), and the rise of palatal harmony in the syllable domain through a series of reciprocal assimilations. Since it is human to read patterns into random events, it would be prudent to look at such arguments with a measure of suspicion. But the number and diversity of phonological processes collaborating to one end do make Jakobson's case persuasive. Others have since argued for similar conclusions. For example, Riad (1992), working in the framework of prosodic generative phonology, has analyzed the major sound changes in North Germanic over the past two millennia as so many stepwise resolutions of an inherent conflict between fixed accent, free quantity, and bimoraic foot structure.

Jakobson further argued that sound change respects principles of universal grammar, including *implicational universals*. The point is quite simple. How could an implicational relation between two phonological properties A and B have any universal validity if sound changes, operating blindly, were capable of changing the phonetic substrate of A and B independently of each other?

Moreover, Jakobson's implicational universals were crucially formulated in terms of *distinctive* features. But purely phonetically conditioned sound changes should not care about what is distinctive in the language (distinctiveness being, by the structuralists' assumptions, a purely structural property imposed a posteriori on the phonetic substance). So what prevents sound change from applying in such a way as to produce phonological systems that violate universals couched in terms of the notion of distinctiveness?

For some reason, Jakobson's work is rarely taken notice of in the literature on sound change, and I am not aware of any explicit attempts to refute it. Perhaps it has simply been rejected out of hand on the grounds that it begs the question by invoking a mysterious mechanism of orthogenesis which itself has no explanation, and that in addition, it throws away the only explanation we have for the regularity and exceptionlessness which are undeniably characteristics of a major class of sound changes. Nevertheless, the existence of sound changes that respect structure and are derived by it in certain ways seems well supported. How can we account for the coexisting properties of exceptionlessness and structure-dependence?

I believe that Jakobson was on the right track in looking to evolutionary biology as a paradigm for historical linguistics. We just need to reject the disreputable version of evolutionary theory that he claimed to be inspired by and replace it by the modern view of variation and selection. In the domain of sound change, the analog to natural selection is the inherently selective process of transmission that incorporates them into the linguistic system. Thus sound change is both mechanical in teh neogrammarian sense, and at the same time structure-dependent, though not exactly in the way Jakobson thought.

We are now free to assume that variation at the level of speech production is conditioned purely by phonetic factors, independently of the language's phonological structure, and to use this property to derive the exceptionlessness property, just as the neogrammarians and structuralists did. The essential move is to assign a more active role to the transmission process, which allows it to intervene as a selectional mechanism in language change. Traditionally, the acquisition of phonology was thought of simply as a process of organizing the primary data of the ambient language according to some general set of principles (for example, in teh case of the structuralists, by segmenting it and grouping the segments into classes by contrast and complementation, and in the case of generative grammar, by projecting the optimal grammar consistent with it on the basis of Universal Grammar). On our view, the learner in addition selectively intervenes in the data, favoring those variants which best conform to the language's system. Variants which contravene language-specific structural principles will be hard to learn, and so will have less of a chance of being incorporated into the system. Even “impossible” innovations can be admitted into the pool of phonetic variation; they will simply never make it into anyone's grammar.

The combined action of variation and selection solves another neglected problem of historical phonology. The textbook story on phonologization is that redundant features become phonemic when their conditioning environment is lost through sound change. This process (so-called secondary split) is undoubtedly an important mechanism through which new phonological oppositions enter a language. But the textbooks draw a discreet veil over the other cases, surely at least equally common, where – in what may seem to be exactly analogous situations – the redundant feature simply disappears when its triggering environment is lost.

The two types of outcome are not just distributed at random. The key generalization seems to be that phonologization will result more readily if the feature is of a type which already exists in the language. We could call this the *priming effect* and provisionally formulate it as follows:

> (14) Redundant features are likely to be phonologized if the language's phonological representations have a class node to host them.

This priming effect, a diachronic manifestation of structure-preservation is documented for several types of sound change, tonogenesis being perhaps the most interesting case. The merger of voiced and voiceless consonants normally leaves a tone/register distinction *only in languages which already possess a tone system* (Svantesson 1989). There is one special circumstance under which nontonal languages can acquire tone by loss of a voicing contrast: in certain Mon-Khmer languages, according to Svantesson, “strong areal pressure to conform to the phonological pattern of those monosyllabic tone languages that dominate the area” (ibid.). It seems, then, that when the voicing that induces redundant pitch is suppressed, the pitch will normally be phonologized only if the language, or another language with which its speakers are in contact, already has a tonal node to host it. On the neogrammarian/structuralist understanding, the priming effect remains mysterious. On our variation/selection model, such top-down effects are exactly what is expected.

Analogous priming effects can be observed in such changes as compensatory lengthening and assimilation. De Chene and Anderson (1979) find that loss of a consonant only causes compensatory vowel lengthening when there is a preexisting length contrast in the language. So the scenario is that languages first acquire contrastive length through other means (typically by vowel coalescence); then only do they augment their inventory of long vowels by compensatory lengthening.[16] Yet loss with compensatory lengthening is a quintessentially regular, neogrammarian type of sound change (in recent work analyzed as the deletion of features associated with a slot with concomitant spread of features from a neighboring segment into the vacated slot). Similarly, total assimilation of consonant clusters resulting in geminates seem to happen primarily (perhaps only?) in languages that already have geminates (Finnish, Ancient Greek, Latin, Italian). Languages with no pre-existing geminates prefer to simplify clusters by just dropping one of the consonants (English, German, French, Modern Greek). In sum, we find a conjunction of exceptionlessness and structure-sensitivity in sound change which does not sit well with the neogrammarian/structuralist scheme. The two-level variation/selection model of change proposed is in a position to make much better sense of it.

The two-level scheme can be related to certain proposals by phonemic theorists. It has often been argued that redundant features help to perceptually identify the distinctive features on which they structurally depend.[17] Korhonen (1969, pp. 333-335) suggests that only certain allophones, which he calls *quasiphonemes*, have such a functional role, and that it is just these which become phonemicized when the conditioning context is lost. This amounts to a two-stage model of secondary split which (at least implicity) recognizes the problem we have just addressed: in the first stage, some redundant features become quasi-distinctive, and in the second stage, quasi-distinctive features become distinctive when their conditioning is lost. If the conditions which trigger the first stage were specified in a way that is equivalent to (14), this proposal would be similar to the one put forward above. Korhonen's suggestion is however based on the direction of allophonic conditioning: according to him, it is allophones which precede their conditioning environment (and only they?) which become quasi-phonemicized. This is perceptually implausible, and does not agree with what is known about secondary split, including tonogenesis. Ebeling (1960) and Zinder (1979) propose entities equivalent to Korhonen's quasi-phonemes in order to account for cases where allophones spread to new contexts by morphological analogy. They do not spell out the conditions under which allophones acquire this putative quasi-distinctive status either. However, the cases they discuss fit in very well with the priming effect, since they involve features which are already distinctive in some segments of the language and redundant in others becoming distinctive in the latter as well.

2.2 The Life Cycle of Phonological Rules

Early generative work on historical phonology thought of sound change as rule addition. One of the most interesting consequences of this idea was that sound changes should be capable of nonphonetic conditioning, through the addition of morphologically conditioned rules, and through the addition of rules in places other than the end of the grammar ("rule insertion"). But of course not just any sort of nonphonetic conditioning is possible. It turned out that the only good cases of rule insertion involved the addition of rules before automatic (transparent) rules, often of a phonetic character, so that an interpretation along the lines of the above structure-preservation story seems more likely. Moreover, this approach by itself does not explain one of the most basic facts about sound change, its phonetic naturalness. Nor, in the final analysis, does it address the question of the relationship between universals and change in a principled way.

By articulating the phonological component into a set of modules with different properties, lexical phonology allows us to think of sound change in a more constrained way that is still consistent with the selection/variation model (Kiparsky 1988). Sound change can be assumed to originate through synchronic variation in the production, perception, and acquisition of language, from where it is internalized by language learners as part of their phonological system. The changes enter the system as language-specific phonetic implementation rules, which are inherently gradient and may give rise to new segments or combinations of segments. These phonetic implementation rules may in turn become reinterpreted as phonological rules, either post-lexical or lexical, as the constraints of the theory require, at which point the appropriate structural conditions are imposed on them by the principles governing that module. In the phonologized stages of their life cycle, rules tend to rise in teh hierarchy of levels, with level 1 as their final resting place (Zec 1993).

In addition to articulatory variation, speech is subject to variation that originates in perception and acquisition, driven by the possibility of alternative parsing of the speech output (Ohala 1986, 1989). Sound changes that originate in this fashion clearly need not be gradient, but can proceed in abrupt discrete steps. Moreover, like all reinterpretation processes, they should be subject to inherent top-down constraints defined by the linguistic system: the "wrong" parses that generate them should spring from a plausible phonological analysis. Therefore, context-sensitive reinterpretations would be expected not to introduce new segments into the system, and context-free reinterpretations (such as

British Celtic *i*^{w} *p*) would be expected not to introduce new features into the system; and neither should introduce exceptional phonotactic combinations.

Dissimilation provides perhaps the most convincing confirmation of this prediction. That dissimilatory sound changes have special properties of theoretical interest for the debate on levels of phonological representation was first pointed out by Schane (1971). Schane marshaled evidence in support of the claim that "if a feature is contrastive in some environments but not in others, that feature is lost when there is no contrast," and argued on this basis for reality of phonemic representations. Manaster-Ramer (1988) convincingly showed that the contrastiveness of the environment is not a factor in such cases, and rejected Schane's argument for the phoneme entirely. However, all his examples, as well as Schane's, conform to a kindred generalization which still speaks for the role of distinctiveness in sound change: *only features which are contrastive in the language are subject to dissimilation*. But in this form, the generalization is a corollary of what we have already said. The reasoning goes as follows. Dissimilation is not a natural articulatory process. Therefore, it must arise by means of perceptual reanalysis. But the reanalyzed form should be a well-formed structure of the language, hence in particualr one representable in terms of its authentic phonological inventory.

The other properties of dissimilation, that it is quantal rather than gradual, and that it is often sporadic, can be derived in the same way. They likewise hold for the other so-called minor sound changes, such as metathesis. Not that minor sound changes are *necessarily* sporadic. On the contrary, they will be regular when the phonotactic constraints of the language so dictate. Dissimilation is regular where it serves to implement constraints such as Grassmann's Law, and the same is true of metathesis (Hock 1985; Ultan 1978): e.g., the Slavic liquid metathesis is part of the phonological apparatus that implements the above-mentioned syllable structure constraints.

The respective properties of major and minor sound changes are summarized in (15):

(15)

	Major changes	**Minor changes**
Source in speech:	Production	Perception and acquisition
Parameter of change:	Articulatory similarity	Acoustic similarity
Gradiency:	Gradient	Discrete
Effect on system:	New segments and combinations	Structure-preserving
Regularity:	Exceptionless	Can be sporadic

Conditions on sound change can then be seen as categorical reinterpretations of the variable constraints that determine the way optional rules apply. Because of the formal constraints on possible structural conditions, obligatory rules cannot fully replicate the complex pattern of preferences generated in language use at the optional stage. Consequently, when a rule becomes obligatory, its spectrum of contextual conditions is simplified and polarized. Thus, this view of sound change explains both why structural conditions on phonological rules retain a gross form of naturalness, and why they nevertheless do not show the intricate microconditioning observed at the level of phonetic implementation.

Not only are phonological conditions on rules derived from phonetic conditions motivated by perception and production, but also the nature of conditions involving morphology, style, and even sex and class can be explained in the same way. For example, some languages of India have undergone sound changes restricted to the speech of lower castes. Such changes are a categorical reflection, under conditions where social boundaries are sharply drawn, of the generally more advanced nature of vernacular speech, due to the fact that the elite tends to stigmatize and inhibit linguistic innovations for ideological reasons (Kroch 1978).

Our conclusion so far is that the neogrammarians were right in regarding sound changes as a process endogenous to language, and their exceptionlessness hypothesis is correct for changes that originate as phonetic implementation rules. They were wrong, however, in believing that sound change per se, as a mechanism of change, is structure-blind and random. The process also involves an integration of speech variants into the grammar, at which point system-conforming speech variants have a selective advantage which causes them to be preferentially adopted. In this way, the language's internal structure can channel its own evolution, giving rise to long-term tendencies of sound change.

3 Naturalness in Sound Change

The study of natural phonology offers a further argument for the structure-dependence of even neogrammarian-type exceptionless sound change, and thereby for the selection/variation view of sound change. In this section, I support this claim by showing the role that underspecification plays in the explanation of natural assimilation rules and vowel shifts – not only of the synchronic rules, but equally, and perhaps in greater measure, of the historical processes that they reflect.

3.1 The Typology of Assimilation

Autosegmental phonology allows assimilation to be treated as the spread of a feature or feature complex from an adjacent position. Coupled with assumptions about underspecification, feature geometry, and the locality of phonological processes, it yields a rich set of predictions about possible assimilation rules. Cho (1990) has developed a parametric theory of assimilation based on these assumptions. The following discussion draws heavily on her work, which, though formulated as a contribution to synchronic phonology, bears directly on sound change as well.

If feature-changing processes consist of feature deletion plus feature filling, we can say that assimilation is fed by weakening rules which de-specify segments for the feature in question, to which the feature can then spread by assimilation from a neighboring segment. The feature-deletion (neutralization) process which on this theory feeds apparent feature-changing assimilation can be independently detected by the default value it produces wherever there is no assimilation (complementarity between assimilation and neutralization).

If we assume that assimilation is spreading of a feature or class node, then it immediately follows that there should be no assimilations which spread only the unmarked value of a feature, since there is no stage in the derivation where only unmarked values are present in the representation. For example, there are two-way assimilations of [±voiced], as in Russian, and one-way assimilations of [+voiced], as in Ukrainian and Santee Dakota, but no one-way assimilations which spread only [–voiced]. Cho's survey confirms this striking prediction for a substantial sample of languages.[18]

(16) (a) Russian: /tak+že/ → ta[g]že "also", /bez tebja/ → be[s] teba "without you"
(b) Ukrainian: /jak že/ → ja[g]že "how", /bez tebe/ → be[z] tebe "without you"

One-way assimilation (spread of the marked feature value) as in (16b) results from ordering assimilation after the assignment of default feature values. Since two-way assimilation applies when default feature specifications have already been assigned, it must involve feature deletion at the target as a prior step, followed by spread to the vacated site. This yields the following additional predictions.

First, two-way assimilation should apply preferentially in environments where neutralization is favored. This seems to be correct: for example, the prevalence of feature neutralization in coda position explains the prevalence of assimilation in coda position (e.g., regressive assimilation in consonant clusters).

Second, in environments where neutralization applies but where no trigger of assimilation is present (for example, in absolute final position), two-way assimilation should be associated with neutralization in favor of the unmarked (default) value. This prediction is also confirmed by such typical associations as (two-way) voicing assimilation with final devoicing, or place assimilation with coda neutralization of place.[19]

Suppose we also allow assimilation to be ordered either before or after *redundant* values are assigned. This gives two subtypes of two-way assimilation: one in which only distinctive feature specifications (e.g., [±voiced] on obstruents) trigger assimilation, the other where redundant feature specifications also trigger assimilation. For voicing assimilation, the first type is represented by

Warsaw Polish (as well Russian and Serbo-Croatian), the second by Cracow Polish:

(17) (a) Warsaw Polish: ja[k] nigdy "as never"
(b) Cracow Polish: ja[g] nigdy "as never"

The theory predicts that one-way assimilation cannot be triggered by redundant feature values (i.e., it must be of the Warsaw type, not of the Cracow type). In fact, the voicing assimilation rules of Ukrainian and Santee (e.g., (16b)) are triggered by obstruents only. It also follow that if a language has both Warsaw-type and Cracow-type assimilation, then the former must be in an earlier level. For example, Sanskrit has lexical voicing assimilation triggered by obstruents and post-lexical voicing assimilation by all voiced segments. For similar reasons, if a language has both one-way and two-way assimilation, then the former must be in an earlier level.

In combination with the formal theory of phonological rules, under-specfication provides the basis for Cho's parameterized typology of assimilation. According to this theory, every assimilation process can be characterized by specifying a small number of properties in a universal schematism:

1 Site of spreading (single feature or a class node)
2 Specification of target and/ or trigger
3 Locality (nature of structural adjacency between trigger and target)
4 Relative order between spreading and default assignment
5 Directionality of spreading
6 Domain of spreading

This approach has a number of additional consequences of interest for both synchronic and historical phonology.

Since codas are the most common target of weakening, and adjacency the most common setting of the locality parameter, it follows that regressive assimilation from onsets to preceding codas will be the most common type of assimilation. Thus, no special substantive principle giving priority to regressive assimilation is required.

Additional consequences follow if we bring in feature geometry. Since the domain of spreading can be limited to a specific node in the feature hierarchy, it follows that assimilation between segments belonging to the same natural class is a natural process. The traditional generalization that assimilation is favored between segments which are already most similar in their feature composition (Hutcheson 1973; Lee 1975) is thus explained in a principled way. "Strength hierarchies" (proposed, e.g., by Foley 1977 to account for the direction of assimilation) also turn out to be epiphenomenal.

An element may be ineligible to spread either because it already bears an incompatible feature specification (whether as an inherent lexical property or assigned by some rule), or because some constraint blocks it from being associated with the spreading feature value. Once the spread of a feature has been so interrupted, further spread is barred by locality. Thus, "opaque" elements need not themselves be specified for the spreading feature; they must only bear the relevant class node.[20]

It seems clear from the work of Cho and others that underspecification is not only relevant for the synchronic analysis of lexical phonology, but plays a role in defining the conditioning of phonetic processes. The difference between marked, default, and redundant feature values – a basically structural difference – constitutes a major parameter along which assimilatory processes vary. We must conclude that a large and well-studied class of sound changes is *simultaneoulsy* exceptionless and structure-dependent.

3.2 Vowel Shifts

The point of this section is similar to that of the last, though this one is offered in a more speculative vein. I argue that vowel shifts are another type of natural sound change whose explanation, on closer inspection, depends on the *structural* status of the triggering feature in the system, specifically on whether the feature is specified in the language's phonological representations or is active only at the phonetic level.

Vowel shifts fall into a few limited types. The most important generalizations about the direction of vowel shifts is that tense (or "peripheral") vowels tend to be raised, lax (nonperipheral) vowels tend to fall, and back vowels tend to be fronted (Labov 1994). How can we explain these canonical types of vowel shifts, and the direction of strengthening processes in general? The attempt to answer this question will reveal another kind of top-down effect.

One of the puzzling questions about vowel shifts is their "perseverance" (Stockwell 1978). What accounts for their persistent recurrence in languages such as English, and their rarity in others, such as Japanese?[21] A simple argument shows that tenseness-triggered raising and laxness-triggered lowering occur only in languages which have both tense and lax vowels in their inventories at some phonological level of representation. Otherwise, we would expect languages with persistent across-the-board lowering of all vowels (if they are lax) or persistent across-the-board raising of all vowels (if they are tense). But there do not seem to be any such languages.

But why would the shift-inducing force of the feature [±Tense] depend on the existence of both feature specifications in the language's vowels? A reasonable hypothesis would be that vowel shifts are the result of a tendency to maximize perceptual distinctness. Consider first the idea that vowel shifts are the result of the enhancement of contrastive features, in this case, tenseness. This hypothesis is undermined by several facts. First, vowel shifts often cause mergers, both through raising of tense vowels (as in English *beet* and *beat*) and through lowering of lax vowels (as in Romance). If the motivation is the maximization of distinctness, why does this happen? Second, even when vowel shifts do not cause mergers, they often simply produce "musical chairs" effects, chain shifts of vowels which do nothing to enhance their distinctness (for example, the Great Vowel Shift). Third, tenseness does not by any means have to be distinctive in order to trigger vowel shifts. In English, for example, tenseness has been mostly a predictable concomitant of the basic quantitative opposition of free and checked vowels, and at some stages it has been entirely that. Yet tenseness is the feature that seems to have triggered the various phases of the Great Vowel Shift. Moreover, those vowels for which tenseness did have a distinctive function do not seem to have shifted any more than the ones for which it did not.

The alternative hypothesis which I would like to explore here is that tenseness can trigger vowel shift if it is present in the language's phonological representations - not necessarily underlyingly, but at any phonological level where it can feed the phonological rules that assign default values for the height features. Vowel shifts can then be considered as the result of suppressing marked specifications of the relevant height feature in lexical representations, resulting in the assignment of the appropriate default value of the feature in question to the vacated segment by the mechanisms discussed above. For example, loss of the feature specification [-High] from a tense vowel will automatically entail its raising by default. The reason why tenseness and laxness activate vowel shifts only if they are both present in the language's phonological representations would then be that, as the theory predicts, only those feature values which are specified in phonological representations can feed default rules, and a feature that plays no role whatever in a language's phonology will not figure in its phonological representations, but will be assigned at a purely phonetic level if at all. This would mean that an abstract distinction at yet another level, that between phonetic and phonological tenseness/laxness, would also be critical to sound change.[22]

Let us see how this approach might work for the Great Vowel Shift. As-sume, fairly uncontroversially, that height is assigned by the following universl default rules.[23]

(18) (a) [−Tense] → [−High]
(b) [] → [+High]
(c) [] → [−Low]

In a language where tenseness plays no role, (18a) is not active, and default height is assigned only by the "elsewhere" case (18b). The canonical three-height vowel system is represented as follows:

(19)

	Distinctive value	Default values (assigned by [18b])
High vowels (i, u)	[ ]	[+High, −Low]
Mid vowels (e, o)	[−High]	[−Low]
Low vowels (æ, ɔ)	[+Low]	

To augment the system with the feature [±Tense], I'll assume the classification of vowels motivated in Kiparsky 1974.[24]

(20)

		−Back −Round	+Back −Round	+Back +Round
+Hi, −Low	+Tense	i	ɯ	u
	−Tense	ɪ	ɨ	ʊ
−Hi, −Low	+Tense	e	ɤ	o
	−Tense	ɛ	ʌ	$ɔ_1$
−Hi, +Low	+Tense	æ	ɐ	$ɔ_2$
	−Tense	a	ɑ	ɒ

Tenseness itself is related to length by the following default rules:

(21) (a) VV → [+Tense]
(b) V → [−Tense]

Now we are ready to lay out the vowel system of late Middle English (ME) (ca. 1400). At this stage, all front vowels were unrounded and all back vowels were rounded. So ME ā, *a* were low nontense *front* vowels, like the [a] of Boston *car, father* and of French *patte* (Dobson 1968, p. 545, 594). The distinction between free and checked nuclei appears to have been basically quantitative (long versus short). Tenseness was distinctive, however, in the long mid vowels (*beet* vs. *beat, boot* vs. *boat*). I will assume that *all other vowels were nontense*. The vowel specifications were accordingly as follows (default and redundant features parenthesized):

(22)

		−Back, (−Round)		+Back, (+Round)	
		Long	Short	Long	Short
(+Hi, −Low)	(−Tense)	ī bite	ɪ bit	ū bout	ʊ but
−Hi, (−Low)	+Tense	ē beet		ō boot	
	(−Tense)	ɛ̄ beat	ɛ bet	ɔ̄ boat	ɔ pot
+Low	(−Tense)	ā bate	a bat		

The default values for the features High and Low are assigned by (18). Tenseness plays no role in the assignment of vowel height. Only the default rule (21b) is active, assigning the feature specification [−Tense] to vowels not lexically marked as [+Tense].

Tenseness was neutralized in short vowels; hence [ɛ] represents both shortened [ē] (*kēp:kɛ̆pt, mēt:mɛ̆t*), and shortened [ɛ̄] (*drɛ̄m:drɛ̆mt, lɛ̄p:lɛ̆pt, clɛ̄n:clɛ̆nliness*), and [ɔ] represents both shortened [ō] (*lose:lost, shoot:shot*) and shortened [ɔ̄] (*clothes:cloth, nose;nozzle, prōtest:prŏtestation*)

The ME diphthongs were:[25]

(23)

ay	*bait*	aw	*law*
ɔy	*boy*	ɔw	*blow*
		ɛw	*dew*
ʊy	*buoy*	ɪw	*pew*

According to the analysis of the historical records by Dobson (1968), the vowel shift took place in three stages, from our perspective consisting of two height shifts with an intervening tensing process:

(24)

Middle English	Raising (≈ 1500)	Tensing (≈ 1650)	Raising (18th c.)
ī	ei		
ū	ou		
ē	ī		
ō	ū		
ɛ̄		ē	ī
ɔ̄		ō	
ā		ǣ	ē
a		æ	

First shift: Raising. In the first stage of the vowel, shift, which Dobson dates to the 15th century, [ē] and [ō] (the only tense vowels of the system according to our assumption) were raised (unmarking of [–High] and default assignment of [+High] by (18b)), and [ī] and [ū] were diphthongized (activation of (18a)). *Second shift: Tensing*. The next phase of the Great Vowel Shift (17th century) was a general tensing of the long vowels: [ɛ] was tensed to [ē], [5] was tensed to [ō], and long and short [a] were tensed to [æ]:[26]

(25)

		–Back, (–Round) Long Short	+Back, (+Round) Long Short
(+Hi, –Low)	(–Tense)	ī bite ɪ bit	ū bout ʊ but
–Hi, (–Low)	+Tense (–Tense)	ē beet ɛ̄ beat ɛ bet	ō boot ɔ̄ boat ɔ pot
+Low	(–Tense)	ā bate a bat	

The tensing process can again be seen as an activation of a default rule, in this case (21a). We have now arrived at system of long and short vowels (25) where *tenseness is entirely predictable*. Yet tenseness in this system feeds the next, third stage of vowel shift, which again raises tense vowels.

Thrid shift: Raising with merger. The second raising of tense vowels (18th century) again implements the default rule (18), which assigns height on the basis of tenseness. But this raising was more restricted, applying only to the long tense front vowels: [ē] was raised to [ē] (loss of [–High]), and [æ] was raised to [ē] (loss of [+Low]). This stage of raising differed from the first in that the resulting vowels merged with existing nuclei (the reflexes of ME /e/ and /ai/, respectively). Moreover, not all dialects underwent this change, and words such as *great, steak, break* retaining the older mid vowel in the standard language are probably from those dialects.

To sum up: the Great Vowel Shift is triggered by both distinctive and nondistinctive tenseness. Evidently it is not the distinctiveness of the feature but its phonological (as opposed to phonetic) status that counts. This supports the idea adopted in lexical phonology that the assignment of phonological default features can take place at several levels of the derivation, including in particular post-lexical phonology.

4 Conclusion

I have defended the neogrammarian hypothesis that sound change is exceptionless and subject only to phonetic conditioning against two potentially serious objections. The first objection, based on lexical diffusion, is answered by the analysis of the phenomenon as a species of nonproportional analogical change proposed and motivated in section 1. The second objection is based on top-down effects in sound change. Structural work in historical phonology in the Jakobsonian tradition supports the position that phonological organization plays a role in sound change, in particular through diachronic "conspiracies" implementing canonical syllable structure. In section 2, I discussed two other types of structure-dependency in sound change: priming effects in secondary split, and maintenance of universal constraints on phonological systems (e.g., the stability of implicational universals, and the failure of cascades of secondary splits to produce giant phonemic systems). Finally, in section 3, I discussed the role in sound change of the status of features as distinctive versus redundant, and phonological versus phonetic, drawing in part on the parametric rule typologies emerging from recent work on natural phonological processes, which make use of abstract properties of phonological representations to explain generalizations in domains where purely physical explanations have hitherto dominated. I argued that all four types of top-down effect can be reconciled with exceptionlessness by giving the transmission process an active selectional role in language change.

1 This paper is in part the result of an exchange with Andrew Garrett and of a reading of portions of Labor 1993 in draft form, though neither Garrett nor Labov necessarily agrees with me, or with the other. I am also grateful to them both as well as to the other participants of a workshop on sound change at Stanford University in February 1993 for valuable comments on a draft of this paper.

2 Such a move is of course legitimate in so far as the exceptions can be identified in some principled way, as when “minor sound changes” such as dissimilation and metathesis are systematically set aside as being of perceptual origin.

3 I exclude here from sound change the “minor” sound changes discussed below in section 2.2 Also, the “no-change” entries in the last line abstract away from *lexical split*, which can result from sound change by the mechanisms discussed at the end of section 2.2 (e.g., *ass/arse*), by analogy (*staff/stave*), and, I would expect, from lexical diffusion as well.

4 However, no commitment to any particular formal evaluation measure need be made at this level. Virtually any theory which characterizes analogy as structural optimization ought to be able to get the same results.

5 See Steriade (1987), Archangeli (1988), and Mohanan (1991) for general surveys from varying points of view. For simplicity of presentation, I will illustrate the point here with segmental features. But everything I say holds equally for other phonological information such as syllabic structure and stress (Kiparsky 1993).

6 This is not how such a rule would actually look. I give it in this old-fashioned form just for simplicity's sake.

7 For two other formulations, see Kiparsky (1982, 1985) and Archangeli (1984), Archangeli and Pulleyblank (1989). The position put forward here is in a sense intermediate between those two.

8 The elimination of MSCs invalidates the objection to underspecification by Christdas (1988), Clements (1985), Mohanan (1991), and others based on the claim that that Morpheme Structure Constraints must be able to refer to default values. The objection is in any case internally incoherent because many of the MSCs cited by these authors require reference to syllable structure assigned by phonological rules, so they couldn't possibly apply to underlying forms. All that these examples show is that level 1 phonological rules in some languages require reference to both feature values. But radical underspecification predicts exactly that because it says that default values can be assigned cyclically, a possibility independently motivated by the cyclic interaction of default and spread rules in harmony systems.

9 The affective words *oops, whoopee*, and *shtup* are the only exceptions I am aware of.

10 Labov (1993) records one token of lax æ in *athlete*; this could be the result of lexicalization of the trisyllabic pronunciation with anaptyctic ə.

11 Another apparently idiosyncratic contrast is reported by Labov in hypocoristic names, where *Frannie, Danny, Sammy* normally have tense A and *Cassie, Cathy* normally have lax æ. This could be accounted for on the assumption that the former are analyzed, by speakers who have this contrast, as derived from monosyllabic bases (*Fran, Dan, Sam*), to which the rule applies regularly on the first cycle, whereas the latter are treated as unanalyzed. So even these seemingly unpredictable cases may well turn out to be rule-governed.

12 For the three-*d* words in [12b], the tensing is now obligatory for Philadelphia speakers of all ages.

13 On the other hand, this derivation of the New York pattern would also explain the relatively high rate of tensing/raising before š compared to other fricatives in New York, by the relatively high rate of post-lexical tensing/raising before š compared to other fricatives in the Northern Cities (Kiparsky 1971).

14 For example, Wanner and Cravens (1980) argue for the lexical diffusion of an intervocalic voicing rule in the Tuscan dialect of Italian.

15 As early as 1886, Kruszewski had cited Darwin on “directed evolution” in order to explain why sound changes, though originating in random articulatory fluctuations, progress in specific directions (“sich in

bestimmter Richtung auf der erwähnten Linie fortbewegen").

16 The only contrary case I know of, where compensatory lengthening is reported to have created distinctive length, is Occitan (Morin 1992).

17 Jakobson, Fant, and Halle (1952, p. 8) note that redundant features may under certain conditions even substitute for the conditioning distinctive features.

18 If the devocing in /bit+z/ → [bits] were a genuine case of assimilation, it would refute the theory. In fact, it appears to reflect a phonetically-based constraint (as far as is known, valid in all languages) which restricts voicing to a continuous portion of the syllable that includes the nucleus (Cho 1990).

19 Place neutralization yields coronals. E.g.,: "For Fante, the pattern of nasal plus consonant may be stated as involving homorganicity with the predominant articulation if any, or otherwise [n]" (Welmers 1973, p. 65). A similar pattern of nasal place neutralization to [-n] (with or without concomitant assimilation) is found in Finnish, Greek, and Italian, and reportedly in Croatian dialects, Avar and Lakk. With debuccalization, the result is a placeless nasal (Sanskrit *anusvāra*), see Ferré (1988) and Paradis and Prunet (1991); apparent neutralization to [ŋ] is via coronal or placeless nasals.

20 The argument of Steriade (1987) that contrastive underspecification is to be preferred over radical underspecification is based entirely on the following important generalization about transparency: a feature spreads only through segments for which the feature in question is redundant, never through segments for which it is distinctive and which have the default value of the feature. But this follows from the assumption that all segments for which a feature is distinctive bear a class node for that feature, together with normal locality considerations. So, contrary to what Steriade implies, her generalization is fully consistent with radical underspecification.

21 It is true that the Okinawan dialect has undergone a kind of vowel shift (M. Matsuda, *in litt.*). However, this was apparently a raising of the *short* vowels *e, o* to *i, u*, their long counterparts remaining unaffected. So on my assumptions, tenseness cannot have been the triggering factor of this change. Rather, I assume that it is a vowel reduction phenomenon, consisting of the neutralization of the distinctive feature [-High], with the neutralized vowels assigned default [+High] by rule (18b) below.

22 The same issue arises in the case of the feature of nasality. According to Schourup (1973) and Ruhlen (1978), whether nasal vowels are raised or lowered depends on whether nasalization is distinctive in the language or not. However, it is not impossible that the relevant distinction is really whether nasalization figures in the language's phonological representations or not.

23 I assume that default rules operate in gradient fashion at the level of phonetic implementation, in this case accounting for the general tendency for lax vowels to be articulated lower than tense vowels.

24 I have left out the front rounded vowels in this version of the chart because they play no role in the English data discussed here.

25 The diphthong [uy] (*buoy, boil, oil*) merged with [ɔy] (*boy, choice, noise*) in most dialects in the ME period. The other old diphthongs were eliminated as part of the vowel shift as follows. ME [ay] merged with ME [ā] and [ɛw] with [Iw] about 1650, earlier in Northern and Eastern dialects (Dobson 1968, p. 594, 778, 798). The diphthong [aw] (*law*) was monophthongized to [ā] in the 17th century (p. 786), and [ɔw] (*blow*) was monophthongized to [ɔ̄], merging with the vowel of *boat* ca. 1600 (p. 805).

26 I am here departing from Dobson's chronology by assuming that long and short [a] were tensed at the same time along with the other long vowels. Dobson (1968, p. 594) thinks that long [ā] was tensed earlier than short [a] was, as early as the 15th century, which would make this part of tensing part of the first shift. Adopting his account would make the first shift more complex but not alter my main point that vowel shift is an unmarking of vowels with concomitant assignment of default values to the vacated features. Since the orthoepic evidence does not seem altogether clear on this point, I have assumed that the tensing processes were concurrent, which gives the simpler schema in (24).

Cite this article

KIPARSKY, PAUL. "The Phonological Basis of Sound Change." *The Handbook of Phonological Theory*. Glodsmith, John A. Blackwell Publishing, 1996. Blackwell Reference Online. 31 December 2007

<http://www.blackwellreference.com/subscriber/tocnode?id=g9780631201267_chunk_g978063120126723>

Bibliographic Details

The Handbook of Phonological Theory

Edited by: John A. Glodsmith
eISBN: 9780631201267
Print publication date: 1996