10 Diachronic and Typological Properties of Morphology and Their Implications for Representation

Joan Bybee
University of New Mexico

This paper presents certain diachronic and typological facts that bear on the issue of the representation of morphologically complex words. It is argued in particular that a gradient model is necessary, and this model has the following properties: Words have varying degrees of lexical strength, depending upon their text frequency; complex words have multiple connections to related words, and parallel phonological and semantic connections constitute morphological relations; lexical connections may vary in strength; both irregular morphologically complex words and high frequency regular complex words are stored in the lexicon, but regular low frequency words may be produced componentially.

INTRODUCTION

As a background to the study of morphological processing it is important to be aware of the range of morphological types found in human languages around the world and the way morphology is created in these languages. A viable theory of morphological storage and processing should be able to predict the direction of change in morphology and be consistent with the range of variation found in the world's languages.

DIACHRONIC PROCESSES AND LANGUAGE TYPOLOGY

The Formation of Affixes

The vast majority of affixes in the languages of the world evolve from independent words by a gradual process of change called 'grammaticization' or 'grammaticalization' (Bybee, Perkins, & Pagliuca, 1994; Heine, Claudi & Hünnemeyer, 1991; Heine & Reh, 1984; Heine & Traugott, 1991). In the gradual progression from a lexical morpheme to a grammatical one, changes occur in the phonological shape of the morpheme, its meaning, and its grammatical behavior. A well-documented instance of this type of change is the development of the future tense in Romance languages such as Spanish and French. A periphrastic construction in Latin consisting of an inflected auxiliary *habere* 'to have' and an infinitive yielded a meaning of obligation or predestination:

amare habeo 'I have to love, I am to love'
love+inf aux+1s

The auxiliary reduces phonologically and comes to appear consistently after the infinitive (where previously it could occur in various places in the clause). In Old Spanish we find the construction indicating future:

amar he 'I will/shall love'
love+inf aux+1s

The auxiliary is written separately from the infinitive because at this stage other morphemes could come between the two; for instance, the object pronoun:

amar lo he 'I will love him'
love+inf him aux+1s

Later this possibility disappears and the auxiliary becomes an actual suffix to the verb:

lo amaré 'I will love him'

In this process the grammaticizing morpheme undergoes phonological reduction (e.g., from *habeo* to *he* to *é*), its position becomes fixed, it fuses with the verb, and the whole construction takes on a more abstract, grammatical meaning.

A similar process leads to the development of derivational affixes. However, in this case the process begins with compounding. If the same element occurs in a number of compounds, it can reduce phonologically and change semantically in such a way that it becomes a derivational affix. For instance, the Modern English suffix *-ly* derives from a noun, which in Old English was *lice* meaning 'body.' The compound *mann-lice* originally meant 'having the body or appearance of a man' whence it generalized to 'having the characteristics of a man,' the modern sense of *manly*. Since *-lic* was used in so many combinations, it lost its stress and reduced to *-ly* by losing its final consonant. Its meaning had already generalized in Old English to sometimes just mean 'pertaining to' as in the form *heofon-lice* 'heavenly.' In Modern English *-ly* is used to derive adjectives, as in *friendly*, and to derive manner adverbs, as in *cleverly*, but it occurs in many other uses as well: Consider *daily, weekly, cowardly, possibly,* and so on. Most derivational affixes in English and other languages can similarly be traced back to independent words where evidence is available.

One interesting aspect of the grammaticization process is that it is unidirectional: Affixes are regularly formed from erstwhile words, while only in the rarest and most special of cases is an affix transformed into a word. (Some possible examples are words such as *pros* and *cons,* but these can also be considered clippings or shortenings, such as *lab* from *laboratory.*) Moreover, the process is not discrete but continuous; grammaticization in the form of semantic change and further phonological reduction and fusion continues even after grammatical status is achieved, and even after affixation occurs. This means that we can categorize morphemes for their 'degree of grammaticization.' Non-affixed forms such as auxiliaries are less grammaticized than affixes; affixes are more grammaticized if they are more reduced (e.g., shorter), cause changes in the stem, or undergo changes caused by the stem. As one instance of this continuing development, consider the Spanish future forms discussed previously. Some time after affixation had occurred, the new suffixes began to condition changes in certain verbs. Thus the combination *veniré* 'I will come' changed to *vendré*; *quereré* 'I will want' changed to *querré*; *teneré* 'I will have' changed to *tendré*. Such changes can be taken to indicate increased fusion between the stem and the suffix.

Another point worth mentioning here is that affixation is not just a matter of form. While it is true that two elements that occur together very frequently will have a tendency to fuse to one another, the formation of a true affix requires that there be a degree of conceptual coherence between the two elements. Thus tense and aspect markers tend to fuse with verbs, while case, gender, and definiteness markers tense to fuse with nouns (Bybee, 1985). In contrast, English contractions such as *I'll, I've, he's, we're*, etc., tend not to produce true affixes because of the lack of conceptual coherence between the subject and auxiliary.

Morphological Typology

It has long been observed that languages tend towards different types in their morphological expression (Sapir, 1921). This typology is largely a matter of the degree of fusion among morphemes, but there is also a measurable difference in the degree of grammaticization of morphemes. Some languages allow no affixation; all words consist of single morphemes. An example of such a language, referred to as analytic or isolating, is Mandarin Chinese. In Mandarin, most words are monomorphemic, although compounds do exist, and a few suffix-like morphemes are developing from compounds in the way described above. Nouns and pronouns do not change in form in subject and object functions, verbs do not agree with nouns, and there are no obligatory categories (inflections) marked on verbs or nouns or any word forms at all.

Synthetic languages, or languages which do have polymorphemic words and inflections, are divided into types according to how much fusion exists among the morphemes of a word. In agglutinative languages, morphemes are more loosely strung together; it is relatively easy to divide a word into morphemes, as there is little phonological fusion at boundaries between morphemes, and morphemes do not vary in their shape greatly. Consider the following example from Turkish.

demizlerimizde 'in our oceans'
deniz - ler - im - iz - de
ocean plural 1st plural in

Fusional languages do not exhibit such neat separability of morphemes, but rather show some of the following characteristics: (i) Several grammatical morphemes may be expressed in a single consonant or vowel; e.g., in Spanish *canté* 'I sang,' the vowel *é* expresses the conjugation class (First Conjugation), the Preterite aspect and First Person Singular; (ii) a number of different variants may occur for a single meaning, as in Latin, where the Nominative Singular could be marked by *-a, -us, -um, -is, ū, ēs* or zero; (iii) the same meaning may be distributed over various parts of the word, rather than residing in a single morpheme, as in the following analysis from Matthews (1991, pp. 170-179) of the Greek verb *elelýkete,* 'you had unfastened':

		(root)	*Perfective*	*Indicative*	*Active*
Past	*Perfective*	*Perfective*	*Active*	*Past Active*	*2nd Plural*
e	le	ly	k	e	te

In this example it is notable that Past is expressed twice while Perfective and Active are indicated three times each. One could easily conclude that there is more to this word than the simple concatenation of meaningful elements.

Sapir (1921) originally argued that languages differ not only in morphological form but also in the types of grammatical meanings that are expressed. It follows that if a language does not have inflection it does not express inflectional meaning. Thus if Mandarin Chinese does not have obligatory categories, it cannot express the same sense of tense as does English, which obligatorily expresses tense. Instead, Mandarin expresses more specific aspectual notions using auxiliaries and compounds. Bybee, Pagliuca, and Perkins (1991) have shown that languages which have less fusion and less grammaticization of form also express meanings that are less grammatical (i.e., more lexical). It appears that part of what is behind morphological typology is a tendency for some languages not to carry grammaticization as far as other languages do. This finding raises the very important issue of what drives grammaticization and what language-specific properties could encourage or inhibit the process.

Languages With Various Types of Morphological Structures

Because grammaticization is an ongoing process at all times, languages often have structures of different morphological types coexisting, often even expressing very similar meanings. Hopper (1991) has pointed out that the renewal of morphology leads to a situation in which newer and older structures create 'layers' in a language. One of his examples concerns the various ways that past tense is expressed in English. The oldest layer still surviving is that represented by the strong verbs, such as *break, broke, sing, sang, know, knew,* etc., which conserve the Proto-Germanic method of forming past tense through vowel changes.[1] Of course, this oldest method is preserved only in the most frequent lexical items. The next method of past tense formation which developed, also before the Old English period, arose from a periphrastic construction in which the past form of the verb *do*, the predecessor of modern *did,* followed the main verb to signal past tense. This auxiliary suffixed to the verb to give modern *-ed*, which, as the productive past tense suffix, has gradually been replacing the older vowel-change method. The newest development echoes the previous one: *did* is used preceding the unmarked verb to signal past tense in questions, negative clauses, and emphatic clauses. Note that the age of the construction correlates with the degree of fusion between grammatical marker and stem. The oldest method shows the greatest fusion, since it is a vowel change in the stem, the next method is more agglutinative as it uses suffixation, and the newest method is the most analytic.

In this particular case the three structures all express the same meaning, but it is more common to find that different structural types correspond to different meanings. In particular, a less grammaticized construction will express a meaning that is more specific and lexical in its content, while a more grammaticized construction will express a meaning that is more abstract. Compare the less grammaticized English *keep on dancing* and *used to dance* with *is dancing* and *danced.*

MORPHOLOGY VERSUS SYNTAX

A tendency exists in structuralist linguistics to view morphology as analogous to syntax—morphemes are strung together to form words just as words are strung

[1] I claimed earlier that most grammatical morphology develops from whole words reducing to affixes. What then of these strong verbs in English? It is very likely that they also have their source in the regular grammaticization process. Affixes could have conditioned these stem changes and later been lost from the verbs, leaving only the stem change to signal past tense meaning.

together to form clauses. The numerous ways in which morphology differs from syntax, however, strongly suggest a different type of processing, in particular, that full words, even multimorphemic ones, can be stored in the lexicon. The reason that morphology resembles syntax, I would argue, is that it is *old* syntax, since affixes were previously separate words (Givón, 1979). The point I want to highlight is that affixes would not develop the properties that distinguish them from independent words unless they were somehow processed differently.

The traditional criteria for determining whether a grammatical morpheme is an affix or an independent word or particle are the following:

(i) affixes occur in a fixed position with regard to the lexical stem; they are not free to change their position or their scope;

(ii) no open class items (other nouns, verbs, adjectives, or adverbs) intervene between the affix and the stem; compare English auxiliaries, which allow adverbs to come between them and the verb: *I have really tried.*

(iii) there is a high degree of interdependence between affix and stem, such that certain stems choose certain affixes: *cats, mice, oxen.*

(iv) there may be phonological fusion between affix and stem in the form of assimilation (as in [kæts] but [dɔgz]) or the coalescence of segments (as in *decide* + *ion* giving [dɪsɪʒən]); and

(v) the stem plus affix combination behaves phonologically like a word, for instance by acting as the domain for phonological processes such as vowel harmony or by being stressed as a single word rather than as two words.

All of these criteria reflect a very high degree of fusion or bonding between stem and affix. The initial bonding and the continued increase in the degree of fusion that is characteristic of affixal development over time indicate that speakers take the stem-affix combination to be a single unit, traditionally called a word, and that this unit has the status of a stored cognitive entity, which we would regard as a stored lexical representation. If morphologically complex words are stored in the lexicon, then some mechanisms for representing their internal complexity, the regularity of certain patterns, and the ability to form new words must be postulated. We now turn to these matters.

MORPHOLOGY IN THE LEXICON

Human beings appear to have (at least) two important capacities that make language possible: (i) the ability to store tens, perhaps hundreds, of thousands of individual lexical items with detailed information about their behavior and meaning; and (ii) the ability to concatenate series of linguistic units to form meaningful utterances. The fact that morphology differs from syntax by exhibiting a greater bonding between units suggests that the first ability is being made use of for storing and processing morphology perhaps to a greater extent than the second ability.

Linguists have objected to the idea that morphologically complex words are stored in the lexicon for two reasons: First, they fear the loss of important generalizations concerning word structure and phonological alternations, and second they argue that in morphologically complex languages the number of words is too great for lexical storage of all words. But these objections are based on an oversimplified view of the lexicon as resembling a dictionary, where words are set down permanently, in an arbitrary order, and each one is isolated from all others. If we view the lexicon as part of the human memory bank, the analogy with the dictionary fails. Rather, three important properties apply to lexical storage: Some stored items are stronger than others, there are multiple relationships among stored items, and generalizations may be formed over properties of stored items, or properties of relationships among items. These general properties of memory can account for much of what has traditionally been viewed as morphological phenomena, and in addition, they can account for aspects of morphological structure that have been traditionally ignored by structural linguists. But before explaining how these properties can capture the generalizations that linguists consider important, let us consider the question of the number of words that are plausibly stored in the lexicon.

Hankamer (1992) argues against the position that *all* possible words of a language are stored in the lexicon on the basis of agglutinative languages such as Turkish, where by his calculations a single verb root has the theoretical possibility of having over a million forms. Sadock (1980) makes similar arguments for languages in which incorporation of a noun into a verb form is a common practice. Of course, not all of the possible words in these languages would actually be semantically or pragmatically plausible, and only some of them would be in common usage. Thus, the model of Bybee (1985) proposes that morphologically complex words may be stored in the lexicon, even if they are regular, but that words of high frequency are more likely to be stored whole, while regular formations of lower frequency are more likely to be formed by combination and not stored. The complex words that the speaker uses frequently and which are entered in the lexicon serve as the basis for the formation of the unstored, less frequent words, in a way that I explain in the next section.

Lexical Connections

Let us now consider the model that follows from the three properties of memory mentioned above: that some stored items are stronger than others, that there are relations among stored items, and that generalizations may be formed over stored items. In Bybee (1985) I proposed a model of morphology based on the notions of *lexical strength* and *lexical connection*. This model has some features of models of lexical storage developed about the same time in cognitive studies (Langacker, 1987) and some properties of processing proposed for connectionist frameworks (Dell, 1991; Rumelhart & McClelland, 1986). I adopt some terminology from these studies in the description of this model.

First, I assume that memory for linguistic units is superpositional; as I argued in Bybee (1985), every time a word or a larger linguistic unit (a phrase or idiom) is processed, it is mapped onto, or superimposed on, some existing representation (unless, of course, it is being heard for the first time). If meaning and phonological form of the word in processing matches a stored representation, then it is mapped onto that representation, which is consequently strengthened. As a result, high frequency words will have greater *lexical strength* than low frequency words; this will account for their relatively greater ease of activation, and other properties as well, as we shall see in the next section.

If, on the other hand, meaning and phonological shape only partially match a stored representation, then the mapping can only be partial, and instead of a direct mapping of the item in processing with a stored representation, the new item goes into storage with the partial mapping to other stored forms represented by *lexical connections*.[2] Lexical connections are of both a semantic and a phonological nature. Thus, synonyms such as *fast* and *quick* will have semantic connections but no phonological ones. Homophones such as *plane* and *plain* will have phonological but no semantic relations. Related words such as *cat* and *kitten* have partial semantic connections and some (unsystematic) partial phonological ones. Parallel phonological and semantic connections, if they represent a pattern found in multiple sets of items, constitute morphological relations.[3] Thus, the word *started* maps only partially onto *start*, with the suffix not matching any material in *start*. However, the suffix itself can map onto other occurrences of the same suffix, creating a complex of morphological relations as in Figure 1:[4]

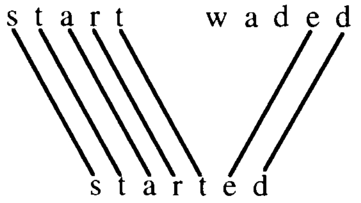

FIGURE 1. Lexical connections among regularly inflected words.

The existence of lexical connections for partial mappings yields an internal morphological analysis of complex words without any other mechanisms being necessary. The analysis of *started* into two 'morphemes' is represented by the connections that word has with the stem *start* and other instances of the suffix *-ed*. Such a representation captures the fact that we only learn that words are morphologically complex by comparing them to other related words. This model, then, allows the storage of whole words as well as information about their internal complexity. It does not insist, as a morpheme-based model would, however, that all words be exhaustively divisible into morphemes and thus allows for the unproblematic representation of complexity even in words where some constituent parts do not actually have morpheme status, words such as *cranberry* (consisting unmistakably of *berry*, but with the non-morpheme *cran* as its other part) or the days of the week such as *Tuesday, Wednesday,* and so on (Bybee, 1988).

In this model, then, morphemes do not have independent status and representation. They are rather seen as epiphenomenal or 'emergent' in the sense of Hopper (1987). Other models with this property are the 'word and paradigm' models of Matthews (1972, 1991) and Anderson (1992). Such models account naturally for the type of inflected word analyzed in the section on Morphological Typology. The Greek word *elelykete,* 'unfastened,' does not display a one-to-one relation between the meanings expressed in it and the form; rather, the notion of perfective is distributed over three elements in this word, as is the notion of active. The greater semantic and phonological fusion exhibited in a single word, the less attractive a strict morpheme-by-morpheme description appears to be.

The current model is quite appropriate for Semitic languages, where lexical roots consist only of consonants and much derivational and some inflectional morphology is expressed by changing the vowel pattern of the word and to a much lesser extent, the consonants of the root themselves. Lexical connections can be made between the consonants of two forms of the same root, as shown in Figure 2, while independently the vowel patterns of words of the same morphological category can be related, as shown in Figure 3.

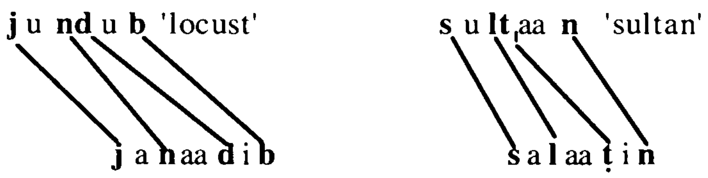

FIGURE 2. Lexical connections between two Arabic nouns and their plurals.

[2] A form that is directly mapped onto a stored form may also form partial connections, particularly if it is morphologically complex.

[3] This statement is a correction of the statement made in Bybee (1985, 1988), where I said that parallel semantic and phonological connections constitute morphological relations. If there were only one pair of items with such connections this would not be regarded as constituting a morphological relation. What is considered morphological apparently needs to apply to multiple sets of relations.

[4] These schematic representations using orthography are not intended to make any statement about the nature of phonological segmentation.

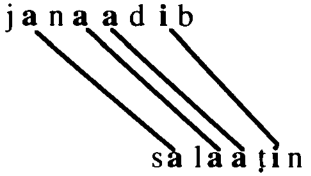

FIGURE 3. Lexical connections between two Arabic plural nouns.

Bentin and Frost (this volume) present evidence that suggests for Hebrew that derivatives of a root are each listed in the lexicon, but that their representations are also interrelated.

Moreover, they find evidence that the vowel pattern and the root are separated in the process of word identification even though neither are listed separately. Such conclusions are totally compatible with the type of representation depicted in Figures 2 and 3. In the section on the Treatment of Patterns in the Lexicon, we discuss the way that recurrent patterns, such as multiple instances of Semitic nouns that are pluralized by the same vowel pattern, are treated in this model.

Lexical connections also easily accommodate agglutinative morphology, such as that illustrated by Turkish earlier, since the constituent parts of words, in this case the suffixes, map onto instances of these same suffixes in other words. The Turkish lexicon will of course look rather different from the Greek or Semitic lexicon, but the difference in structure will be entirely due to the nature of the linguistic material that is entered in the lexicon. The basic principles of lexical strength and lexical connection remain the same.

Considerable evidence exists for differential strengths of lexical connections. On the phonological side, lexical connections can be computed simply from shared phonological features, although the possibility that some shared features create stronger lexical connections than others cannot be ruled out. (Thus, a general feature such as syllabic will be less important to morphological relatedness than a more specific one, such as velar or round.) With semantic features, the number of identical matches will of course be an important determinant of morphological relatedness, but here the evidence is good that sharing some features produces stronger morphological relations than sharing others. In particular, shared features that are more 'relevant' to the stem, in the sense of Bybee (1985), comprise stronger relations than those that are less relevant. Thus, for verbs, aspect and tense distinctions are more relevant to the verb and produce larger meaning changes than person/number agreement. We expect, then, that verb forms in the same tense or aspect category are closely related despite differences in person and number. While this hypothesis awaits experimental testing (but see Stanners, Neiser, Hernon, & Hall, 1979, and Bybee & Pardo, 1981, for applicable experimental results), there is considerable linguistic evidence for it: A strong cross-linguistic pattern shows morphophonemic differences aligned with tense and aspect categories, rather than person/number categories. This can be illustrated with Spanish examples, where morphophonemic differences in the verb tend to coincide with the important aspectual distinction of perfective/imperfective:

Imperfective (Past)			*Perfective*	
		tener 'to have'		
ten-ía	ten-íamos		tuv-e	tuv-imos
ten-ías	ten-íais		tuv-iste	tuv-isteis
ten-ía	ten-ían		tuv-o	tuv-ieron

In a substantial set of high frequency verbs, stem changes produce the result that there is one common stem for the imperfective and one for the perfective. The person/number forms within each aspect are very similar to one another, being very closely related. On the other hand, since person/number agreement is a minor semantic dimension for verbs, forms sharing the same person/number, such as first person *tenía* and *tuve*, are not as closely related.

Lexical Strength

The proposal that lexical items differ in their relative lexical strength is based very simply on the idea that much-used items are more ingrained or entrenched in memory than lesser used items. This notion accounts for a number of psycholinguistic, diachronic, and typological facts about morphology. I have already mentioned that lexical strength can account for the relatively greater ease of access of frequent words over infrequent ones. In the following I mention other phenomena that can be accounted for with a notion of lexical strength.

First, there is a universal tendency for morphological irregularity to be restricted to the highest frequency forms of a language. Thus, irregular verbs tend to have such frequently used meanings as 'come,' 'go,' 'have,' and 'be.' Or, if there is a large number of irregular verbs, their meanings will not so much be predictable as their relatively high frequency. Thus, among the 30 most frequent verbs of English, 22 are irregular according to Francis and Kučera (1982). Nouns with irregular plurals are commonly the words for man, woman, child, or common livestock animals. Also, irregular plurals tend to reside in nouns in which the plural is much used—nouns designating objects that come in pairs or groups (*mice, feet, teeth, oxen, geese*) (see Tiersma, 1982, and discussion following).

Much morphological irregularity develops naturally as the byproduct of phonological change. For example, a general voicing of fricatives (such as *s, f,* and *th*) between vowels in Middle English gives us an alternation in singular/plural pairs such as *wife, wives, leaf, leaves,* and *house, hou*[z]*es*. Again in Middle English, the shortening of vowels before certain consonant clusters yielded irregularity in *sleep, slept, keep, kept, weep, wept,* and *leap, leapt,* among others. An irony long noted by historical linguists is that regular sound changes create irregular morphology. A further irony is that re-establishment of the morphological regularity undoes the phonological regularity. That is, it is common for the morphological regularity to reassert itself and for new regular forms to appear (formed on the basic form of the paradigm, see following discussion). For instance, new regular past tense forms *weeped* and *leaped* are considered acceptable in current English.

One of our interests in this diachronic process of regularization lies in the fact that infrequent words tend to regularize before frequent ones. Thus, while *weeped* and *leaped* are acceptable as past tense forms, the highly frequent **keeped* and **sleeped* clearly are not (Hooper, 1976). In fact, the entire pattern of regularization of strong verbs in the last millennium in English shows that infrequent verbs regularize and the frequent ones maintain the vowel change as the indicator of past tense and past participle status (Bybee, 1985, pp. 119-120). This diachronic pattern of regularization then leads to the synchronic pattern (which is cross-linguistically valid) in which only the relatively high-frequency forms tend to be irregular.

The notion of lexical strength determined by frequency accounts for this pattern in the following way: If frequent words have stronger representations than infrequent ones, they are more easily accessed and there is no need to create new regular forms. If, on the other hand, infrequent irregulars are only weakly represented, they are not so easily accessed and are thus more likely to be replaced by newly formed regulars. (More discussion of the treatment of regulars and irregulars in this model follows.)

The other relevant fact about the regularization process, which can also be accounted for by lexical strength, is the fact that the direction of change in morphological regularization is usually predictable. When there is an alternation or allomorphy in a morpheme—that is, when a morpheme has two variants—such as *house, hou*[z]*es* or *weep, wept*, regularization entails the loss of one of the allomorphs or variants, accomplished by the replacement of one variant for the other. (This is also known as 'analogical leveling' in historical linguistics.) Interestingly, we are often able to predict which variant survives and replaces the other—it is usually the variant in the category member that is the most basic, the least marked, and the most frequent (Kuryłowicz, 1968; Mańczak, 1980).

Thus, in nouns we expect the variant used in the singular to replace the one used in the plural; so many people say *hou*[s]*es* for the plural, but no one tries *hou*[z] for the singular noun. In languages with case marking the nominative is usually the basic member, and in European gender systems, the masculine is basic. For verbs, the present indicative is more basic than the past or future or other moods (see Greenberg, 1966). Thus, for English we find the regularization of the past tense form in *weeped* and *leaped* and not a change in the base or present form giving **wep* and **lep*.

While some (Jakobson, 1957; Kuryłowicz, 1968) would argue that markedness or basicness is a purely structural dimension, Greenberg (1966) has shown that a major correlate of markedness is frequency. That is, the basic or unmarked member of a category is the most frequently occurring member of the category. Moreover, Tiersma (1982) has shown that the frequency criterion can predict apparent exceptions to the rule that the least marked member survives in leveling, while the structural account cannot. In West Frisian (a dialect of a West Germanic language, spoken primarily on islands off the coasts of The Netherlands) singular and plural forms of many nouns have vowel alternations: *hoer, hworren,* 'whore,' *koal, kwallen*, 'coal,' and these tend to regularize with the stem of the singular replacing that of the plural: *hoer, hoeren* and *koal, koalen*. However, some nouns regularize in the opposite direction, with the stem of the plural replacing that of the singular: thus *earm, jermen,* 'arm,' becomes *jerm, jermen* and *goes, gwozzen* becomes *gwos, gwozzen*. Tiersma argues that nouns which more frequently occur in texts in the plural because their referents more commonly occur in the real world in pairs or groups have the plural as their "least marked" or "basic" form. The nouns that regularize on the basis of the plural are 'arm' and 'goose' as shown above and 'animal horn,' 'stocking,' 'tooth,' 'wood shaving or splinter,' 'thorn,' and 'tear.' Thus, it is actually frequency which determines the member that survives in regularization and not a structural relation.

Since members of paradigms differ in their token frequency, they also differ in their lexical strength; the more frequent member(s) will be easier to access than the less frequent ones, as shown in Sereno and Jongman (1991). If the less frequent members of a paradigm are not accessible, a speaker would have to create a new form on the basis of the one that is available, regularizing on the basis of the most frequent form of the paradigm. This would be particularly likely to happen if the whole paradigm is low in frequency.

The differential lexical strength of members of paradigms leads to a hierarchical structure for paradigms in which more frequent members dominate the less frequent members. We return in the next section to a discussion of paradigm-internal relations.

Further evidence for lexical strength is found in the creation and maintenance of suppletive paradigms. I am using the term "suppletion" in its narrowest sense, in which it refers to synchronic inflectionally related forms that come from different roots historically. In English the examples from verbs are *go, went* and *is, be, were*; in French we find *va, allons*, etc., 'go,' as well as *être, suis*, etc., 'be.' In adjectives suppletion occurs in *good, better, best; much, more, most; bad, worse, worst*; cf. Spanish *bueno* 'good' and *mejor,* 'better, best,' *mal,* 'bad,' and *peor,* 'worse, worst.' The first fact to observe about suppletion is that it occurs only in the most frequent paradigms (Rudes, 1980). For our purposes, the way suppletion arises is of considerable interest: In order for a suppletive paradigm to develop, it is necessary for a form that once belonged to one paradigm to split away from that paradigm and join another. For instance, *went* was formerly the past tense form for *wend*, meaning to go turning or winding along. Now *went* has no semantic or morphological relation with *wend* but rather is associated with *go*. In order for that to have happened, *went* had to have lexical autonomy from *wend*, which would be possible if it became very frequent (eventually more frequent than *wend*) and had a high degree of lexical strength. The possibility of a split of a non-basic form of a paradigm from the basic form indicates that it must have had its own lexical representation despite the strong phonological relation between the two. The fact that only frequent paradigms have suppletion suggests that frequency is important in determining lexical representation. Suppletion also helps us understand the relation between lexical strength and lexical connections, which we turn to in the next section.

Interaction of Lexical Strength with Lexical Connections

The notion of lexical connection is based on the idea that linguistic units (like other units of perception and memory) are often understood and remembered in terms of other linguistic units. However, high frequency units are available enough in the input to be remembered as autonomous items. Members of high frequency paradigms, then, will all have a high degree of lexical strength and be less dependent upon connections with other members of the paradigm, while low frequency paradigms will be characterized by stronger lexical connections among the members.

The evidence for this inverse relation between lexical strength and lexical connections come from diachronic splits among related items, both in inflection, where it results in suppletion, and in derivation, where it results in divergent meaning among historically related forms, such as *awe, awful,* and *awfully.* As mentioned above, it is clear that inflectional suppletion is heavily determined by frequency. Splits among derivationally related items are much more common, but frequency appears to play an important role in this process as well. Infrequent derived forms will tend to maintain a close relation with the base from which they derived, but frequent derived forms will diverge in both meaning and form and tend to become autonomous (Bybee, 1985, pp. 88-89). Words beginning with the prefix *pre-* in English show this trend clearly. Pagliuca (1976) studied the 323 words with this prefix listed in the Shorter Oxford English Dictionary and, using the definitions provided there, rated them as having transparent or opaque meaning according to whether or not their sense was a simple combination of the meaning 'before' with the stem's meaning. A strong association between frequency and opacity emerged, as did an association with the extent of vowel reduction, as shown in Table 1.

These facts suggest that phonological fusion of prefix to stem and reduction of the prefix (phonological opacity) are related to the development of unpredictable meaning (semantic opacity), and both of these factors are related to token frequency.

TABLE 1.

	Percent of Words with		
Vowel Quality	Average Frequency	Predictable Meaning	Example
[iy]	5.74	74.76	predecease
[i]	2.54	59.52	predestine
[ɨ]	49.80	3.30	prediction
[ɛ]	81.32	2.89	preface

Since related forms may maintain a close relation or diverge gradually over time, the lexicon contains a range of types of relations among forms; on one end of the scale are the forms with the strongest lexical connections, i.e., the semantically and phonologically transparent words in pairs such as *happy, happiness* or *pure, purity*. Going down the scale, lexical connections weaken because of lessened phonological or semantic similarity, or the high frequency of derived forms. *Opaque, opacity* have a lesser phonological similarity, *recite, recital* have a lesser semantic similarity (one does not necessarily recite at a recital), and *face, facet* differ phonologically and also semantically, or at least their range of usage is so different that people rarely view them as related; *awe* and *awful* are phonologically transparent and not radically semantically divergent, but their frequency disparity weakens their connectedness (*awful* is three times as frequent as *awe* according to Francis & Kučera, 1982). A similar scale can be applied to inflection (see Bybee, 1985, pp. 91-96).

In this model, then, regular and irregular morphological relations, productive and nonproductive relations among words are represented using the same mechanisms. The differences among them, from which we derive our notions of regularity and productivity are due to the type frequency of the various morphological relations and the ease with which language users can construct generalizations over these relations.

The Treatment of Patterns in the Lexicon

Given the superpositional mapping of items in processing onto stored items, inflectional affixes will accumulate considerable lexical strength. The higher type frequency of English plural *-(e)s* over irregular plural *-en* accounts for the former's greater availability for new formations. The allomorph *-en* occurs in the very frequent word *children*, but given the irregularity of this word (vis à vis its singular counterpart) and its very high frequency, any lexical connections it might have to the suffix in *oxen* would be very weak and would not contribute to the strength of this allomorph. It is the type frequency of affixes that determines their relative productivity, not their token frequency (see MacWhinney, 1978, and Bybee, 1985, pp. 132-134).

Lexical connections also provide a means for representing minor lexical patterns that arise because small groups of words share some salient characteristics. Among English strong verbs there is one class that exhibits some productivity, given a verb of the appropriate phonological shape: Many past tense verbs end in the vowel [ʌ] plus a velar and/or a nasal, e.g., *stung, strung, hung, stuck, struck, dug,* etc. This partial similarity in the rhymes of these verbs paired with the meaning 'past tense' produces a series of connections describable as a schema, which then is applicable to other items, such as *sneak* yielding *snuck,* or *drag* yielding *drug* in some American dialects.

In Bybee and Moder (1983) we pointed out that the formation of such a schema depended upon there being a critical mass of words exhibiting the pattern (the minimum being around six or eight, see Bybee & Pardo, 1981) and that the participating words not be of such high frequency that they form no connections, or only form weak ones, for without lexical connections no schema could emerge.

Thus, regular and irregular patterns are treated the same way in this model. The difference between them is an automatic consequence of the number of items that participate, the relative frequency of the items and, in the case of minor patterns, the degree to which participating items resemble one another. Approaches that treat regular and irregular patterns as governed by different *types* of processing, i.e., rules vs. schemas (Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992) or as processing at different levels (Kiparsky, 1982) fail to account for the fact that regular patterns have wider applicability, while in the current model this is accounted for by the strength of the regular pattern, which is totally derivative of its applicability or type frequency.[5] The traditionally cited evidence that regular patterns are best accounted for with disembodied symbolic rules, while irregular patterns may be lexical, is the overgeneralization of regular patterns found in children's speech and the rarity of generalization of irregular patterns. However, Marcus et al. (1992) have found in a massive study of the generalization of regular past tense in English that overgeneralization of regular past tense to irregular verbs occurs in only 2.5% of irregular productions during the period in which children are overgeneralizing. This low level of generalization of the regular pattern does not suggest rule-governed behavior but rather is more consistent with the treatment of even regular patterns as lexical generalizations.

[5] A case is made in Clahsen and Rothweiler (1992) that the productive German participle ending *-t* does not have a higher type frequency than the suffix *-en,* but this case depends upon counting the German verbs with separable prefixes as different types. This is comparable to counting English *break up, break down, break in two,* etc., as different verbs rather than considering them all to be instances of the verb *break.* If such instances are counted as examples of one type, then the suffix *-t* does have a substantially higher type frequency than *-en.*

The Treatment of Regular Inflected Forms

While many researchers agree that there are good reasons to list irregular morphological forms in the lexicon, as I have just mentioned, regular formations are usually treated as derived by combinatory rules. In the model being described here it is proposed that high frequency forms are treated the same, whether they are regular or irregular. Even regular forms may develop a greater lexical strength due to a high level of usage. On the other hand, the lexical patterns that emerge from the lexical connections yield a means by which lower frequency regular forms may be produced by combination. Thus, both lexical access and combinatory generation are available in this model, with frequency determining the method to be used, which is the reason that Losiewicz (1992) refers to the current model as a 'dual-access frequency-dependent' model.

Stemberger and MacWhinney (1986, 1988) demonstrate that in naturally occurring and experimentally induced speech errors involving English verbs, high-frequency regular items have a lower error rate than low-frequency items. They interpret this to mean that at least high-frequency inflected regulars are stored in the lexicon. However, their evidence does not bear on the question of whether or not low-frequency regularly inflected forms are stored in the lexicon.

Losiewicz (1992) presents experimental results indicating that high-frequency inflected forms are stored lexically while low-frequency forms are not. Her experiments were based on a previous finding by Walsh and Parker (1983), who reported that morphemic /s/ in English words such as *laps* is significantly longer than non-morphemic /s/ in words such as *lapse.* Losiewicz hypothesized that if high-frequency inflected forms were lexically stored but low-frequency forms were not, then the suffix (in this case past tense /d/) would be longer in the latter words than in the former. This is exactly the result that she obtained; when comparing verbs such as *needed / kneaded* and *covered / hovered,* she found a difference in length in the predicted direction for every subject and every verb pair.

Losiewicz interprets this result in the current framework in which both whole words and their connections to other words are stored in the lexicon. She says:

> For high frequency multi-morphemic words, the most highly primed, and activated, item would be the whole word representation. For low frequency words, the multi-morphemic word would not have a strong lexical representation (i.e., would not be highly primed), so a construction process of stem + affix would be the most readily available, and fastest, processing route. (p. 50)

As for the reason for the length difference in the suffix, my own interpretation of this is that an affix is more highly fused to the stem in a frequent combination than in an infrequent one.[6] Thus, the attached affix in the stored representation of a word would be shorter than the stronger, more autonomous representation of the affix.

The Representation of Paradigms

A paradigm is a set of inflectionally related words sharing the same stem. In English, paradigms are relatively simple: Each count noun has two forms, a singular and a plural, and since most of these are regular, the plural form need not be strongly represented. For verbs there are several forms: *play, plays, playing,* and *played*, and for irregulars sometimes there is a difference between the past tense and the past participle form. Still, this amounts to only five forms for each verb, with four being much more common. But other languages have much more complex paradigms for verbs, and sometimes for nouns as well. How are these paradigms accommodated within the current model?

Paradigms are represented as clusters of highly connected words. The strongest words may be thought of as bases around which other words cluster. Especially in languages with complex morphology, there may be more than one strong form per paradigm. Evidence from acquisition and morphophonemic patterning and change in Spanish and Portuguese suggests that in these languages the Present tense of a verb has a base or strong form (usually the third singular form), and the Preterite may have two strong forms (the first and third singular forms). In Bybee and Brewer (1980) we showed what emerged as the base forms in Preterite paradigms in diachronic restructuring were also the most frequent forms in these paradigms.

The following paradigm is the standard Spanish Preterite for First Conjugation and the result of largely regular phonological changes applied to the Latin Perfect:

	Singular	Plural
1st	canté	cantámos
2nd	cantáste	cantásteis
3rd	cantó	cantáron

In various dialects of the Iberian Peninsula, changes have occurred in paradigms of this sort. The most common change is the change of first plural to *cantémos*. This change can be analyzed as the formation of a new first plural on the basis of the stronger first singular form *canté*, with the suffix *-mos*, which is perfectly regular in every first plural form, added to it. Another common change is the construction of a new second singular on the basis of the third singular, again a low frequency form being rebuilt on the basis of a high frequency form.

[6] This claim perhaps reminds the reader of Zipf's (1935) demonstration that frequent words strongly tend to be shorter than infrequent ones. Zipf was apparently unaware that the mechanism that creates this correlation is gradual phonetic reduction and fusion.

These changes show that both the first and third singular are strongly represented in these Preterite paradigms, and the lower frequency forms are either weakly represented as connected to the stronger forms or they are created on-line by applying schemas to the stronger forms. The diachronic evidence suggests, then, that a single paradigm, even of a regular stem, can have multiple strongly represented forms. In the case just discussed, in addition to the representation of the Preterite, the Present, Imperfect, and other tenses will also have to be represented. Moreover, the evidence presented in the section on Lexical Strength concerning the role of frequency in determining lexical strength shows that different paradigms, even if they are structurally the same, can have a different organization in the lexicon. For instance, in most noun paradigms the singular will be the strongest member, but in others, such as the nouns denoting 'arms,' 'geese,' or 'horns' the plural could be the strongest form.

Losiewicz (1992) points out that in this model, since parts of words may be connected or strengthened by repetition, stems will be highly primed as well as the frequent words that contain them. This accounts for the fact that the frequency of both the stem and the inflected forms of a word contribute to its latency in lexical decision tasks (Burani, Salmaso, & Caramazza, 1984; Nagy, Anderson, Schommer, Scott, & Stallman, 1989). Moreover, while some restructuring can be analyzed as the addition of affixes to existing words, as in the case of *cantémos* produced from the word *canté*, there are cases in languages in which stems seem to play a role in the creation of new forms.

SUBSTANCE DETERMINES STRUCTURE

In the model being described here the substance of words—their actual phonological and semantic substance—determines not only their structure but the larger structure of the lexicon and morphology into which they fit. There are no predetermined modules or levels of description; all the structure that is built up in the lexicon emerges precisely from the words or phrases that are stored there.

The lack of modules or strict components separating rules from representations means that the model is capable of handling various sorts of gradient phenomena as well as languages of different morphological types. I have already pointed out that regular and irregular patterns are handled in the same way, with the only difference between them being the number of items involved, which determines how strong or entrenched the pattern is.

Similarly, inflectional and derivational morphology are treated the same way; the differences between them emerge from the fact that inflectional forms tend to be more closely related semantically, and they tend to fit into highly entrenched patterns (see Bybee, 1985). The fact that derivational processes at times do have the properties of inflection, that is, they can also be highly productive and make small meaning changes, and the fact that some high frequency inflected forms are very irregular and can even split apart from related forms means that we cannot draw a strict line between inflection and derivation and proposals such as Anderson's (1982, 1992), which assign inflection and derivation to different components of the grammar, cannot be maintained.

Another gradient phenomenon that this model can accommodate well is the gradual nature of grammaticization. As grammaticizing elements reduce and begin to develop into affixes, we must allow for a gradual passage from independent to dependent status. The current model, which does not insist that units be strictly categorized as 'in' the lexicon or 'out of' the lexicon, would allow a newly developing affix to be stored with its stem if it is a high frequency combination; a sufficient number of such combinations would lead to the generalization that the unit has become an affix. Further fusion between stem and affix serves as evidence that the new word is taken to be a coherent semantic, phonological, prosodic, and indeed, lexical unit.

Finally, the typological differences among languages in this view derive from the differences in substance that give rise to differences in structure. The amount of morphological complexity in a language will be measured by the number of related words and the complexity of their connections to one another. An agglutinative language will be characterized by highly entrenched regular affixes with few variants and by less overlap in connections at boundaries. A fusional language will have more complex relations and more variations on relations, including a wider range of patterns and more irregular forms that do not participate in dominant patterns.

ACKNOWLEDGMENT

I would like to thank Carol Fowler and Beth Losiewicz for valuable comments on an earlier version of this chapter.

REFERENCES

Anderson, S. (1982). Where's morphology? *Linguistic Inquiry, 13,* 571-612.

Anderson, S. (1992). *A-morphous morphology.* Cambridge: Cambridge University Press.

Burani, C., Salmaso, D., & Caramazza, A. (1984). Morphological structure and lexical access. *Visible Language, 18,* 342-352.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form.* Amsterdam: John Benjamins.

Bybee, J. (1988). Morphology as lexical organization. In M. Hammond & M. Noonan (Eds.), *Theoretical Morphology.* New York: Academic Press.

Bybee, J., & Brewer, M. (1980). Explanation in morphophonemics: Changes in Provençal and Spanish preterite forms. *Lingua, 52,* 271-312.

Bybee, J., & Pardo, E. (1981). On lexical and morphological conditioning of alternations: A nonce-probe experiment with Spanish verbs. *Linguistics, 19,* 937-968.

Bybee, J., & Moder, C. (1983). Morphological classes as natural categories. *Language, 59,* 251-270.

Bybee, J., Pagliuca, W., & Perkins, R. (1991). Back to the future. In E. Traugott & B. Heine (Eds.), *Approaches to grammaticalization.* Amsterdam: John Benjamins.

Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world.* Chicago: University of Chicago Press.

Dell, G., & Juliano, C. (1991). Connectionist approaches to the production of words. *Cognitive Science Technical Reports CS-91-05.* Urbana, Illinois: The Beckman Institute.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage.* Boston: Houghton Mifflin.

Givón, T. (1979). *On understanding grammar.* New York: Academic Press.

Greenberg, J. (1966). *Language universals.* The Hague: Mouton.

Hankamer, J. (1992). Morphological parsing and the lexicon. In W. Marslen-Wilson (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press.

Heine, B., Claudi, U., & Hünnemeyer, F. (1991). *Grammaticalization: A conceptual framework.* Chicago: University of Chicago Press.

Heine, B., & Reh, M. (1984). *Grammaticalization and reanalysis in African languages.* Hamburg: Helmut Buske Verlag.

Heine, B., & Traugott, E. (Eds.). (1991). *Approaches to grammaticalization.* Amsterdam: John Benjamins.

Hooper, J. (1976). Word frequency in lexical diffusion and the source of morpho-phonemic change. In W. Christie (Ed.), *Current trends in historical linguistics.* Amsterdam: North-Holland.

Hopper, P. (1987). Emergent grammar. *Proceedings of the Thirteenth Berkeley Linguistic Society Meeting,* 139-157.

Hopper, P. (1991). On some properties of grammaticization. In E. Traugott and B. Heine (Eds.), *Approaches to grammaticalization.* Amsterdam: John Benjamins.

Jakobson, R. (1957). Shifters, verbal categories and the Russian verb. Reprinted in *Roman Jacobson, Selected Writings, III.* The Hague: Mouton.

Kiparsky, P. (1982). Lexical phonology and morphology. In I.-S. Yang (Ed.), *Linguistics in the morning calm.* Seoul: Hanshin.

Kuryłowicz, J. (1968). The notion of morpho(pho)neme. In W. Lehmann & Y. Malkiel (Eds.), *Directions in historical linguistics.* Austin: University of Texas Press.

Langacker, R. (1987). *Foundations of Cognitive Grammar* (Vol. I). Stanford: Stanford University Press.

Losiewicz, B. (1992). *The effect of frequency on linguistic morphology.* Doctoral dissertation, University of Texas, Austin.

MacWhinney, B. (1978). *The acquisition of morphophonology.* Monographs of the Society for Research in Child Development, *43* (Serial No. 174).

Mańczak, W. (1980). Laws of analogy. In J. Fisiak (Ed.), *Recent developments in historical phonology.* The Hague: Mouton.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T., & Xu, F. (1992). *Overregularization in language acquisition.* Monographs of the Society for Research in Child Development, *57* (4, Serial No. 228).

Matthews, P. (1972). *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation.* Cambridge: Cambridge University Press.

Matthews, P. H. (1991). *Morphology: An introduction to the theory of word structure* (2nd ed.). Cambridge: Cambridge University Press.

Nagy, W., Anderson, R., Schommer, M., Scott, J., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24,* 262-282.

Pagliuca, W. (1976). *PRE-fixing.* Unpublished master's thesis, SUNY, Buffalo, NY.

Rudes, B. (1980). On the nature of verbal suppletion. *Linguistics, 18,* 655-676.

Rumelhart, D., & McClelland, J. (1986). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sadock, J. (1980). Noun incorporation in Greenlandic. *Language, 56,* 300-319.

Sapir, E. (1921). *Language.* New York: Harcourt Brace.

Sereno, J. A., & Jongman, A. (1991, January). *Inflectional morphology in the mental lexicon.* Paper presented at the Annual Meeting of the Linguistic Society of America.

Stanners, R., Neiser, J., Hernon, W., & Hall, R. (1979). Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior, 18,* 399-412.

Stemberger, J., & MacWhinney, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition, 14,* 17-26.

Stemberger, J., & MacWhinney, B. (1988). Are inflected forms stored in the lexicon? In M. Hammond & M. Noonan (Eds.), *Theoretical morphology.* New York: Academic Press.

Tiersma, P. (1982). Local and general markedness. *Language, 58,* 832-849.

Walsh, T., & Parker, F. (1983). The duration of morphemic and non-morphemic /s/ in English. *Journal of Phonetics, 11,* 201-206.

Zipf, G. K. (1935). *The psycho-biology of language.* Boston, MA: Houghton Mifflin.