5

Formal Universals as Emergent Phenomena: The Origins of Structure Preservation

Joan L. Bybee
University of New Mexico

5.1 INTRODUCTION

All explanations for linguistic phenomena, both universal and language-specific, must necessarily have a diachronic dimension, since all linguistic phenomena have histories which determine their present conventionalized state. With respect to language universals—more appropriately called "cross-linguistic similarities" since there are so few absolute universals—I have argued that an explanation is not valid unless it can be demonstrated that the explanatory principle is actually at work in the mechanism of change that brings about the cross-linguistic pattern (Bybee 1988b). Taking the role of diachrony one step further, one could argue that since there are so few absolute universals, identifying the mechanisms of change behind cross-linguistic patterns will lead us closer to an understanding of the factors that produce cross-linguistic patterns, and these factors, I would maintain, are the only true universals of language in the sense that they operate in all languages at all times.

Thus, the focus for establishing the explanations for cross-linguistic similarities should be on the mechanisms of change (Bybee et al. 1994; Bybee 2006b). Identifying the causal mechanisms in change requires a detailed look at all the properties of a change—including its directionality, gradualness, spread through the community and through the lexicon—as these properties may give clues to the mechanism involved. For instance, the lexical diffusion of a phonological change gives important clues as to its causes: change taking place in phonetic environments affects high-frequency words before low-frequency, pointing to the automatization (with practice) of the neuromotor sequences involved in the production of the word; in contrast, change that leads to the regularization of paradigms affects lower-frequency items first, pointing to the mechanism of analogical change used when low-frequency forms are not readily accessible (Hooper 1976b; Phillips 1984; Bybee 2001).

A central tenet of Usage-Based Theory is that structure is created as language is used. In the preceding example, both neuromotor practice and analogy are processes that occur in individual usage events. With multiple applications of a mechanism within and across individuals, a change might progress to the extent that it is noticed by linguists and by speakers.

What are the stages in a "usage event" where change might occur? These include the selection of expressions, lexical access by the speaker, articulatory production, perceptual decoding, lexical access by the hearer, categorization, assignment of meaning and inference-making. All of these operations have certain inherent tendencies towards change, especially upon repetition. With repetition by different speakers, these tendencies can develop into a noticeable linguistic change (Bybee et al. 1994; Bybee 2001; Pierrehumbert 2001; for the speaker as the locus of language change, see Keller 1990/1994 and Croft 2000).

As change is initiated and carried out by the same mechanisms across languages, we find very similar paths of change in related and unrelated languages at different times. This has been noted in phonology (Foley 1972; Mowrey and Pagliuca 1995; Blevins 2004a), where, for instance, a common path of change for voiceless labial [p] is cross-linguistically documented as [p] > [pf] > [f] > [h] > zero. In the grammatical domain, such paths of change are also amply documented: e.g., constructions signaling movement towards a goal become futures, verbs meaning 'finish' become perfects and pasts, a coordinate clause can become subordinate, a verb becomes an auxiliary, and so on (Lehmann 1982/1995; Heine and Reh 1984; Heine et al. 1991; Hopper and Traugott 1993; Bybee et al. 1994).

The Greenbergian theory of language universals (Greenberg 1969, 1978a, 1978b) views language as a complex system. The synchronic cross-linguistic patterns are not the end point of universals research, but just the starting point: synchronic patterns are the result of movement along these common paths and underlying the paths are certain recurring mechanisms of change, which have the following properties:

1. mechanisms of change are universal in the sense that they can be found operating in all languages at all times;
2. they are relatively few in number;
3. they involve neurocognitive tendencies that manifest themselves as language is produced and processed;
4. they apply during individual usage events; and
5. the cumulative effect of their application over multiple usage events creates grammar.

This view is consonant with the theory of complex systems, in which the systematic structure of language is considered to be continually evolving through the ongoing application of processes during multiple usage events. Grammar (the cognitive organization of language) is thus said to be "emergent" rather than fixed. The ability to create language systems through categorization, analogy, neuromotor automatization, semantic generalization, and pragmatic inferencing derives from the innate neurocognitive capacities of human beings. These are largely domain-general capacities that happen to be used to create language. The hypothesis is that there is no need to posit innate linguistic universals, but rather that the similarities that exist across languages can be explained through the interaction of a small number of mechanisms of change.

The complex system view contrasts with that of Kiparksy (this volume), who distinguishes between patterns created by change and generalizations written into Universal Grammar (or innate generalizations). Note that Kiparsky's theory considers some cross-linguistic generalizations to be universals of grammar, while the argument to be pursued here is that the deeper level of explanation requires understanding the mechanisms of change.

The explanatory power of diachronic typology is also demonstrated in the chapters in this volume by Hopper and by Kuteva and Heine. Hopper demonstrates that an understanding of a well-established grammatical pattern in many languages (verb serialization) can be fruitfully studied in languages where it is only a minor tendency (such as English) and that a thorough, discourse-based analysis sheds light on the origins of the construction type. Kuteva and Heine show that given the set of mechanisms behind grammaticization, both generalizations and exceptions can be explained. While the mechanisms are applicable in all languages at all times producing the common paths of change as illustrated above, these mechanisms also sometimes produce other outcomes, making it possible to have other, minor paths of change as well, depending upon their interaction and the type of linguistic material they apply to.

In this chapter, I will illustrate the relationship among synchronic universals, paths of change, and mechanisms of change with respect to the phonological changes that create the structural tendency known as Structure Preservation in Lexical Phonology. The outline of this explanation is given in Bybee (2001: 214–215), but here it is worked out in more detail.

5.2 SUBSTANTIVE AND FORMAL UNIVERSALS

Substantive universals are those cross-linguistic tendencies that involve either phonetic or semantic substance, while formal universals are those tendencies that involve grammatical form or the structure of the grammar (Chomsky and Halle 1968). Paths of change can be categorized in the same way. Paths that specify changes in phonetic substance, such as the reduction of a voiceless labial stop shown above, are substantive. Within the framework of grammaticization, examples of substantive universals are those paths of change involving meaning, such as the generalization that anteriors (perfects) become perfective or past. A parallel formal universal involves the cline discussed in Givón (1979) and Hopper and Traugott (1993), by which a content item becomes a grammatical word, then a clitic, and then an affix. When these substantive and formal paths of grammaticization operate simultaneously, the result is a perfective marker that is an affix. The operation of these two paths accounts for the fact that with few exceptions, the perfective and past are marked with affixes (Bybee and Dahl 1989; Bybee et al. 1994).

However, formal universals could also refer to the form of the grammar, as in properties such as modularity. Such properties are usually considered to be given innately as a starting point for language acquisition (Kiparsky, this volume). However, it is also possible that the general structure of modularity is emergent from the nature of change. That is, certain recurring, parallel paths of change create patterns that are largely modular. Under this view, there would be transitional phases between modules, i.e., exceptions to the strict separation of levels. Given that there is ample evidence that such exceptions exist—phonological alternations dependent upon morphology and syntax, as well as morphological and syntactic alternations dependent upon phonology—examining an emergentist view of such separations becomes a necessary endeavor.

The main focus of the current chapter is the principle of Structure Preservation, which deals with the distinction between contrastive and non-contrastive segments and has been formulated as a structural universal of language (Kiparksy 1985). By examining a case which creates difficulties for this principle, I show that this proposed structural universal is in fact emergent from the parallel development of three unidirectional paths of change, propelled by certain mechanisms of change, which are universals in the sense that they apply in all languages at all times.

5.3 A FORMAL UNIVERSAL: STRUCTURE PRESERVATION

Structure Preservation is a principle formulated in Lexical Phonology (Kiparsky 1985), though it reflects a principle recognized in earlier structuralist theories. The principle states that only contrastive sounds or features take part in morphologically or lexically conditioned alternations; or, stated differently, alternations that are restricted to the word level involve only contrastive features. Segments or feature combinations that are non-contrastive must be introduced by postlexical rules, which are automatic and phonetically conditioned and often apply across word boundaries.

This principle correctly captures a strong tendency in the languages of the world for alternations conditioned either lexically or morphologically (or both) to involve contrastive features and segments. Consider for example two alternations that English /k/ enters into: in some words with Latinate affixes /k/ alternates with /s/, as in *electri[k], electri[s]ity; criti[k], criti[s]ism*. This alternation applies at the word level: it does not occur when two words come together; it is unproductive, lexically restricted and at least partially morphologically conditioned. In contrast, English /k/ also has a palatal variant [c] before a front vowel, as in *key, kiss, came*. This variant appears automatically (that is, the process that creates it is productive), and it is not lexically or morphologically restricted. In principle it could apply when two words come together, as in *break even*, though I know of no phonetic studies that show that this is the case. This sort of situation—where phonemes alternate when there are lexical or morphological restrictions and non-contrastive elements alternate in purely phonetic environments—is typical of the phonologies of the languages of the world. Kiparksy designates it as "Structure Preservation" because the lexical phonological rules do not introduce any feature combinations that are not already present in the lexicon. That is, a lexical phonological rule could not introduce a palatal stop into the English lexicon.

A principle with the same effect was discussed in American structuralism under the rubric of "separation of levels". The phonemes of a language together with their allophones could be arrived at using only phonetic information (Hockett 1942). Once the phonemes were established by phonetic principles such as complementary distribution, then alternations among phonemes in words could be discovered.

Early in this discussion Pike (1947, 1952) correctly noted that using only phonetic information to predict the occurrence of allophones was impossible both in terms of procedure and in terms of theory. He noted in particular that the behavior of phonemes at junctures (boundaries) could only be predicted on the basis of grammatical and lexical information, not purely phonetic information. He brings up the contrast between *nitrate* [najt^{h}rejt] and *night rate* [najt˺rejt], in which the allophone of /t/ that is used depends upon knowing that the /t/ in the latter phrase occurs at the end of a word. This case is not particularly a problem for Kiparsky's formulation as long as word boundaries can block the postlexical rule of aspiration.

Counter-examples to Structure Preservation have also been discussed (see below). We are presented, then, with a typical dilemma in linguistic theory: a strong tendency is evident in the grammars of all languages encountered; it seems to represent a basic organizing principle of language and yet, if it is canonized as a structural principle, counter-examples or exceptions quickly come to light. In the face of exceptions, researchers try to revise the principle or reanalyze the counter-examples. A question that rarely arises, however, is why grammars would have such an organizing principle. I suggest that if we take explanation as the primary goal and set about trying to understand why this strong tendency exists, and how it arises in languages, we can explain not only the general tendency but the exceptions as well, and gain further insight into the nature of grammar.

A famous counter-example to the Structure Preservation principle (discussed extensively in the structuralist and the generativist literature) is the alternation between German [x] and [ç] (Moulton 1947; Leopold 1948; Hall 1989). The basic facts are these (examples from [Hall 1989]): [x] occurs after back vowels; [ç] occurs after front vowels and /n/, /r/, and /l/.

(1)	Buch	[bux]	'book'	siech	[zi:ç]	'sickly'
	Koch	[kɔx]	'cook'	Pech	[pɛç]	'bad luck'
	nach	[nax]	'after'	Köchin	[kœçɪn]	'cook (fem)'

However, the diminutive suffix *-chen* is always [çən]:

(2)	Kuhchen	[ku:çən]	'little cow' (*Kuh* + *chen*)
	Tauchen	[taoçən]	'little rope' (*Tau* + *chen*)
	Pfauchen	[pfaoçən]	'little peacock' (*Pfau* + *chen*)

The invariant form of the diminutive despite the preceding vowel produces phonemic contrasts with the following words:

(3)	Kuchen	[ku:xən]	'cake'
	tauchen	[taoxən]	'to dive'
	pfauchen	[pfaoxən]	'to hiss'

In addition, assimilated borrowings use the palatal fricative in word-initial position.

(4)	Chirurg	[çirʊrk]	'surgeon'
	Chemie	[çemi:]	'chemistry'
	Cholesterin	[çolɛsteri:n]	'cholesterol'
	Fotochemie	[fo:to:çemi:]	'photochemistry'

The dilemma is that one would like to analyze the velar and palatal fricatives as allophones of the same phoneme, with the palatal being the mere output of a postlexical rule (if you choose the velar as the underlying phoneme), but that pesky diminutive suffix makes such an analysis impossible. Moreover, borrowed words with the palatal fricative in word- and syllable-initial position create additional problems. In pregenerative structuralism, Moulton (1947) argued in favor of a segmental juncture phoneme preceding the /x/ to condition the palatalization. However, in the diminutive suffix, since no pause is present, this juncture has a zero realization. Leopold (1948) argues that such an analysis is circular since we only know that the juncture is there because the [ç] appears. Similar discussions in Lexical Phonology (Hall 1989 and MacFarland and Pierrehumbert 1991) also lead to lack of consensus.

This case lends itself nicely to a diachronic explanation: the alternation started out as phonetically conditioned, as the older form of the suffix *-ichiin* had the front vowel conditioning context within the suffix. Apparently the palatal variant was established before the first vowel was lost, so that it remained despite the loss of its conditioning environment. Now [ç] has gradually become established as an independent element (or phoneme) in the diminutive suffix and has also been recruited for use in loanwords. Moreover, in some dialects the palatal variant is now prepalatal [ʃ], indicating that the phonetic distance between the two originally predictable variants has also increased. The "exception" is really a kind of intermediate case, and as such has a diachronic explanation, but does this contribute to our understanding of the synchronic principle? I will argue in the remainder of the paper that indeed it does. I will argue that the general principle is not a synchronic organizing principle of grammar, rather a general tendency that results from the coevolution of phonological changes along several parallel paths. I argue, then, that both the general principle and the exceptions to it have diachronic explanations.

5.4 THREE UNIDIRECTIONAL PATHS OF CHANGE IN PHONOLOGY

Three well-documented universal paths of change occur in parallel and lead to the synchronic situation that is described as Structure Preservation. First, phonetically conditioned sound change creates alternations that gradually acquire morphological or lexical conditioning (Vennemann 1972; Hooper 1976a; Dressler 1977, 1985).

(5) phonetic conditioning > morphological or lexical conditioning

Second, what starts as a small phonetic change tends to continue to change phonetically over time, leading to a greater distance between the original sound and the resulting one. Thus the two alternating sounds grow more different from one another (Hooper 1976a; Janda 1999).

(6) small phonetic change > larger phonetic change

For instance, a [k] before a high front vowel might move forward to a palatal position. The extent of palatalization might increase until the sound in that context becomes an alveo-palatal affricate. Such changes are documented in Romance languages, where Latin /k/, going through stages such as [tʃ], [ts], ends up as [s]. This created the alternation discussed above, between /k/ and /s/, that was borrowed into English along with the French words.

Third, simultaneous with the preceding developments, productive phonetically conditioned alternations between two sounds are likely to become unproductive. This path of change is related to the previous two in ways that will be discussed below. An example would be the loss of intervocalic voicing of fricatives in English; this process created the *wife*, *wives* alternation, but now is no longer productive, as voiceless intervocalic fricatives are allowed in English (e.g., *classes*).

(7) productive processes > unproductive

These paths of change together result in Structure Preservation, since as a change begins to take on morphological and lexical conditioning, the new variant tends to grow more distant from its source, producing a larger phonetic change, one that could be phonemic. Simultaneously, its tendency to become lexicalized and to cover a larger phonetic distance leads to the loss of productivity and the ability of the new sound to occur in contrast with the original sound. The actual mechanisms behind these paths of change are discussed in the following sections.

5.5 A MODEL OF SOUND CHANGE

This section proposes a series of steps by which sound change takes place and presents a model that accounts for the phonetic gradualness of sound change, the lexical gradualness of sound change, and the eventual result that only contrastive elements occur in morphologically and lexically conditioned alternations (Bybee 2001). The model is usage-based, in the sense that cognitive representations are affected by usage events and are emergent from them. In this model, a principle such as Structure Preservation is not in itself an organizational principle of language, but rather the result of the interaction of the more basic mechanisms of change that are operative when language is used.[1]

In keeping with the usage-based viewpoint, sound change is viewed as the result of the reduction or retiming of gestures that occurs with automation of production in language use (Browman and Goldstein 1992; Mowrey and Pagliuca 1995). Sound change is manifested early on as variation in casual speech. Such variation is influenced by the phonetic context and eventually results in allophonic variation which can become quite stable. In this view, sound change is largely, if not wholly, phonetically conditioned and reductive (see the authors mentioned above and Bybee 2001 for more discussion). Note that it is phonetically conditioned sound change that creates allophones of phonemes; it follows then that in general the distribution of allophones can be stated in purely phonetic terms.

Since sound change occurs as language is used, sound change takes place in actual production units, i.e., words and phrases. The evidence for this claim is the fact that articulatorily motivated sound change takes place earlier in high-frequency words than in low-frequency words (Fidelholtz 1975; Hooper 1976b; Phillips 1984, 2001; Bybee 2000b, 2002).[2] In order to account for this lexical diffusion phenomenon, the immediate effects of sound change are registered in an exemplar representation (Johnson 1997; Pierrehumbert 2001, 2002). Exemplar representations allow a cluster of phonetic variants for a word and this cluster is constantly being updated as new variants are experienced. Cole and Hualde (1998), and Booij (to appear) argue that the fact that the effects of sound change are never reversed provides evidence for the hypothesis that sound change has an immediate and permanent effect on the memory representation of words. These researchers point out that when a sound change or the alternation it sets up has ceased to be productive, the change is not undone, as one might expect in a theory in which underlying forms remained unchanged and only surface forms are affected by the sound change (i.e., where sound change is rule addition). Thus, Booij points out that long vowels created by lengthening in open syllables in

[1] Blevins (2004: 244ff.) also notes that Structure Preservation can be derived from phonologization. My account here differs from hers both in fleshing out the details and also in the mechanisms of change that are proposed.

[2] As Hooper (1976b), Bybee (2001), and Phillips (1984, 2001) have shown, changes that affect low-frequency words first are not due to articulatory reduction, but result from other mechanisms of change.

Dutch do not shorten again when this rule becomes unproductive; rather, they stay long.

Other evidence that the results of productive processes are registered lexically is that the phonetic form of existing words can be used in the creation of new words, a phenomenon which would not be possible if only phonemic forms were stored in memory. Steriade (2000) (also arguing against a strict distinction between phonetic and phonological features) notes the difference in the medial coronal consonants in *fatalistic*, which has a flap, and *positivistic*, which has a [t]. The difference corresponds to the pronunciation of the base word, *fatal* with a flap, and *positive* with a [t]. This distinction suggests that the mental representation of *fatal* has a flap in it, as does the experimental evidence of Connine (2004).[3] Similarly, in compounds such as *night rate* the [t] is phonetically the same as it would be if it were word-final. Of course, one can derive this effect by placing a word boundary in the compound, but what it really means is that the compound is formed by using the phonetic shape of the word *night* rather than some more abstract phonemic shape.

Immediate registration of sound change in words also accounts for the tendency for phonetic change to become lexically and morphologically conditioned, as we will see in section 5.7.

If a change is occurring in a number of words, the general neuromotor routine that governs the gestural sequence is gradually changing, too. This accounts for the automatic nature of phonetically conditioned alternations, that is, the fact that they apply to new or nonce words. The general neuromotor routine itself is not static, but allows for a range of variation and may be biased towards lenition or anticipation of gestures. (See Pierrehumbert 2002.)

5.6 CATEGORIZATION OF PHONETIC VARIANTS

According to Miller (1994), phonetic variants are categorized by phonetic similarity and organized around a best exemplar, or the variant that speakers judge to best represent the category. Speakers can make such judgements appropriate to different phonetic contexts (e.g., English [t] after [s] vs. English aspirated [t^h]), suggesting that phonetic categories may correspond more to "allophones" than to "phonemes". Over multiple instances of exemplar categorization, a continuous parameter with a bimodal distribution can sharpen and separate into distinct categories. Wedel (2006) points out that in an exemplar model that includes both perception and production and models sound–meaning correspondences, overlapping or intermediate stimuli tend to be lost because their categorization is less consistent. Stimuli or tokens close to the centers of

[3] Steriade accounts for this phenomenon by using a constraint labeled Paradigm Uniformity. My account needs no such constraint; registering the variants in memory storage has the desired effect. See Garrett (this volume) for a different critique of such proposed Paradigm Uniformity constraints.

categories are more consistently classified than tokens near the boundaries between categories; as a result two nearby categories tend to diverge from one another. In addition, since marginal and infrequent members of categories tend to be lost over time, categories that once had overlapping members can evolve into distinct categories with no overlaps.

Given these categorization effects, the range of phonetic categories used in a language is dynamic and changeable, but not infinite, giving rise to a set of phonetic categories for allophones, many of which are also linked to specific phonetic environments, and thus are considered allophonic in phonological theory. Related advantages are the resulting limited set of general neuromotor routines that are used repeatedly in different words. As Lindblom (1992) and Studdert-Kennedy (1987, 1988) point out, if each word had its own unique set of gestural features there would be a strict limit on the number of words a language could have. In order to acquire and maintain an unlimited lexicon, a constrained set of gestural configurations must be reused in the words of a language. Presumably the same would hold for the perceptual configurations. Both production and perception are made more efficient by the use of a constrained set of units for all the words of a language.

For our purposes, what is most interesting here is that new categories can be formed if phonetic variants in different contexts start to differentiate. In the formation of new contextual categories, intermediate variants tend to be lost. For example, though the American English alveolar flap was originally a variant of the /t/ or /d/ categories, it has now formed a distinct phonetic category that is contextually restricted. In these contexts the new best exemplar is the flap and a full [t] or [d] does not occur in natural speech. The current range of variation for this category contains the flap and further weakened versions of it, but excludes full [t] and [d]. One can of course produce an aspirated [t^h] in a word such as *butter*, but that is done by accessing a different category.

5.7 SOUND CHANGE HAPPENS TO WORDS

The lexical storage unit that is relevant for the phonetic categories of the language is the word or phrase. This is also the unit of production to which neuromotor routines apply. Thus words tend to have constrained ranges of variation unless they are of very high frequency, in which case they may have variants specific to certain phrases.

I have argued that sound changes that take place at word boundaries show the tendency for a word to have a small range of variation: alternations created at word boundaries tend to be resolved in favor of the variant that occurs in preconsonantal position (Bybee 2000a, 2001). For example, in the reduction of Spanish syllable-final /s/ to [h], word-final /s/ at first shows variation according to the phonetic environment, with [s] occurring before vowels and [h] before consonants. Later stages show [h] extended to the majority of word-final tokens, even those with a following vowel.

Thus *más o menos* becomes [mahomenoh] 'more or less'.[4] The representation for the word *más* at one point had a large range of variation, occurring with final [s] and final [h] and many variants in between. Since the [h] variants were more frequent, as the following word would begin with a consonant twice as often as a vowel, the more marginal [s] variants were lost and the final [h] became established as the best exemplar of the word-final category.

Since such examples are common and variations at the level of the word according to phonetic context are not common (though they do occur in special constructions or phrases, as in French liaison [Bybee 2001]), I take such cases as evidence that there is a strong tendency to keep the phonetic variation in an individual word down to a small range. Only high-frequency words such as *don't* encompass wide ranges of variation, but in these cases, the variants are restricted to certain phrases, and it can be shown that each phrase is itself behaving like a word. Thus the *don't* in *I don't know* is in a different item of storage than the *don't* in *we don't smoke*, where *don't* is functioning as a separate word (Bybee and Scheibman 1999).

Since words are the units within which sound changes are established, words containing the same morpheme in different phonetic contexts provide the locus for alternations to develop. Using the Spanish example again, there are a few nouns that originally ended in [s], whose plurals would add *-es*. Thus the singular form of the noun *voz* [bos] 'voice' would become after the change [boh] while the plural *voces* would become [boseh] retaining the [s] in a position before a vowel. As the singular and plural are distinct words, this "variation" is not resolved in the same way as variation within a single word. Rather the two allomorphs are retained unless analogical change manages to level them.

Postulating words as the units of representation in memory and a tendency for words to have a narrow range of variation explains how word-level phonological alternations develop. It also explains how phonetically conditioned alternations become lexically and morphologically conditioned, the universal path shown in (5). As phonetic variants become established in words during sound change, particular morphemes take on different forms in different phonetic contexts, which are different morphological and lexical contexts as well. When a new phonetic category is established and intermediate variants are lost, the resulting alternation is associated with certain morphemes and/or stems as much as it is associated with certain phonetic conditions. Given further changes, such as an increase in the phonetic distance between variants, the loss of productivity of the original phonetic routine, or the loss of conditioning environment, only morphological and lexical conditioning will remain viable. (See section 5.8.)

As applied to the case of the German velar fricatives, the diminutive *-ichiin* always has the fricative after a palatal vowel; thus it was always produced as [ç] once this variant entered the language. In all the words with the diminutive suffix, the [ç]

[4] Examples from transcripts of Cuban speakers in the 1970s collected by Tracy Terrell.

was firmly established. Thus when the palatal vowel was lost, the [ç] remained. Its association with this particular suffix had been long established.

5.8 FURTHER CHANGES

Once a new phonetic category is established for certain phonetic contexts and represented lexically, the further changes mentioned above can occur. First, the phonetic change itself may continue to progress, as in the case of the German palatal fricative continuing to become more fronted. Such a change could be the result of the continuation of the articulatory trend that originally set the change in motion, or it could be related to the perceptual consequences of the new categorization and transgenerational reinterpretation of the variation (Janda 1999). In either case, the phonetic distance between the original variants will continue to grow. As mentioned above, the establishment of a new phonetic category also means the establishment of a new neuromotor routine. Thus there is a neuromotor routine for producing [ç], which begins to be possible even after back vowels.

Second, the new, independent set of variants and their associated neuromotor routine can be used in new contexts, as in loanword adaptation, where, for example, the German palatal fricative is used in word-initial position. The use in new combinations has the potential for creating more instances of contrast.

Third, the productivity of the original alternation is lost as the neuromotor routines are revised and the routine for producing [ç] or [ʃ] is no longer tied to the presence of a preceding front vowel. This leaves the door open for new instances of [x] to occur after front vowels as well. Such new instances would undoubtedly assimilate to the front vowel, but not to the extent that the older reflexes did.

If the establishment of new "phonemes" corresponds to the establishment of new categories that have the potential for contrast, then the change is covert: it actually happens long before exceptions develop. Thus phonetic categories that are considered predictable in traditional analysis may have already achieved the phonetic distance and the lexical or morphological associations to become phonemic when the occasion arises. Not only is the German palatal fricative such a case, but also vowel length in English, which is used as a perceptual clue to the voicing of final consonants, even though it is still "predictable" (Bybee 2001).

5.9 THE EXPLANATION FOR STRUCTURE PRESERVATION

In the preceding sections we have established that the convergence of several factors that naturally occur in change explains the tendency for Structure Preservation to hold; indeed, it explains the general phonetics/phonology distinction, which must be viewed as a continuum. The way these factors interact is as follows.

First, sound change, realized as gradual phonetic change, takes place in words and permanently affects their representation. Thus variants are associated with particular words, phrases, or morphological categories. Second, marginal or infrequent variants of words are lost, giving the phonetic categories a limited range of variation. Third, phonetic change continues to progress, taking the changed variants farther away from their original source. This entails the establishment of new neuromotor routines that are not necessarily dependent upon the phonetic context. It also qualifies the new variants perceptually for phonemic contrast, should the occasion arise. This scenario, then, explains how and why "word-level" phonology develops and why such phonology usually involves segments and features that are used contrastively elsewhere. However, it also explains how and why intermediate cases develop.

As Greenberg (1969: 186) says in his description of this dynamic theory: "It is not so much that the 'exceptions' are explained historically, but that the true regularity is contained in the dynamic principles themselves."

5.10 MECHANISMS: PROCESSES THAT ARE CONSTANTLY IN OPERATION AS LANGUAGE IS USED

The mechanisms behind the paths of change just discussed are reviewed here.

(8) a. Repetition of sequences that make up words and phrases leads to automatization of these units and gestural reduction.
b. Cognitive representations are affected by language use; experience with language is recorded in memory; thus the effects of sound changes are registered in the phonetic representations of words immediately.
c. Phonetic variants are categorized during usage events based on phonetic similarity.
d. Repeated instances of categorization can sharpen the differences on a continuum, leading to the split of one continuum into more than one category.
e. Phonetic change in a certain direction tends to continue.

Note that none of these mechanisms that create the structure of the phonology has to be stated as a constraint. No constraints need to be formulated because the structure that evolves is a natural consequence of multiple applications of the processes that human beings use to produce and decode speech. I submit that if we begin to think realistically about the processes activated during language use, explanations for many structural phenomena will emerge.

It should be noted that structural theories, such as American structuralism and Lexical Phonology, propose no explanation for the structural properties they have identified. A "principle" such as Structure Preservation or separation of levels is simply a property of grammars and in those theories requires no further explanation, since structure is assumed. However, a usage-based emergent grammar seeks a higher level of explanation. It is a principle of such theories that structural properties—or more appropriately, tendencies—arise as language is used and find their explanations in the nature of the categorization and processing capacities of the human brain.