CHAPTER 13

The Diachronic Dimension in Explanation

Joan L. Bybee

As a sign system that makes use primarily of symbols rather than icons, human language is underlaid by a large number of conventionalized elements and relations. These conventionalized aspects of language must be learned and are passed on from generation to generation with minimal alteration. The language user does exert some influence over these conventions, manipulating them for expressive purposes, but their apparent rigidity has attracted much attention and evoked the metaphor that language is a system, is governed by a set of rules, and so on. The sound–meaning correspondence of lexical items is a typical example of a conventional relation, as is the set of formal and semantic relations that constitute inflectional morphology. Some aspects of grammar are subject to varying degrees of conventionalization. For instance, word order and ellipsis may be freer at the sentence level but conventionalized at the level of discourse.

In principle, conventions may be of any sort, as long as they are understood by all relevant parties. One property of conventions that has been commented on particularly with regard to language is that they may be partially or wholly arbitrary. Of course, conventions are not by necessity arbitrary, but in language it appears that they are largely so, as evidenced by the fact that each language has its own set of arbitrary pairings of meaning and sound. This observation was made by Saussure, and the framework that grew up from the acceptance of this notion predicts that the similarities among languages are due entirely to the way linguistic systems are structured – for example, the existence of contrast, of complementary distribution, of phrase structure rules and transformations, or whatever structural devices one wants to propose – while the content of these structures is arbitrary. However, the steady flow of research on empirically based language universals over the last thirty years has revealed that language or grammar is not as arbitrary as the structuralist program would predict. Rather, many similarities obtain cross-linguistically which involve not just the structure of language, but also its substance.

The existence of similarities among languages that refer to the *substance* of grammar – categories and elements such as noun and verb, subject and object, singular and plural, constructions such as passive and causative, phonological processes such as palatalization and nasalization – all point to a much richer set of general principles governing language than one would expect of an arbitrary, conventional object. However, all of these elements are part of the *conventional* aspects of language. Their cross-linguistic similarity suggests that general principles govern the way in which conventions come to be established. This means that if we are to explain similarities across languages, then we must explain what factors govern the establishment of one set of grammatical conventions rather than another.

Greenberg (1957) made this point in outlining a new program for general linguistics, one that could deal with statements pertaining to all languages, by saying that whether a particular typological pattern was frequent or rare was 'the resultant of two factors, one of origin, the other of survival' (p. 89). This statement emphasizes that synchronic states must be understood in terms of the set of factors that create them. That is, we must look to the diachronic dimension to learn how the conventions of grammar arise if we are to know why they take the particular form that they do. Unfortunately, Greenberg's suggestion is not always followed: too often we find linguists stop short of this level of explanation, being instead satisfied with the formulation of general principles offering a summary of the cross-linguistic patterns without suggesting how they might come into existence. To see the place of diachronic considerations in the explanation of language universals, let us consider a typical strategy for progressing toward explanation, illustrated by work in word order studies.

1 Towards Explanation

Often the first step towards explanation in language universals research is the statement of an *empirical generalization*, such as those formulated by Greenberg (1963), on the basis of the examination of a large number of languages, for example:

> With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.

Of course, such statements are descriptive only, and represent what needs to be explained.

The second step is to reach for a higher level of description by formulating a *principle* that ranges over several empirical generalizations. In formulating such principles, one must make some assumptions, which are theoretical in nature, about how and why the separate empirical generalizations are similar to one another. For instance, Greenberg noted that the orders of constituents in SOV and VSO languages were largely mirror images of one another, so that one could postulate some link or 'harmony' among co-occurring word orders. Vennemann (1973) suggested that this link may be found in the 'operator'–'operand' relation – that is, the linear order of modifier and modified elements is consistent in a language. Similarly, principles proposed by others, such as Hawkins (1979, 1983) and Dryer (1988), make use of this observation about grammatical relations between ordered constituents.

The formulation of such principles is an important step in theory building because it shows how a number of apparently diverse grammatical conventions are similar to one another. Many linguists seem content to have reached this stage in their theory building, because such general principles are capable of making predictions about new phenomena, for example the word orders in some language not yet studied. But it is important to point out, as Lass (1980) does, that prediction and explanation are in an asymmetrical relationship: the fact that we can predict some of the linear orders in a language if we know one does not necessarily mean that we can explain why these orders tend to correlate. On the other hand, if we can explain why the ordering of different constituents in a language tend to correlate, then we can also predict some ordering relations on the basis of others.

Vennemann (1972) proposes that an innate predisposition exists which allows the speaker/hearer to grasp the operator–operand relation and linearize pairs of elements in a consistent way in his or her language. Vennemann refers to this as an analogical process which can be represented as a single rule of linearization. Similarly, Hawkins argues that the explanation for his Cross-Category Harmony principle is that it allows the formulation of more consistent rules of ordering in a language and therefore results in a simpler grammar. These proposals are intended to answer the 'why' question.

My own intuitions about explanation are not satisfied by such principles, however, unless they provide answers to the 'how' question – how do such generalizations arise in language? What are the mechanisms that bring the state of affairs about? Perhaps seeking mechanisms or causes in language-specific synchronic studies may not be possible or interesting (Itkonen 1983), but in language universals causal factors are linguistic changes that create particular synchronic states, and the existence of massive cross-language similarity in synchronic states implies powerful parallels in linguistic change. Moreover, the identification of mechanisms that bring about synchronic states serves as a test of the principle formulated in that the validity of the principle as explanatory can only be maintained if it can be shown that the same principle that generalizes over the data also plays a role in the establishment of the conventions described by the generalization. In the current case, then, we would want to know how the analogical principles of harmony are manifested through linguistic changes.

Both Vennemann and Hawkins claim that the analogical principle (Vennemann's version is the Natural Serialization Principle, and Hawkins' the Cross-Category Harmony Principle) operates in linguistic change to create word order correlations. Hawkins puts it as follows: 'In the word order co-occurrence preferences defined by CCH [= Cross-Category Harmony], I see a strong internal motive for any language either to remain within, or to move toward, a preferred type' (p. 646). In fact, Hawkins quite explicitly claims that the existence of an implicational universal of the form 'if P, then Q' means that if a language develops a structure P, it implies the prior existence or simultaneous acquisition of Q. Of course, this follows logically, but it represents a prediction and not necessarily an explanation. The explanation must tell us what the relation is between P and Q and how the development of one influences the other. We still must seek the mechanism of change to see if these reveal a causal relationship between one structure and the other.

One might hope that the existence of the correlations is in itself proof enough that the orders of different constituents are related analogically. Unfortunately this is not so, since there are other possible explanations for at least some of the correlations. Consider the very strong correlation between the ordering of adpositions, as either prepositions or postpositions, and the ordering in a noun–genitive construction. One of the strongest correlations found by Greenberg (1963), confirmed in a larger sample by Hawkins, is stated in Greenberg's Universal 2:

> In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes.[1]

This correlation follows from both Vennemann's and Hawkins' principles, since the adposition in an adpositional phrase and the possessed noun in a genitive phrase are both the heads, or operands:

operand	*operator*
adposition	noun
noun	genitive

Now since the ordering in phrases such as these is usually fixed in a given language, that is, highly conventionalized, it is necessary to ask how these orders came to be fixed in such a way that these correlations obtain. That is, does the ordering in an adpositional phrase exert some influence over the ordering in a genitive phrase (or vice versa) and if so how is this manifested in the development of such constructions? To answer these questions, we must ask how these constructions arise in languages, and when we do, we see at least two types of strong relations between adpositional phrases and genitive phrases.

First, it often happens that adpositions develop out of genitive phrases (Greenberg 1963: 99; Vennemann 1973: 32). For instance, prepositional phrases in English such as *inside the house* and *outside the house* derive from the use of the nouns *inside* and *outside* in genitive constructions, that is, *inside of the house* and *outside of the house*. The preposition *of* is now optionally deleted, making *inside* and *outside* prepositions. Similar developments may be observed in other languages (for instance, Abkhaz, Basque, Bihari, Buriat and Kui) where a genitive marker remains in adpositional phrases. Consider the following example from Buriat (Poppe 1960).

ger-ei xazuu-da
house-poss side-loc
'at the side of the house; by the house'

Note that in Buriat the order in a genitive phrase is GN, so the resulting adposition will be a postposition, unlike English, where the adpositions are prepositions. Thus if a language that has a productive order NG produces new adpositions through the genitive construction, it will produce prepositions, while a language with a productive order GN will produce postpositions.

Interestingly enough, we can find causality moving in the opposite direction as well. A frequent source of the genitive marker is an adposition, as in English where the more recently developed genitive *of* is a preposition. Given that the genitive marker appears between N and G if it is an adposition, a language with prepositions will development a new genitive construction in the order NG, while a language with postpositions will develop a new genitive with the order GN.

Both of these frequent diachronic developments contribute heavily to the correlation of adpositional and genitive phrase orders.[2] Yet in neither case do we find analogy in the form of rule simplification playing a role. One grammatical order is not established on analogy with or to harmonize with another order, rather a new grammatical construction develops in a language out of constructions that already exist and shows ordering consistent with the construction from which it developed.

Considering these diachronic sources for the correlation raises a whole new sheaf of questions, and impels us to ask how and why one construction is formed out of another. Ultimately, we are brought back to the synchronic plane where we must ask what cognitive processes are behind the development of genitive markers from other types of adpositions, and what motivates the development of new adpositional phrases from nouns in genitive constructions. Thus, having consulted the diachronic domain, we find that the questions we want to ask are very different from the ones we were considering when only synchronic cross-linguistic generalizations were being considered.

This example is meant to illustrate what I consider to be the necessary third step following the formulation of a *principle*: testing the principle to see if it can be shown to be actually involved in the diachronic processes that lead to the states described by the principle. I have given two examples that show developments leading to states describable by a principle that are the result of processes that are independent of the principle. This evidence does not falsify the principle, but it does diminish its explanatory power. I would suggest that further work on universals of word order take into account to a greater extent how the correlating structures develop historically if the goal of such work is the explanation of language universals.

2 Principles as Constraints on Change

Principles of the type we have just been discussing (that is, Cross-Category Harmony or Natural Serialization), which define syntactic typologies, have also been invoked as constraints on possible diachronic changes. For instance, Hawkins (1979, 1983) discusses this possibility: 'All languages in their evolution are constrained by implicational universals such as have been defined, and can change only relative to the co-occurrence possibilities which these permit' (Hawkins 1979: 647). That is, implicational universals may be used to predict linguistic change in the sense of setting the upper limits of such changes. However, such constraints cannot be invoked as explanations unless the mechanism by which the universals constrain change can be explicated.[3]

An unfortunate tendency exists nevertheless to invoke typological facts as *explanations* for particular historical changes. A very specific case is found in Fleischman (1982), where she discusses the question of why the Romance future formed from the Infinitive plus postposed *habeo* underwent fusion to produce a synthetic construction, while the Perfect from preposed *habeo* plus Past Participle did not fuse. She says that since *habeo* has person/number suffixes, fusion of preposed *habeo* would result in internal person/number inflection:

> With preposed *habeo*, fusion would have (a) phonetically reduced or obliterated altogether the person-number information, which was carried by the final syllable of *habe-o*, *-es*, *-et*, etc., and (b) created the anomalous situation, for languages such as Latin and its offshoots, of having prefixed or infixed inflections. (p. 115)

Fleischman presumably believes that it is normal for the fusion of grammatical morphemes with stems to take place in the position in which they develop, unless otherwise blocked. The possibility of preventing the loss of information is often invoked as an explanatory factor in change, but this cannot be a very powerful force, if it is a force at all, since language change does in fact in so many instances bring about the loss of morphological information. The other potential preventive is more interesting in the present context, since it is an appeal to typological principles. Not only would it be anomalous for Latin and her daughters to have person/number markers closer to the verb stem than a tense/aspect marker, it is an extremely rare situation in any of the world's languages (although it does occur, for example in Athapaskan languages, such as Navaho Bybee 1985b). But this typological fact cannot explain why affixation did not occur in this particular case, for it is certainly not possible for the speakers who were tending toward this change to test out the results, compare them with existing typologies and decide against the change. Moreover, this type of 'explanation' has the relationship between diachrony and universals reversed. Since the order of grammatical morphemes, and particularly affixes, is a matter of convention, we must look to the manner in which such orders are established to explain cross-linguistic patterns. We cannot use cross-linguistic patterns to explain why a particular change is or is not implemented.[4]

Now consider a similar example from Comrie (1980). Comrie considers the problem of the development of subject agreement suffixes in an SOV language, where the pronouns from which they evolved could occur either before or after the verb. However, only the pronouns that occurred after the verb were affixed, not the preposed pronouns. Among the three potential explanations Comrie suggests is the following:

> Since the languages in question are already exclusively or overwhelmingly suffixing, preference is given, in deriving affixes, to that word order which produces further suffixes rather than introducing prefixes, i.e. in accordance with the existing patterns of the language.

This statement implies that speakers in some sense 'decide' on suffixes rather than prefixes, as more appropriate for their language. This would mean that the typology of a language is manifested in individual grammars in such a way that it may come into play during the creation of new structures. Is this what Sapir calls 'the structural "genius' of the language' (Sapir 1921: 120)? It may be, but exactly what it is, how it is manifested in the grammar and how it functions in language change has not been investigated.

Again I would argue that we must seek instead causal mechanisms. In this case, Comrie actually offers two other potential explanations which are causal in nature, given the assumption that grammatical material tends to reduce and fuse. The first is that the pronouns in postposed position are unstressed, while the pronouns that occur before the verb are stressed. Thus, as Comrie observes, the postposed pronouns are more likely to reduce and fuse. The second factor is that only the postposed pronouns are consistently adjacent to the verb. The subject pronouns in initial position are separated from the verb by an object or other complement in many cases. Both of these factors point to the greater likelihood of fusion for the postposed pronouns. Both contribute to the causal mechanism; neither refers to the resulting state.

These cases illustrate that typologists are well aware of the fact that cross-linguistic generalizations must be considered together with linguistic change, but they are not always consistent about the proposed relationship between the two. Causal and non-causal explanations are often invoked indiscriminately. My suggestion is that complete explanations must specify a causal mechanism; thus we cannot explain change with reference to preferred types, but we must explain common types by referring to the factors that create them.

3 Identifying Causal Mechanisms

Most of the attempts at explaining grammatical phenomena that are frequent cross-linguistically proceed by identifying certain factors that show the phenomena in question to be beneficial from the language user's point of view. These might be called synchronic explanations; they are explanations based on processing ease, on iconicity, on cognitive or semantic factors, or on typical discourse structure. In order for these factors to qualify as explanations, a causal connection between the factor and the grammatical phenomenon must be demonstrated: that is, it must be shown that the factor appealed to as explanation actually contributes to the creation of the particular grammatical convention. Let us consider several cases to see what this would entail.

3.1 *Processing*

Consider first a processing explanation proposed by Cutler, Hawkins and Gilligan (1985) for the greater frequency cross-linguistically of suffixes over prefixes. They cite psycholinguistic evidence that the beginnings of words are more salient than the ends, and that stems are processed before affixes. They propose the following: 'the stem favors the most salient beginning position of the word, and the affix the less salient end position, because in the compositional process of determining the entire meaning of a word from its parts, the stem has computational priority over the affix' (p. 748). As Hall (this volume) points out, this may well be true, but it does not qualify as an explanation until it can be shown that the computational process referred to actually contributes to the formation of suffixes, or impedes the formation of prefixes. It must be remembered that the actual process of forming affixes begins gradually as lexical material reduces and grammaticizes. This fact creates two problems for the proposed principle. First, since affixes can be demonstrated to begin as stems, this principle would seem to predict that affixes cannot develop at all, since their development entails the reduction of a stem. Second, grammatical morphemes (hereafter, *grams*) are usually fixed in their position long before they actually fuse with a noun or verb. Thus it is important to ask whether non-bound grams also tend towards postposing, and if so whether the processing principle described above contributes to this positioning. Another possibility is that postposed material tends to fuse more often than preposed material. In this case, the processing principle might act as a retardant to fusion, which is in itself a primarily phonological process. Alternatively, preposed grams may more often be separated from their semantic hosts by other lexical material (as in the case cited above from Comrie 1980), which prevents their fusion. If this is so, then processing order has nothing to do with affix order. The viability of a processing principle as an explanation depends upon a complete understanding of the factors involved in the creation of grammatical structures, and a demonstration that processing ease is one of those factors.

3.2 *Iconicity*

With the recent interest in universals and non-arbitrariness in grammar has come the suggestion that some grammatical and lexical structures are iconic. Haiman (1983) proposes the 'distance principle';[5]

> The linguistic distance between expressions corresponds to the conceptual distance between them.

The 'linguistic distance' has to do with expression units which range from most distant to least distant as follows:

two separate words with the possibility of intervening material;
two separate words, but always contiguous;
a stem and affix;
a single lexical unit.

Haiman proposes several ways of viewing conceptual distance. We will only consider one, that is, 'two concepts are close to the extent that they are perceived as inseparable' (p. 783). This principle makes very specific predictions with regard to the structure of causatives. To state the prediction informally: if a language has more than one way to express causation and these differ in their 'linguistic distance' or degree of fusion, 'then the conceptual distance between cause and result will correspond to the formal distance between cause and result' (p. 783).

This may be illustrated with the well-known 'kill' vs. 'cause to die' examples from English. The lexical expression of causative implies very strongly that the cause and result take place at the same time and place, with possible physical contact, while the periphrastic expression implies the opposite. Consider the following pairs that Haiman offers as examples:

(1) I caused the tree to fall.
I felled the tree.

(2) I caused the chicken to die.
I killed the chicken.

(3) I caused the cup to rise to my lips.
I raised the cup to my lips.

Examples similar to these, as well as examples illustrating causatives that appear in other expression types, may be cited from many different languages.

Haiman's principle of distance in this manifestation is very similar to the 'relevance' principle that I propose in Bybee (1985a) and (1985b). My hypothesis is that the degree of fusion of grammatical material with lexical depends upon the semantic relevance of the grammatical morpheme to the lexical: the extent to which the grammatical meaning directly affects or modifies the semantic content of the lexical morpheme. With this principle even finer degrees of fusion or 'distance' between linguistic units, such as the ordering of affixes, the existence of allomorphy, irregularity, stem change and suppletion may be predicted.

As general principles concerning the relation of meaning to form, these principles are far-reaching and seem to grasp some essential quality in the structure of language. However, despite making correct predictions, they do not qualify as explanations. They certainly tell us what to expect in the expression of particular notions, but to the extent that these principles deal with conventionalized structures, such as affixes or lexical items, they would need also to explain how these structures become conventionalized, and in that respect these principles fail to be explanatory. Haiman does not claim that iconicity provides an 'explanation' for the structures he discusses; he prefers instead to use the term 'iconic motivation'. I argue in Bybee (1985a) and (1985b) that we need evidence that semantic relevance (or distance in Haiman's terms) actually influences the diachronic process of fusion that leads to the formation of affixes.

This argument is made in the following way: Two elements may become fused if they occur next to one another frequently. This means that at a stage in which the positioning of these elements is not yet conventionalized, if speakers nevertheless place them together very frequently, they are likely to become fused. The motivation for placing them together in the speech stream may very well be their semantic closeness to one another. If this is so, then the proposed principle, when coupled with other principles in a general theory of affix formation, may be said to be explanatory.

The argument might also be made that lexicalization, which is the formation of new autonomous lexical items, is influenced by relevance. Thus if two elements together form a semantic complex that can be taken to be unitary and discrete from other semantic complexes, the two elements might together comprise a single lexical item. Questions of categorization patterns in the lexicon relate to the general psychological issue of how categories are formed and structured (Berlin and Kay 1969; Rosch 1978). Research in this area indicates that the nature of linguistic categories depends a great deal upon how we perceive reality. If this is so, then it is not clear that the phenomenon actually belongs in the domain of iconicity.

Thus until we understand better the way that fusion occurs, how affixes are formed and what factors encourage or impede their formation, how new lexical items are formed and become autonomous, we cannot demonstrate that the relevance or distance principles are explanatory.

3.3 *Economy*

In the same article, Haiman (1983) also discusses 'economic motivation' which is the principle that the 'simplicity' of words or expressions is an index of their familiarity or frequency. He cites Zipf's 'principle of least effort' (Zipf 1935) which is intended to explain why the more frequently used words of a language tend to be shorter than the less frequently used words. He also cites work by Givón and Bolinger which suggests that the more familiar, predictable information is signalled in shorter units than the less familiar, less predictable information.

Haiman's discussion is largely ahistorical so no mechanism is proposed to explain why the more familiar or frequent expressions are generally shorter. Of course, where speakers have a choice – for instance, in how many lexical items to include in a phrase or clause – they are probably motivated by economy. However, in the conventional aspects of language, the speaker is limited by the established elements. If pronouns are shorter than full nouns, if auxiliary verbs are shorter than main verbs, if 'horse' is shorter than 'elephant', the speaker must use them all the same. So what is it that motivates the economy in grammar? The choice of the phrase 'economic motivation' suggests that familiar words grow shorter *in order to* make familiar conversation more economic. But this does not tell us what the mechanism behind the principle is.[6]

While one might expect that research into the implementation of such an obvious question would have progressed quite far, unfortunately the role of frequency of use in conditioning phonological change has been neglected while structural factors in change have received primary attention. However, the casual observation that frequent words and phrases undergo reductive change at a faster rate than infrequent ones (Schuchardt 1885) has been documented by numerous examples in the work of Mańczak (Mańczak 1978), and by comparison of frequency counts with differential reduction by Fidelholtz (1975), Hooper (1976) and Pagliuca (1976). If frequent words undergo phonetic reduction at a faster rate than infrequent ones, then Zipf's correlation is created. The question, then, is what *causes* reductive sound change to progress more quickly in frequent words? If reductive and assimilatory sound change is caused by a sort of physical economy of articulation, why isn't that economic motivation equally applicable in infrequent words and phrases? In the end we must admit that the 'principle of least effort' labels a correlation or at best describes the outcome of change, it does not provide an explanation for the facts of phonetic reduction.

3.4 *Discourse*

Discourse-based explanations for grammatical phenomena typically do involve a diachronoic dimension and do identify a causal mechanism, thus coming closer than the other principles we have discussed to constituting valid explanations. The structure of such explanations is as follows: A certain configuration of syntagmatic, grammatical or semantic elements is shown to be frequently occurring in discourse, due to the way that information flow is typically structured. These frequently chosen patterns, it is argued, become rigidified or frozen into syntactic rules in some languages. Thus what is optimal discourse structure in one language is grammatical rule in another.

A good example of this sort of explanation is found in Du Bois (1985), which treats the discourse basis of ergativity. Du Bois finds that natural discourse in Sacapultec is structured such that a clause, whether it be transitive or intransitive, rarely contains more than one full noun phrase.[7] In the intransitive clause this one noun phrase is the subject, while in the transitive clause it is the object. The subject of the transitive verb occurs very rarely in discourse. Du Bois argues that this strong discourse tendency gives rise to the 'absolutive' case – it is the case of the noun phrase most frequently occurring in discourse. The rare case – that of the subject of the transitive verb – is a category apart, that is, the ergative case.[8] Du Bois argues further that the 'preferred argument structure' derives from the general preferred information flow of Sacapultec discourse, in which new protagonists are introduced by intransitive (usually motion) verbs and thereafter referred to only by agreement morphology. Only the objects of the verbs to which they are agents receive full noun phrase coding.

The causal factor of this type of explanation is frequency in discourse. The mechanisms that must be understood are discourse structuring and the process of grammaticization, whereby a frequent grammatical structure changes from being preferred in a certain context, to being obligatory.[9] At our present state of knowledge, then, discourse explanations seem to involve fewer unknown factors than processing factors, iconicity or economy.

4 Diachrony in Explanation: the Semantics of Futurity

4.1 *The phenomenon*

In this section we take up an explanation for a much observed language 'universal' in which diachrony plays a very important role. The phenomenon in question is a largely semantic one, which makes it appear somewhat different from the universals involving structure which we have been discussing. However, I will argue in the conclusion that the role of diachrony is similar in all these cases.

The phenomenon in question is the strong tendency for morphemes whose primary function is to signal future time reference to also express, directly or as secondary meanings, various modality senses. This tendency has been observed for specific languages, and cross-linguistically by Fries (1927), Ultan (1978), Fleischman (1982), Chung and Timberlake (1985), Dahl (1985) and undoubtedly many others. Most of these authors have assumed some cross-linguistic similarity among what are called 'Future' morphemes in the languages of the world, but this cross-linguistic similarity has been made more precise in the study of Dahl (1985). Dahl reports on the results of a questionnaire survey of sixty-four languages, in which native informants translated more than two hundred sentences designed to cover the major uses of tense and aspect morphemes in the languages of the world. Dahl measured the overlap in the uses of tense and aspect morphemes and postulated a small number of prototypical cross-linguistic categories. A category he labeled FUTURE was one of the most common of these. The uses of this category are defined by the sentences in the questionnaire that most commonly took future marking in the languages of the sample. One very common use of future morphemes in his data is the prediction use: specifically, in sentences in which the speaker is making a prediction about future time, as in Dahl's number (36):

(4) [It's no use trying to swim in the lake tomorrow]
The water BE COLD (then).[10]

This use probably represents what most linguists would identify as the 'pure' future sense, without modal overtones. This prediction sense is, according to the analysis done by Coates (1983) on both spoken and written corpora of British English, the most common use of the three future markers of English, *will*, *shall* and *be going to*, as illustrated by these examples, some of which come from the corpora studied by Coates and some from that used by Wekker (1976):

(5) I think the bulk of this year's students *will* go into industry. (Coates 1983: 170)

(6) I've given him sedation and he'*ll* be all right for a bit.
His beauty *will* be permanently spoiled, but I don't suppose it was ever very much. (Wekker 1976: 61)

(7) We *shall* no doubt live to see stranger things. (Wekker 1976: 44).

(8) (in reference to taking 'just water') Otherwise I *shall* end up like the song The Seven Drunken Knights. (Coates 1983: 186)

(9) Within a few years at the present rate of development, Paris *is going to* look like London, and London like New York. (Wekker 1976: 125)

(10) I think there'*s going to* be a storm. (Coates 1983: 201)

It is important to note that future morphemes in clauses that are not predictions, but do make future time references, are not nearly as common, and in fact, in some cases are not grammatical. Thus in English, *will* does not occur in *when* clauses, even where the clause refers to future time:

(11) When you (*will) see him, give him this message.

(Cf. Dahl's questionnaire, where only seven out of forty-seven languages with future markers use a future in *when* clauses.)

Another extremely common use of morphemes labeled as futures occurs in cases where either prediction (by the speaker) or intention (of the subject of the clause), or both, are signaled, as in the second clause of the following examples from Dahl's questionnaire:

(12) [Said by a young man]
When I GROW old, I BUY a big house.

(13) [The boy is expecting a sum of money.]
When the boy GET the money, he BUY a present for the girl.

These two examples received the highest number of future markings in the languages of Dahl's sample. (Forty-two out of forty-seven languages use the future in these examples.) Note that in (12) with *will* in English ('When I grow old, I'll buy a big house') the most salient interpretation is that this is a statement of the speaker's intentions, and not a simple prediction. The example (13) with *will* ('When the boy gets the money, he'll buy a present for the girl') could be interpreted as a prediction made by the speaker, or as a statement of the subject's intentions. The intention uses of futures, then, represent one case of the oft-cited overlap of future with modal senses.

The data on other modal senses of futures is often more impressionistic. The cross-linguistic survey conducted by Ultan (1978) and the diachronic survey of Fleischman (1982) suggest the following non-temporal uses of future morphemes, some of which are also non-modal: desire, intention, obligation, necessity, habitual, general truth, characteristic behavior, imperative, optative, hortative and supposition.[11] Some of these senses are more closely related to one another than others are. I propose to break them down in the following way, in order to discuss them in groups:

Desire, intention, obligation and necessity are called agent-oriented modalities because they predicate certain conditions on an animate, usually human, agent. (*Desire* in this case does not refer to expressions of the speaker's desire; that is referred to as *optative*.) The corresponding expressions in English would be *want to*, *is going to* and *have got to*, as in the following examples.

(14) She wants to practice her Spanish with you. (*desire*)

(15) She's gonna apply to the graduate school. (*intention*)

(16) She's gotta help her mother on Saturday. (*obligation*)

(17) I gotta get eight hours of sleep or I'm wrecked. (*necessity*)

Obligation and necessity are very closely related: obligation is socially imposed while necessity is physically imposed.

Habitual, general truth and characteristic behavior can be regarded neither as tense nor as modality notions. They are related to each other, however, since they all signal that the same situation holds on different occasions. Examples of future morphemes used in this way are:

(18) Boys will be boys.

(19) Whatever you say to him, he will not answer.

Imperative, polite request, optative and hortative are related in that their function is to get the addressee to do something. They can be regarded as marking a speech act of a certain type, one which either imposes an obligation on the addressee or expresses the speaker's wishes.

Supposition is the term used by Ultan and Fleischman for expressing the epistemic notion of probability. A future marker is sometimes used for stating propositions that are probably true in the present, for example:

(20) *English* (on hearing the phone ring)
That'll be John now.

(21) *Spanish*
Tendrá veinte años. She is probably twenty years old.
have – future twenty years

Since future markers fulfilling these same non-tense functions may be found in unrelated languages, some general and non-specific explanation must be sought. Some authors have informally proposed *principles* which attempt to 'explain' the overlap of future and modal semantics by referring to the uncertainty of future events. Consider Ultan's statement:

> The reason for the preponderance of modal applications of future tenses must lie in the fact that most modal categories refer to differing degrees of uncertainty, which correlates with the element of uncertainty inherent in any future event, while past tenses generally refer to completed, hence, certain, events. (pp. 105–6)

Chung and Timberlake hold a very similar view, as seen in the following:

> Situations in the future are inherently uncertain as to actuality. Any future event is potential rather than actual. . . . The future is thus a semantic category where tense and mood merge. (p. 243)

Both of these statements associate future with epistemic modalities of possibility and probability, by noting that the future is uncertain. The problem with this view is that it is too simplistic to say that the past is certain and the future uncertain. Some events in the future are quite certain (for example, 'The sun will rise in the east tomorrow'), while some events in the past may be uncertain in the sense of being unknown and even unknowable (for example 'All human languages developed from a common proto-language'). Moreover, natural languages provide us with various means of expressing uncertainty about the past and present in the form of evidentials or epistemic modals such as *may* and *might* (for example, 'I might have left that book at the office').

Another problem with the view that futures are associated with modalities because the future is inherently uncertain is that many of the non-tense uses of futures do not imply uncertainty at all. The characteristic behavior or general truths use of English *will*, as in 'Water will boil at 100 degrees centigrade', does not give a sense of uncertainty, nor in fact does the prediction sense, illustrated in (5)–(10), where the interpretation is that the speaker actually believes that the event will take place. To see that this is so, consider the same sentences with a better indicator of uncertainty, such as *may* or *might*, in them rather than the future morpheme.

While the statements of Ultan and Chung and Timberlake associate future with epistemic modality, the following statement by Dahl 1985 attempts to associate futurity with the agent-oriented modalities of intention and obligation, as well as with the epistemic modalities:

> Normally, when we talk about the future, we are either talking about someone's plans, intentions or obligations, or we are making a prediction or extrapolation from the present state of the world. As a direct consequence, a sentence which refers to the future will almost

> always differ modally from a sentence with non-future time reference. This is the reason why the distinction between tense and mood becomes blurred when it comes to the future. (p. 103)

While Dahl associates future with a wider range of modalities, his statement, like the others, does not explain how or why future morphemes acquire modal uses. Moreover, none of these statements is able to correctly predict *which* modalities will be associated with future tense.

4.2 *Diachronic lexical sources*

One suggestion (as outlined in Bybee and Pagliuca 1987) is that the semantics associated with futurity may be accounted for in a general theory of the development of grammatical morphology.[12] Grammatical morphemes (henceforth, grams) develop from lexical material, either single lexical items, such as the Old English verbs *willan* or *sceal*, or polymorphemic sequences such as *be going to*. We propose that the original lexical semantics, in conjunction with very general principles of change, determines the course of development, and that there are certain universal paths for the development of grams. The evidence for this is the fact that in many unrelated languages the lexical sources for grams are the same or very similar. In the case of futures three very common sources are:

1 *Desire*: an auxiliary verb with an original meaning of 'want' or 'desire', or less commonly a derivational desiderative morpheme, which in turn has as its source a main verb meaning 'want' or 'desire'.[13]

2 *Movement towards a goal*: a verb meaning 'movement towards a goal' or a movement verb in construction with an allative adposition such as *to*, or less commonly a derivational andative morpheme, which has as its source a verb meaning 'movement towards a goal'.[14]

3 *Obligation*: a verb meaning 'to owe' or 'to be obliged', or more commonly a construction with a copula or possession verb and a non-finite main verb, such as English *to have to*, or *to be to*.[15]

With their original meanings, these constructions have very specific semantics, which in each case requires a human, or at least an animate agent. Their development into future grams requires a loss of some specific semantic features, which allows an extension to contexts in which the agent is not human or animate.

Since these three distinct semantic complexes all eventually develop into expressions of future time reference, their paths must converge at some point. Our study of the history of *will*, *shall* and *be going to* in English (which represent the three most common sources of futures) reveals that the convergence of paths of development begins early as each of these constructions is used to state the intentions of a first person subject. *Shall* is frequently used during the Old English period in its original meaning to express obligation, and also to state an intention by the speaker. Consider the following example from *Beowulf*.

(22) Ic thæm godan *sceal*, for his mod-thræce, madmas beodan. (Beowulf, line 384)
I shall offer the good (man) treasures for his daring.

Similarly, *will*, which is not frequent in Old English, becomes more frequent in Middle English and is used both in its original meaning of 'want' and to express the speaker's intention. Consider these examples from *Sir Gawain and the Green Knight*.

(23) I *wyl* nauther grete ne grone. (line 2157)
I will neither cry nor groan.

(24) Now *wyl* I of hor seruise say yow no more. . . . (line 130)
Now I will tell you no more of their service.

Be going to is much more recently developed as an expression of intention or future, dating from the seventeenth century (Scheffer 1975). One of its current uses is the expression of intention, as in the following examples from the corpora examined by Coates (1983: 199).

(25) Listen, my dear, I asked you to marry me, didn't I?
And *I'm going to* do my very best to make you happy.

(26) We're not *going to* let you walk home on your own.

These examples show that grams from all three lexical sources develop a use which expresses the speaker's intention before they progress to the point of expressing prediction or future time reference. But the convergence at this point is only partial: for while they may all express intention, they each have other uses that are not shared, uses associated with their lexical meaning. In addition, there may be different implications in the expression of intention. An intention may have an external motivation, in a social obligation or physical necessity, or it may originate internally in the goals and desires of the agent. Although it is difficult to know exactly how to interpret *shall* and *will* in the older texts, there may be, along with the expression of intention, a flavor of obligation for *shall* and desire for *will* retained from their original lexical senses. Consider the following example from *Sir Gawain*, where *shall* and *will* are juxtaposed in the same sentence:

(27) And I *schal* erly ryse, on hunting *wyl* I wende. (lines 1101–2)

If the choice of *shall* and *will* is not random, then it may be that their positions in this sentence are governed by the expression of obligation and desire respectively. That is, the speaker *wants to* go hunting, and consequently he *has to* get up early. (In this context, hunting is a sport, not a necessity.)

Grams that are used to express intentions are not necessarily future markers. As mentioned above, the defining use of a 'future' is to make a prediction in a future temporal frame. A prediction is a type of assertion made by the speaker, in which the future marking has propositional scope. A prediction will be free of agent-oriented meaning, such as obligation, intention or desire. A marker of prediction can be used in a sentence with a non-animate subject.[16] As illustrated above in sentences (5)–(10), the Modern (British) English *shall*, *will* and *be going to* all qualify as 'futures', since they may all be used in predictions about future situations. Since the Middle English period, then, these future grams have continued to develop by losing more and more of their original lexical meaning, expanding their scope to include the whole proposition and extending to clauses that have inanimate subjects.

Despite the fact that both *shall* and *will* have undergone a long history of development and have reached the stage of expressing prediction, they both retain some remnants of their original lexical meaning in specific contexts. For instance, in the following examples, *will* expresses the willingness of the agent. Willingness is not as strong as desire, but it is related to desire in that it refers to an internally motivated disposition.

(28) Give them the name of someone who *will* sign for it and take it in if you are not home. (Coates 1983: 171)

(29) If he *will* meet us there, it will save a lot of time.

In negative sentences, *won't* often has the sense of 'is unwilling to' or 'refuses to'.

(30) He *won't* eat meat, even if it's offered to him.

Shall can be found in the somewhat archaic expression of obligation in decrees and laws (Coates 1983), but in colloquial British English it is restricted to the first person, where it is used to state intentions, make predictions (see examples above) and in questions to ask the addressee's will.

(31) *Shall* I ring at 11 p.m. one night (English time) in the week after you get back? (Coates 1983: 186)

We argue in Bybee and Pagliuca (1987) that this use is a direct descendant of the obligation sense of *shall*, since it does not speak of internal motivation, but asks for external motivation. Note that *will* is inappropriate in such a sentence, since *will* there could only be interpreted as 'do you predict that I . . .' or 'do I want to . . .' Note further that in the history of English *shall* does not develop any readings related to desire, and *will* does not develop any readings related to obligation.[17]

4.3 *Universal paths*

Since the same three lexical sources for future grams appear in so many unrelated languages, and since the product of their semantic development is so similar cross-linguistically (see Dahl 1985), it seems safe to assume that the paths of development for future grams are universal. If this is so, then the 'explanation' for the existence of desire, obligation and intention readings for futures is diachronic: desire and obligations are older meanings 'glimmering through' (as Fries 1927 puts it), and intention is a use that predates the development of the pure prediction use, but remains after prediction develops. This theory is also specific about the possible combinations of modal nuances that may co-exist in a particular gram. That is, a future gram may express obligation, intention and prediction, or desire, intention and prediction, but not desire and obligation, since the latter two imply two distinct lexical sources.[18]

Before discussing the other non-temporal uses of future grams, let us consider the nature of the explanation just offered. We have identified what might be called an 'immediate cause' for the existence of readings related to obligation, desire and intention in future grams. Furthermore, by postulating that paths of development for futures are universal, we are able to predict linguistic changes as well as the synchronic range of uses for future grams.

Identifying an immediate cause is necessary to ensure that the search for explanation is on the right track, but it is not the end of the process. Rather, behind the immediate causes are a further set of more general and more interesting questions. In particular it is known that the type of semantic change found in grammaticalization is not restricted to future grams, but can be found in the development of all grammatical morphemes (Givón 1979, Lehmann 1982, Traugott 1982, Heine and Reh 1984, Bybee and Pagliuca 1985). Such change is characterized by three interrelated processes:

1 The gradual loss of specific components of meaning in some contexts, such as the loss of the desire sense of *will*.

2 The generalization of meaning or function with the result that the gram may appear in more contexts. This process is tied closely to the loss of specific meanings, since more specific meanings usually imply more co-occurrence restrictions, for example if *will* means 'desire' it may only occur with an animate agent. If it loses this sense, then it may appear in contexts with inanimate agents as well. Such generalization of function corresponds to an increase in text frequency.

3 Often, grams developing from auxiliary verbs increase their scope from verb phrase scope, which the agent-oriented senses of desire, obligation and intention have, to propositional scope, as found in the prediction sense of futures (for other examples see Bybee and Pagliuca 1985 and Traugott 1982).

These general principles of change can predict a large part of the course of development of futures, as well as that of other grams, such as pasts, perfectives, imperfectives, demonstratives, definite and indefinite articles, and so on. However, these principles are themselves in need of explanation. Why do certain lexical items undergo these related semantic changes and develop grammatical characteristics? To answer this question we are drawn back to the synchronic plane to investigate the way language is used. This part of the investigation has not been undertaken thoroughly as yet, so I will only make a few speculative remarks on some possible directions such an investigation might take.

First it should be observed that there are certain functions that language is called on to perform quite commonly. For instance, linguists who study narrative propose that nearly every sentence of a narrative can be assigned to the foreground of the narrative or to the background (Hopper 1982). For our purposes, spoken face-to-face interaction is more relevant, and here we would hypothesize that the expression of intentions and the making of predictions are very common functions. Some support for this hypothesis comes from Coates (1983), who gives a tally of the frequency of 'modal meanings' (that is, meanings expressed by modal verbs in English) in her corpora of both spoken and written British English. Her figures show indeed that prediction is the most frequently expressed of these 'modal meanings', and that intention is also very frequent, compared to other meanings. That does not imply, of course, that a language must express these functions with grammatical markers. In fact, many languages express these functions with an unmarked verb form. It does mean, however, that linguistic material expressing intention or prediction could have a very generalized range of use.[19]

A second factor is the well-known tendency for people to 'speak as though', that is, to use language metaphorically rather than literally. For instance, Fleischman (1982: 59) comments on the use of *have to* in English to state an intention as though it were born of an obligation, even though it is not:

(32) What are you doing tonight? Oh, I have to go to a party.
Later today? I have to go jogging at six . . .

Such uses generalize the function of the phrase *have to* and weaken its obligation sense. Another relevant instance of this phenomenon is the use of modal verbs with inanimate objects:

(33) This door doesn't want to open.

A third tendency instrumental in the semantic changes accompanying grammaticization is the tendency to take what one may usually infer from the meaning of an expression to be its actual meaning. Neither *have to* nor *want to* (taken as examples because they are semantically similar to the lexical sources of *shall* and *will*) literally express intention, yet if I say I want to do something, or I have to do something, and it is within my powers to do it, the hearer will infer that I intend to. Since in most cases this inference will be correct, it can become an obligatory inference of the phrase, and as the desire and obligation components weaken, it can become the main function signaled. Similarly, the relation between intention and prediction depends upon inference. Since there is a fairly high level of agreement between our stated intentions and acts carried out, one can usually take a stated intention as predictive.

I would suggest then that function, metaphor and inference are the synchronic phenomena that need to be investigated in order to understand the overlap of modality with future meaning. We reached this conclusion by first identifying the causal mechanisms that bring about the overlap of modal and future meanings. Note that the direction of the investigation is quite different than it would had been if we had tried to investigate the 'uncertainty' of the future more directly. In fact, I would claim, such an enterprise would not have been fruitful at all.

4.4 *Probability*

Let us turn our attention now to the use of futures to express the epistemic modality of supposition or probability. The particular sense of probability that futures express is a probability about a situation existing at the moment of speech, in other words, in present time. For example:

(34) *English*:
A commotion in the hall . . . 'That *will* be Celia', said Janet. (Coates 1983: 177)

(35) *Spanish*:
Ya tú comprenderás cómo nos reímos. 'Now you probably understand (future) how much we laughed.' (Moreno de Alba 1970)

Judging from Dahl's questionnaire data, this is not an extremely common use of futures. It occurs in seven of his forty-seven languages with futures (15 percent). Of these, four are Indo-European languages, and the other three are each from different families.

This use is semantically very close to the prediction use of futures. It does in fact constitute a prediction, but not a prediction about future time, rather a prediction about the present in cases where direct evidence is not available. Besides this semantic relation, there are two indicators that this use develops out of the prediction use of futures. First, it does not seem to be specific to any one lexical source of futures, since it occurs with a desire-derived future in English (as in the example above) and in Greek, with the obligation-derived future in Spanish (again see above) and Italian, and with the future derived from *ir a* 'to go to' in Spanish (Moreno de Alba 1970, no example provided) and a movement-future in Sotho. Second, it appears to be a relatively late development, compared to the development of intention and prediction. This use of *will* in English is documented by the OED from the fifteenth century, while the intention and prediction uses begin in Old English. In the Mexican Spanish corpus studied by Moreno de Alba, the older synthetic future was found in this use in three times as many cases as the newer periphrastic go-future.

4.5 *Imperative*

In some languages the future and the imperative (or optative or hortative) have the same form, or the gram used for future is also used for imperative (for example, in Atchin, Alyawarra, Danish, Maidu, Motu, Nimboran and Yagaria). Thus it could be said that this is another case of mood or modality overlapping with future. These particular mood functions for futures are all directive speech act indicators: an imperative is a speech act by which the speaker assigns an obligation to the addressee. An optative is an expression of the speaker's wish or will, while with a hortative the speaker urges the addressee to action. The use of futures in optatives and hortatives is not so easy to verify, and may involve the use of an additional morpheme along with the future. Let us concentrate on the use of future grams in imperatives.

While philosophers often associate imperative with the deontic modality of obligation, an important distinction must be made for our purposes. The agent-oriented modality of obligation (for example, *must* or *should*) states that an obligation applies to the agent in the clause.

(36) The students must get the permission of the instructor before registering for this course.

An imperative marker, on the other hand, has the whole proposition in its scope, and signals that the speaker is performing a certain type of speech act. The use of the future for direct commands is easy to verify in English (*You will go to bed!*) and in other languages, but this type of imperative is usually secondary, and not the primary means of commanding. Such a use constitutes an indirect speech act, that is, a prediction is made in the second person, which has the force of an imperative, given the social context and intonation. This type of imperative use of the future should be derivable from any of the lexical sources of future, since it is an adaptation of the prediction sense. Another potential source is specific to obligation-derived futures. An imperative gram may derive from the second person of an agent-oriented obligation marker, such as *must* in *You must go to bed*. This path of development is documented in Tamil and Malayalam (Subrahmanyan 1971). If the same obligation marker in other persons developed into a future, then a situation would be created in which a future is used in the imperative. I know of no such documented cases, however.

4.6 *Characteristic behavior*

The aspectual functions attributed to future grams are listed by Ultan as gnomic or general truth, and customary or habitual event. A closer examination of the actual cases, however, reveals that the term 'characteristic behavior' (as used by Fleischman 1982) is a better description and covers both of the categories that Ultan seeks to establish. Gnomic or general truth statements that involve permanent states do not use the future (at least not in English):

(37) Elephants have long trunks.
(38) *Elephants will have long trunks.

Rather the type of timeless statements that take the future involve a change of state, or a characteristic behavior.

(39) The arctic hare will turn white in winter.
(40) Water will boil at 100 degrees centigrade.

The alleged cases of the future used to mark habitual action also fit better under the rubric of 'characteristic behavior'. For instance in Hausa, the movement-derived future *za* is sometimes used in statements such as the following (Kraft and Kraft 1973):

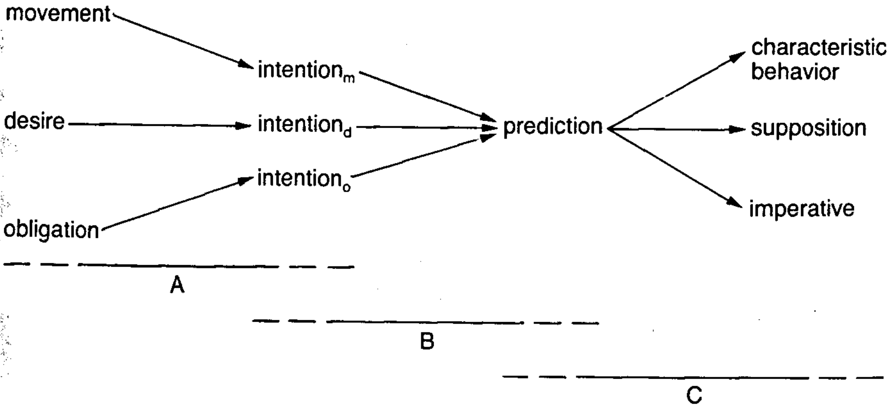
(41) Hausa
(Some men can really tell a tale)
ai wani sa'ì zā sù fi mātā à wurimmù
well some time fut 3pl surpass women for us
Even sometimes they'll surpass the women.

Note that the sentence has a non-specific subject. A true habitual could be used with a specific agent.

The relationship between future and characteristic behavior once again has to do with prediction or predictability, and we argued in Bybee and Pagliuca (1987) that this use is an extension of the prediction use. This hypothesis is supported by Dahl's data, which show an exceptionless implicational relation between characteristic behavior and prediction uses. All the languages with characteristic behavior uses in his sample also have prediction uses (16 out of 47), but languages with prediction uses to not necessarily have characteristic behavior uses, in fact 23 out of 47 do not. Moreover, futures from any source may have the characteristic behavior reading: the English and Persian desire-derived future has such a use, as does the Spanish obligation-derived future (Moreno de Alba 1970), and the Hausa and Hindi movement-derived future.

4.7 *Explanation*

In this brief survey I have not discussed all the uses ever attributed to future grams, but I have covered the ones that are most frequently mentioned, and the ones that are verifiable across languages. As we argued in Bybee and Pagliuca 1987, it is possible to show, on the basis of data from a wide range of languages, that certain uses or semantic nuances are retentions from the original lexical meaning of a future, while others develop out of the prediction sense and may appear in futures of any lexical source. Consider the following diagram that plots the most commonly occurring paths for the development of the semantics of future:



A, B and C are arbitrarily chosen synchronic states. The range of use of a future gram at any stage, such as A, B or C, is partially determined by the history of that gram. The history of a gram follows certain general characterizable processes of change, that is loss of specific lexical meaning in most contexts with retention in some contexts, loss of co-occurrence restrictions, increase in frequency and a shift to propositional scope.

This theory is explanatory in a number of respects. It explains why it is often difficult to find a single abstract meaning to characterize all of the uses of a grammatical morpheme. It attributes the cross-linguistic similarity of grammatical meaning to similar paths of development, and general principles of historical development; it explains cross-linguistic differences in the meaning of grams with reference to differences in lexical source or differences in the extent of development undergone by particular grams. It predicts possible combinations of uses for language-specific grams. It allows the reconstruction of the lexical source of grams on the basis of their meaning. Finally, it points to a certain set of dynamic processes in synchrony that have to be studied further to explain the nature of grammatical meaning.

The implications of the futures case for explanations of language universals is that formulating broad principles to cover cross-linguistic generalizations, such as 'the future is uncertain, thus future tense and modality overlap', does not necessarily lead our investigation in the appropriate direction. However, identifying causal factors helps us to uncover the relevant set of diachronic and synchronic phenomena that will lead to richer and richer explanations. The relation of synchronic generalizations to diachronic processes is the same for the case of the semantics of future morphemes as it is for the more structural phenomena discussed in the first two sections. In each case, if we can identify the factors involved in the establishment of the grammatical conventions, then we can approach a valid explanation. Thus to understand how the ordering of adpositions relates to the ordering of noun and genitive, we must look into how such orders are established when new adpositions arise and when new possessive constructions arise. To find out why suffixes are more common than prefixes, we must refer to the fact that the position of a new affix is determined by the position of that same element before it becomes an affix, and ask what determines its position originally. If we wish to explain why some pairs of elements (such as verb and causative) are more fused than others, we must understand the fusion process itself and what governs it. To understand why frequent words are shorter than infrequent ones, we must investigate how frequent words reduce. To understand case marking patterns, we must investigate how case markers arise, as well as how nouns in different grammatical roles are used in discourse.

I have argued, then, that in order to reach valid explanations for language universals it is necessary to attend to the causal mechanisms operational in the establishment of grammatical conventions, and find the general dynamic principles behind the causal mechanisms. This view requires that explanation – that is, linguistic theory – have both a diachronic dimension and a synchronic dimension. The diachronic dimension plots paths of development of grammatical phenomena across time, and the synchronic dimension fills in the small steps along these paths by referring to the way in which language users manipulate the linguistic conventions they have inherited for their conceptual and communicative purposes.

Notes

I am grateful for comments and suggestions on earlier versions of this paper to the following, none of whom necessarily agrees with all my conclusions: John Corcoran, Julie Gerhardt, Jack Hawkins, Jürgen Klausenburger, William Pagliuca and Revere Perkins. Some of the data used here is from the GRAMCATS sample of eighty languages, constructed by Revere Perkins, with data analyzed by Perkins, William Pagliuca and Roula Svorou. Work on the GRAMCATS project is supported by NSF grant #BNS 8318262. This support is gratefully acknowledged, as is the support of the Netherlands Institute of Advanced Study during the 1983–4 academic year.

[1] Since Hawkins (1979, 1983) reformulates the universals to be exceptionless, his statement of this correlation is somewhat different.

[2] Adpositional phrases also develop from VO constructions, especially in languages with verb serialization (Givón 1975). In a VO language these constructions will become prepositional, while in an OV language they will be postpositional. Adpositions also develop from adverbs which come to take an object; the factors governing the position of this object have not been investigated, but the placement of the object might be determined on the basis of the VO order, in which case the analogical principle might be involved.

[3] There is a use of the notion 'possible state' in historical linguistics that is legitimate, but this is in reconstruction. It is considered justifiable only to reconstruct language states that correspond to documented types.

[4] Fleischman does not say exactly what she means by fusion in this case. Apparently she has in mind the fact that the postposed auxiliary is written bound and the preposed one is not.

[5] This is a case of diagrammatic iconicity, in which the relation between two elements on the level of meaning is paralleled by the relation of the elements representing them on the level of expression.

[6] The mechanism that Zipf proposes for the shortening of frequent words is 'clipping', the process that gives us *fridge*, *lab*, *auto*, etc. Historical documentation shows that clipping is an extremely restricted process which applies only to lexical morphemes (and probably only to nouns), and that the real mechanism behind the shortening of frequent words is phonetic reduction.

[7] William Pagliuca (personal communication) points out that this 'typical' information flow is also a matter of cultural convention, pushing back our attempts at explanation one step further.

[8] The tendency of the absolutive case to be unmarked, while the ergative is marked, follows from the fact that a marker will arise only in cases where the role of the noun phrase is not the usual one. Thus the following statement by Du Bois is unnecessary: 'a redundancy avoidance principle is best served when the one zero morpheme available in the paradigm is "assigned" to agree with the syntactic category which will in any case be represented by a full noun phrase: the absolutive' (pp. 352–3).

[9] Of course many grammatical 'rules' are not categorical, but rather have an extremely high probability of application. Still there is a mechanism that continues to increase the frequency of a structure until it becomes an almost categorical choice.

[10] In Dahl's questionnaire the examples were given in English, but the verbs occurred without English inflections in order not to bias the choice of tense and aspect in the translation. I am grateful to Östen Dahl for making available to me his data on the use of future morphemes in each of the forty-seven languages that had them.

[11] Omitted from this list are *types* of futures, such as immediate future (or imminence) vs. remote future.

[12] Meillet 1948, Givón 1979, Traugott 1982, Lehmann 1982, Heine and Reh 1984.

[13] A few languages which have desire-derived futures are English, Central Sierra Miwok, Serbo-Croatian, Mandarin, Chukchi, Modern Greek and Swahili.

[14] Movement-derived futures seem to be the most common. A few languages in which they are attested are Southern Sierra Miwok, Haitian Creole, Isthmus Zapotec, Logbara, English, Hausa. See Ultan (1978), Heine and Reh (1984) and Bybee and Pagliuca (1987) for more examples.

15 Obligation-derived futures seem to be the least common cross-linguistically. The only examples we have found of a verb meaning 'to owe' that becomes a future are the Germanic cognates of *shall*. Futures derived with copulas or possession verbs which originally have obligation senses may be found in the Eastern Kru languages, the Western Romance languages and Korean.

16 In Modern British English, *shall* occurs only with first person subjects in spoken discourse.

17 In Bybee and Pagliuca (1987) we also discuss the meaning of *be going to* in terms of its historical source. Remnants of the original lexical semantics of *will* are evident in the way children use *will*, especially as opposed to *be going to*, in negotiatory contexts (Gee and Savasir 1985).

18 Comrie (1985) is too quick to reject the diachronic when he says: 'Finally, one might observe that expressions of future time reference frequently derive diachronically from modal expressions, e.g., of desiderativity, such as English *will*. However, this diachronic relation says nothing of the synchronic status of such forms' (pp. 45–6).

19 A generalized function leads to higher frequency. Frequency plays an important role in grammaticization as it appears to be linked both to the rapid phonological reduction of grams and to their semantic reduction.

References

Berlin, B. and Kay, P. (1969) *Basic Color Terms: Their Universality and Evolution*. Berkeley, Calif.: University of California Press.

Bybee, J. L. (1985a) 'Diagrammatic iconicity in stem-inflection relations'. In J. Haiman (ed.), *Iconicity in Syntax*. Amsterdam: John Benjamins.

(1985b) *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins.

and Pagliuca, W. (1985) 'Cross-linguistic comparison and the development of grammatical meaning'. In J. Fisiak (ed.), *Historical Semantics and Historical Word Formation*. The Hague: Mouton.

(1987) 'The evolution of future meaning'. In *Papers from the VIIth International Conference on Historical Linguistics*, ed. by A. G. Ramat, O. Carruba and G. Bernini. Amsterdam: John Benjamins.

Chung, S. and Timberlake, A. (1985) 'Tense, aspect and mood'. In T. Shopen (ed.), *Language Typology and Syntactic Description*, vol. 3. Cambridge: Cambridge University Press.

Coates, J. (1983) *The Semantics of Modal Auxiliaries*. London: Croom Helm.

Comrie, B. (1980) 'Morphology and word order reconstruction: problems and prospects'. In J. Fisiak (ed.), *Historical Morphology*. The Hague: Mouton.

(1985) *Tense*. Cambridge: Cambridge University Press.

Cutler, A., J. A. Hawkins and G. Gilligan (1985) 'The suffixing preference: a processing explanation'. *Linguistics*, 23, 723–58.

Dahl, Ö. (1985) *Tense and Aspect Systems*. Oxford: Basil Blackwell.

Dryer, M. S. (1988) 'Universals of negative position'. In M. Hammond, E. Moravcsik and J. Wirth (eds), *Studies in Syntactic Typology*. Amsterdam: John Benjamins.

Du Bois, J. W. (1985) 'Competing motivations'. In J. Haiman (ed.), *Iconicity in Syntax*. Amsterdam: John Benjamins.

Fidelholtz, J. L. (1975) 'Word frequency and vowel reduction in English'. *Proceedings of the Chicago Linguistic Society*, vol. 11.

Fleischman, S. (1982) *The Future in Thought and Language*. Cambridge: Cambridge University Press.

Fries, C. C. (1927) 'The expression of the future'. *Language*, 3, 87–95.

Gee, J. and I. Savasir (1985) 'On the use of WILL and GONNA: towards a description of activity types for child language'. *Discourse Processes*, 8, 143–75.

Givón, T. (1975) 'Serial verbs and syntactic change: Niger-Congo'. In C. Li (ed.), *Word Order and Word Order Change*. Austin, Tex.: University of Texas Press.

(1979) *On Understanding Grammar*. New York: Academic Press.

Greenberg, J. H. (1957) 'Order of affixing: a study in general linguistics'. In J. H. Greenberg (ed.), *Essays in Linguistics*. Chicago: University of Chicago Press.

(1963) 'Some universals of grammar with particular reference to the order of meaningful elements'. In J. H. Greenberg (ed.), *Universals of Language*. Cambridge, Mass.: MIT Press.

Haiman, J. (1983) 'Iconic and economic motivation'. *Language*, 59, 781–819.

Hall, C. J. (this volume) 'Integrating diachronic and processing principles in explaining the suffixing preference'.

Hawkins, J. A. (1979) 'Implicational universals as predictors of word order change'. *Language*, 55, 618–48.

(1983) *Word Order Universals*. New York: Academic Press.

Heine, B. and M. Reh (1984) *Grammaticalization and Reanalysis in African Languages*. Hamburg: Helmut Buske.

Hooper, J. B. (1976) 'Word frequency in lexical diffusion and the source of morphophonological change'. In W. Christie (ed.), *Current Progress in Historical Linguistics*. Amsterdam: North-Holland.

Hopper, P. J. (ed.) (1982) *Tense-Aspect: Between Semantics and Pragmatics*. Amsterdam: John Benjamins.

Itkonen, E. (1983) *Causality in Linguistic Theory*. London: Croom Helm.

Kraft, C. H. and M. G. Kraft (1973) *Introductory Hausa*. Berkeley, Calif.: University of California Press.

Lass, R. (1980) *On Explaining Language Change*. Cambridge: Cambridge University Press.

Lehmann, C. (1982) *Thoughts on Grammaticalization*. Cologne: Arbeiten des Kölner Universalien Projekts, 48.

Mańczak, W. (1978) 'Irregular sound change due to frequency in German'. In J. Fisiak (ed.), *Recent Developments in Historical Phonology*. The Hague: Mouton.

Meillet, A. (1948) 'L'évolution des formes grammaticales'. In A. Meillet (ed.), *Linguistique Historique et Linguistique Générale*. Paris: Champion.

Moreno de Alba, J. G. (1970) 'Vitalidad del futuro del indicativo en la norma culta del español hablado en México'. *Anuario de Letras*, 8, 81–102.

Pagliuca, W. (1976) 'PRE-fixing'. MS, SUNY at Buffalo.

Poppe, N. N. (1960). *Buriat Grammar*. Indiana University Publications in Uralic and Altaic Series, vol. 21. Bloomington, Ind.: Indiana University, and The Hague: Mouton.

Rosch, E. (1978) 'Principles of categorization'. In E. Rosch and B. B. Lloyd (eds), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.

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

Scheffer, J. (1975) *The Progressive in English*. Amsterdam: North Holland.

Schuchardt, H. (1885) [1972]. 'On sound laws: against the Neogrammarians'. In T. Vennemann and T. H. Wilbur (eds), *Schuchardt, The Neogrammarians and the Transformational Theory of Phonological Change*. Frankfurt-am-Main: Athenäum Verlag.

Subrahmanyan, P. S. (1971) *Dravidian Verb Morphology*. Tamilnadu: Annamalai University.

Traugott, E. C. (1982) 'From propositional to textual and expressive meanings: some semantic-pragmatic aspects of grammaticalization'. In W. Lehmann and Y. Malkiel (eds), *Perspectives on Historical Linguistics*. Amsterdam: John Benjamins.

Ultan, R. (1978) 'The nature of future tenses'. In J. H. Greenberg, C. A. Ferguson and E. A. Moravcsik (eds), *Universals of Human Language*, vol. 3. Stanford, Calif.: Stanford University Press.

Vennemann, T. (1972) 'Analogy in generative grammar: the origin of word order'. Paper presented at the 11th International Congress of Linguists, Bologna.

Vennemann, T. (1973) 'Explanation in syntax'. In J. Kimball (ed.), *Syntax and Semantics 2*. New York: Academic Press.

Wekker, H. C. (1976) *The Expression of Future Time in Contemporary British English*. Amsterdam: North Holland.

Zipf, G. K. (1935) *The Psycho-Biology of Language*. Boston, Mass.: Houghton Mifflin.