SYNTACTIC INNOVATION: A CONNECTIONIST MODEL

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF LINGUISTICS

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Whitney Tabor

September 1994

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Elizabeth Closs Traugott
(Principal Advisor)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Paul Kiparsky

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

David E. Rumelhart
(Department of Psychology)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Thomas Wasow

Approved for the University Committee on Graduate Studies:

Abstract

This thesis uses the continuous representation space of a Connectionist network to make predictions about syntactic innovation in natural language. It proposes to replace current, categorically organized accounts of linguistic structure with a metrically organized model in which innovation is treated as a quantitative interpolation process.

Current theories of grammar, following Chomsky 1957, are based mainly on data gathered by introspection. Their generalizations tend to be sweeping. The inventories of representational and derivational devices they posit are large. Such theories are not very useful for making constrained predictions about historical grammar change for there is no domain-independent structuring of the representation space that reveals which changes are probable and which are not. Nevertheless, recent work in the field of *grammaticalization* (e.g., Traugott and Heine 1991, Hopper and Traugott 1993) indicates strong constraints on grammar change. These constraints are especially apparent in sequences of closely-spaced historical texts, for the distributional properties of such texts change gradually. The gradualness is evident both at a categorical level, in that successively emerging types tend to be similar to one another, and at a quantitative level, in that constructions appear and disappear via long periods of probabilistic alternation, with the probabilities varying gradually.

I propose to unite probabilistic and categorical gradualness in a single mechanism. I replace categories with clusters and let quantitative differences in usage rates manifest themselves as contrasting distances from a prototype. The theory is implemented in a recurrent Connectionist network trained on the next-word prediction task studied by Elman 1990 and 1991. Three empirical phenomena

provide evidence in its favor: linked frequency changes in grammatically related constructions, quantitative changes which anticipate events of categorical re-analysis, and emergence of hybrid structures during periods of transition. Case studies of degree modifier *sort/kind of* and of future *be going to* are presented. The model permits simplification of the theory of language change by replacing *reanalysis* and *analogy* with a single mechanism. It also permits simplification of the theory of language structure by allowing the same interpolative mechanism to handle both normal productive syntax and the hitherto problematic hybrid cases.

Acknowledgements

Elizabeth Traugott
Paul Kiparsky
Tom Wasow
Dave Rumelhart
Joan Bresnan
Josep Fontana
Dan Rosen
Hinrich Schütze
Andrew Garrett
Peter Sells
Ivan Sag
Rowland Tabor
Kajsa Tabor
Lesley Tabor
Thomas Heckman
Michelle Murray
Emma Pease
Gina Wein
Jon Mara
Victoria Hand
David Hooper
Soledad Maria Baylosis

Contents

List of Figures

Chapter 1

Introduction

1.0   Overview

This thesis is about language structure. It investigates language structure by examining language change at a very detailed level in relatively continuous sequences of historical texts. It is driven by the hypothesis that if we seek a structural model that makes diachronic prediction easier, then we will also learn something useful about synchronic representation.[1] Its central claim is that evolutive language change is gradual at the level of grammatical representation. Although this claim is not plausible under standard, category-based models of linguistic structure, it is tenable if the quantitative properties of representative corpuses are taken into account. The historical texts that form the basis of the study are English texts from the Middle English period onward.[2] Case studies

of Degree Modifier *sort of* and *kind of* and of Auxiliary *be going to* are presented. The model is formalized as a recurrent Connectionist network trained on word-prediction as in Elman 1990 and 1991. The primary advantage of the network's underlyingly continuous representation over traditional discrete models is that it makes more constrained predictions about structural innovation or *reanalysis*.

Chapter 1 motivates the analysis and synopsizes the remainder of the thesis. Chapter 2 reviews evidence from the linguistic historical literature in favor of a representation that is sensitive to the quantitative properties of corpuses. Chapter 3 presents the Connectionist model and analyzes its predictions formally. Chapters 4, 5, and 6 investigate, respectively, three distinguishing empirical predictions of the model, which I refer to as *Frequency Linkage*, *Q-Divergence*, and *Hybrid Structures*. Chapter 7 considers implications for synchronic and diachronic linguistics.

From here, Chapter 1 proceeds by motivating the continuous representation on the grounds that it makes prediction easier (§1.1–1.3). Some remarks are then made on the historical context in which non-continuous models arose in linguistics, on the problems that attend them, and on solutions other people have proposed (§1.4). Evidence for the continuous model is previewed in §1.5 and then the model itself is introduced (§1.6–1.7). I call it the *Restrictive Continuity* model to emphasize the fact that it permits intermediate behaviors while continuing to make the appropriately restrictive predictions that linguistic categorical models make. The last part of the chapter synopsizes the three predictions made by the Restrictive Continuity model which distinguish it from other accounts, in particular from the "Competing Grammars" models of the Variationist School (Kroch 1989a, 1989b) (§1.8).

1.1   Continuity and predictability

For prediction, it is sometimes helpful to choose a representation that varies continuously with time. Even if the entity you are studying evolves in a serpentine and inscrutable manner, so that you have little hope of modeling every detail of

---

[1] By *synchronic representation* I mean the representation (or *grammar*) of a language at a particular point in time. By *diachronic prediction*, I mean prediction of a closely-spaced, ordered series of synchronic language behaviors. I model such a series of behaviors by positing a corresponding series of grammars and sampling the outputs of these grammars.

[2] I take the texts from each period to reflect the mental representations of individual speakers from that period. In most cases, the sample representing a particular point in time is drawn from the writing of one person. In some cases, the sample is a mixture of documents written by different people. Even in the latter cases, for the phenomena I consider, it seems plausible to assume that the corpus distributions are equivalent to individual distributions. The only matter that hinges on this point is whether the linguistic representation system derived by the current study is a model of the same phenomenon as the large number of linguistic analyses that are based on the outputs of individual speakers. This assumption receives some independent support from the fact that the proposed representation makes better predictions than

other current theories on certain *hybrid structure* data generated by indidual speakers—see Chapter 6 and Chapter 7, Section 2.

its evolution, you may nevertheless be able to make reasonable short-term predictions if you can find a system in which chronologically sequential behaviors are placed near each other. In that case, the structure of your representation will correspond to a set of generalizations about what the sequential behavior of your entity is likely to be.

Consider, for example, a downhill ski area. Imagine the problem of watching an arbitrary skier from a distance and trying to predict her future location on the basis of her present location at each moment. Downhill ski areas typically consist of a set of skilifts organized radially around a lodge (Figure 1.1). Skiers can ride a lift to its top and then ski down to the bottom again. Because their skis are heavy and their boots are rigid, the skiers cannot traverse or climb uphill easily, so they generally ski down in a eye-shaped swath of terrain issuing from the top of whatever lift they most recently rode. Thus we could form a basic representation of a ski area in terms of lift names. If one knows the name of the lift a skier is currently riding or most-recently rode, then one can be reasonably confident that the skier will be somewhere in the neighborhood of that lift in the near future. Indeed, this lift-name representation is commonly used by ski-patrol members in planning their daily activities and describing locations of injured skiers.

But the lift-name representation has certain short-comings from the standpoint of the prediction task. First of all, we might want to be able to pinpoint a skier's location more specifically. Moreover, the representation is essentially useless in making predictions about change in representational status. For although some skiers tend to stick to one lift for long periods of time, others switch impulsively from one to another. In the general case, short of having elaborate information about the psychologies of particular skiers, it is hard to predict what lift somebody is going to ride on the basis of the lift they just rode. Since all the lifts are roughly equally accessible from any ski-slope, there are no physical constraints on the order in which a skier is compelled to ride the lifts.

One might say that, without knowing people's intentions, the prediction task is impossible. But this is not quite true. The lift loading areas are in slightly different locations from one another at the bottom of the bowl. If one watches a skier carefully as she is making her final descent into the lodge-area, one can often tell which lift she is heading for. This suggests an alternative

Figure 1.1: Schematic of a Ski Area
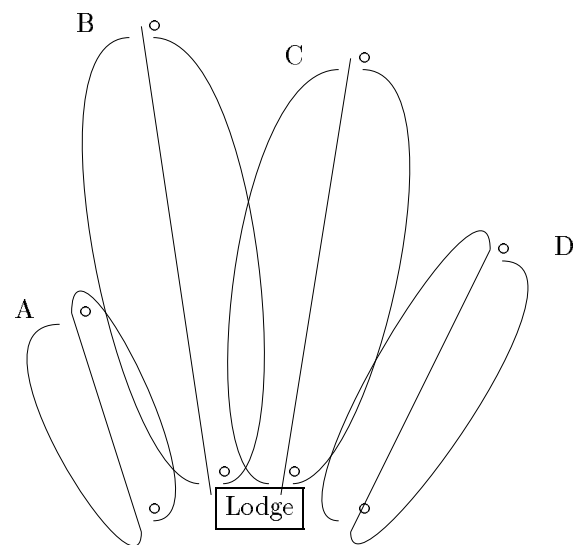
representation in which skiers, ski-slopes, and loading-areas are located in a cartesian coordinate system and the relationships of skiers to loading-areas are taken into account in making predictions. Indeed, this is a case in which the switch to a continuous representation system makes the evolution of an entity (the position of a skier) easier to predict. Note that this switch also permits a refinement of the short-term predictions within lift-zones, for even though we may not be able to predict every turn a skier will make, we can be confident that she will never traverse ground at beyond a certain speed and so we can constrain our predictions about the short-term future of her location on the basis of her current location at any point.

Yet the cartesian-coordinate representation has shortcomings as well. First of all, it takes more memory to encode, for even if we only approximate continuity, we must have fine gradations along the coordinate axes in order to distinguish effectively among the four lifts at their bases. But also, it does a worse job than the lift-name notation of constraining future locations on the basis of present ones in the sense that it makes no distinction between the regions that people actually ski on and the regions between the eye-shaped swaths

where they don't. In this light, it would seem best to have a representation that combined the advantageous aspects of the the two types.

Current linguistic theory, with its word and phrase categories and parametric-switches, is much more like the lift-name representation than it is like the cartesian-coordinate representation. For making short term (decade- or half-century-hence) predictions about the evolutionary path of a language, it would be desirable if we could find a representation analogous to that of the cartesian-coordinate system for skiers. But we don't want to abandon the symbolic representation entirely, for it gives us an efficient way of describing great ranges of synchronic data and distinguishing possible behaviors from impossible ones. Here also, then, we might wish for a representation that combines the virtues of both systems. The first aim of this dissertation is to show that by switching to an analog of cartesian-coordinates, we can put much stronger constraints on local properties of paths of syntactic change than has previously been thought possible. The analogs of the coordinates are facts about the frequencies with which words and phrases occur in various contexts in large corpuses of natural language. I will provide evidence that languages really do something like "heading for another lift" when they are about to develop a new grammatical characteristic. Choosing a representation system in which these harbinger behaviors can be encoded not only makes it possible to have a more constrained model of change, but also provides new insight into the synchronic structure of language. A second aim is to provide evidence that certain Connectionist representations combine the virtues of the discrete and continuous systems inasmuch as they provide both parsimonious description and good short-term prediction ability even in cases where radical changes of category are imminent.

1.2 The relationship between diachronic data and synchronic representation

Sometimes linguists concerned with the synchronic structure of language note that a child learns its language natively without knowledge of its history, and take this fact as evidence that we need not take diachronic data into account in building a theory of synchronic structure. This may be true in some ideal sense.

If we can correctly guess the right theory of the language acquisition device then we will probably be able to tell that it is better than various wrong theories on the basis of synchronic data alone. But this is a big "if". Moreover, it is often the case with other natural systems that we can learn much about their structure by looking at how they change over time, especially if we pay attention to cases in which there are *perturbational forces* which distort their structure by impinging differentially on them. When elements change in a correlated manner we can hypothesize that they are linked to one another structurally. Of course, there is always the possibility that correlations are coincidental, but if we observe the same types of correlations across a variety of languages, or we observe persistent correlation among two elements over a long period of time, then we may be more confident in the significance of their correlation (see Lightfoot 1979, pp. 15–16).

1.3 The feasibility of predicting language change

Predicting structural change in language is certainly not an easy thing to do. Weinreich, Labov, and Herzog 1968 suggest that it will be quite difficult to make predictions about the onsets (or "actuations") of grammar changes if we do not have a well-developed model of the social forces that impinge on languages. This makes structure-based prediction seem pretty hopeless. Nevertheless, short-term prediction is generally easier than long-term prediction. Thus if we adopt a representation that refers to quantitative properties of language use, we can examine language at a much more fine-grained level and the feasibility of making short-term predictions may be increased.

Moreover, there may be a close relationship between short-term diachronic prediction and the type of synchronic prediction that we are already quite good at. Synchronic prediction is the kind of prediction that linguists (and children) engage in when they examine a subset of the forms of a language and then make predictions about what other forms (with which corresponding meanings) will be grammatical in the language. Whether our effectiveness in synchronic prediction bodes well for the prospect of doing diachronic prediction depends on a subtlety about the nature of language structure. If language structure is rigidly modular in the sense that a force impinging upon one part of the language affects all structurally related parts equally, then our ability to make synchronic

predictions does not bode well for our ability to make diachronic predictions. For if a new construction emerges, our structural knowledge lets us know that all structurally related constructions must be emerging simultaneously; but this isn't a diachronic prediction, since nothing we can't observe in the present has been predicted. On the other hand, if language structure is only semi-rigid in the sense that a force producing ongoing change in one part of the language causes a buildup of potential energy in related parts, then prediction may be easier. For we can project the ongoing change into the future and identify a point at which the buildup of potential energy in the related parts is sufficient to bring about a change in them. It is worth noting that diachronic trends of both a quantitative and qualitative nature have been observed,[3] and it is imaginable that such energy build-up is associated with either type.

Most current models of grammar assume that structure is rigid in the sense just described. Before examining counter-evidence, it will be useful to consider how the rigid model arose in the structuralist tradition fathered by Saussure, and to consider ways in which people have tried to adapt it to handle change and variation data.

1.4   Saussure and Co.

This is what Saussure said:

> The first thing that strikes us when we study the facts of language
> is that their succession in time does not exist insofar as the speaker
> is concerned. He is confronted with a state. That is why the linguist
> who wishes to understand a state must discard all knowledge of
> everything that produced it and ignore diachrony. He can enter
> the mind of speakers only by completely suppressing the past. The
> intervention of history can only falsify his judgment. It would be

---

[3]By a qualitative trend, I mean a succession of related structural changes (e.g., the spread/loss of a morphological marker across words (e.g., Schupbach 1984) or the spread/loss of a case-marking system across clause-types (e.g., Harris 1985)). It may be reasonable to think of a qualitative trend as a quantitative trend under an abstraction, although sometimes there does not seem to be a single grammatical type which can serve as the domain in which the qualitative trend progresses (e.g., when Nouns evolve first into Adpositions and then into Inflectional Case-Markers—(e.g., Hopper and Traugott 1993, pp. 106–8.)

absurd to attempt to sketch a panorama of the Alps by viewing them simultaneously from several peaks of the Jura; a panorama must be made from a single vantage point. The same applies to language; the linguist can neither describe it nor draw up standards of usage except by concentrating on one state. When he follows the evolution of the language, he resembles the moving observer who goes from one peak of the Jura to to another in order to record the shifts in perspective. [Course in General Linguistics, Wade Baskin's translation, pp. 81–82]

Saussure was surely right in one respect. By proclaiming thus he galvanized interest in seeking system in the synchronic states of languages, whose coherence prior linguists had often missed. Though Saussure himself seems to have considered syntax a matter of "will" and hence of *parole* and hence not of system, the effort he spawned to seek system in synchrony was a crucial prerequisite to our perceiving the large-scale regularities that we now call "syntax".

But it is often said that Saussure's insight created a paradox for historical linguistics. Thus Weinreich, Labov, and Herzog 1968:

> The bulk of theoretical writing in historical linguistics of the past
> few decades has been an effort to span the Saussurean dilemma, to
> elaborate a discipline which would be structural and historical at
> the same time. [p. 98]

What is the "Saussurean Dilemma"? It seems to have to do with the fact that if we observe historical change either in texts or in the nonwritten output of speech communities we have a sense that change happens gradually. This impression is strengthened if we focus attention on what Andersen 1972 calls "evolutive innovations" and distinguish these from "contact innovations" (see also Andersen 1973, 1975, 1989). Yet grammars, with their discrete categories and rules covering large numbers of cases, do not seem very amenable to gradual mutuation.

Andersen 1973 proposes to allow abrupt representation change, but to ensure gradual observed change by positing "adaptive rules" which initially mask the differences between grammars that are radically different. Over the course of

time, the adaptive rules are incrementally removed, thus creating the appearance of gradual structural change. But the adaptive rule approach seems dubious in the case of morpho-syntactic change because the particular rules required seem to violate well-motivated constraints on syntactic transformations. For example if an element is gradually changing from being generated syntactically to being generated lexically, as happens frequently, the adaptive rules must divide underlyingly unitary lexical items into separate syntactic constituents, often thoroughly revising phrase structure in the process (see discussion in Chapter 7, Section 2.1.2).

Weinreich, Labov, and Herzog 1968 claim that the Saussurean Dilemma can be resolved if we permit competence (*a la* Chomsky 1965) to accomodate grammar-mixture:

> The solution...lies in the direction of breaking down the identi-fication of structuredness with homogeneity. Nativelike command of heterogenous structures is not a matter of multidialectalism or "mere" performance, but is part of unilingual linguistic competence. [p. 102]

They suggest a *variable rule* formalism in which linguistic rules can be as-sociated with probabilities that determine their rates of application in various contexts. Recently, their suggestion has been followed up with explicit and thor-ough modelling efforts using abstract variable parameters rather than variable rules: Kroch 1989a, 1989b, Santorini 1989, 1992, 1994, Pintzuk 1991, Taylor 1992, Fontana 1993. If one claims that two different grammars can both help to generate the speech of an individual speaker and one posits that a probabilis-tic choice is made between them with every utterance, then change can be as gradual as one likes, for an arbitrarily small bound can be placed on the rate of change of the value of the probabilistic variable. Moreover, it is automatically predicted that behaviors uniquely associated with one grammar should change in unison. This prediction is largely, though not completely borne out as the work of Kroch and his followers has shown (see Chapter 4). Moreover, this "compet-ing grammars" account seems to have a distinct advantage over the Adaptive Rules approach in proposing to get by with the typologically-motivated mech-anisms of Universal Grammar alone—no special adaptive machinery is needed.

And yet there is also reason to be suspicious of this approach as an end-all solution to the Saussurean Dilemma. It doesn't put any constraints on the nature of *reanalysis*,[4] i.e., on structural change, except by virtue of the restrictiveness of Universal Grammar itself. This, unfortunately, makes it a not-very-powerful theory of change, given the immense variety of language behaviors that Universal Grammar needs to encompass (cf. Plank 1984, p. 1). Moreover, there is reason to believe that reanalysis can be non-trivially constrained on the basis of grammatical information alone.

Many researchers in Grammaticalization, the study of the diachronic evolu-tion of grammatical structure,[5] would say that to have a properly constrained theory of change, you need to take into account *semantic* information. They mean that you need to have a theory of what kinds of situations are similar *in the conceptual world apart from language* and that if you connect such a theory with a model of the mapping between language and the conceptual-world-apart-from-language, then you can state much stronger constraints on which changes are likely to happen. This position is often opposed to the program which seeks a theory of the formal properties of language in isolation from their meanings (i.e., an "autonoumous" theory of structure). I argue here for a compromise between the two viewpoints: constraints stated in terms of the properties of language alone can be much strengthened if we let the continuity of change at the level of *corpuses* influence the design of our theory of formal structure.[6] Moreover, the resulting theory consists of elements whose structure is much more like that of the semantic/conceptual entities that grammaticalization the-orists say we should be paying attention to. For people who are interested in

---

[4]Here, by *reanalysis*, I mean "grammar change" where "grammar" is construed as having to do strictly with categorical properties of language use. I take it that the grammar includes both the lexicon and the syntax so reanalysis can affect both these areas. Under this interpretation, reanalysis is opposed to "quantitative change", which only affects *language use*. But I must note that one of my primary aims in this thesis is to make it plausible that "grammar" as a theory of mental representation ought to be redefined so that quantitative as well as qualitative properties of language use are pertinent. Consequently, in the end, I will propose that *reanalysis* is this standard sense is not a useful theoretical mechanism. It is only good as an approximate abstraction (See Chapter 7, Section 2).

[5]See Cahpter 2, Section 2.

[6]I take the corpuses of relevance to be the bodies of speech that language-users are ex-posed to during the courses of their lives. For historical periods we cannot observe these bodies directly but surviving written records often provide a reasonable approximation of their important characteristics.

semantic properties for their own sakes, this strategy may yield useful tools. The problem with the current pure conceptual-similarity approaches is that it is hard to be explicit about the nature of conceptual structure. Corpuses provide a rich, relatively untapped source of new data on the problem. It is data about which it is relatively easy to make disprovable claims, and testing those claims is becoming easier by the moment, as more and more corpuses are being computerized and tagged with linguistic annotations.

1.5 Evidence for Continuity in Corpus Structure

I'll outline now several kinds of evidence in favor of the claim that corpuses contain much of the information we need for providing a constrained theory of structural change. Chapter 2 reviews this data in more detail.

(a) In two of the case-studies in which Kroch's competing grammars model has been tested (Kroch 1989, Fontana 1993), varying the value of underlying real number probability gives an accurate fit to the data for a certain window of time. But it seems to be necessary to posit a reanalysis at the forward edge of that window in order to model subsequent changes. Moreover the reanalysis is not an arbitrary one. Instead it seems to be a natural consequence of the previous distributional shift in the sense that the shift has reduced the available evidence for assigning the relevant element to its original category and increased the evidence for assigning it to a new category. This suggests that reanalysis is not arbitrary, but in fact is related to *quantitative* properties of language behavior (see Chapter 2, Section 1).

(b) Work on grammaticalization processes during the past 20 years (e.g., Givon 1971, Hopper and Traugott 1993) indicates that certain transitions defined in terms of a mixture of syntactic and semantic/pragmatic properties are very common, especially in languages of certain families. To give a few examples: adpositions have developed from serial verbs with meanings like "be at", "give", "accompany" in languages of the Niger-Congo group (Lord 1973, Heine and Reh 1984, Carlson 1991); subjects have developed from topics in Philippine languages (Shibatani 1991; cf. Givon 1979 (p. 209–11), Kroeger 1993); future

markers have developed from verbs meaning "come" or "go" in languages all over the world (e.g., Perez 1990, Bybee, Pagliuca, and Perkins 1991). One may interpret these data as evidence that reanalysis ought to be jointly constrained by a combination of syntactic and semantic properties. If one believes that semantic constraints are part of grammar, then this is evidence that the theory of grammar can put non-trivial constraints on reanalysis. I argue below, in fact, that not only should these semantic constraints be considered part of grammar, they should be considered within the purview of an augmented theory of syntax whose concern is governing the distributions of words in corpuses in quantitative (as well as qualitative) terms. On this view, reanalysis can be constrained on the basis of syntactic theory (see Chapter 2, Section 2).

1.6 The Restrictive Continuity model

What sort of grammar would permit us to have a more constrained theory of reanalysis? Instead of thinking of grammar as a model of the *inventory* of types of forms that a native speaker finds correct, I will think of it as a model of the quantitative *distribution* of those types in a corpus of natural usage. Thus, it will make a difference in the representation when a speaker uses one particular form where several are possible. In contrast with Labov, Kroch, and the Variationists generally, where a change in frequency implies a change in the annotation on a variable rule or parameter, I will suppose that the choice to use a form actually changes the set of categories that form the grammar of the language. What sort of change in the categories do I have in mind? We do not want to lose the ability to describe the many categorical systematicities which permit current grammatical descriptions to compass languages so efficiently. Suppose, then, that we associate words in context with regions of a *metric-space*, that is, a space in which a measure of distance is defined. Let us assume that this distance-measure takes on a continuum of real values. We can place same-category instances of words in clusters, thereby recapitulating the structure of a categorizing grammar, but allow word-instances within clusters to be slightly removed from one another depending on the relative frequency with which the types to which they belong are used in the grammatical environments associated with the region (Figure 1.2). This proposal is a variety of

prototype theory, in which categories are structured around a central member (or "prototype") and elements are characterized in terms of how much they resemble the prototype (e.g., Rosch *et al.* 1976). If we assume that evolutive change is gradual at the level of this representation, we put a strong constraint on the set of new grammars that a language can embody within any small increment of time: it can only move from one state to a nearby state; such nearby states will be related to each other as small frequentistic distortions in corpus distribution. Moreover, incremental changes in the rate of use of a form can cause its representation to stray away from a region associated with one set of grammatical behaviors and toward a region associated with another. As a consequence, certain mounting frequency distortions, if they progress far enough, will lead to categorical innovations. These are the cases in which we will, thinking in terms of canonical grammatical descriptions, say that a "reanalysis" has occurred. But because we are predicting that the model can only make such a transition by passing through a contiguous set of intermediate states, we have a much more constrained theory of reanalysis.

Figure 1.2: Categorical Representation vs. Metric-Space Representation.
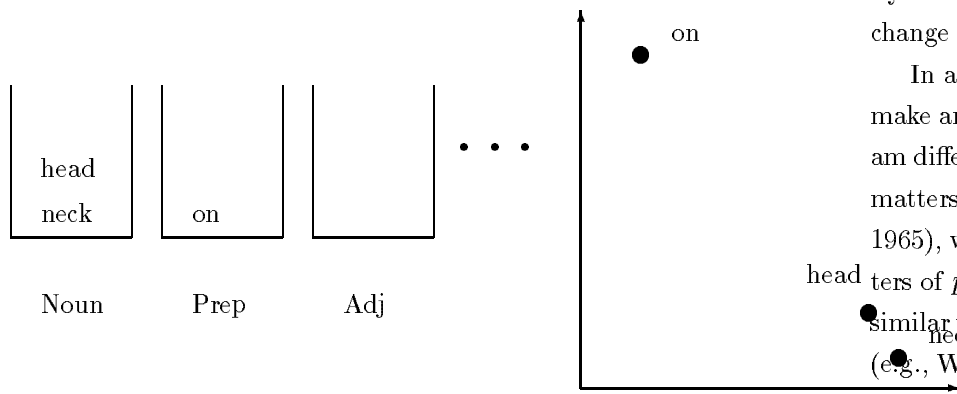
Can it really work, this proposal to capture all the categorical systematicities that generative grammars are generally concerned with and also allow change to be gradual at the level of representation? I believe it can if we adopt a suitable model of language structure. A recurrent Connectionist network trained on the

word-prediction studied by Elman 1990 and 1991 shows evidence of having the desired properties. I describe the Connectionist model briefly in Section 7 and examine it in detail in Chapter 3.

For ease of reference, I'll refer to the hypothesis that change is gradual at the level of a restrictive representation as the "Restrictive Continuity Hypothesis". In fact, all historical linguists who posit that language change is structurally constrained in some fashion can be characterized as invoking some kind of continuity hypothesis, for they are attempting to define a representation system in which states of a language are "adjacent" to one another in the sense that the language will or can transit between them. What is different about the current approach is the attempt to define grammar in such a way that nearby states are always nearby by the same criterion, a *single distance metric* which takes on real-number values. Contrast this with the case in which frequency-changes involve one part of the grammar (the probabilistic variable), lexical category changes involve another part (the lexicon), and major syntactic changes involve yet another (the syntactic parameters). There is nothing *a priori* that says such a taxonomic approach to predicting change cannot work, and indeed it may be easier to make it get the facts right. The advantage of the Restrictive Continuity model is that it makes expressing significant generalizations about language change easier.

In assuming that speaker's representations are sensitive to the choices they make and hear people make in situations where multiple options are available, I am differing from Saussure who thought that matters of "will" or "choice" were matters of *parole*, not of *langue*. Similarly, I am differing from Chomsky (e.g., 1965), who has suggested that matters of "will" in this sense in syntax are matters of *performance*, not *competence* and hence not of grammar. My proposal is similar to the variable-rule and -parameter hypotheses of the Variationist school (e.g., Wolfram 1969, Labov 1969, Cedergren and Sankoff 1981, Kroch 1989) in the sense that it also wants to treat quantitative information as something we can make formal predictions about. But as I noted above, it diverges from these hypotheses in proposing that the quantitative information be treated in the same manner as qualitative information (0 out of 100 potential instances is here taken to differ only in degree from 1 out of 100 potential instances). This difference is crucial because it puts a much stronger constraint on which

quantitative distributions can be associated with which traditional qualitative "grammars" at particular times.

1.7   A Connectionist Implementation

A Connectionist network studied by Elman 1990 and 1991 is useful in formalizing the Restrictive Continuity hypothesis. This section summarizes the main interesting properties of the network and motivates the use of this network instead of certain other frequency sensitive models. Chapter 3 gives a more in-depth introduction to the type of Connectionist model studied in this thesis. See Rumelhart *et al.* 1986 for a psychologically and linguistically motivated introduction to Connectionist networks in general. See Hertz, Krogh, and Palmer 1991 for a survey of the architectural and mathematical characteristics of the main types of Connectionist networks.

1.7.1   Elman's Simple Recurrent Network trained on word-prediction

Connectionist networks consist of *nodes* and links between the nodes, called *connections*. Each node is associated with a number called its *activation* and each connection is associated with a number called its *weight*. In one kind of network-processing, called *relaxation*, each node's activation is reset according to a function whose value depends on the activations of connected nodes, scaled by the weights along the connections. In another kind of processing, called *training* or *learning*, the weights are adjusted in a systematic, incremental fashion in order to get the nodes to take on certain desired activation values when relaxation occurs.

A network with the architecture shown in Figure 1.3, called a *Simple Recurrent Network*, can be trained to do sequence-prediction. The input units are set to activation patterns which represent successive elements of a sequence (e.g., a sequence of words in a sentence). It is taken as desired that when element *i* is represented on the *input layer*, element *i+1* should be represented on the *output layer*. The connections feeding from the hidden layer to the *context layer* are

designed to make the context activation at time *t* equal to the hidden activation at time *t-1*. This context layer plays a crucial role in making predictions about sequences in which next-elements are not predictable from immediately preceding elements alone. They can be used as a repository of information about past-states of the network so that, in principle, arbitrarily long-distance dependencies can be encoded. In practice, it turns out to be more difficult to get a Simple Recurrent Net to learn longer distance dependencies than shorter-distance ones, under standard training techniques. This feature is actually a desirable property from the standpoint of using the network to predict certain hybrid structures that emerge in processes of language change (see Chapter 6).

Figure 1.3: Elman 1990 and 1991's Simple Recurrent Net.
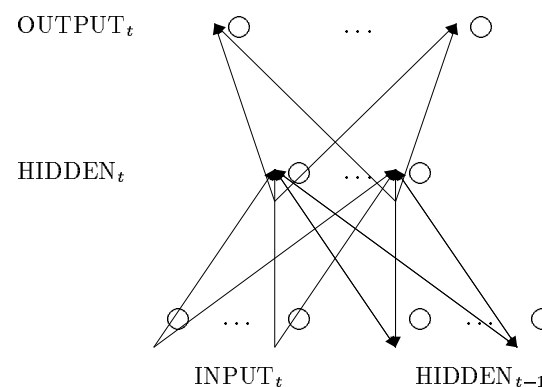
Elman 1990 and 1991 trained a Simple Recurrent Network to do sequence-prediction for some sequences resembling natural language sentences. In one experiment, he used a simple grammar to generate a set of "sentences" according to the patterns in (1).

(1) Noun Intransitive-Verb
     Noun Transitive-Verb Noun

The Nouns and Verbs fell into certain semantically-defined classes and were appropriately matched (e.g., agentive verbs only occurred with animate subjects, only nouns referring to fragile entities could occur as the object of the verb *break*, etc.). 10,000 such sentences were generated at random in accordance with these

restrictions and then strung end-to-end to form one long symbol-sequence which was used to train a Simple Recurrent Network. To perform perfectly on the prediction task, the network would have had to store a representation of the entire corpus in all its detail. This, of course would not be a very interesting behavior from the standpoint of understanding how people represent language since it is clear that we do quite a bit of *productive* recombining of elements we have heard before (e.g., Chomsky 1957). Interestingly, Elman's network extrapolated from the cases observed in the corpus to a larger set which was quite consistent with human productive intuitions. In particular, it learned to distribute activation over successor-words on the output layer as a probability-distribution, where the probabilities reflected the semantic and syntactic constraints of the simple grammar that Elman used to generate the corpus.

But even this behavior is not terribly interesting, since the corpus Elman used contained a great deal of information about the structure of the grammar that was extrapolated and the grammar itself was not a very complex grammar. What is more interesting is that under certain conditions, a network of this sort fails to learn aspects of the statistical structure of a grammar whose output it has been trained on. It doesn't fail to learn altogether, but it cuts corners or *smooths* the data in certain ways. These extrapolations reflect structure that the network is bringing to the task in virtue of its architectural constraints and learning procedure. In this sense, a network can offer distinctive hypotheses about the nature of natural language representation from the grammar that produces the data it gets trained on. Moreover, these distinctive hypotheses turn out to be useful in making predictions about how languages will change, so it is reasonable to believe that studying them can tell us something about social/psychological reality. Chapter 3 discusses these smoothing propensities of the network in more detail. Chapters 4, 5, and 6 present empirical evidence in favor of the particular smoothing tendencies of the kind of network studied here.

1.7.2   The Change Model

The network representation is naturally designed for thinking about change as a continuous process. Its states range over the set of weight values that can be

assigned to its connections. This set of values is called its *weight space.* The weights take on real number values, so we can think of a state of the network as a point in R^n where n is the number of connections in the network. In fact, the weights can take any real value so two network states can be arbitrarily close together. Thus we can model gradual grammar change as gradual change in the position associated with network in its weight space.

Of course, there are many ways that a point can move around continuously in R^n. It seems to make most sense to think about the network's representation as changing in response to various external pressures that impinge on language—e.g., sociological pressures on members of a speech community who share a common language (see Eckert 1988 and discussion in Chapter 7, Section 2.1.3), technological pressures which create a need for new ways of talking, pressures which arise as a result of the nature of language transmission (Chapter 7, Section 2.1.4), pressures that arise because of the way communicative needs "wear out" the current language (see Lehmann 1985 and Chapter 7, Section 2.2), etc. A natural way of simulating these pressures is to interpret them as generating new input-target pairs for the network and to model weight-change as Connectionist learning. I employ this strategy throughout the thesis. Connectionist learning is introduced in Chapter 3, Section 2. I do not, in fact, try to measure the external pressures on language directly, but I collect data from the history of the language which indicate certain persistent trends in its formal characteristics. It seems likely that these trends reflect persisting external pressures. Often the trends in particular characteristics have a number of additional formal correlates. Therefore I take the object to be to try to predict the correlates. These often seem to stem not from correlations in the language-external world but rather from facts about how the language itself is structured.[7] This is the sense in which my focus is primarily on language structure and only secondarily on language change. But this way of focussing on structure is especially useful in making predictions about change.

---

[7] e.g., if if a transitive verb becomes successively more preposition-like by becoming increasingly restricted to embedded environments, then it may also become more preposition-like by beginning to participate in Prepositional Phrase-specific processes (see Lord 1973 and Chapter 2, Section 2.2).

1.8 Predictions made by the Connectionist Restrictive Continuity model

The Connectionist model makes three predictions that non-quantitative models do not make. The first, which I call "Frequency Linkage", is related to a prediction of Kroch's Competing Grammars model. The other two, which I call "Q-Divergence" and "Hybrid Structures" present problems for Kroch's model, as well as for the non-quantitative models, and hence provide a clear way of distinguishing the Connectionist model from Competing Grammars. Chapters 4, 5, and 6 concentrate respectively on the three predictions. As a preview, I outline them here.

1.8.1 Frequency Linkage

The first prediction is that changes in the relative frequencies of grammatically-related forms should be correlated. It has often been noted that new grammatical constructions come into a language gradually: there is an old construction which serves a function invariably at first; then a competitor arises with low frequency; over the course of time the competitor's frequency increases, usually describing an S-shaped curve as a function of time (e.g., Graudina 1964, Bailey 1973). Kroch 1989b cites a number of studies showing that when one subdivides the environments of use of the two competing constructions into grammatical classes which cross-cut the distinction between the competing forms, then frequency changes of the new construction in the different grammatical environments occur in parallel. This is not to say that the frequencies of the new construction are the *same* across environments, but only that they covary across time. For example, Noble 1985 shows parallel rising curves for the development of *have got* vs. *have* in British English in the environments of abstract and concrete Noun Phrase complements in recent British English (see Sections 4.1.1 and 4.2). Hiltunen 1983 shows parallel developments in the ordering of the English verb with respect to various particles across main and subordinate clauses during the period OE–EME. In a related vein, Pintzuk 1991 argues that changes in the the base-position of the verb (clause medial versus clause final) occurred in parallel in matrix clauses and subordinate clauses during OE.

Fontaine 1985 shows parallel falling curves for the use of subject-verb inversion in Middle French across sentences with three different types of subjects: pronoun subjects, full NP subjects, and pro-drop subjects.

Given the pervasiveness of this *frequency linkage* effect, we should want any model of the quantitative properties of language structure to predict it. The Restrictive Continuity representation embedded in the Change Model does so in the following way: words that belong to the same grammatical category often have similar distributional behaviors. The restrictiveness of the representation (the *hidden unit space*—see Chapter 3) makes it imperative that items with similar behaviors be assigned similar representations, or else much of the structure in the data cannot be successfully encoded. Change is modelled as the process of training a network in one grammar-embodying state to adopt a different one. Since change by training is approximately continuous, changes in one part of the representation space must be accompanied by corresponding changes in nearby parts. Therefore, similarly-distributed elements are predicted to change similarly.

Kroch's Competing Grammars model also predicts the Frequency-Linkage effect but it does so in a slightly different way. On the Grammar Mixture model, frequency-change is assumed to be due to change in the probabilities associated with competing grammars. Therefore, if two constructions are generated by the same grammar, changes in their frequencies must be correlated. This model makes predictions that are empirically indistinguishable from those of the Restrictive Continuity model in cases where the distributions of like-classed items are very similar. However, it makes contrasting predictions in cases where the theory of grammar underlying Kroch's model (a version of Principles and Parameters Theory—e.g., Koopman 1984, Chomsky 1986) classes constructions together but their corpus distributions are measurably dissimilar. In this case, the Competing Grammars model still predicts that changes in the forms will be perfectly correlated. The Restrictive Continuity model predicts that changes will be correlated *to the degree* that the representations are alike (Chapter 3, Section 8 and Chapter 4). I argue in Chapter 4 that the Restrictive Continuity model makes superior predictions in the case of quantitative developments in English periphrastic *do*.

1.8.2 Q-divergence

It is not very surprising to learn that when words change category over time, they often undergo a set of correlated changes. A common type of *grammaticalization*,[8] involves a content word (like a verb or noun) giving rise to a function word (e.g., an auxiliary, preposition, complementizer, etc). For example, when the Kwa verb, *kpèlú*, 'accompany' became a preposition meaning 'with', it took on a variety of behaviors characteristic of prepositions: acceptance of Instrument and Manner arguments, participation in a PP-fronting rule, participation in PP/NP-conjunction (Lord 1973—see discussion in Chapter 2, Section 2). A somewhat different set of cases, where the same property obtains, are content-word to content-word transitions. For example, when *floor*, an old Germanic noun took on verbal meaning in ME 'to make a floor' (OED 15th cent), it took on the typical syntactic behaviors associated with verbs: taking subjects, direct objects, combining with *-ing* and *-ed*.

Nor is it very surprising to learn that in the grammaticalization cases some of the correlated changes may be quantitative in character. Grammatical elements generally have much higher frequencies than lexical elements (Zipf 1943) so when a word changes from being a lexical element to being a grammatical one, its frequency tends to rise, not only because it has two uses where before it had one, but because its new role is a high-frequency role.

What is perhaps surprising is that *prior* to the appearance of the first qualitative signs that a word has changed category, it may undergo changes in its quantitative distribution which foreshadow the qualitative change. I call this phenomenon *Q-divergence* for "quantitative divergence" and present evidence for its existence in Chapter 5. Two case-studies provide the primary evidence: (i) prior to the emergence of the expressions *sort of* and *kind of* as Degree Modifiers in English (e.g., *The weather is kind of windy.*), the frequencies with which they appeared in now-ambiguous contexts like *That is a kind of ominous-looking cloud.* increased significantly; (ii) prior to the first appearances of *be going to* as a future marker (e.g., *It is going to rain.*), there was an increase in its frequency as a motion verb in now-ambiguous contexts like *I am going to bring Fred a tarp.* These developments suggest that the quantitative properties of word-behaviors

---

[8]See Chapter 2, Section 2.

can influence their qualitative classifications.

The Competing Grammars model has no way of predicting correlations of this nature for it holds that quantitative and qualitative change are independent of one-another. It also makes no claims about correlations between lexical changes except where these interact with the parameter-settings. Thus, under the Competing Grammars model, though *going* may be classified as a motion verb by a given grammar, this implies nothing about whether it is also classified as an auxiliary verb or not. Even if *going* happens to be classified as both a motion verb and an auxiliary verb, no correlations in the way the two items change are predicted. Thus Q-divergence is a missed generalization under this model.

The Restrictive Continuity model, on the other hand, predicts Q-divergence effects because it treats quantitative and qualitative information homogeneously. If two types of elements cooccur in some context (e.g., Noun-Prep sequences and Degree Modifiers), but with different frequencies, then the Restrictive Continuity model takes this quantitative contrast as one of the contrasts that signals the difference between the two elements. Consequently, if a particular element belonging to one of these classes changes its quantitative distribution in the direction that makes it more like the members of the other class, then the model predicts correlated rises in rate at which the element exhibits behaviors associated with the other class. This follows from the restrictiveness of the representation: it cannot allow elements to take on arbitrary combinations of features because there is not enough "room" in the representation space (See Section 3.8).

So far, this account says nothing about the quantitative changes *preceding* the qualitative changes. In fact, all behaviors are modelled quantitatively in the Restrictive Continuity model, so to reconstruct a notion of qualitative change, we need a way of categorically distinguishing different quantitative behaviors. A natural method is to define a probability threshold such that usages that are below the threshold count as ungrammatical and usages that are above the threshold count as grammatical. This turns out to be easy to do under the Connectionist representations of at least the simple corpuses I have experimented with: in most cases, a single number can be found such that all grammatical usages have probability greater than this number (Chapter 3, Section 6.2 and

Chapter 5). This threshold must be positive since the network assigns positive probability to all behaviors, including ungrammatical ones. Consequently, it is possible for the correlations described in the previous paragraph to consist of correlated changes in below-threshold and above-threshold probabilities. When a trend in the above-threshold range is associated with a trend that brings an ungrammatical element across the grammaticality threshold, we will perceive a quantitative trend preceding a qualitative development (See Figure 1.4). In this way the Restrictive Continuity model predicts Q-divergence.

Figure 1.4: Quantitative change anticipating qualitative change.
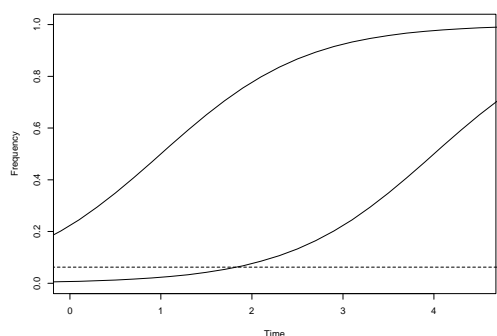
1.8.3   Hybrid Structures

The third prediction made by the Restrictive Continuity model has to do with certain hybrid structures that emerge during periods of transition. By "hybrid structures" I mean constructions that appear to be patched-together from parts of independently-existing syntactic or lexical units. Perhaps the most famous examples of such hybrids are the cases of morphological double-marking which involve the simultaneous presence of several synonymous formatives. Thus, Middle English (ME) eventually adopted *breþren* over *broþre*, *breþre*, and *broþren*; Modern English (ModE) children's speech has examples like *sanged* for the past tense of *sing* as well as *singed* and *sang*. Similarly, when preposition-stranding was starting to occur in English *wh*-relatives and questions, resumptive prepositions like those in (2) appeared [Allen 1983: 230].

(2) a. Til that the knight <u>of</u> which I speke <u>of</u> thus,...
   'Until the knight that I spoke of thusly.' [Ch F Frank. 807]
   b. And eek <u>in</u> what array that they were <u>inne</u>.
   'And also what array they were in.' [Ch Pro. 41]

Such hybrids are not something the Competing Grammars model can easily predict, for it posits that competing grammars are independent of one another. At best, the Competing Grammars paradigm can include a third grammar in the competition to generate these cases. But then there is no explanation for the fact that such "third-grammars" persistently show up in conjunction with periods of transition.

On the Restrictive Continuity model such behaviors are not only generable, but expected precisely during periods of transition. They are generable because, under continuity, an element may be assigned a representation that is intermediate between two independently motivated categories of the language. If the two categories are sufficiently similar, such an element can do a reasonable job of satisfying both of their contextual conditions simultaneously. Hybrids are expected to appear only during periods of transition because it is at such times that the parent categories are distributionally similar-enough that an intermediate element can span them. This contrasts with the situation of simple ambiguity, in which there is mixed evidence for the classification of a word, but the environments of its alternative uses bear no systematic relationship to one another.

1.9   Summary

The central argument can be summarized as follows:

Observation: Although grammar change looks abrupt, corpus change is relatively gradual.

Hypothesis ("Continuity"): Evolutive grammar change is actually gradual as well. In particular, the grammar assigns similar representations to elements with similar quantitative distributional structure.

Corollary: Historical "reanalysis" as traditionally construed, where an element changes category, or a new category is introduced into the language, is the long-term outcome of gradual change in syntactic representation.

Question: Can we have a representation in which grammar change is gradual but grammars are still restrictive?

Answer: Yes. A Recurrent Connectionist network trained on word-prediction has these properties, at least for simple data-sets.

In the next chapter, I review evidence from the literature on historical linguistics which motivates the Restrictive Continuity model.

Chapter 2

Evidence for Continuity

This chapter examines evidence from the linguistic historical literature which motivates my claim that evolutive structural change is most efficiently described with an underlyingly continuous grammatical representation. For each of two types of theories (Competing Grammars Models and Grammaticalization Models), I examine the phenomena that the theories make predictions about and show that there are systematicies which they fail to account for. These systematicities turn out, in each case, to be reflected in the structure of corpuses, provided we take into account information about the frequencies with which words and phrases occur. In each case, the proposed theories are hard-pressed to characterize the nature of the similarity between historically adjacent elements which are nevertheless uncontroversially related and have manifestly similar quantitative distributional characterisitics. Consequently, the data provide evidence for a model, like the Connectionist model proposed in the next chapter, in which elements with quantitatively similar distributions give rise to similar representations.

2.1 Evidence for continuity (1): Competing Grammars studies

Kroch's (1989a, 1989b) studies of the history of English periphrastic *do* provide the initial motivation and, to-date, the most thorough justification of the Competing Grammars model of diachronic variation. The central idea behind this model is to predict diachronic frequeny-variation by assuming that parameters of Universal Grammar can be reset on a probablisitic basis. The model makes constrained predictions about the variation data because it holds that when two construction-types are governed by the same parameter-setting, their relative frequencies must change in parallel (a Frequency Linkage effect—see Chapter 4). A number of people have followed up Kroch's work by arguing that certain long-term diachronic trends are most efficiently described under the Competing Grammars hypothesis: Santorini 1989, 1992, Pintzuk 1991, Taylor 1992, Fontana 1993. Among these studies, Fontana's is centrally concerned with an event of diachronic reanalysis so it is of particular relevance to the discussion here. In the following sections, I review Fontana's findings. Kroch's work is also of relevance to reanalysis but since it is directly concerned with Frequency Linkage, I take it up in Chapter 4.

2.1.1 Fontana 1993 on Spanish object clitics

Fontana 1993 is concerned with some dramatic behavioral changes that Spanish object clitics have undergone during the past nine centuries. At least five properties separate Old Spanish (OSp) object clitics from their counterparts in Modern Castilian Spanish (ModCSp) as indicated in Table 2.1. For brevity, in the remainder of this section, I will use the term "clitic" to mean "object clitic".

Clause-initial clitics (Contrast 1) can occur in Modern Spanish when the verb is the first heavy phonological element in the sentence. An example is given in (3).

(3) ModCSp: lo   vió   Juan
             it[CL]   saw   Juan
             'Juan saw it.' [F: 2]

Figure 2.1: Contrasts in the behavior of object clitics in Old Spanish and Modern Castilian Spanish [based on Fontana 1993].

	Old Spanish (OSp)	Modern Castilian Spanish (ModCSp)
1.	Almost no clause-initial clitics.	Clause-initial clitics allowed.
2.	Words can intervene between a clitic and its following verb in declarative sentences. ("Interpolation")	The verb immediately follows the clitic in declarative sentences.
3.	Clitics almost always encliticize	Clitics pro-cliticize to V in declarative word-order.
4.	Verbs can host enclitics when in initial position even in declarative sentences.	Verbs only host enclitics in infinitives, gerunds, and imperatives.
5a.	Almost no "Clitic Doubling" on Indirect Objects.	Frequent Clitic Doubling on Indirect Objects (Mandatory for some verbs)
5b.	Almost no Clitic Doubling on Direct Objects.	Frequent clitic doubling on Direct Objects in some dialects.

Examples of OSp "Interpolation" (Contrast 2), in which a word intervenes between a clitic and its following verb, are given in (4). All such sentences are ungrammatical in ModCSp.

(4) a. 13th c.   por que te   assi   encerreste
                because   yourself[CL]   thus   locked
                'Because you locked yourself up this way' (EE-II.3r) [F: 41]

    b. 13th c.   assi como les   dios auie prometido
                so   as   them[CL]   god   had   promised
                'As God had promised them' (GE-I.60v) [F: 42]

The direction-of-cliticization claim (Contrast 3) cannot, of course, be substantiated for every historical sentence since we have only spelling and poetic rhythm information to judge by. The former is only a weak indicator since archaic spelling systems can be very persistent and the latter is not available for a great many utterances. Nevertheless, spelling indicating pro-cliticization is rare in OSp and grows more common over the centuries.

Examples of (probable) encliticization to V in a declarative clause (Contrast 4) from earlier Spanish are given in (5).

(5) a. 15th c. E   gano-la   Yñigo Lopes de Mendoca
                and won-it[CL] YL de M
                'And Yñigo Lopes de Mendoca won it.' (Atal) [F: 232]

    b. 16th c. y   propuso-lo   en Consejo
                and proposed-it[CL] in council
                'And he proposed it to the council' (CV.25) [F: 232]

In regard to Contrast 5, the phenomenon standardly known as "Clitic Doubling" in the linguistic literature on Spanish is actually the doubling of direct object-clitics shown in (6), which is characteristic of certain dialects spoken in Central and South America.[1]

(6) ModRPSp:[2]

    Yo lo   voy   a comprar el diario   justo antes de subir.
    I   it[CL]$_i$ go.1sg to buy   the newspaper$_i$ just   before of coming-up
    'I'm going to buy the newspaper just before coming up.' (Suñer 1988) [F: 221]

Here, I follow Fontana in referring to both Indirect Object doubling (7) and Direct Object doubling (6) as "Clitic Doubling" (see also Suñer 1988 and Franco 1991, 1993).

(7) ModCSp:

    Tuvimos que venirnos porque le   parecia a tu padre...
    had.1Pl for return   because to-him[CL]$_i$ seemed to your father$_i$
    'We had to come back because it seemed to your father...

    que iba   a llover
    that was-going to rain
    ...that it was going to rain.' [F: 278]

Fontana provides a series of quantitative tables showing that certain roughly monotonic frequency-shifts occurred during the course of the 12th through the 16th centuries (pp. 236, 244, 246, 249, 252). He argues that the frequency-shifts reflect two distinct grammar-competitions going on approximately simultaneously:

(i) Loss of topicalization-position to Spec(IP).

(ii) Loss of $I^0 \rightarrow C^0$ movement.

As a result of these two shifts taking place, he claims, a third change became increasingly probable.

(iii) Object clitics became classified as lexical formatives (in his terms, X^0 categories) affixed to V, rather than syntactic projections (X^{Max}).

The structural analyses associated with frequency shifts (i) and (ii) are illustrated in Figures 2.2 and 2.3. In each case, the outgoing grammar had the movement rule; the incoming grammar lacked it.[3] One can think of the use of a movement rule as being controlled by a boolean parameter-setting: when the parameter has one value, movement is obligatory; when the parameter has the other value, movement is prohibited. Thus the variation between using the movement rule and not using it can be modeled, in each case, as a probabilistic "competition" between two grammars.

The logic of Fontana's claim that shifts (i) and (ii) encouraged shift (iii) is as follows: Shift (i) often left the clitic in sentence-initial position where it had no host to its left to encliticize onto; consequently it procliticized to whatever was on its right; the verb was increasingly on its right because of shift (ii); since procliticization is phonologically similar to affixation, and argument-marking on a verbal head is a structure that Universal Grammar provides (as evidenced by the fact that it occurs in many other languages), the situation was ripe for a reanalysis of the clitics as affixes.

---

[1]Fontana notes that DO-doubling also occurs in some 16th century Castilian texts (Fontana 1993: 264).
[2]Modern River Plate Spanish.

[3]And perhaps had some other rule to replace it. For example, Fontana make the standard assumption that a constituent can be made topical by adjunction to CP.

Figure 2.2: Topicalization to Spec(IP).
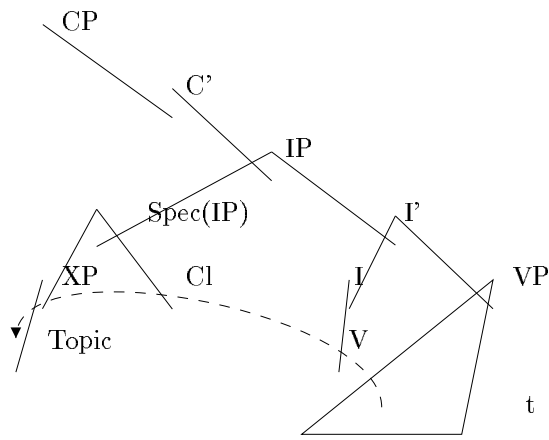
Figure 2.3: I^0-to-C^0 Movement.
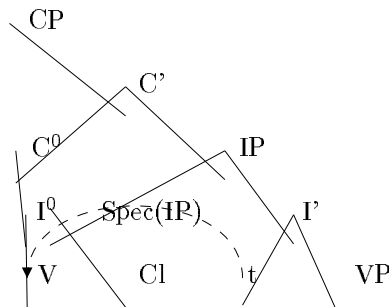
In short, Fontana's account supports the notion that frequency-shift can influence structural change. This suggests that there is a Q-divergence effect involved in the Spanish clitic case. But should we believe his account? There are several potential concerns that need to be addressed.

First, How do we know that the clitic pronouns have really become affixes in ModCSp? In fact, the issue is controversial: Fontana argues for treating the elements as agreement markers on the grounds that (i) they mandatorily participate in Doubling in various (di-)transitive clauses in various dialects and

(ii) the IO-Doubling and IO+DO-Doubling dialects can be viewed as successive stages on a path toward generalized object agreement marking, which seems to be the historical trend in this case (pp. 277–80—see also Suñer 1988, Franco 1991, 1993), and indeed is attested in several languages (Chafe 1976; Givón 1976, Bresnan and Mchombo 1987). However, other researchers (e.g. Di Sciullo 1990) note that (a) the Verb-Clitic order used in imperatives, gerunds, and infinitives suggests a verb-movement analysis, which makes the morphological treatment suspect; (b) the phenomenon of *clitic climbing*, in which the clitic attaches to the highest verb in a verb complex, suggests NP-movement, which also casts some doubt on the affixal account; (c) and the fact that in DO-Doubling languages, the clitics double definite NPs but fail to double indefinite NPs, makes them seem more like full arguments and less like agreement markers. Fontana counters by noting that (a) there are other languages in which what look like morphological classes seem to alternate their positions within the word (pp. 281–2); and (b) there are other languages in which what looks like object-agreement morphology appears either on the main verb or on the tensed verb when there is a sequence of verbs (pp. 284–8); and (c) there are other languages in which what appears to be a morphological agreement system has gaps partially conditioned by definiteness features (pp. 283–4). What is ignored in this exchange, and what seems worth noting is that these facts may be most efficiently captured using a representation system that permits intermediacy: Di Sciullo's problems for the affixal treatment involve a divergence from the simplest, most canonical lexical behavior (mandatory fixed-position marking on the main verb) *in the direction* of the distributional behavior that typified the OSp clitics, for (a), as noted above, the clitics alternated between pre- and post-verbal position in OSp and (b), the tendency for clitics to occur in second position implied a tendency for them to be adjacent to the tensed verb, rather than the main verb, in sequenced verb constructions (Fontana 1993: 285–8). Moreover, it seems likely that the restriction to definite NPs reflects an incompatibility between ad-pronominal apposition and indefinite reference:

(8) John sees it, the dog.
     # John sees it, a dog.

This incompatibility must have made the historical distribution of these pronoun + appositive constructions quite definite-heavy so if we make the plausible assumption that the modern Doubling constructions descend from such appositives, we can again say that the modern distribution is divergent from the canonical lexical distribution precisely in the direction of the historical state.

At any rate, despite the existence of some disagreement about the current status of the current "clitic" pronouns in Spanish, to substantiate Fontana's claim about quantitatively-conditioned reanalysis, we need only show that some pertinent reanalysis has occurred since the OSp time. This is less problematic, for we can cite properties 2, 3, and 5a from Table (2.1) above, which constitute categorical differences between the Modern clitic pronouns and their ancestors: interpolation is now ungrammatical in all dialects (Contrast 2); the pronouns can no longer encliticize onto a preceding phrasal host with declarative word-order (Contrast 3); and, the clitics not merely optionally but obligatorily double certain classes of objects in certain dialects (Contrast 5a).

One may also wonder whether the Q-divergence effect that Fontana seems to have observed may not be a Q-divergence effect after all but may, in fact, be describable within the Competing Grammars framework as it stands. In particular, one might ask, Could the switch from X^{max} to (something like) X^0 of the object pronouns be modeled as a parametric correlation with one of the other trends that Fontana observes? In this case, the rise in the use of the X^0 behaviors would be counted as a frequency-linkage effect (assuming the slopes of the relevant relative-frequency curves turn out to be the same—see Chapter 4).

This account won't work because Topicalization to Spec(IP) and I^0-to-C^0 movement are not always bidirectionally correlated with non-affixal clitic behaviors. For example, regarding Topicalization, it seems to have been possible to have a V-2 construction that was also a Clitic Doubling construction (9) (although Fontana suggests that this example, with its intervening caesura, may involve a right-dislocated object and hence not be true Doubling).

(9) 12th c.

Gran Iantar le   fazen | al buen canpeador
big meal   him[CL]$_i$ prepare | to-the good "canpeador"$_i$
'They are preparing a big meal for Cid, the good fighter.' (PMC) [F: 263]

Correspondingly, it may also have been possible, in earlier Spanish, to have a construction in which Spec(IP) topicalization failed to occur and in which an object clitic simultaneously failed to behave lexically. A case of this sort would be a verb-initial clause with an overt object (direct or indirect) which does not belong to one of the types that sanction I^0-to-C^0 movement[4] and in which there is no clitic present. Such a case could not involve Spec(IP) Topicalization, because if it did, the verb would not be initial. Such a case could not involve a lexical agreement-marker clitic because then the presence of the overt object would mandate the presence of the clitic. It seems probable that such cases will have occurred in the transition period when there were many verb-initial sentences occurring. Indeed they are predicted to be possible under Fontana's analysis although I have not been able to find one among his examples. Of course, it would be possible to claim, of such a case, that it involved a lexical object clitic that happened not to be a mandatory agreement marker and hence was optional. In other words, the theory would involve an additional parameter which distinguished two types of lexically attached pronominal elements: those that are mandatory agreement markers and those that are optional agreement markers. But if this is the case, then Spec(IP) Topicalization would not be controlled by the same parameter as Doubling, so one could not treat the correlation between the demise of the one and the rise of the other as a frequency linkage effect anyway.

Nor was affixation linked parametrically with the failure of I^0-to-C^0 movement: it was possible, in earlier Spanish, to have I^0-to-C^0 movement with Clitic Doubling (10). This example does not plausibly involve right-dislocation because the subject (*el senor Tamurbeque*) is after the doubled object.

(10) 15th c.

---

[4]The types that sanction I^0-to-C^0 movement are Questions, Imperatives, the Narrative Inversion construction, and possibly one other hard-to-define environment (Fontana 1993, pp. 170–7).

   & vencio-lo   al turco   el senor tamurbeque

   & defeated-CL$_i$ the turk$_i$ the lord Tamurbeque

   'And Tamurbeque defeated the Turk' [F 93: 268]

Correspondingly, it seems to have been possible to have sentences where I^0-to-C^0 movement failed to occur but in which a clitic failed to behave as an agreement marker. An example of this type is (11) which has the discourse properties that permit I^0-to-C^0 movement,[5] but nevertheless I^0-to-C^0 movement has not occurred. But nor is there a clitic, despite the presence of an object (*el Alhanbra*).

(11) 15th c.

   E   en este año tomo el rrey Ysquierno de Granada   el Alhanbra...

   and in that year took   the king Ysquierno de Granada the Alhambra

   'And that year king Y. from Granada took the Alhambra...' [F 93: 255]

This isn't a particularly surprising sentence from the perspective of ModCSp because the object is a direct object and direct objects are never doubled in that dialect. Nevertheless, given the possiblity of Direct Object Doubling evidenced by other dialects, this example makes it clear that whatever parameter controls Doubling is not the same parameter that controls I^0-to-C^0 movement. It should be noted as well that the mere fact of dialectal contrast with respect to the Doubling of Direct and Indirect Objects makes it problematic to predict the rise of Doubling in general as a frequency-linkage effect.

In sum, affixation was not categorically linked with either loss of Topicalization to Spec(IP) or loss of I^0-to-C^0 movement, so it cannot be predicted as a correlate of one of these changes under the Competing Grammars model.

Finally, we might want to ask, Could we not assume that there was alternation between using object clitics that were maximal projections and using object "clitics" that were affixes right from the beginning, and that this pair of competing grammars played out its competition through the same period that loss of V-2 and loss of I^0-to-C^0 were occurring? We could, but then all the

---

[5] In particular, it is a Narrative Inversion environment. The speaker could presumably have said *E tomo en este año....*

correlations would be treated as coincidences so this would be an undesirably stipulative account.

In sum, although Fontana's treatment of ModSp dependent object pronouns as X^0 categories is controversial, the observation that the loss of Spec(IP) Topicalization and of I^0-to-C^0 movement coincided with the emergence of a new role for the clitic object pronouns is not in doubt. Moreover, this correspondence cannot plausibly be treated as a frequency-linkage effect in the Competing Grammars sense because the occurrence of affixation is categorically (though not quantitatively) independent of the first two developments. Thus, as Fontana has suggested, the case does provide evidence that quantitative distributional shifts can encourage structural reanalysis, or in other words, evidence for Q-divergence effects.

2.1.2 Why not a Probabilistic Reanalysis Model?

One might also ask: Does the existence of quantitative influences on structural change really argue against a standard, categorical model of grammatical structure? Could we not simply augment the standard model with a probabilistic account of when reanalyses are more and less likely to occur in order to predict correspondences like the one Fontana has observed? For example, if one supposes that various parameter-setting configurations of Universal Grammar overlap in their coverage of form-meaning pairings, then one can posit that the likelihood of a particular learner's choosing any given parameter configuration to handle one of the structurally ambiguous cases increases with the relative frequency with which that learner encounters the construction. One could then suppose that the emergence of a new analysis in an ambiguous environment will be correlated with its emergence in other, non-ambiguous environments.

One would thus say that in earlier Spanish, as sentences with a non-doubled object clitic preceding the verb and nothing preceding the clitic (12) became more and more common relative to sentences in which Spec(IP) topicalization occurred, speakers became more and more likely to analyze the clitic as an affix rather than a clitic.

(12) ObjectClitic Verb ... (no overt object NP)

The more they chose the affixal analysis in cases like (12), the more they tended

to use an affixal analysis elsewhere. Consequently, Doubling constructions appeared, and rose in relative frequency. I'll call this the *Probabilistic Reanalysis* account since it augments the traditional reanalysis model with probabilistic information.

2.1.2.1 Form-based Relative Frequency

To make this hypothesis explicit, it is necessary to define *relative frequency*. Relative frequency, in turn, depends on the notion of an equivalence class, that is a set of instances whose cardinality forms the denominator in the calculation of the frequency ratio. In Competing Grammars studies, equivalence-classes are usually defined in terms of optional transformations: if a transformation is optional, then all the derivations in which it could occur form an equivalence class (e.g., Kroch 1989b posits that in ME, V^0-to-I^0 movement was optional for main verbs [see Chapter 4]); Fontana 1993 posits that I^0-to-C^0 movement was optional during much of the history of Spanish). But Competing Grammars researchers sometimes imply that semantic or pragmatic interchangability can also define an equivalence class (e.g., Fontana 1993 assumes that two different ways of topicalizing elements—Spec(IP) Topicalization and Clitic Left-Dislocation—defined an equivalence class in earlier Spanish [Chapter 5]). Unfortunately, neither the notion of transformational equivalence nor the notion of semantic/pragmatic equivalence seem to be pertinent to all cases of innovation. For example, in Chapter 5, I discuss the development of Degree Modifier *sort of* and *kind of* from the earlier Noun+Preposition usage of these collocations. It is not clear that either construction is a transformational variant of anything. And although certain Degree Modifier usages of *sort/kind of* are similar semantically to behaviors of words like *rather* and *somewhat*, these words are not universally interchangable (see Chapter 5, Section 5.2.1). Similar problems arise in determining semantic similarity for future *be going to* (see Chapter 5, Section 5.2.2) where again, transformational optionality does not seem to play a role.

Fortunately, there is an alternative notion of equivalence which provides a more robust way of making predictions about innovation: we can take all the instances of a phonological string to form an equivalence-class and compute the ratio of the number of times the string occurs in some particular syntactic

environment to the total number of instances of its occurrence. I'll refer to this kind of relative frequency as *form-based* relative frequency. *Form-based* relative frequency has the advantage that it is quite easy to measure. Moreover, if we choose forms that are words or phrases in the language under study, then form-based relative frequency statistics should be constrained by the tendency for there to be a one-to-one mapping between word/phrase form and meaning. Some researchers have suggested that this tendency plays an important role in language change (e.g. Bever *et al.* 1976).

2.1.2.2 Problems with Probabilistic Reanalysis

One might propose to use the above-described Probabilistic Reanalysis model to make predictions about form-based relative frequencies. Unfortunately, this approach does not work very well. I'll now explain why.

We can distinguish two varieties of Probabilistic Reanalysis: one, called *Actuative Reanalysis* in which frequency-change of a conditioning construction enhances the chance of a speaker's adding a new structure to her grammar but has no implications for the rate at which the new structure is used; and another, called *Regulative Reanalysis* in which frequency change in the conditioning environment is monotonically related to frequency change in the new environment. The latter hypothesis is rather similar to the frequency-linkage effects posited by the Competing Grammars model but it differs in not requiring the slopes of the correlated frequency curves to be identical (see Chapter 4).

The problem with Actuative Reanalysis is that it makes no predictions about frequency trends, although these are very commonly observed in conjunction with events of reanalysis. Although Fontana does not provide quantitative data on the rise of clitic Doubling, it is clear that there is a general trend from virtually no Doubling to virtually categorical Doubling in certain constructions across the history of Spanish, and this trend is at least roughly correlated with the loss of Spec(IP) topicalization and the loss of I^0-to-C^0 movement. In Chapter 5 I present other evidence for correlations among frequency trends in conjunction with reanalyses.

The problem with Regulative Reanalysis is that it implies that change in the relative frequency of the emerging unambiguous construction is monotonically

correlated with change in the relative frequency of the ambiguous construction. A hypothetical episode of the type implied by these assumptions is plotted in Figure 2.4.[6] Is the form of this picture consistent with what happened? Unfortunately for the Probabalistic Reanalysis account, it is not. As I noted above, we have the information that clitic Doubling was rare in Old Spanish and has become obligatory in certain constructions in modern Spanish. Thus the form-based relative frequency curves for the case at hand must actually look something like Figure 2.5. Note that the slope of the curve for the unambiguous construction is first positively correlated and then negatively correlated with the slope of the curve for the ambiguous construction. It is this "succession of lobes" character of the frequency data that the Regulative Reanalysis account cannot predict.

The Restrictive Continuity account, by contrast, is ideally suited to characterizing succession-of-lobes frequency data for it models reanalysis by positing continuous progress of an element's representation[7] through a succession of regions in a representation space. Since the mapping from the representation space to behavior is also continuous, and each region of the space is significantly associated with some particular set of categorical behaviors, the tendency for the transiting item to exhibit those behaviors waxes as the item's representation approaches the region in question and wanes as it moves away again. I return to this point and show lobed frequency data and predictions for the history of *be going to* in Chapter 5.[8]

2.2 Evidence for continuity (2): The Grammaticalization literature

Grammaticalization, as a field of study, is concerned with processes in which the grammatical characteristics of certain linguistic elements (usually morphemes,

---

[6] I have assumed, for the sake of specificity, that the relative frequency of Doubling constructions is always half that of ambiguous Cl-V constructions, although the point holds under any theory that takes the frequency of Doubling to be a monotonic function of the frequency of the ambiguous construction.

[7] In this case, the representation of each object pronoun.

[8] See also the discussion of Craig 1991's quantitative data on Rama postpositions and preverbs in Section 2 below.

Figure 2.4: Relative frequency correlations predicted by the Probabilistic Reanalysis Model (assuming Regulative Reanalysis).
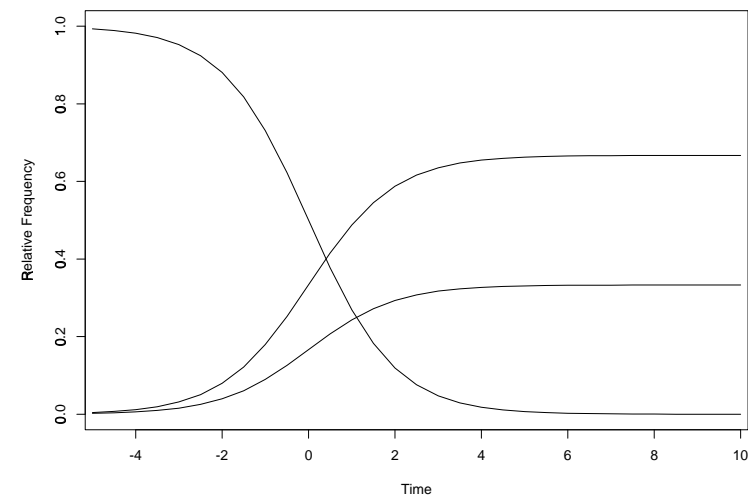
words, or phrases) change across time. The examples I mentioned in Chapter 1 are typical: development of serial verbs into case markers (Lord 1973, Carlson 1991), development of topics into subjects (Shibatani 1991), development of motion verbs into future markers (Perez 1990, Bybee, Pagliuca, and Perkins 1991). In fact, Fontana 1993's claim, discussed in the previous section, that Spanish object clitics are developing into agreement markers, is consistent with a tendency for pronouns to develop into agreement markers that has been observed elsewhere (Lambrecht 1981: non-standard French; Chafe 1976: Iroquoian; Givón 1976, Bresnan and Mchombo 1987: Bantu). These category changes tend, by and large, to have certain properties which can be expressed as directional tendencies: concrete meanings are replaced by more abstract meanings, world-referential meanings are replaced by discourse-referential or "subjectified" meanings,[9], low-frequency elements become high-frequency elements, content words become function words, syntactic units become morphological units, phonologically independent forms become dependent, phonologically full

---

[9] See Traugott 1989, Forthcoming.

Figure 2.5: Relative frequency correlations predicted by the Restrictive Continuity Model.
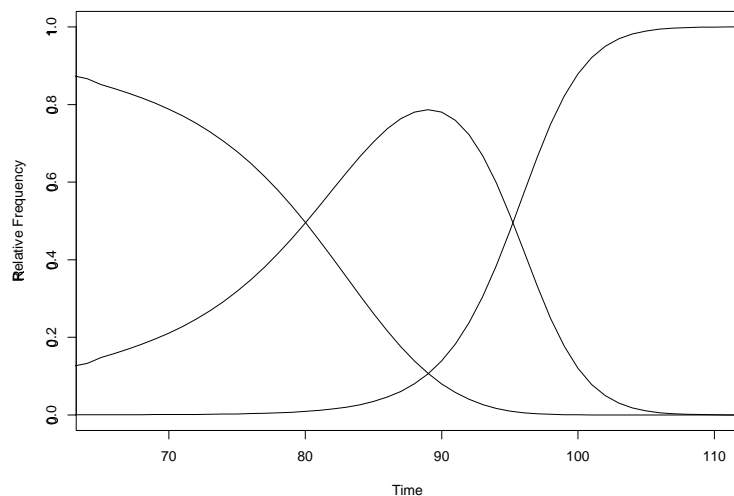
forms erode. These widespread directional tendencies have been subsumed under the general heading "Unidirectionality", and may reflect a universal law, although there seem to be sporadic exceptions (Janda 1980, Campbell 1991; see Hopper and Traugott 1993: 48–50 for further discussion).

The data on grammaticalization provide three kinds of evidence bearing on the Continuity Hypothesis: (a) evidence that syntactic theory as currently conceived is a better basis for a constrained theory of reanalysis than is sometimes assumed but is not an adequate basis (Sections 2.1-2.2); (b) evidence that adding semantic information can go a long way toward making the account appealingly restrictive, but that it is hard to characterize the semantic information in an explicit-enough way to make a predictive theory of it (Section 2.3); and (c) evidence that certain corpus-distribution changes are correlated with reanalyses and that therefore the theory can use this information (which is more easily described in explicit terms and more easily embedded in a general theory) to constrain its predictions (Section 2.4).

2.2.1   Lexical Predispostion

Consider the somewhat arbitrary sample of grammaticalization transitions in Figure 2.6.

Figure 2.6: Some Transition-types in Grammaticalizations.

Change	Language Group	Source
Verb [of saying] > Complementizer [embedding]	**Kwa** (Niger-Congo)	Lord 1976
Verb [of locational existence] > Prep [locative]	**Kwa** (Niger-Congo)	Lord 1973
Verb [locational existence)] > Postposition [locative]	**Senufo** (Niger-Congo)	Carlson 1991
Verb ['meet, assemble'] > Preposition [comitative]	**Kwa** (Niger-Congo)	Lord 1973
Verb ['know'] > Auxiliary ['can']	**English** (IE)	Lightfoot 1979
Noun ['time'+Dat] > Compl. [*while*]	**English** (IE)	Traugott and König 1991
Noun ['way'+Dat] > Preposition ['because of']	**German** (IE)	Hopper and Traugott 1993
Noun ['mind'] > Adjective→Adverb suffix	**Romance** (IE)	Lausberg 1962 (H&T 93)
Noun ['step'] > Negation Marker	**Romance** (IE)	Möhrer 1943, etc.
Adjective ['full'] > Noun→Noun suffix	**English** (IE)	Marchand 1966
Article > Gender marker	(Bantu)	Greenberg 1978
Postposition > (Inflectional) Complementizer	**Rama** (Chibchan)	Craig 1991
Preposition > Conjunction ('and')	**Kwa** (Niger Congo)	Lord 1973
Preposition > Conjunction	**To'aba'ita** (Austr.)	Lichtenberk 1991

This table casts doubt on the simple-minded hypothesis that significant constraints on reanalysis other than those entailed by Unidirectionality can be stated in terms of basic lexical-class predispositions to change in certain ways. There are many diverging paths that originate at "Verb" and "Noun", and there don't seem to be any natural generalizations about them. This *lexical predisposition* hypothesis is further vexed by the fact that some grammaticalization

processes involve reanalyses of multiple elements that do not plausibly correspond to lexical constituents or even to lexically-headed phrasal consituents (Figure 2.7).

Figure 2.7: Grammaticalization cases in which a non-constituent is reanalyzed as a constituent.

Noun + Prep > Degree Modifier	**English** (IE)	Tabor 1994 and Chapter 5
Object Clitic + Verb > Inflected Verb	**Spanish** (IE)	Fontana 1993
Verb + Copula > Past tense verb	**Polish** (IE)	Andersen 1987
Postposition + Verb > Relational verb	**Rama** (Chibchan)	Craig 1991

Moreover, there is evidence that a single element at one point in the history of a language can spread in parallel into several different domains in a process called "polygrammaticalization" (Craig 1991, Givón 1991b, Lord 1976; see review in Hopper and Traugot 1993). If "polygrammaticalization" is at all common, then lexical class information alone is only a very weak predictor of future grammatical development.

2.2.2   Syntactic Conditioning

Much greater predictive ability can be achieved if we take into account the syntactic environment in which each transition occurs.

For example, in Yoruba (a Kwa language of the Niger-Congo group), the form *kpèlú* can be used as either a verb meaning 'be included among' (13) or a preposition meaning *with* (14).

(13)   ìwe   náà '   <u>kpèlú</u>   àwõ tí   mõ rà
       book the SUBJ be-included-among those that I   buy
       'The book is included in those that I bought.' [Lord 1973: 280]

(14)   mō wà n'ibè <u>kpèlú</u> ākĩ
       I   be there with   Akin
       'I was there with Akin.' [Lord 1973: 281]

Lord 1973 argues that the verb usage is historically prior and gave rise to the preposition usage. The following discussion is a review of relevant points from her paper.

In support of her claim that *kpèlú* has true verb status, Lord notes that it shares four distinguishing characteristics with other verbs:

(i) It can occur alone as the sole argument-licenser in a clause.

(i) A non-pronoun subject preceding it ends in a high tone (13).

(ii) It undergoes a "focus-placement transformation".

(iii) It takes tense-aspect markers.

But in other contexts, *kpèlú* shows all the signs of being a preposition:

(i) It can introduce Instrument and Manner arguments:

(15) a.   ó   gé   ērã   kpèlú òbē
          he cut meat with   knife
          'He cut meat with a knife.' [L 73: 281]

     a.   ó   gé   ērã   kpèlú ès`ɔ
          he cut meat with   care
          'He cut meat with care.' [L 73: 282]

(ii) It can be fronted in question-formation (not possible with verb phrases):

(16)   sé kpèlú ɔwɔ   nī   w ó kí   ī
       Q   with   respect that they greet him
       'Was it with respect that they greeted him?' [L 73: 283]

(iii) The conjunction, *àtī* 'and', can be used to conjoin noun phrases with noun phrases, prepositional phrases with prepositional phrases, and *kpèlú*-phrases with prepositional phrases or other *kpèlú*-phrases, but not verb phrases with verb phrases.

Despite the bimodal character of the synchronic distribution, there is compelling evidence that the two forms are historically derived from a single source:

(i) The two words are homophonous.

(ii) Three other Kwa languages have verbs and prepositions which are similar phonologically, semantically, and syntactically to each other and to the Yoruba forms:

Ewe	kpé	'meet, come in contact'
	kple	'with'
Gã	kpè	'meet', 'collide'
	kè	'with'
Fon	kple	'assemble', 'bring together'
	kpôdô…kpan	'with'

Several researchers have, in fact, argued that Gã *kè* is a verb (e.g., Trutenau 1973) because, like verbs, it takes object pronouns and zero-marking with an inanimate third-person singular referent, but Lord notes that it has several properties that distinguish it from verbs: it does not occur without another verb in a clause and it does not inflect.

Such meaning/form coincidences would be unlikely if the words did not have a common ancestor. Although it is possible that each form originated independently in one language and was borrowed by all the others, this would involve positing six coincidental borrowings of vocabulary items with relatively core meanings, an unlikely scenario. It also seems much more likely that the hybrid behaviors of Gã *kè* reflect its historical antecedents than that its marked behavior is merely coincidentally similar to that of an existing class. Lightfoot 1979, reviewing these Kwa facts, notes that although synchronic similarities can be used to argue for a historical relationship between forms, they do not in-and-of themselves provide evidence bearing on the direction of a change. Here, however, we can cite the prevailing unidirectionality of grammaticalization as evidence in favor of Lord's claim that the Kwa change was from verb to preposition rather than vice versa.

Given this summary of the Kwa situation, we may ask: What properties of the situation were crucial in permitting the reanalysis? It is probably not coincidental that the Kwa languages are serial verb languages with VO word-order. Lord 1976 notes: "these languages tend to prefer using verbs whenever they can to express what other languages use casemarkers, adverbs and adjectives

for" (p. 179). For example, in Twi, the verb *ma* 'give' can be used to mark a beneficiary argument of another verb:

(17) ɔ   yɛ   adwuma <u>ma</u>   me
he does work   give me
'He works for me.' [Lord 1973: 270]

In Ewe, the verb *le* 'be-at' can be used to mark a locative argument of another verb:

(18) me fle   agbalẽ <u>le</u>   keta
I   buy book   be-at Keta
'I bought a book in Keta.' [Lord 1973: 271]

Moreover, Lord 1973 cites evidence comparable to the data on Yoruba *kpèlú* that the Yoruba preposition *ní* 'at' derives from a verb meaning 'be', Carlson 1991 cites evidence for similar developments in the Senufo languages which also make heavy use of serial verbs, and Li and Thompson 1973 argue for a similar development in Mandarin, which also employs serial constructions. These facts, combined with the fact that the prepositions in languages which do not employ serial constructions (e.g., Germanic, Romance) do not show signs of being historically derived from verbs, indicate that serial verb syntax may have been a crucial conditioning factor for the reanalyses.

(19) Concerning/regarding the new data

Moreover, the Senufo case studied by Carlson 1991 involves a development not from verb to preposition but from verb to postposition. Predictably, the Senufo languages have OV word order, so it is highly likely that the about-to-be-reanalyzed verb followed the argument it licensed. [10] In light of these observations, we may posit (following Lord 1976) that the reanalyses from Ewe *be-at* to *at* took the form:

---

[10]For example, the Senufo language, Supyire has a verb *fùn* which means 'consider taboo, abstain from, refrain from' (20). Cebaara has a postposition *fùn*, which means 'without' (21).

(20)
Sàànɔgɔbíí   màha sòòŋi   fùn.
Saanogos.DEF HAB terrapin.DEF refrain-from
'The Saanogos refrain from eating the terrapin. (taboo)' [Carlson 1991: 213]

(21)

(22)
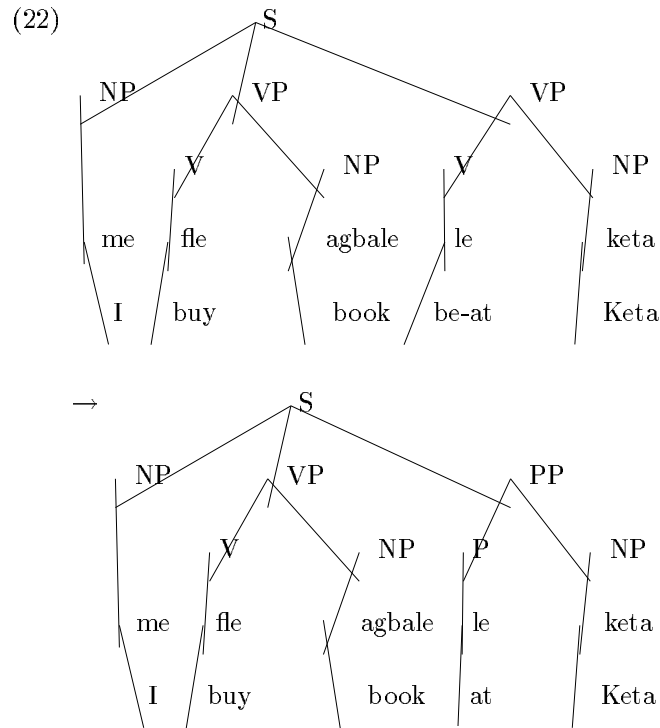
It is tempting, on the basis of this picture, to hypothesize that reanalysis is constrained to re-label the nodes on phrase-markers while preserving their branching structure. If such a claim were tenable, it would mean that syntax could put a very strong constraint on reanalysis. Unfortunately, the constraint is not tenable under current opinions about syntactic analyses, for evolutive change seems to be able to give rise to categorically new phrase structures (e.g. Schwegler 1983, 1988 and Chapter 7, Section 2.3 on French *pas* 'NEG2' < *pas* 'step'; Lightfoot 1979, Warner 1990, Lightfoot 1991 on the English modals; Fontana 1993 and Chapter 2, Section 1.1 on Spanish object pronouns) and also seems to be able to permit quite arbitrary re-bracketings of constituents (e.g., English *a [sort-of red] brick* < *a sort [of red brick]* (Tabor 1994 and Chapter 5)). Nevertheless, the correlation between serial verb-structure and Verb >

Wi   i   Senárì nyuu tèè lè fùn.
S/he PROG Senari speak fault without
'S/he speaks without fault.' [Carlson 1991: 213]

Adposition reanalysis and the correlation between basic VP word order and the positioning of the newly-created adpositions relative to their objects suggests that syntactic information puts some strong constraints on which reanalyses a particular language is susceptible to.

Lightfoot 1979 observes these the Kwa Verb > Preposition developments, along with a Verb > Auxiliary development (see Schacter 1974) and certain Verb > Complementizer developments (see Lord 1976) in the same family. He concludes (against my claim about syntactic conditioning):

> Actual changes seem to involve an extremely wide range of formal characteristics and it is by no means clear that there are any limitations, other than those imposed by the theory of grammar [p. 228]

But each of the "wide range of formal characteristics" he refers to is associated with a syntactic setting which is almost ready-made to engender the reanalysis in question. The adpositions develop from verbs that introduce locative, beneficiary, instrumental, or other potentially thematic arguments of verbs, just as adpositions generally do. The auxiliary verb probably developed from an ancestral intransitive verb meaning 'go' which occurred in the SVO Kwa proto-language between the subject and more-contentful verbs in various serial constructions (cf. (23), just as auxiliary verbs in SVO languages normally do.

(23)   Kofi a-ko-pase
        Kofi PERF-go-walk
        'Kofi has gone for a walk.'

The complementizers probably descend from a Kwa proto-verb meaning 'say' which occurred after verbs with meanings like 'believe', 'think', 'want', 'be afraid', etc. and before a finite clause, and probably had coordination semantics (see Lord 1976). In each case, the ancestral distribution appears to have strongly resembled independently motivated canonical properties of the new type that the reanalysis produced (adposition, auxiliary, and complementizer, respectively). The hypothesis that these three cases of resemblance are coincidental seems much less probable than the hypothesis that the likelihoods of particular reanalyses are enhanced by distributional similarity. Thus the data argue

for rather than against the claim that reanalysis is significantly constrained by syntactic structure.

2.2.3 Semantic Conditioning

But is syntactic structure, as it is currently portrayed, capable of serving as the basis for a maximally constrained theory of reanalysis? Without making the absurd claim that evolutive change is deterministic (which would be tantamount to assuming that dialectal-divergence is never evolutive) we can put even stronger constraints on reanalysis if we are willing to examine *semantic* conditioning factors.

For example, we have seen evidence that two distinct constructions of the form [NP V NP V NP] were reanalysed as [NP V NP P NP]: the first case involved a verb meaning something like 'meet; assemble' becoming a comitative preposition; the second case involved a verb meaning something like 'be at' becoming a locative preposition. Surely it is not coincidental that 'meet' became 'with' and 'be at' became 'at' rather than the alternative pairing. We might posit that reanalyses of this sort are constrained to preserve the thematic role assignments of the verb/adposition arguments.

Craig 1991's study of Rama (Chibchan) postpositions and complementizers provides additional evidence for thematic constraints on reanalysis. On the basis of their phonological and semantic similarities, she argues that a number of complementizers are historically derived from postpositions in contemporary Rama:

(24) a. **PostP:** naas sii   <u>ba</u>   aa   taak-iikar
         I   water   PostP(for)   NEG   go-want
         'I don't want to go for water.' [C 470]

     b. **Compl:** tiiskama ni-sung-<u>bang</u>   taak-i
           baby   1-see-SUB(to)   go-TNS
           'I am going to look at the baby.' [C 470]

(25) a. **PostP:** ipang <u>su</u>   an-siik-u
         island   PostP(on)   3PL-come-TNS
         'They came to the island.' [C 470]

b. **Compl:** nais   tum-ting-atkut-<u>su</u>   y-aakir-i
          right so   dark-happen-ASP-SUB(upon)   3-stay-TNS
          'Upon getting dark, he stays.' [C 470–1]

(26) a. **PostP:** nah altawa-i   naing taata <u>kang</u>
          I   afraid-TNS   my   father   PostP(from)
          'I am afraid of my father.' [C 470]

     b. **Compl:** nah kaafi ngu-atkut-<u>ka</u>   kalma   ni-sku-ut
           I   coffee   drink-ASP-SUB(when)   clothes   1–wash-TNS
           'When I have drunk up my coffee, I will do the wash.' [C 470]

She notes that her hypothesis receives independent support from Genetti 1986 (see also 1991) who finds evidence for very similar relationships between postpositions and subordinators in the unrelated Tibeto-Burman languages (Figure 2.8). The hypothesis that the three shared semantic transitions are coincidental seems much less plausible than the hypothesis that semantic information puts some constraints on reanalysis. And again it looks like some kind of thematic constancy is involved.

Figure 2.8: Parallel semantic developments in Rama and Tibeto-Burman languages (from Craig 1991—see also Genetti 1986)

Rama [Craig 1991]		Tibeto-Burman [Genetti 1986]	
Postpositions	Subordinators	Postpositions	Subordinators
bang, kama 'goal'	*bang, kama* 'purpose'	dative	purpose
		allative	purpose
ka(ng) 'ablative'	*ka* 'time, condition'	ablative	when/while/after, because, non-final
ka(ng) 'ablative'	*kata* 'counterfactual'		
su 'locative'	*su* 'time, after/upon'	locative	if/although, when/while/after
		ergative/instrumental	because/when/while

Of course thematic role constraints are now taken to be quite within the purview of syntactic theory, so the claim that uniquely semantic information is needed is perhaps not convincingly substantiated by these examples.

More compelling evidence comes from the history of English Degree Modifier

sort of and *kind of.* In Chapter 5, I provide evidence that the Degree Modifier usage (e.g. *It kind of withers.*) developed from the Noun-Preposition usage (*this kind of lupine*) in environments like (27) which are now ambiguous between the two interpretations.

(27) c. 1675 ... upon a <u>kind of</u> large Pin-cushion cover'd with a course and black woollen stuff. Boyle, *Electricity and Magnetism* p. 18

In this case, it seems especially significant that the division between the set of objects that we would be willing to call "genuine large pin-cushions" and those that we would classify as "nearly-large pin-cushions" is vague, or at least changeable from circumstance to circumstance, so it seems reasonable to suggest that semantic-equivalence or near-equivalence was a crucial conditioning factor in the reanalysis. Such an account makes a host of accurate predictions that a purely syntactic account cannot make, for there are many noun phrases of the form [Det [N [of [Adj N]]]] whose "N of" subsequences have not undergone reanalysis as degree modifiers and show no signs of doing so. For example, we predict that *slice of* in *a slice of ripe watermelon* is not likely to be reanalysed as a Degree Modifier because there is no particular similarity between slices of ripe watermelon and nearly ripe watermelons (or somewhat ripe watermelons, barely ripe watermelons, etc.). And in this case, the thematic role component of syntactic theory does not seem to offer much help.

The point can be made again with a rather different example: the development of epistemic uses of modal verbs from their root (or "deontic") uses (see Goossens 1982, Shepherd 1982, Traugott 1989, Sweetser 1990). For example the use of *must* to convey moral or social imperative (28) preceded and very probably led to the use of *must* to convey logical imperative (29).

(28) You must be home by 10. (Mom said so.)

(29) You must have been home last night.

Similarly, the use of *may* to convey permission (30) preceded and very probably led to the use of *may* to convey possibility (31). (31).

(30) John may go. (His visa is current.)

(31) That may be true.

Sweetser 1990, following Talmy 1982, notes that we can assimilate the earlier usage to the later one in each case if we employ an abstract representation involving forces and barriers. Thus *must*, in either sense, can be understood as referring to a "compelling force directing the subject towards an act" while *may* in both cases, conveys an "absent potential barrier". The change can thus be analyzed as metaphorical transfer from the socio-physical world to the world of reasoning. This analysis allows us to assimilate this case to numerous other cases of metaphorical transfer from the socio-physical domain to the world of reasoning (Lakoff and Johnson 1982, Sweetser 1990). It also predicts that we should never see a modal conveying obligation give rise to a modal conveying possibility and that we should never see a modal conveying permission give rise to a modal conveying necessity. As far as I know, this prediction is accurate (e.g., Shepherd 1982). Thus there appear, in this case as well, to be strong semantic constraints on which reanalyses can and cannot occur, and again, thematic role information does not seem particularly relevant.

These considerations seem to lead to the conclusion that what we need for the constraining of reanalysis is a combination of syntactic and semantic information. But the development of an explicit semantic theory of constraints on reanalysis promises to be a very difficult task. We must ask, for example, In what representation space do all the correspondences in Figure 2.9 obtain?

Figure 2.9: A sample of semantic transitions associated with processes of grammaticalization.

	Semantic Transition	Ref.	Example here
a.	to bring X and give to Y → to bring X for Y	Carlson 1991	[(17)]
b.	to do X accompanying Y → to do X with Y	Lord 1973	(13)–(14)
c.	to do X and be-at Y → to do X at Y	Lord 1973	(22)
d.	to be at place Y → to happen at time Y	Perez 1991	§5.2.2
e.	a type of X-ish Y → a somewhat X-ish Y	Tabor 1994	§5.2.1
f.	to do X, refraining from Y → to do X without Y	Carlson 1991	(20)–(21)
g.	to be obliged to Y → to necessarily Y	Sweetser 1990	(28)–(29)

(The example and section numbers refer to this thesis.)

Building such a representation on the basis of our introspective awareness

is difficult not only because of the subjective nature of some of the notions involved (benefiting, avoiding, obliging), but because we perceive the different cases in such different terms. There is an evident similarity between mapping (f) and mapping (g)—both cases involve a shift from a constraint induced by a human judgment (root meaning) to a constraint that is taken to be an objective property of the world (epistemic meaning). But there are all sorts of incidental properties of the relationship between the two cases that we have to factor out in order to realize this: in the source domains, (f) refers to two actions while (g) refers to only one; (f) implies a squelched desire while (g) suggests a squelched aversion; (f) *presupposes* a judgmental evaluation while (g) *refers to* one; change (f) involves a shift in syntactic subcategorization (from VP to NP) while change (g) involves no such shift. All of this filtering requires an extensive amount of "human pre-processing" which is hard to make explicit and must be performed separately for every case. For these reasons, it would seem desirable to choose a level of description at which the terms of comparison over elements are more easily defined. The distributions of vocabulary in corpuses may constitute such a level. It is particularly encouraging that all of the above-listed transitions appear to have been relatively gradual at the level of corpus metamorphosis, for this indicates that similarities in diachronically adjacent corpuses are very likely to coincide with the semantic similarities which are so tricky to define.

Moreover, although a very refined semantic theory might in principle be able to make maximally accurate predictions about reanalysis, our current semantic apparatus (even with its heavy reliance on human preprocessing) has trouble making needed distinctions. For example, if we claim that it was the semantics of the proto-Kwa verb 'meet' that made it suitable for reanalysis as a preposition meaning 'with', we must then explain why current Yoruba *bá*, which also means 'meet' shows all the signs of being a verb but none of the signs of being a preposition (Lord 1973: 291).

English Degree Modifier *sort/kind of* can also be cited in this regard. If the similarity between the meanings "a type of red brick" and "a somewhat-red brick" was what licensed the reanalysis of *sort/kind of*, then why hasn't *type of* itself been reanalysed (*They type of gave us a lesson in etiquette.*)? And in what sense is *a crude sort of prehensile thumb* different enough in meaning from *a crudely prehensile thumb* that *crude sort of* has not been reanalysed as

an adverb?

Or, if Marchand 1966 is right in asserting that the suffix *-ful* which derives measure units from container nouns in English came into being by reanalysis of phrases like *a spoon full of sugar*[11] we may ask why the same thing has not happened to *a cup packed with sugar* or *a bag stuffed with leaves*. One might suppose that the presence of productive morphology on *packed* and *stuffed* makes it unsuitable for lexicalization (*cup-packed-s*, *bag-stuffed-s*), but such a constraint hasn't prevented lexicalization in *notwithstanding* (cf. *notwithstood*) and *used to* (pronounced [justə] when used as an auxiliary). Nor does this explanation account for the fact that *a squirrel dead on the road* has not been reanalyzed as parallel with *a squirrel-corpse on the road* or that *Receive this gift free with your next purchase.* has not been reanalyzed as parallel with *Receive this free-gift with your next purchase* (*squirrel-deads*, *gift-frees*).

A final example of failed grammaticalization where semantic differences are hard to tie down is the case of *be going to* in English which has developed from a motion verb into a future marker. Here we may ask why *be walking to, be travelling to, be heading to*, etc. have not (yet) gone the same route (e.g., *It is going to rain* but *It is heading to rain*) (see Chapter 5, Section 2.2).

2.2.4 Distributional Conditioning

It is, in fact, common in languages in which grammaticalization processes have taken place for the grammaticalized element to exist side-by-side with a more contentful word whose meaning resembles the meaning the grammaticalized element had in its "youth". Such situations are part of the common phenonmenon of *layering* whereby several elements go sequentially down the same grammaticalization path in a single language over a period of time (see Schwegler 1983, Hopper 1991, Hopper and Traugott 1993). Moreover, it has often been observed that grammaticalization is accompanied by "abstraction" of the meaning of the

---

[11] That *spoonful* is a noun and not an ellipsed form of *spoon full of X* is evidenced by the fact that it takes plural marking (*spoonfuls*).

grammaticalizing element.[12] These observations suggests a way in which semantic theory might be refined to make appropriate predictions about the cases just mentioned. Presumably what we want to say about cases like Yoruba *bá*, English *type of, cup packed, walking to* and the like is that their meanings are not sufficiently abstracted yet to permit their reclassification. Unfortunately, it seems to be just about as hard to make robust judgments about relative degrees of abstraction as it is to make robust judgments about truth-conditional near-equivalence, so although this observation gives some insight into what the essential conditions for reanalysis are, it is hard to use it as a basis for formalization.

But here again there is a correlate at the level of corpus structure that we can take advantage of. As Bybee and Pagliuca 1985 point out, there is often a close relationship between generality of meaning and generality of distributional behavior:

> A more general morpheme has a more general distribution since it can be used in more contexts, and on the other hand, it is more general in that it lacks certain specific features of meaning. [p. 63]

The term "more contexts" is a quantitative evaluation. It is a property that can be measured in corpuses or grammars. Moreover, it need not be treated as a simple, one-dimensional property for we can ask in each case, Which additional contexts?. If we treat each context as a unique dimension, then there is a very rich range of relationships that particular corpuses (and their generating grammars) can bear to one another.

Although the grammaticalization literature has not tended to focus on quantitative properties of change, a number of studies report quantitative correlates of semantic and categorical changes. I'll remark on a few of them here in order to make it plausible that a useful kind of information is available, and can well be used as the basis for a theory.

Andersen 1987 examines the evolution of an Old Polish copula into a set of person/number/past-tense markers in the modern language. He notes that a set

---

[12]Traugott 1989 and Forthcoming argues that it is inaccurate to describe the changes as merely involving semantic information loss, for loss of content-word information is invariably accompanied by increased saliency of pragmatic inferences. She calls this process "pragmatic strengthening".

of independent copular forms alternated in Old Polish with a set of clitic copulas which had a strong tendency to occur in the second (Wackernagel) position typical of many clitics. The independent copular forms dropped out of use early on (1500s) and the clitic forms proceeded to become increasingly strongly associated with the verb. This increasing bondedness took several forms: an increasing tendency for the reduced copulas to occur immediately following the verb, regardless of whether this meant the copulas were in the Wackernagel position (Figure 2.10); a change in the interaction of the post-verbal copulas with a penultimate stress rule: singular copulas were counted as part of the verb from the standpoint of the stress rule from the 1500s onward, but it was only in the 1700s (as evidenced by changes in poetic rhyming practice and the comments of grammarians) that plural forms began counting, sometimes, for the stress rule as well. That this development is indeed a morphologically-conditioned development and hence involves a morphosyntactic reanalysis rather than a phonologically-conditioned change is evidenced by the fact that the reduced copulas only affect stress when attached to verbs and not when attached to other kinds of words, like adverbs (32) [See Andersen 1987, pp. 24–33, discussion in Hopper and Traugott 1993, pp. 136–7.]

(32) a. Wcz'oraj-em   prz'yszed-l
         yesterday-1stSgPast arrive
         'I arrived yesterday.'

     b. Wcz'oraj przysz'edl-em
         yesterday arrive-1stSgPast
         'I arrived yesterday.'

What these observations indicate is that there was a correlation between a quantitative development (the rise in frequency of adjacency between the verb and reduced copula) and a categorical innovation (the change in the interaction of the reduced copulas with the stress rules). Such observations suggest a representation, like the Restrictive Continuity representation discussed in the next chapter, in which quantitative change can lead to categorical change.

Craig 1991 presents some quantitative data on the distribution of certain verb-noun relational markers in Rama, which are related to the clause-markers

Figure 2.10: Increasing Cohesion between Verb and Reduced Copula during the History of Polish (from Rittel 1975: 91 via Andersen 1987, p. 29).

	Total Examples	Deviations from Wackernagel's Rule		Agglutination to Verb	
1500s	580	12	2%	130	23%
1600s	1303	64	4%	649	49%
1700s	1439	62	4%	994	68%
1800s	1988	308	15%	1395	80%
1900s	3325	503	15%	2817	84%
expository prose	525			525	92%

discussed in Section 2.2.3 above. Though purely synchronic, her data constitute another case of the lobed frequency contrasts which indicate a metrically-structured representation (compare Section 2.1.1 above).

Craig argues that certain cognate relational markers in Rama appear in three distinct syntactic environments: they appear as independent postpositional elements with an adjacent overt NP (33); they appear as "clitic preverbs" without an overt NP present (34); and they appear as "lexical preverbs" with an overt NP present (35).

(33) Nsu-suluk u   angka nsu-uung-i
      our-finger PSP/with can't   1Pl-make-SUB
      'With our fingers we can't do it.' [C 465]

(34) ungi yaadar tkua yu=nsu-uung-kama
      pot   thing   hot   CliticPV/with=1Pl-make-SUB
      'For us to do hot things in the pot with (it).' [C 465]

(35) ngulkang banku yu-an-siik-u   kaing
      wild pig now   LexPV/with-3Pl-come-Tns Disc.
      'They brought the wild pig now.'

The second two types are distinguished from the first on two counts: (a) they are marked with a different form (*yu* versus *u* in the examples shown); (b) while

the postpositional type need not occur immediately to the left of the verb, the preverb types can only appear in that position. The cliticized preverbs are distinguished from the lexicalized preverbs by the fact that they occur without an overt NP object present, that they occur with a wide variety of verbs (i.e. productively), and they have compositional semantics. The lexicalized preverbs, on the other hand, require the presence of an object, are more limited in the range of verbs they occur with, and tend to have idiosyncratic semantics.

Given these bases for distinguishing the types, Craig draws up the table shown in Figure 2.11, using a set of narrative texts in Rama. Although these data are based on a set of simultaneously-existing forms, it is reasonable to suppose that the synchronic forms range across a set of behaviors that constitute different points along a diachronic path along which they are travelling: such *layering* of forms at different stages is normal in grammaticalization (Givón 1984, Hopper 1991, Hopper and Traugott 1993: 124-6) and the trend from adposition to inflection is also common (see Hopper and Traugott 1993: 106–8). A graph of Craig's data under this interpretation appears in Figure 2.12.[13] Note again the series-of-lobes character of the data. These data too, then, suggest a continuous representation: the frequencies of certain types (postpositions and clitics; clitics and lexical preverbs) tend to be positively correlated, as though they are nearby in the representation space, while the frequencies of other types (postpositions and lexical preverbs) tend to be negatively correlated, as though they are representationally distant. Moreover, the "nearness" relationships (essentially: lexical-item is next to clitic which is next to affix) are ones for which even current, non-metrically-structured grammatical theories often introduce rudimentary devices for encoding intermediacy (e.g., Inkelas 1989).[14]

---

[13] The different markers are evenly spaced along the horizontal axis for lack of a better assumption about how to interpret them as corresponding to points in time.

[14] There is a question in this case, as to whether a distributionally-based model of structure could detect the appropriate similarity relationships. From the schematics given in Figure 2.11, one might think that the postpositions are more distributionally similar to the lexical preverbs than they are to the clitic preverbs, since the former two types both cooccur with an overt NP, and the latter does not. However, it appears from Craig's presentation that the postpositions need to occur immediately adjacent to their objects while the lexical preverbs may be separated from the NP by other words (e.g. (35)). Moreover, the postpositions and clitic preverbs both have compositional and nearly identical semantics. This makes it probable that they share a great many cooccurrence behaviors, that is, behaviors of cooccurrence with particular verb types and (ellipsed) object types. Under an appropriately-designed representation system, these correlations may well overshadow the factor of cooccurrence with the NP object present.

Figure 2.11: Text counts of Rama postpositions, clitic preverbs, and lexical preverbs. (from Craig 1991: 463)

	[NP PSP] (Postposition)	0 [PV-Verb] (Clitic Preverb)	NP [PV-verb] (Lexical Preverb)
ba(ng)- PURPOSE	0% (0)	39% (13)	61% (20)
u/yu- ASSOC/INST	31% (35)	51% (57)	18% (20)
a(ak)ya- DATIVE	49% (21)	39% (17)	12% (5)
su/su- LOC	89% (34)	11% (4)	0% (0)
ki/ki- LOC	98% (90)	2% (2)	0% (0)

Figure 2.12: Lobed Rama relational-marker data.

Marker [1 = ki/ki, 2 = su/su, 3 = a(ak)ya, 4 = u/yu, 5 = ba(ng)]
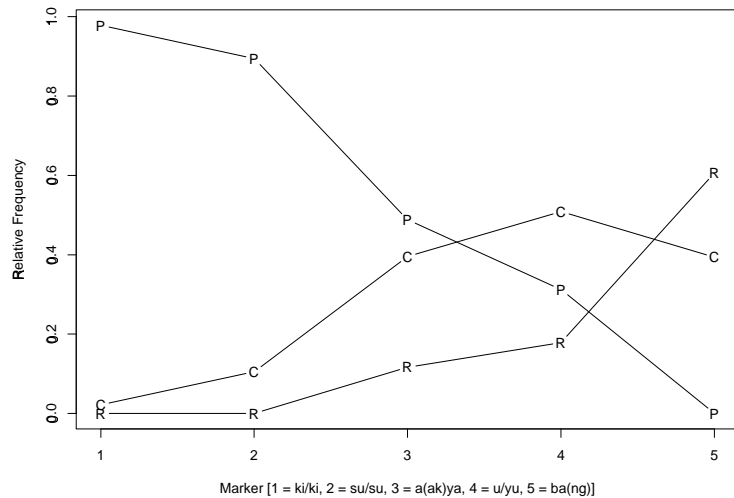
Thus several examples of quantitative data collected to adumbrate grammaticalization studies indicate the relevance of quantitative distributional information to the prediction of reanalyses. In the next chapter, I describe a

Connectionist model which has both the sensitivity to quantitative contrast and ability to make appropriate qualitative distinctions that this chapter has argued we need.

Chapter 3

A Connectionist Model of Grammar Change

Figure 3.1: Nodes and Connections.

3.1   Connectionist Networks

Inspired by biological neurons and synapses, Connectionist networks are clusters of nodes with directed connections between them (Figure 3.1). Associated with each node is a number called its *activation*. Associated with each connection is a number called its *weight*. Information about the degree of activation of each node is passed out along the connections issuing from it, with the connection weights scaling the strength of the signal. At successive time steps nodes revise their activations according to an *activation rule* that often says something like: "Take a high activation value if the sum of your scaled inputs is high; take a low value if that sum is low." The sigmoid activation rule shown in Figure 3.2 works well in many problems and is used in the simulations described in this thesis, except where noted otherwise. Appropriately designed networks have the appealing property that for any set of starting activations, repeated application of the activation rule causes the network to *converge* on a state in which the activations are stable.

3.2   Feedforward Networks

A simple case of this kind is the *feedforward network* (Figure 3.3). In a feedforward network, some subset of nodes is designated as input, another subset as output, and no path of successive directed connections loops back on itself. Usually, input units are units whose values are determined by the environment in which the network exists (hence they correspond to perceptual organs in the biological case). Output units are units whose values are transmitted to the environment (hence they correspond to motor neurons in the biological case). For behaviors of living organisms, we tend not to think of inputs and outputs as numerical vectors but rather as various categorically-defined behaviors (e.g., perception of the color red in the upper left region of the visual field, production of the phoneme [b]). But we can easily convert such behaviors to vectors by letting the elements of the vector correspond to particular features of the environment (e.g., production of a voiced phoneme) and letting the magnitude

Figure 3.2: Net Input and Sigmoid Activation Rule

The net input to unit i is given by

$$(36) \quad net_i = \sum_j a_j w_{ij} + \beta_i$$

where a_j is the activation of unit j, w_{ij} is the weight from unit j to unit i (0 if the units are not connected), and β_i is a bias term which is treated like a weight coming from a unit that is always on.

The activation of unit i is then given by the sigmoid activation rule:

(37)

$$a_i = f(net_i) = \frac{1}{1 + e^{-net_i}}$$

A graph of the sigmoid is given below.
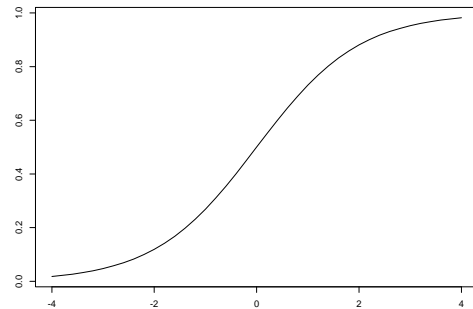
of the activation of each element correspond to an estimate of the probability of the presence of the feature.

In the case in which all nodes are either inputs or outputs, it is convenient to organize the weights into a matrix, placing the weight from unit j to unit i in cell ij of the matrix. Similarly, if the connections in a feedforward network are nicely organized into successive layers, with the nodes on one layer connecting

only to nodes on the next, then the weights can be specified as a series of matrices, one corresponding to each pair of successive layers. Usually the first layer is interpreted as the input layer and the last as the output layer, and the intervening layers are called "hidden" layers because the values on their units are not directly manipulable by the environment. A simple and useful case is the 3-layer network, with one input layer, one hidden layer, and one output layer (Figure 3.3).
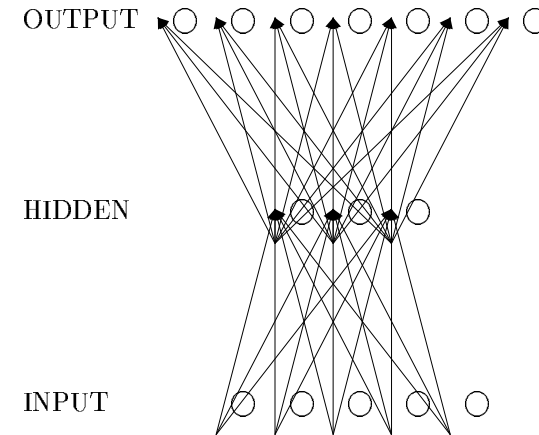
Figure 3.3: A 3-layer, feedforward net.

It turns out that if the behavior one is studying can be interpreted as a function (one output corresponding to each input) and one has a representative sample of the inputs and outputs of the function, then there are algorithms which can usually get the network to approximate the function by exposing it to a sample of input/output pairs. This process is called *training* or *learning*. The learning algorithms often work by letting the network begin in a state in which it embodies an arbitrary hypothesis about the relationship of inputs and outputs and then, on the basis of the examples, incrementally adjusting the weights on the connections in such a way that the function the network actually embodies converges on one that fits the observed mapping. An advantage of the nets with hidden units is that the algorithms tend to make use of the hidden units to form representations that capture important generalizations about the mapping. In this sense, the representations adopted in the hidden layer are

analogous to the representations that linguists posit in order to give efficient descriptions of languages. Many of the problems formal linguists address can be conceived of as problems of choosing a compact representation such that, for each linguistic item, its representational instantiation is predictable from observable facts about it, and moreover, can be used to predict other facts about it (e.g., from information about the meaning of a verb, determine its argument-structure, and from its argument structure, predict which sequences of words, with which associated meanings, it will occur in; or, from information about the identity of a word, choose its underlying form, and based on this form, predict how it will be pronounced in various contexts).

3.2.1 Learning by Backpropagation of Error

The learning algorithm I use here, "backpropagation", is a gradient-descent process. To understand gradient descent, imagine being blind, out of water, and standing on top of a ridge. How could you quench your thirst? Not being able to see, you could not plan a route to a creek or lake in a valley bottom. But if you felt the shape of the ground around you, you could find a way to move a little bit in a direction that would take you downhill, and if you repeated this process over and over again, you would probably end up near water. One way of describing the style of this solution is to note that the land altitude at any point on earth defines an *error function* with the following properties: (1) if, for some pair of coordinates the error-function is at a minimal value then water is (almost) certain to be present; (2) if the error function is at a non-minimal value, then water is not so likely to be present. The error-function is continuous so you can always get from one place to another by taking minimal steps. Therefore, continually moving a little bit downhill is a good way of finding water.

We can also define an error function for the function-induction problem. Suppose we are trying to induce the function $g : \mathcal{I} \to \mathcal{O}$, where $\mathcal{I}$ and $\mathcal{O}$ may be spaces of several dimensions. For any function $\hat{g} : \mathcal{I} \to \mathcal{O}$, let the error of $\hat{g}$ on input $x \in I$ be given by:

(38)

$$E = (g(\vec{x}) - \hat{g}(\vec{x})) \cdot (g(\vec{x}) - \hat{g}(\vec{x}))$$

$$= \sum_i (o_i - t_i)^2$$

where o_i is the value of the ith output dimension of $\hat{g}$ and t_i is the value of the ith output dimension of g. For L a list of elements from $\mathcal{I}$, let the error of $\hat{g}$ on L be the sum of the errors of $\hat{g}$ on each element of L:

(39)

$$E_L = \sum_{x \in L} (g(\vec{x}) - \hat{g}(\vec{x})) \cdot (g(\vec{x}) - \hat{g}(\vec{x}))$$

$$= \sum_{x \in L} \sum_i (o_i - t_i)^2$$

This is called the *sum of squares* error function. It has the property that if L includes every element in $\mathcal{I}$ then it is zero if and only if $g = \hat{g}$. If L includes only a representative sample of I, a low value for E may still mean that $\hat{g}$ is close to g. It is in this case that the function induction algorithm is really useful because it constitutes a kind of interpolation mechanism: having observed only some of the behaviors of g, we can make good guesses about other behaviors of g by finding a $\hat{g}$ with low error and examining its behavior.

For the layered, feedforward nets described above, the function $\hat{g}$ is given by

(40)

$$\hat{g}(\vec{x}) = f(W_n \ldots f(W_2 f(W_1 \vec{x})) \ldots)$$

where W_i is the weight matrix from layer i to layer $i+1$ and f is the multidimensional extension of the activation function, (37) (i.e., $f(\vec{x}) = (f(x_1), f(x_2), \ldots)$).

For a given network architecture, the different functions it can embody (i.e., the different $\hat{g}$'s) are determined by the different possible values of its weights. Thus, to lower the error associated with the network, we should perform gradient

descent in the space of weight-settings (or "weight space"). Equations (36), (37), (39), and (40) imply that E has continuous partial derivatives with respect to the weights at every point in weight space, so we can compute a total derivative which points in the direction of steepest descent and use it to choose a descending path. This is what the *Backpropagation Algorithm* does (Rumelhart, Hinton, and Williams 1986):

Backpropagation Algorithm. Given sample L from function g, adjust the weight vector by adding vector $\Delta \vec{w}$, given by:

(41)

$$\Delta w_{ij} = \sum_p \epsilon \, \delta_{pi} a_{pj}$$

where p indexes elements of L, Δw_{ij} is the weight change on the weight from unit j to unit i, ϵ is a small positive constant called the *learning rate*, and a_{pj} is the activation of unit j when element p is the network input. For output units, δ_{pi} is given by

(42)

$$\delta_{pi} = f'(net_{pi})(t_{pi} - o_{pi})$$

For hidden units, δ_{pi} is:

(43)

$$\delta_{pi} = f'(net_{pi}) \sum_k \delta_{pk} w_{ik}$$

where k indexes units that the hidden units project to, and f' is the derivative of f.

The recursion in the definition of the δ's implies that the δ's for higher layers must be computed prior to the δ's for lower layers. Since the network is feedforward, the recursion always bottoms out. The algorithm is called *backpropagation of error* because this recursion process is a way of taking an error-signal at the

output level and working systematically backward through the net to figure out to what degree each weight is responsible for that error, and consequently how to change it to get an improvement.

The algorithm is guaranteed to find a minimum of the error function if the learning rate, ϵ, is small enough, and the list of example inputs and outputs is presented to the network often enough.[1] For some kinds of problems, we may not wish to present the patterns over and over again but rather just present them once. For example, if we want to treat a language-learning network as a model of a language-learning child, it is not very plausible to present one batch of patterns over and over again. Fortunately, the network can often find a path to a minimum if weight-update is performed after presentation of every example, p. All the simulations described here were done with pattern-by-pattern update. In this case, we can eliminate the summation symbol from equation (41).

3.2.2 A Syntax Example

To see how a 3-layer network trained with backpropagation makes use of its hidden unit layer to create a compact representation, it is useful to consider an example. It is generally acknowledged that Nouns and Determiners constitute two distinct categories in many natural languages. Moreover some languages, English among them, encode a contrast between Singularity and Plurality in both their Nouns and Determiners. These distinctions reflect a number of systematic contrasts in the behaviors of the elements in question. For example, in English, nouns follow adjective modifiers and determiners, they conjoin primarily with nouns, if they are singular they combine only with singular determiners, if they are subjects and singular, they take only singular verbs. Determiners, by contrast, precede adjectives and nouns, they conjoin only with other determiners, etc.[2]

---

[1] In fact, there is a small possibility that the network wil converge on a saddle-point of the error function. This is a very rare occurrence. However, it is very possible that the minimum found by the network will not be the global minimum of the error function, even when there is a $\hat{g}$ that matches g at every point. Such cases have not proved to be a serious problem in the simulations I consider here.

[2] There is, of course, a certain circularity involved in referring to the classes we are motivating when we enumerate the contrasts that constitute the motivation. But this circularity does not create an inconsistency (it only makes the task of discovering the membership of the

Suppose we wanted a network to use the behavioral information I have just described to set up a representation that efficiently captures the major grammatical contrasts involved. We can formulate the problem as a function-induction problem by choosing, arbitrarily, a representation for each word we are interested in and calling this its "input" representation. We can then define the "output" representation for each word as being a vector of features identifying behaviors that the word exhibits in grammatical constructions at least some of the time.[3] To make the representations numeric, we can let the number 0 stand for presence of a feature and 1 stand for its absence. Thus for the outputs, I've recorded, in Figure 3.4, a 1 for every feature that sometimes occurs in conjunction with a given word and a 0 for every feature that never occurs. For the input representations, I've coded each word as a vector with a value of 1 on one dimension and 0s on all others. Thus we'll need (at least) V input dimensions, where V is the number of words in the vocabulary (Figure 3.5). This input-coding scheme has the property that all inputs are orthogonal and equally distant from one another. Since the network is fundamentally attentive to distance-relationships, using these orthogonal, equi-distant input representations is a good way of avoiding encoding any prejudices we might have ahead of time about how the individual words are related to one another.[4]

I trained a 3-layer network with a two-unit hidden layer (Figure 3.6) by repeated presentation of these input-output pairs. I started the network in a state in which all of its weights had random values close to zero. Consequently, by Equation (37), the hidden unit activations were all fairly close to 0.50. Figure 3.7 shows a graph of the activation of Hidden Unit 1 versus the activation of Hidden Unit 2 at the start of training for each of the words in the vocabulary. We can think of this graph as a picture of the space in which the network is

Figure 3.4: Output Representations for a Simple Syntax Problem

Word	Output Representation												
	A	B	C	D	E	F	G	H	I	J	K	L	M
person	1	1	1	1	0	0	0	1	0	0	0	1	0
people	1	1	1	1	0	0	0	0	1	0	0	0	1
house	1	1	1	1	0	0	0	1	0	0	0	1	0
houses	1	1	1	1	0	0	0	0	1	0	0	0	1
box	1	1	1	1	0	0	0	1	0	0	0	1	0
boxes	1	1	1	1	0	0	0	0	1	0	0	0	1
joy	1	1	1	1	0	0	0	1	0	0	0	1	0
joys	1	1	1	1	0	0	0	0	1	0	0	0	1
load	1	1	1	1	0	0	0	1	0	0	0	1	0
loads	1	1	1	1	0	0	0	0	1	0	0	0	1
bird	1	1	1	1	0	0	0	1	0	0	0	1	0
birds	1	1	1	1	0	0	0	0	1	0	0	0	1
tree	1	1	1	1	0	0	0	1	0	0	0	1	0
trees	1	1	1	1	0	0	0	0	1	0	0	0	1
fish	1	1	1	1	0	0	0	1	1	0	0	1	1
sheep	1	1	1	1	0	0	0	1	1	0	0	1	1
this	0	0	0	1	1	1	1	0	0	1	0	1	0
these	0	0	0	1	1	1	1	0	0	0	1	0	1
the	0	0	0	0	1	1	0	0	0	1	1	1	1
some	0	0	0	1	1	1	1	0	0	1	1	1	1
each	0	0	0	1	1	1	1	0	0	1	0	1	0
one	0	0	0	1	1	1	1	0	0	1	0	1	0
a	0	0	0	0	1	1	0	0	0	1	0	1	0
lots-of	1	0	0	0	1	1	1	0	0	0	1	0	1
every	0	0	0	0	1	1	1	0	0	1	0	1	0
many	0	0	0	1	1	1	1	0	0	0	1	0	1

Key:
A.   Follow Adjective modifiers
B.   Follow Determiner modifiers
C.   Conjoin with Nouns
D.   Precede Prepositional Phrase complements
E.   Precede Adjective modifiers
F.   Precede head Nouns
G.   Conjoin with Determiners
H.   Take Singular Determiner modi
I.   Take Plural Determiner modifie
J.   Modify Singular Nouns
K.   Modify Plural Nouns
L.   Agree with Singular Verbs
M.   Agree with Plural Verbs

constrained to form its representation. It is a bounded, two-dimensional region.[5]

---

classes in an unknown language a bit more difficult), for we can simply take "Determiner" and "Noun" in the motivation-statements to be short-hand for the sets of words that ultimately receive these classifications. These sets can be defined without making reference to any abstract structure beyond the level of the word.

[3] I come to the matter of taking into account the rate at which each behavior occurs in Section 5.

[4] Of course, such picayune orthogonality is very memory-wasteful, so it is only practical in small problems. A natural approach to larger problems is to use small, partially-adequate sets of information to bootstrap the input-coding assignments and then let the network modify them in the process of fitting the large data-set.

[5] The space is bounded because the sigmoid function limits the activation values to the

Figure 3.5: Input Representations for a Simple Syntax Problem

Word	Input Representation
person	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
people	0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
house	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
houses	0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
box	0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
boxes	0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
joy	0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
joys	0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
load	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
loads	0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
bird	0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
birds	0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
tree	0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
trees	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
fish	0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
sheep	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
this	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
these	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
the	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
some	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
each	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
one	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
a	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
lots-of	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
every	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
many	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1

The network will form a representation in this region by moving inputs with similar outputs to nearby regions in the hidden unit space and inputs with far-apart outputs to distant regions in the hidden unit space. To see this, note that the weight update-rule (Equation (41)), which was applied every time an input pattern was presented to the network, mandates that weights coming from the

range (0,1) for each unit.

Figure 3.6: Network for Simple Syntax Problem

Figure 3.7: Hidden Unit Locations at the Beginning of Training.

input layer to a particular hidden unit should be increased or decreased if doing

one of these things will make the error lower on the current pattern. Since the input from only one word can be on at a given time, the weight adjustments in the input→hidden block always have the effect of moving the hidden-unit representation for exactly one word in one direction or another.

Moreover, note that equation (41) implies, in this network, that the weight change from input unit i to hidden unit h is always
(44)

$$\Delta w_{hi} = \epsilon \, f'(w_{hi}) \sum_j (t_{ij} - o_{ij}) w_{jh}$$

$$= \epsilon \, f'(w_{hi})(\vec{t_i} - \vec{o_i}) \cdot \vec{w}_{\_h}$$

where j indexes output units, $\vec{t_i}$ is the target when unit i is the sole "on" unit at the input layer, $\vec{o_i}$ is the actually produced output under the same circumstances, and $\vec{w}_{\_h}$ is the vector of weights projecting from hidden unit h to the output layer. If we compare the inputs for two words, i and i', we can note that at the beginning of training, all weights in the network are approximately equal so $w_{hi} \approx w_{hi'}$ and $\vec{o_i} \approx \vec{o_{i'}}$. Thus the only term that can make a substantial difference in the direction of weight change between the two words is the target, $\vec{t}$. Therefore, words with similar targets will change input→hidden weights in approximately similar ways and words with different targets will change differently. Moreover, since the activation function is monotonic, the activation change at the hidden layer always points in roughly the same direction as the weight change.[6] This means that at the beginning of training, words in the same classes will move in essentially the same directions.

And since these conditions obtain initially, they continue to obtain as learning proceeds, as long as the distances between outputs and targets remain large compared to the distances between words with similar targets. The result is that during the course of training, the network plows all the hidden unit locations for words belonging to a single class into the same region of hidden-unit

space and it plows the locations for different classes into different regions (see Figure 3.8).

Figure 3.8: Trajectories of the word-representations in hidden unit space during training. (The representations start in the middle of the space and move toward the periphery as training proceeds.)
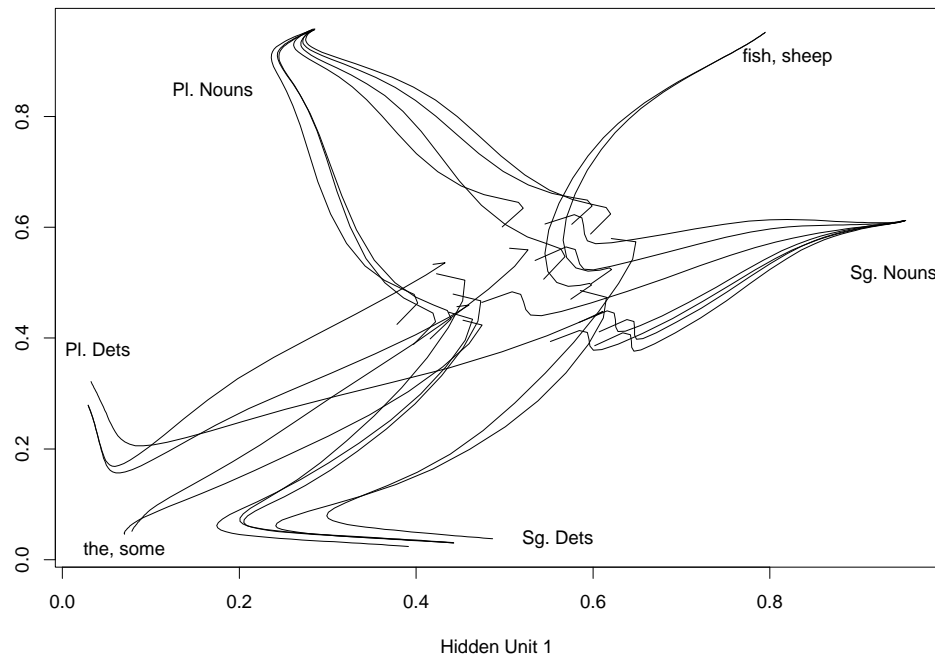
Moreover, these regions are different quadrants of a rotated space of the same dimensionality as the hidden-unit space. If we draw in the axes for this new space,[7] it becomes clear that the representation the network has adopted is essentially a parametric one along the lines of what linguists normally propose in such a case: one parameter encodes the distinction between Nouns and Determiners, another one codes the distinction between Singular and Plural.

---

[6] In fact, near the beginning of training, the weights are near 0 so the activation function is close to linear, so the direction of the activation change is essentially identical to the direction of the weight change.

[7] We can do this in a systematic way by computing the directions of the first and second principle components of the distribution of hidden unit locations at the end of training.

3.3 Recurrent networks

One property of the model developed in the previous section seems less-than-desirable: the ambiguous elements, <sheep>, <fish>, <the>, and <some>, are assigned intermediate representations. This means we cannot consistently use the network's word-representations to recover information about the contexts in which the words were used. We also run a danger of assigning two words the same representation even when they occur in contrasting environments. This means we may not reliably be able to recover behaviors from representations (e.g., if one word is ambiguous between being a singular noun and a plural determiner (e.g., *awl/all*), while another is ambiguous between being a plural noun and a singular determiner). These observations indicate that we would do well to build context into the representation somehow.

We might try using the same tack with the contexts that we used with the lexical items in the first place: assign orthogonal vectors to each context and let the network discover hidden-unit representations that are useful in predicting the behaviors. A network architecture for this approach is shown in Figure 3.9. But what do we mean by "context"? In the general case, "context" seems

Figure 3.9: A network architecture with an input vector for encoding word-identity and another one for coding context-identity.
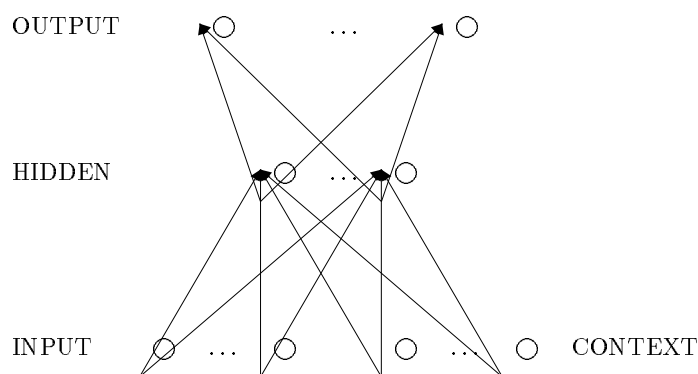
to be a large batch of information about what was happening in the world at the point at which a word was spoken along with some information about what words cooccurred with it and in what order. But to represent all this

information with orthogonal input vectors would require such a huge input space that no existing computer could perform an experiment with the network. Even if we were to restrict our attention to linguistic context, perhaps keeping the context associated with each word relatively small by taking into account a few surrounding words, the number of input units required for any real problem would be huge. Moreover, the network would have no ability to generalize to new contexts on the basis of contexts it had seen before. In the case of the orthogonal word-codings this lack of generalization ability does not eliminate the interest of the network as a historical model since there are many interesting changes other than word-coinages. But in syntactic and morphological change, almost *everything* of interest involves some kind of a change of context. For these reasons the orthogonal context approach is rather unappealing.

But there is another, much less costly way of letting the network induce contextual representations, one which, moreover, permits generalization across contexts. We can let the hidden unit representation code not merely a word but a *word in context*. A good way of doing this (suggested by Elman 1990 and 1991) is to feed words to a network in the order in which they occur in real language, and to train the network on the task of predicting, for each word when it appears on the input layer, what word is coming next. If the hidden units are to encode information about what words have come previously at any point, they must have access to such information. One way of giving them such access is to let them receive activation from each other (and from themselves) every time a word is presented on the input layer. This way, the hidden unit state at time step t can serve as a representation of prior-context for time-step t+1 (Figure 3.10). The connections that loop back on each other are called *recurrent connections* and the resulting network is called a *recurrent network*.

Backpropagation can be used to train recurrent networks as well as feedforward networks. This is because for any recurrent network, there is a corresponding feedforward network which can be derived by "unfolding the recurrent network in time" (see Rumelhart, Hinton, and Williams 1986). This method can be fairly costly for large training sets because at each time-step, weight-adjustments have to be calculated for every instantiation of the recurrent net at every previous time step, so the total number of weight adjustments needed grows with the square of the number of exemplars (rather than linearly, as in the

Figure 3.10: A network with recurrent connections in the hidden layer

OUTPUT

HIDDEN

INPUT
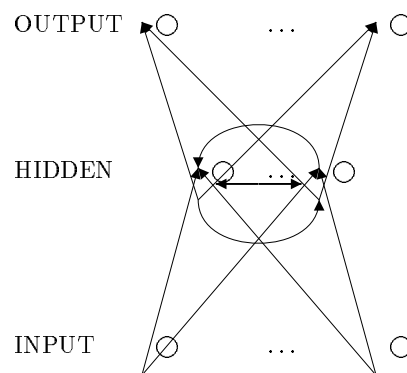
simple feedforward case). But it works well in many cases to use a "truncation of the gradient", taking into account only recent transitions in estimating the path of steepest descent.

3.4 Elman 1990's recurrent network for discovering grammatical structure

Elman trained a recurrent net with a truncated gradient on the task of predicting next-words in corpuses. The architecture he used is called a Simple Recurrent Network. I introduced it in Chapter 1, Section 7. The diagram of its archicture is repeated here in Figure 3.11. Elman trained the net with simple, artificial corpuses intended to reflect certain properties of the grammar of English.

As I noted in Chapter 1, Elman 1990 found that a Simple Recurrent Network trained to do word-prediction on a set of sequences generated by a simple template-grammar can learn to predict the sequences accurately.

In the template-grammar experiment, Elman noted something of interest about the trained network's structure. The hidden unit layer can be thought of as a sort of parameter tool-shop which the network is free to put to use in ways that help it solve the word-prediction task. Once the network was trained, Elman studied the hidden unit activation patterns associated with particular inputs using the technique of *hierarchical clustering*. Hierarchical clustering is

Figure 3.11: Elman 1990 and 1991's Simple Recurrent Net.

OUTPUT_t

HIDDEN_t

INPUT_t HIDDEN_{t-1}
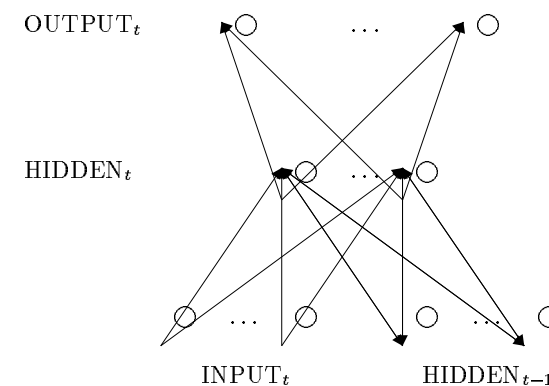
a method of probing the cluster structure of a set of points. It works as follows: a notion of distance between clusters of points is defined; initially, every point forms a separate cluster; then the (or some) pair of clusters closest together are joined to form a new cluster; the process is repeated until all the data have been joined into one big cluster.[8] A tree-like graphical representation of the structure discovered by the clustering algorithm is called a *dendogram*.

Elman was interested in the network's representation of the relationships among words. Consequently, he fed a sample of sentences to the trained network and recorded the hidden-unit locations associated with each input token in context. He then averaged all the hidden-unit locations associated with tokens of individual words and produced a hierarchical clustering of these averages. Figure 3.12 shows the result. The interesting property of this dendogram is that its hierarchical structure roughly reflects the kind of hierarchical structure that linguists have posited for the English lexicon. There is a fundamental division

---

[8] There are various common ways of measuring the distance between clusters: one can take the distance to be the *minimum* distance between points from each cluster; one can take it to be the *maximum* distance; one can take it to be the *mean*. Elman 1990 does not report which method he used, but I have found in my own experiments that using the *maximum* method, which tends to produce compact, relatively spherical clusters, works well in the sense that it produces linguistically sensible results. Therefore, the clustering results I report on elsewhere in this thesis use the *maximum* method (also sometimes called the *compact* method). Finch 1993 reports a number of plausible results using the *mean* method.

between Nouns and Verbs. Within the verbs, there is a division between transitives and intransitives. Within the nouns there is a division between animates and inanimates. This analysis reveals two things:

(i) Something like the abstract structure which linguists have found to be relevant for a large variety of language tasks is extractable from a sample of distributional behavior.[9]

(ii) The network encodes certain kinds of hierarchical distributional structure by means of proximity relationships in a continuous metric space (i.e. a set on which a real-valued distance-measure is defined).

Point (i) indicates that the network and its training regime can do a reasonable job at modelling those aspects of language structure that linguists have found to be especially relevant for making predictions about language behavior (both synchronic and diachronic). Although this point is only very weakly and impressionistically supported by Elman's results because the simplification involved in the approximation to natural language is so great, it seems encouraging that Finch 1993 has been able to produce similar clustering results using "real" (newspaper corpus) data. Point (ii) implies that there can be additional information in the network's representation that is absent under a purely categorical analysis: certain clusters can be relatively near other clusters even if they are not clustered with them except at the highest (trivial) level; these "extra-categorical" adjacencies give rise to predictions about the conditions under which reanalyses can occur. For one thing, quantitative distributional changes can move one element closer to a new category without initially affecting the hierarchical clustering. This is what gives rise to Q-divergence effects (See Chapter 5). Moreover, although I don't explore the matter in this thesis, the extra-categorical adjacencies give rise (under the language-evolution-as-learning paradigm) to predictions about which category changes can happen directly and which must go through intermediate categorical stages. This may be a fruitful area for future research.

---

[9]Elman reports that, in his experiment, even clustering the tokens of word-uses without first averaging across same-input instances produces a dendogram with the same initial structure as that shown in Figure 3.12. I have found that this property does *not* obtain in all cases: if the same word can be used in very distinct contexts, the lower-level clusters will be by context, not by input identity.

Figure 3.12: Hierarchical clustering of hidden unit locations for word-prediction experiment from Elman 1990.

Elman 1991 found that a similar network can encode information about subject-verb agreement and its interaction with relative-clause structure. In each case, the inputs and outputs to the network are bit vectors with only one bit "on". Each uniquely represents a word. As I also noted in Chapter 1, a recurrent network trained with Backpropagation on such data does not actually learn to make perfect predictions about successor words. Instead it learns to distribute activation on the output layer as a probability density function roughly

reflecting the transition-probabilities implicit in the grammar. This means that the activations on the output layer all lie between 0 and 1, they sum to 1, and each activation approximates the probability of encountering the word it corresponds to as the next word, given the context. Of particular relevance to the model of syntactic innovation proposed here is the fact that the network assigns *positive* activation to every transition in every context. Therefore, every possible sequence of the vocabulary is assigned some positive likelihood of occurrence. Consequently, the standard practice of mapping the contrast between grammatical and ungrammatical sentences into a contrast between generable and non-generable sequences will not suffice here. Instead, it makes sense to define grammaticality in terms of a likelihood threshold: sequences whose likelihood is above a certain positive value θ are considered grammatical. Those whose likelihood is below θ are ungrammatical. In Section 6.2, I explain how the likelihood threshold can be determined. The advantage of conceiving grammaticality in these terms rather than as a binary contrast between 0 probability and positive probability is that there can be correlated change in below-threshold and above-threshold likelihoods. Some of these trends, if they persist long enough will result in the likelihoods of particular constructions crossing the theshold. Thus they are useful in making predictions about changes in the categorical structure of the language. I give examples of such threshold-crossing events in Chapters 5 and 6.

3.5   The conceptual value of induced representations

For certain linguistic tasks, the network representations seem to make similar predictions to the linguistic theories without making exactly the same predictions. Although the network representations are relatively well-understood at the level of the algorithmic function-fitting process outlined above, they are not as well understood at the level of the types of entities that linguists commonly work with (e.g. "Noun", "Verb", "Phonological Rule", "Movement Rule", "Grammar Module" etc). Inasmuch as the networks seem to be able

to match the linguistic predictions on many accounts and also offer some improvements (they accomplish learning from positive-evidence without extensive pre-structuring [Elman 1990, 1991], they have suggested ways of uniting certain linguistic laws under a more general theory [e.g., Prince and Smolensky 1993, Dell *et al. 1993*, Hare *et. al* (in press)] a valuable research strategy would seem to be to take a network that is performing well on some linguistic task and to study its representation in the hope of gaining some insight about linguistic theory. It is this strategy that I have employed in this thesis.

3.6   A network-based model of language change

If we take a particular network architecture to be a model of Universal Grammar, then a language-state is a weight-setting (or position in weight-space) of that network architecture. We can thus model a language change episode as a series of ordered pairs of times and weight-settings.

Like many of the generative linguists who have studied change, my interest is in seeing what the study of change can tell us about the structure of language. I noted in Chapter 1, Section 2 that it is often the case with structured systems that we can learn about their structure by looking for correlations among aspects of them that are changing. If two distributionally similar words or phrases change simultaneously in the same way, we are justified in suspecting that they are linked in the grammar. If we see the same two types changing together in many diverse situations, and there are few or no exceptions to their correlatedness, then our suspicions may be confirmed. Or if we can observe that the two words/phrases that change have done so simultaneously after a long period of time when they did not change, then our suspicions may also be confirmed (e.g., when *sort of* became a Degree Modifier, *kind of* did so as well). The coincidence of their simultaneous change would be highly unlikely if they were not linked in the grammar.

Such conclusions are not nullified by the discovery that some functional pressure operates simultaneously on different parts of the language like the tines of a rake simultaneously moving several independent pebbles. For words and phrases are not inherently structured so as to be susceptible to deformation by the same forces unless it is because of the similarity of their sounds. But if the operation

of the functional pressure is conditioned purely by sound-properties, then it will correlate purely with sound properties and will not be able to select particular words and phrases which are arbitrarily sound-related for distinctive change. In fact, this kind of change, called *sound change*, is well studied, and is distinguishable on a number of empirical counts from morphological and syntactic change (see Kiparksy 1992). On the other hand, conclusions reached about the structural significance of diachronic correlations *will* be put in doubt if it can be shown that clearly unrelated functional pressures are responsible for the simultaneity.

Given these reflections, we should like our model of grammar to predict persistent correlations in the behaviors of distributionally similar but phonetically dissimilar elements.

In the next section, I motivate a model of diachronically-correlated behavior based on Connectionist learning.

3.6.1 Short-term change as Connectionist learning

A generative parametric grammar makes predictions about diachronic correlations because all elements subsumed under a single parameter setting are required to change in tandem. For example, Kroch 1989b's model predicts that when MainVerb→Infl movement is being lost in English during the 15th and 16th centuries, it must be lost simultaneously in all clauses; this predicts correlated behaviors of periphrastic *do* in a variety of syncatic environments (cf. Lightfoot 1991).

The correlation-predictions such models make stem purely from the the *structural range* implied by the theory. Nothing is said about how languages must transit from one state to another; only the set of states they can occupy is constrained.

The Connectionist model, because of the dimension-reduction constraints imposed by its hidden-layer representation, makes analogous structural-range predictions (see Section 8). But there is another way that structure may constrain change: the *transition mechanism* may be subject to some constraints that make it possible for only certain states to follow others. I find the "Rubik's Cube" a useful analogy here: the 8 corner-blocks and 8 edge-blocks of a Rubik's

Cube are under strong structural-range constraints in that they must always remain corner-blocks and edge-blocks respectively. However, within that general framework, they are capable of occupying almost arbitrary positional relationships relative to one another. Nevertheless, this does not mean that arbitrary transitions can occur: the structure of the cube requires that all change happen by successive rotations of cube-faces by 90 degrees. This constraint gives rise to certain persistently-observable diachronic correlations. For example, if two cubes on face F make a rotation R at a given time then all cubes on face F make rotation R at that time-step. By observing these diachronic correlations, we can learn something about the internal structure of the cube that we can't learn from the structural-range constraints on corner-blocks and edge-blocks alone.

I make a simple assumption about the nature of the transition mechanism in language change:

> **The Transition-by-Learning Assumption:** Structural properties of evolutive language change are accurately simulated by Connectionist learning.

This is a very weak assumption about the nature of the transition mechanism. It is equivalent to the claim that essentially arbitrary forces can impinge on the grammar, so long as they operate to change relationships among the behaviors as they are coded at the output.[10] The only significant constraint the assumption imposes is that the response to those forces be small at any time step.[11] The interest of the assumption is that it interacts with the structural constraints imposed by the representation system to make strong predictions about the paths of change that languages can follow.

---

[10] The fact that I am employing input and output codings that only make reference to word- and morpheme structure means that phonological and semantic entities are not available to be loci of forces. That phonological structure is left out is not a great cause for worry at present, since most of the phenomena I examine aren't plausibly strongly influenced by phonological factors. Moreover, it is not difficult to let phonological information come into play by structuring the input representations along phonological lines. That semantic structure is not explicitly coded is also, I believe, not a cause for worry because, as I argued in Chapter 1, corpuses seem to contain the information we call "semantic". This is an empirical claim and so must be evaluated in light of the empirical effectiveness of the model.

[11] This assumes that the learning rate is small. I am taking a small learning rate to be a necessary attribute of proper "Connectionist learning".

What are these predictions? Suppose we take a network trained on some data-set to be a model of the grammar of a language at some time t. We can then create a new data-set by altering the first data-set in some way. When we train the network on the new data-set, it simulates the effect of exerting some differentiating force on a part of the grammar during the time period immediately following t. Because of the structure of the representation system, the language may not be able to adjust only the part of the grammar that the force is impinging on. It must also adjust nearby parts (this follows from the constraint that the representation must be continuous and change must be gradual—see Section 8). Similarly, if we impose a force that operates on just one aspect of a word's behavior, then because of the restrictiveness of the representation, we predict that certain correlated changes will occur in other aspects of the word's behavior (Section 8). In this manner, we predict the local diachronic correlations which I am calling Frequency Linkage and Q-Divergence. Thus far, I have only been able to experiment with small artificial corpuses that are meant to capture a few significant properties of particular language states. Nevertheless, the predictions the network makes about these "toy" problems correspond appropriately to the observed facts of language chagne. Examples of the correlations predicted are the parallel rise of periphrastic *do* in a variety of syntactic contexts (Kroch 1989, Chapter 4); rise of *have got* in a variety of "semantic" contexts (Noble 1985, Chapter 4); spread of *sort of* to Degree Modifier environments in conjunction with the expansion of *sort of*+Adjective+Noun constructions; spread of *kind of* to Degree Modifier environments in conjunction with spread of *sort of* to Degree Modifier environments (Chapter 5); spread of *be going to* to Raising Verb environments in conjunction with its saturation of Equi Verb environments; acceptance of dummy subjects by future *be going to* in conjunction with acceptance of general inanimate subjects (Chapter 5); emergence of certain blend constructions in association with structural transitions (Chapter 6).

On the other hand, although the model predicts short-term diachronic correlations, it doesn't predict eternal correlatedness between the changing elements. This is because forces may act to differentiate the structure of once-similar elements over time. In Chapter 4, I show how this may have happened in the case of affirmative declarative periphrastic *do*.

In the descriptions of experiments given below, I refer to the process of training a "blank" network to get it to embody the structure of some initial language-state as *Initial Training*. I call the process of training additionally on a distorted corpus *Post-Training*.

3.6.2   Grammaticality as a likelihood threshold

The word-prediction network introduced in Sections 2.3 and 2.4 is designed to make predictions about the frequencies with which forms occur in different contexts. For this reason it appears to be more a model of language *performance* than a model of *competence* in the sense of Chomsky 1965. One property that is generally taken to distinguish performance-models from competence models is that competence models are supposed to tell which sentences are grammatical and which are not. In the case of the performance model at hand, in fact, there is a straightforward way of extracting predictions about grammaticality: we can model the distinction between grammatical and ungrammatical sentences as a *likelihood threshold*.

> **Def.** <u>*L*-value of a string</u>. For $\vec{s} = (\vec{s}_0, \vec{s}_1, \ldots \vec{s}_n)$, a sequence of vectors successively presented to a network as current-word/next-word pairs let
>
> $$p_i = \frac{\vec{s}_i \cdot \vec{o}_i}{\sum\limits_k o_{vi}}$$
>
> for $i \in 1, \ldots, n$, where $\vec{o}_i$ is the observed output vector when $\vec{s}_i$ is the target, o_{vi} is the output on unit v under the same conditions, and v ranges over output units. The *L*-value of the string $\vec{s}$ is given by
>
> $$L(\vec{s}) = ln \ p_1 \cdot p_2 \cdot \ldots \cdot p_n$$

The quantity L gives the natural log of the probability of observing the string if the network is treated as a sentence-generator: that is, when the network is in some specified initial state, the output activations (which are all positive) are

normalized to sum to 1 so they can be interpreted as a probability distribution over the vocabulary items; one vocabulary item is then chosen according to this probability distribution and presented as the occurring "next word" on the input layer, and the process is repeated. To provide a consistent, semi-realistic measure of comparison across strings, the specified initial-state of the network is, in every case, taken to be the state that resulted after a sample sentence from the training data was presented to the network in its entirety. In other words, the network was at the juncture between the end of one sentence and the start of a new sentence at the beginning of the test.[12]

Based on L, we can define grammaticality in terms of a threshold θ which we will choose by examining strings that people judge to be "near grammatical".

> **Def.** For W, a weight-setting, and $\vec{s}$, a string, $\vec{s}$ counts as a *grammatical string* if $L(\vec{s}) > \theta$.

If the model is a good model of the language, then strings that are grammatical in the language will be grammatical in the model and the strings that are ungrammatical in the language will be ungrammatical in the model.

It is desirable to define grammaticality independently of human judgments. A definition that works somewhat well is to let θ be slightly less than the minimum of L over all the sentences in the training corpus. This definition only yields good results, however, if all the training-corpus sentences are grammatical, i.e. no noise in the data. Therefore, it is even more desirable to find a way of defining grammaticality in terms of properties of the network representation. I leave this as a matter for future research.

In the experiments described in Chapters 4, 5, and 6, I am generally interested in showing that the model predicts some kind of categorical change in the distributional behavior of the language. Consequently, the experiments all take the form of showing that when a once-trained network is post-trained on a corpus with some particular distortion, the correlated changes predicted in the manner described in the previous section include transitions by certain strings

---

[12]Since the sentences are all generated independently of one another in all the simulation corpuses described in this document, the network state at the end of one sentence tends to resemble rather closely its state at the end of another. This means it doesn't matter too much which sentence we pick to "prime" the network for testing a sequence, so long as we pick a grammatical one.

from having below-threshold L-values to having above-threshold L-values. In this way, the model predicts "grammar changes" in the traditional sense of the term.

3.7 Continuity in the Change Model

In Chapter 1 I suggested that it is desirable to have a model in which change is (relatively) continuous at the level of representation because such a model makes short-term prediction easier. Moreover, when combined with knowledge about the structure of the domain, such a model may enhance the possibility of anticipating radical structural changes, and thus make long term prediction easier as well. We would like, therefore, to know if the Connectionist model just proposed has such a continuity property.

To formalize the notion of continuous change at the level of representation, it is useful to assume that there is exactly one representation corresponding to each point in time as the language evolves. We can then refer to the *function* that maps from time to representations and ask if this is a continuous function. Function continuity is normally defined by defining a measure of *distance* in both the input and the output spaces.

> **Def.** For a set (or "space"), S, a function $d : (S, S) \to \mathcal{R}$ is said to be a *distance measure* if it satisfies three conditions:
>
> (i) $d(x, x) = 0$ for all $x \in S$
>
> (ii) $d(x, y) = d(y, x)$ for all $x, y \in S$ (Symmetry)
>
> (iii) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in S$ (Triangle Inequality)

This definition picks out some of the main properties of our intuitive notion of *shortest-path*: the length of the shortest-path from a location to itself is zero, and this happens to be the shortest path of all (follows from the combination of (i), (ii), and (iii)); if you travel someplace by the shortest-path, then the shortest-path back has the same length; if you travel someplace by the shortest-path, then a route to that place via some intermediate location can never be shorter.

Def. A set (or "space"), M, with an associated distance measure, d, is called a *metric space*—denoted (M, d).

We can now define continuity.

Def. A function f from metric space (M_1, d_1) to metric space (M_2, d_2) is said to be *continuous* if for any point p in M_1, choosing an input close to p in M_1 guarantees that you'll get an output close to $f(p)$ in M_2. Or, in "epsilon-delta" terms, given point p, for any $\epsilon > 0$ there exists a $\delta > 0$, such that $d_1(p, q) \leq \delta$ implies $d_2(f(p), f(q)) \leq \epsilon$.

In the proposed model of the language evolution function, the input space is time. In this space, distance can be defined in the normal way: we'll say that the distance between two times is the absolute value of the difference between them. The output space is the set of weight-settings that can be associated with the Connectionist model. For the sake of easy comparison between weight-settings, we can think of a weight-setting as a vector. Thus, if the network has N weights, then a weight-setting is a point in N-dimensional space. A natural measure of distance in weight-setting space (or "weight space") is Euclidean distance:

Def. If v_i and v_j are vectors in R^N, the *Euclidean distance*, d_E between them is

$$d_E(v_i, v_j) = \sqrt{(v_{i1} - v_{j1})^2 + \ldots + (v_{iN} - v_{jN})^2}$$

where v_{ab} denotes the value of the bth dimension of vector a.

For one, two, and three-dimensional spaces, this definition corresponds to our intuitive notion of distance.

The most basic property we desire of the model (suggested by the Ski-Area analogy) is the refinement-of-observation property: that the representation-system be able to instantiate states that are quite close together. Since the weights are allowed to take on any real value, the network-state can be at any location in R^N, and these locations can, in fact be arbitrarily close together so this property is trivially satisified.

We also want to know if the representation changes continuously with time. In fact, in real simulations, the representation does not change continuously with time, for the learning rule, (41), specifies discrete changes in the weight-values. Nevertheless, there is a continuous function in weight space which corresponds to the limiting behavior of the network as the learning rate, ϵ, approaches 0 (Rumelhart, Hinton, and Williams 1986). We can think of this limiting behavior as the ideal model and the simulations as approximations to the ideal model. In this sense, the change model has the evolutive continuity property which I ascribed (by hypothesis) to language in Chapter 1. If it turns out that network evolutive continuity is a good model of language evolutive continuity, then the network representation should be useful in making predictions about language change.

Before examining the properties of the change model itself, it is worth remarking on a closely-related but distinct model of change that the network also forms the basis of: we can hypothesize that a grammar at a given time t has the structure of a network trained on a sample of data generated by speakers living at time t. I will call this the "Position Model" of language change and contrast it with the "Transition Model" described above, which is based on the Transition-by-Learning assumption. About the Position-Model, we can also ask, Is the mapping from time to representations a continuous function? I have not yet tried to answer this question empirically, but it seems like an interesting direction in which to pursue this research, so I will use the remainder of this section to suggest some ways of going about it.

It will be necessary to train networks on large, real corpuses. For this, it will certainly be difficult to use the word-prediction paradigm in Section 4 because of the large number of vocabulary items involved. However, a preliminary process of assigning low-level grammatical tags by using some of the automatic clustering mechanisms that are currently being proposed (e.g., Brown *et al.* 1990, Brill *et al.* 1991, Schuetze 1993a, Finch 1993) could well reduce the number of distinct items the network would need to represent without removing the diachronic structure of interest.

Moreover, in the case of real corpuses, we have no possibility of examining behavior in the limit (as is done in the mathematical evaluation of continuity) for we can only make discrete observations. Nevertheless, the observation that

corpuses evolve continuously is not an illusion. It seems, rather to reflect an intuitive notion of relative continuity that applies to discrete functions: if successive values of a function are more-or-less close-together relative to the range of values the function assumes, then we are more-or-less inclined to call it "continuous". Shepard and Carroll 1966 suggest the measure shown in (45) as a way of formalizing the notion of (inverse-)continuity (see also von Neumann 1941).

(45)

$$Z = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}\left(\frac{y_{i+1}-y_i}{x_{i+1}-x_i}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Here, the y_i's are the output of the function and the x_i's are its input; $\bar{y}$ is the mean of the y-values so the denominator of Z is simply the sample variance of the y-values. The numerator measures how far-apart successive y-values are relative to the spacing between the corresponding x-values on average, so the whole formula is large if the function makes relatively large y-jumps for small x-jumps. In this way Z corresponds in a not-implausible way to our intuitive sense of relative inverse continuity for discrete functions.

Below I describe several metrics for comparing grammars induced from different corpuses. It is easier to estimate distances between grammars by examining their interactions with corpuses than it is to examine directly the relationships between the grammars themselves. Thus it is desirable to be able to compute a measure of relative continuity on the basis of distances between grammars alone without assuming knowledge of the structure of the grammars themselves. For such cases, the following is a natural generalization of Z:

(46)

$$Z' = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}\left(\frac{d(\vec{y}_{i+1},\vec{y}_i)}{x_{i+1}-x_i}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}d(\vec{y}_i,\bar{y})^2}$$

Here, $d(\vec{y}_i,\vec{y}_j)$ is the distance between the entities $\vec{y}_i$ and $\vec{y}_j$. For certain metric

spaces and sets of distances, there may be no set of entities with the specified distances, so the mean, $\bar{y}$ will not exist. If such cases arise in real situations, we can assume provisionally that the lack of a mean is due to noise in the data. We can then compute an approximate mean and use this in the evaluation of Z'. It is not easy to compute a mean (exact or approximate) for a large set of points based on the distances between them alone. Nevertheless, the following approach may work reasonably well: choose an arbitrary set of points in some high-dimensional space, one for each unknown point and move them incrementally toward or away from one-another until their distance relationships closely approximate the known distance-relationships; compute the mean of these points. If the minimal error is close to 0 then we may take our provisional assumption that the discrepancy is due to noise to be justified. We can then compute the mean of the $\hat{y}_i$'s and use this as an approximation of $\bar{y}$.

Before examining some distance measures, there is a question that begs to be asked: What will be the the point of finding out how continuous language change is according to the Z-measure during various intervals of time? If continuity is "relative", as this measure takes it to be, then it is hard to see how we could arrive at a definitive answer to the question, "Is language change continuous or is it not?" The worth of measuring relative continuity lies not so much in the possibility of answering this question, but rather in the possibility of evaluating the claim that, under the Connectionist representation, linguistic change is continuous enough that strong constraints can be put on short-term prediction. If it turns out that even under the Connectionist representation, language states vascillate as much in short periods of time as they do over long periods of time, then we we should be skeptical of the prospects of using it as a basis for a constrained theory of change. Of course no one will be surprised if it turns out that, according to a network-based measure of continuity, languages often go through long periods when they change very little. This is evident from the textual data in an intuitive way and it is also suggested by the fact that speakers of modern languages can fairly easily interpret states of their own languages from many centuries preceding. What will be interesting is to find out if even at points where it is evident that dramatic structural change occurred and where current theories essentially throw up their hands and say

arbitrary "reanalysis" must have created a different grammar, the network-based measures show anticipatory and ensuing trends. For example, Lightfoot 1979 and Kroch 1989b have both proposed models of the evolution of English grammar which maintain that at certain points (especially around the mid-16th century), radical structural changes occurred in the grammar. At these points, then, measurement of the continuity of the network representation may be revealing.

So how can we measure distances between corpuses? The model under consideration here suggests a class of answers to this question that all have the same basic form: train a network separately on several different corpuses (assumed, for simplicity, to be formed from the same vocabulary) and define a measure of distance between the weight-settings of the network. The obvious tack of measuring the distance between the points in R^n corresponding to the different weights will not work because in separate trainings of a network, the hidden units may take on different roles. One way of getting around this problem is to compare the *behaviors* of the different trained-networks instead of comparing the networks directly.

3.7.1 Behavior-based corpus-comparison metrics

In each case, I'll assume that we have two weight-settings, W_i and W_j which have been arrived at by training a single network architecture, A, on two different corpuses. For the purpose of testing the Continuity Hypothesis, these corpuses should be samples of the usage of a single language at two different points in its history. The first metric is based on comparing the behaviors of the two networks on a single, chosen corpus, which we can think of as a sort of *origin* in the space of corpuses. Therefore I call it the *Origin Metric*.

Def. <u>The Origin Metric</u>. For C_0, a corpus, and W_i and W_j, weight-settings obtained by training networks on corpuses C_i and C_j respectively, the distance between C_i and C_j is given by:

$$d_O(C_i, C_j) = |E(W_i, C_0) - E(W_j, C_0)|$$

where $E(W, C)$ is the total error that results when W is tested on corpus C.[13]

This metric has the advantage that it is simple and easy to compute, but it has the disadvantage that radically diverse behaviors are lumped together on isoclines since there are many different grammars that can produce the same error on a single corpus, so it cannot provide a very convincing demonstration of evolutive continuity.

If we assume that the training corpuses include a "period" symbol which marks precisely the end of every sentence, and sentences are (roughly) independent events, we can think of corpuses as consisting of sets of sentences, which in turn are ordered sequences of words. Given these assumptions we can make reference to the notion of a grammaticality threshold, as defined in Section 6.2 above, and define a metric in terms of the agreement on grammaticality judgments between two weight-settings.

Def. <u>The Threshold Metric</u>. For θ, a likelihood threshold, and W_i and W_j, two weight-settings obtained by training networks on corpuses C_i and C_j respectively, the distance between C_i and C_j is given by:

$$d_T(C_i, C_j) = 1 - \frac{||G(W_i) \cap G(W_j)||}{||G(W_i) \cup G(W_j)||}$$

where $G(W)$ is the set of finite strings of elements drawn from the vocabulary followed by a period which network W classifies as above threshold, and $||S||$ denotes the cardinality of set S.[14]

---

[13] That this measure is a distance metric follows from the fact that absolute value is a distance metric.

[14] The only difficult part of proving that d_T is a distance metric is proving the Triangle Inequality. The following argument is the best I have been able to make, although surely there is a simpler one!

Let $A, B,$ and C be sets. I will show that,

(47)

$$\left(1 - \frac{||A \cap B||}{||A \cup B||}\right) + \left(1 - \frac{||B \cap C||}{||B \cup C||}\right) \geq 1 - \frac{||A \cap C||}{||A \cup C||}$$

By rearrangement of terms, we have,

(48)

$$\frac{||A \cap B||}{||A \cup B||} + \frac{||B \cap C||}{||B \cup C||} \leq 1 + \frac{||A \cap C||}{||A \cup C||}$$

In the simple case in which $A \cap B = \varnothing$ and $B \cap C = \varnothing$, this equation is clearly satisfied since the left hand side is 0 and the right hand side is always greater than or equal to 1.

To make the notation simpler for the other cases, I let $a = ||A - (B \cup C)||, b = ||B - (A \cup C)||, c = ||C - (A \cup B)||$ where $X - Y$ is the set of elements in set X that are not in set Y. Let ab denote $||(A \cap B) - C||$. Let abc denote $||A \cap B \cap C||$. We can write

(49)

$$ab + abc + cb + abc = ab + cb + abc + abc$$

Suppose either $A \cap B \neq \varnothing$ or $B \cap C \neq \varnothing$. Then $ab + cb + abc > 0$ so we can write

(50)

$$\frac{ab + abc}{ab + cb + abc} + \frac{cb + abc}{ab + cb + abc} = 1 + \frac{abc}{ab + cb + abc}$$

which means that for the special case in which $A \cup C = B$, the Triangle Inequality is satisfied. Now suppose C contains some objects which are not in A or B. Then $c > 0$. From (50), we have

(51)

$$\frac{ab + abc}{ab + cb + abc} + \frac{cb + abc}{ab + cb + abc + c} \leq 1 + \frac{abc}{ab + cb + abc + c}$$

based on the fact that if $M \geq N$, $x \geq y > 0$, and $c > 0$ then,

(52)

$$\frac{M}{x} - \frac{M}{x + c} \geq \frac{N}{y} - \frac{N}{y + c}$$

By the same argument, we can conclude, on the basis of (51) that

(53)

$$\frac{ab + abc}{ab + cb + abc + a} + \frac{cb + abc}{ab + cb + abc + c} \leq 1 + \frac{abc}{ab + cb + abc + c + a}$$

which leads to

(54)

$$\frac{ab + abc}{ab + cb + abc + a + ac} + \frac{cb + abc}{ab + cb + abc + c + ac} \leq 1 + \frac{abc + ac}{ab + cb + abc + c + a + ac}$$

based on the fact that if $x \geq y > 0$ and $a > 0$ then,

(55)

$$\frac{x + a}{y + a} \geq \frac{x}{y}$$

and

For some grammar-induction systems, G will be infinite for some corpuses in which case this distance measure will not be very useful. But in my experience, the networks described here tend to assign probability less than q for some $q < 1$ to all transitions so G(W) is always finite. In practice, the number of sentences deemed grammatical may be unmanageably large, but we can replace $G(W)$ in the definition above with a sample of sentences generated by the network according to the procedure described in section 6.2. In this case, we will be comparing the propensity of two networks to generate the same sentences. If my claim (Chapter 4) is correct that new constructions become grammatical by successive minor saltations (rather than simultaneously, as per Kroch 1989a and 1989b), and the network does a good job of modelling the sentences people deem grammatical, then the Threshold Metric should detect continuity in corpus change.

A shortcoming of behavior-based metrics is that they do not facilitate dissective analysis of representational change. Yet, in many instances, we may be interested in looking at how certain elements are changing status relative to others in the same language (e.g., in the case of the *reanalysis* of a particular word, while other words maintain their states). To do dissective analysis, we would like to be able to compare specific features of one representation to those of another across time. Extraction of *Principal Components* is one method of comparing representations across networks trained in different settings.

3.7.2 A structure-based metric

The Principal Components of a set of points are the square roots of the ordered eigenvalues of the matrix of covariances among the dimensions in terms of which

(56)

$$\frac{x}{y + a} < \frac{x}{y}$$

This last observation also allows us to conclude, based on (54), that

(57)

$$\frac{ab + abc}{ab + cb + abc + a + ac + b} + \frac{cb + abc}{ab + cb + abc + c + ac + b} \leq 1 + \frac{abc + ac}{ab + cb + abc + c + a + ac}$$

which means that for all remaining conditions, the Triangle Inequality is satisfied.

the points are represented.[15] To perform Principal Component analysis is to assume that the points are distributed in a symmetric hyper-ellipsoid, and to discover a set of perpendicular coordinate axes such that one of them (the first Principal Direction) is alligned with the longest dimension of the ellipsoid, another (the second Principal Direction) is alligned with the second-longest, etc. The Principal Components measure the stretch of the ellipsoid along each of the Principal Directions. Weigend 1994 finds it useful to compare different weight-settings in a 3-layer network by recording the hidden unit locations associated with a representative sample of points, computing Principal Components, and comparing same-ordinal components pairwise. This suggests the following metric:

> **Def.** <u>The Principal Components Metric</u>. For corpuses, C_i and C_j and corresponding weight-settings W_i and W_j learned in the same network architecture, the distance, d_{PC} between the corpuses is the distance between the vectors of principal component magnitudes for the clouds of hidden-unit locations obtained when W_i is tested on C_i and W_j is tested on C_j.

One advantage of this metric is that we don't have to reconstruct the points that gave rise to the distances in order to compute the means since we know what they are in the first place. I suspect that because the Principal Components identify the dimensions of variance in the very space that the network uses to form its representation, they should reflect in quite a sensitive way, properties of the representation that make important distinctions with respect to the task. Consequently, they should provide a fairly convincing test of whether the representation is changing continuously.

---

[15]The *matrix of covariances* among a set of points, $x_1, \ldots, x_n$, has, as entries,
(58)
$$cov_{ij} = \frac{1}{n-1} \sum_p (x_{ip} - \bar{x}_i)(x_{jp} - \bar{x}_j)$$
where x_{ip} is the ith dimension of point p.

If we treat the set of Principal Component vectors as a basis for the hidden unit space, we may be able to make meaningful comparisons between the representations of individual items in different networks. Thus we can ask, How has the status of word w_i with respect to the grammar changed with time? If the assumption about the ellipsoidal character of the distribution is accurate, and both the network and classical linguistic models are doing a good job of expressing the correlational structure in the corpus data, then the principal components may correspond to interpretable linguistic contrasts. This will make it easier to answer to questions like, Has word w_i become nearer to word w_j in a grammatically-relevant sense?, which are pertinent to evaluating the hypothesis that Q-Divergence is significantly correlated with grammar change (see Chapter 5).

Unfortunately, the ellipsoid-distribution assumption underlying Principal Component analysis may not be accurate in all cases of interest here. Although the hidden unit representations tend, as training time goes to infinity, to cluster on orthogonal dimensions for many of the linguistic data-sets I have examined, there seem to be cases in which components are relevant in one part of the space but irrelevant in another. For example, Number may be a feature of some categories (e.g., Determiners, Nouns, Verbs) but not others (e.g., Prepositions); a case-contrast may pertain to some grammatical functions but not others. In such cases, even though the representations permit encoding of the relevant distinctions, the language does not make use of them. It seems likely that such asymmetries can play a significant role in historical change (e.g., when a new case-allignment system spreads across verbs (e.g., Harris 1990)) so it will be helpful to have a way of detecting such asymmetries in the distribution. For this, the methods described in the next section, again based on proposals of Shepard and Carroll 1966, may prove useful.

3.8 Restrictiveness in the Change Model

One of the desiderata identified in Chapter 1 is that we would like our grammatical representation to be restrictive. Choosing a continuous representation that is no more restrictive than one's descriptive representation generally makes

for a very unenlightening theory. For example I have adopted here a very unconstrained descriptive representation: corpuses consists of sets of "sentences", which are sequences of words drawn from a fixed vocabulary. If we claim merely that the distribution of observed sentences must change gradually with time, it will be (i) implausible, since even encountering identical sentences across generations seems to be a rarity and (ii) terribly underconstrained, since even if we start the model in a plausible state, arbitrary combinations of word-strings will be permitted to emerge and grow prominent. Clearly we prefer a representation that puts constraints on cooccurrence of particular sentences. For the most part, this is what linguists mean by a *restrictive* theory of *syntax/lexicon*, although it is often proposed that some of the restricting work be done by other components of grammar (e.g., phonology, semantics, pragmatics).

Just as we can talk about the restrictiveness of a theory of change in terms of its diachronic predictions, we can talk about the restrictiveness of a theory of grammar in terms of its synchronic predictions. Linguists have identified two kinds of synchronic predictions that it seems desirable for the theory of grammar to make: *typological* and *extrapolative*.

> i. (Typological) The theory of grammar should say which properties of sentences will be correlated across the languages of the world.
>
> ii. (Extrapolative) On the basis of a small, well-chosen sample of sentences from a language (e.g., those a child is exposed to), the theory of grammar should be able to say which other sentences are also in the language.

Presumably these two desiderata will pinpoint fairly similar theories if sufficiently probed. Connectionist networks which learn from examples are naturally suited for theory-development in pursuit of desideratum (ii). Consequently, I shall focus on that desideratum here.

An effective strategy is to choose a representation that reduces the redundancy in the set of observed behaviors. It turns out that Shepard and Carroll 1966's measure for relative (inverse-)continuity, (referred to as Z in the previous section), provides a basis for finding good redundancy-reducing representations.

They are concerned with the problem of having a set of points that are distributed in some high-dimensional space, but happen to have underlyingly low-dimensional structure. For example, a circle is an essentially two-dimensional object, but a circle can exist in three-dimensions, four-dimensions, etc. People generally have a hard time interpreting even low-dimensional objects in four-dimensional and higher spaces, especially if they are not alligned in a natural way the coordinate axes in terms of which the objects are perceived. But if the objects themselves are really of lower dimensionality, this difficulty is an unnecessary impediment. We can make use of our three-dimensional intuitions to make inferences about the higher-dimensional object, provided we know how to map from the lower dimension to the higher. Thus it is desirable to find a way of discovering lower-dimensional representations for higher-dimensional objects.

Consider, now, this dimension-reduction problem in the context of a partial data-sample for some object. We might, for example, have a scatter of discrete points around the perimeter of a circle. Even in this situation, if the circle is sampled thoroughly and randomly enough, people can quite easily perceive its structure in lower dimensions. Thus, we might desire to perform dimension reduction for discrete samples and interpolate between the points. Here, the notion of the inverse-continuity measure, Z, comes in handy. We want to find a set of points in a lower dimensional space which preserves the relationships among the points in the higher dimensional space. One way of doing this is to start with a function mapping an arbitrary set of points in the lower dimensional space to the known set of points in the higher-dimensional space and to adjust the lower-dimensional points by performing gradient descent on a generalized Z-measure, which computes relative inverse-continuity for functions from m-space to n-space. Carroll proposes the following generalized measure, which he calls κ:

(59)

$$\kappa = \frac{\sum\limits_{i \neq j} \sum\limits_{j \neq i} \frac{d_{ij}^2}{D_{ij}^4}}{\left[\sum\limits_{i \neq j} \sum\limits_{j \neq i} \frac{1}{D_{ij}^2} \right]^2}$$

Here d_{ij} is the Euclidean distance between $\vec{y}_i$ and $\vec{y}_j$ (the higher-dimensional points) and D_{ij} is the Euclidean distance between $\vec{x}_i$ and $\vec{x}_j$ (the lower-dimensional points).

Carroll is mainly concerned with showing that the algorithm can reduce the dimensionality of objects themselves in a satisfactory manner and he describes several experiments in which it does so (reducing a circle to a line, reducing the surface of a sphere to a planar object—in the latter case, the algorithm found something like the azimuthal equi-distant projection used by cartographers).

The point of interest here is that the task addressed by the algorithm appears to be similar to the problem of designing a grammar-induction device which can make good inferences on the basis of limited data-sets. In particular, both tasks involve finding a reduced-dimension representation that preserves continuity in the sense of placing inputs with similar output behaviors near each other in the representation. And for the 3-layer feedforward network models, a close comparison can be made: Carroll's points in a high-dimensional space correspond to target-values in the network's output space. The chosen lower-dimensional space corresponds to the hidden unit space in which the network is constrained to find a representation. The strategy of starting with arbitrary points in the input space is analogous to the strategy of starting with arbitrary weights on the connections from input to hidden units. One important difference is that while Carroll's algorithm works by holding the y-values fixed and adjusting the x-values, the network adjusts both the x (hidden) and the y (output) values simultaneously.

We would like to know, of course, if the backpropagation learning algorithm does the same thing as Carroll's continuity-improving algorithm. I have not yet had a chance to analyze the relationship between the algorithms explicitly, so I cannot give an analytic answer to this question. Nevertheless, the discussion in Section 2 surrounding the question of how the network learns to parameterize a data-set indicates that the network is doing something very similar to continuity-improvement. Much linguistic data, and the example data discussed in Section 2 have a largely discrete character, so one may wonder if the network performs similarly to Carroll's algorithm for problems of continuous-object reduction. (This is, in fact, what I am claiming it must to do in order to account

for the graded character of language change.) Slight evidence in favor of an affirmative answer was provided by the words that were ambiguous between Singular and Plural in the example simulation in Section 2. These occupied intermediate locations in the output space and were assigned intermediate locations in the representation space. Further examples related to these are discussed in Chapter 6. To make a direct comparison with Shepard and Carroll's data, I also tested the network on mapping problems where the outputs were sampled from continuous objects in high dimensional spaces and the hidden unit space was smaller (e.g., straight-lines from 3 dimensions to 1 [need to try also some curves and higher-dimensional objects]) and found that it performed similarly.

Assuming its correctness, the interpretation of Connectionist dimension-reduction as continuity-minimization provides a nice way of understanding how the network model predicts Q-divergence effects in lexical reanalysis, cases of which are the focus of Chapter 5.

Lexical Q-divergence is the phenomenon whereby a quantitative change in some aspect of a word's behavior against the backdrop of a (relatively) static language is correlated with other changes in the word's behavior, which may be categorical in nature. Figure 3.13 gives an illustration of how this effect arises in terms of mappings from lower-dimensional to higher-dimensional representations. There is a set of labelled inputs, $word_1$, $word_2$, $word_3$,, etc. which, in the process of initial-training have come to be associated with particular locations in the hidden unit space (portrayed here as one-dimensional). Each of these hidden-unit locations maps to an output which forms a (relatively) continuous object in some higher dimensional space (portrayed here as a (one-dimensional) curve in three dimensions). Moreover, the mapping is an especially good solution to the continuity-maximization problem. If, in post-training, we impose a small change on one dimension of the output associated with one element, say $word_2$, by including a target that is distorted along just this dimension, the network will try to adjust the weights to accomodate this change. But any adjustment it makes in the mapping from hidden unit representation to output is quickly erased because the current mapping is near an continuity maximum for the remaining, unchanging points (which are still being presented in their original form), so every time one of them is encountered, it tends to revert back to its original shape. On the other hand, an adjustment in the mapping

Figure 3.13: Q-Divergence Effects as a Result of Dimension-Reduction.
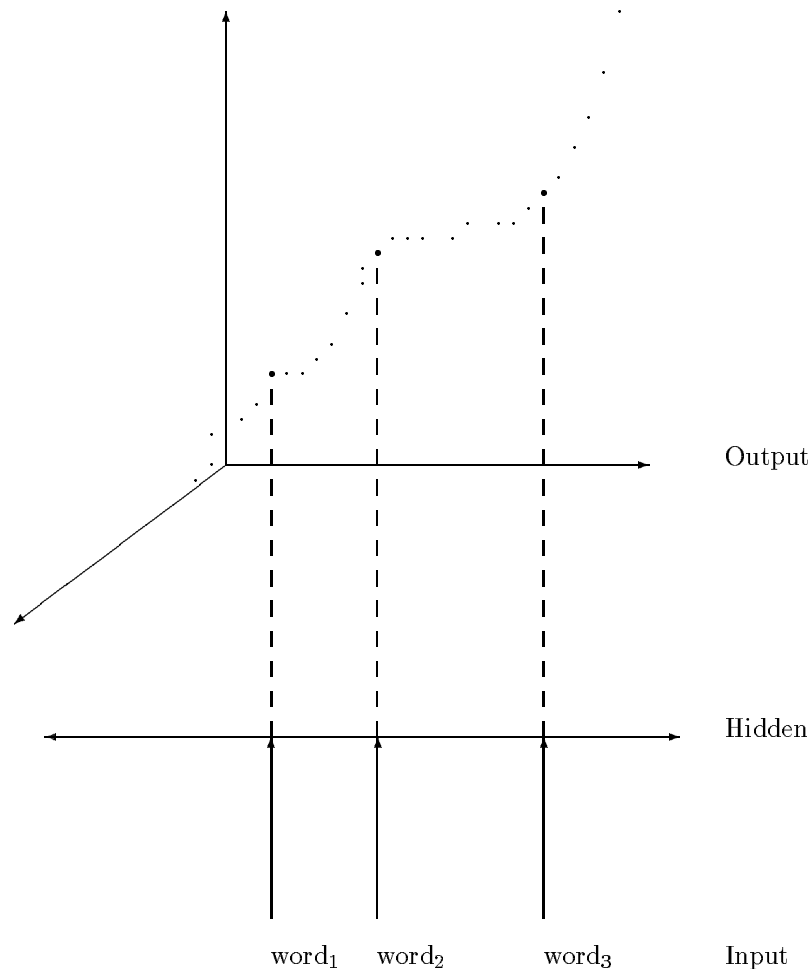
output-curve winds around obliquely relative to the axes, change in the single dimension of the representation space produces change in all three dimensions of the output space. Thus we predict a correlation between the perturbation (which was along just one of the output dimensions) and other changes in the perturbed item which were not imposed. Since there is no qualitative distinction between categorical and frequency changes in this model (categorical change is simply frequency change that crosses a certain threshold), a frequency-change in one dimension can correspond to a categorical change in another. If contextual information is included in the model (as in the recurrent network), then the input↔hidden correspondences vary from context to context. Nevertheless, the same *weights* mediate between an input and the hidden units in every context. Consequently a change in these weights induced by a perturbation in one context can produce correlated changes in other contexts as well. This is how the network predicts lexical Q-divergence effects.

Frequency-linkage effects, on the other hand, have to do with changes in the mapping from hidden to output units.[16] Consider, for simplicity, a network with no biases on the units. Suppose we have two input patterns p_1 and p_2 and we want to know whether imposing a force which changes p_1's behavior on some output dimension, o_i, will produce a correlated change in p_2's behavior on that dimension. If we assume that p_1 and p_2 exhibit a variety of behaviors in other contexts which do not change during the time that the force impinges on o_i, then the weights from the input layer to the hidden layer will remain stable. Consequently, the force will have an effect only on the hidden→output

---

[16] Whether a particular force induces frequency-linkage or lexical Q-divergence (or a mixture of the two) depends on its nature. If it acts to change the behavior of a single content-word in a majority of its contexts then the representation of just that word will change and lexical Q-divergence effects will be observed. If it acts to change the behaviors of a majority of words in a single context, then the representation of the context will change and other words occurring in the same context will be swept along. This is the frequency-linkage effect. If a force impinges on a word that itself is largely or solely responsible for marking a context (i.e., a function word), then the network will change the representation of both the word and the context. In this case, since the word codes a pervasive distinction, it may command an independent dimension in the representation space so the change in its lexical representation may not produce assimilation to another lexical type. On the other hand, the change in its contextual representation will affect all the subcontexts (and words) associated with that context so frequency-linkage across them will be observed. Such appears to have been the case with the rise of periphrastic *do* and the *have > have got* switch in British English (See Chapter 4).

from $word_2$'s input-label to the hidden representation has consequences only for the behavior associated with $word_2$. Consequently, the algorithm *will* make progress in adjusting the connections between that label and the hidden units if doing so will allow it to better fit the perturbation. Moreover, since the induced

weights. Suppose that, upon presentation of input p_1, the change in the vector of weights projecting from the hidden layer to output unit o_i is $\Delta \vec{w}$. Then the output-values associated with patterns p_1 and p_2 will change in parallel if the change in net input upon presentation of p_1 is the same as the change in net input upon presentation of p_2. Since there is no input→hidden weight-adjustment, the change in net input is just the product of the weight-change and the hidden-activation. So we will have parallelism if

(60)

$$\Delta \vec{w} \cdot \vec{a}_1 = \Delta \vec{w} \cdot \vec{a}_2$$

where $\vec{a}_j$ is the hidden activation associated with input pattern p_j. This suggests defining

(61)

$$C = |\Delta \vec{w} \cdot (\vec{a}_1 - \vec{a}_2)|$$

where we can think of C as a measure of the degree to which the elements p_1 and p_2 will exhibit frequency-linkage effects. We will observe perfect same-slope behavior if the weight-change is orthogonal to the difference between the hidden unit representations of the two elements ($C = 0$) and the degree of correlation will diminish the more the weight change is alligned with the difference between hidden unit representations ($C > 0$). If we model the differential force on the behavior of p_1 as learning under the Backpropagation Rule, then the weight change is given by,

(62)

$$\Delta \vec{w} = \epsilon \ f'(net_i)(t_i - o_i)\vec{a}_1$$

$$= Q \ \vec{a}_1$$

28 where Q is a scalar. In other words, the weight-change associated with the presentation of input p_1 is alligned with the hidden unit vector $\vec{a}_1$. Thus, if all the hidden unit vectors are nearly the same length, which seems often to be the case when there is a large number of hidden units, the weight change will have the strongest effect for hidden unit activation states that point in the same direction as $\vec{a}_1$. Formally,

(63)

$$C = |\Delta \vec{w} \cdot (\vec{a}_1 - \vec{a}_2)|$$

$$= |Q \ \vec{a}_1(\vec{a}_1 - \vec{a}_2)|$$

$$= |Q' \ \left[\frac{\vec{a}_1 \cdot \vec{a}_1}{||\vec{a}_1||^2} - \frac{\vec{a}_1 \cdot \vec{a}_2}{||\vec{a}_1||^2}\right]|$$

where $Q' = ||\vec{a}_1||^2 Q$. If $\vec{a}_1$ and $\vec{a}_2$ are the are nearly the same length then

(64)

$$C \approx |Q' \ (1 - cos \ \theta)|$$

where θ is the angle between $\vec{a}_1$ and $\vec{a}_2$. Equation (64) says that inputs will undergo correlated weight change if their hidden unit representations are exactly the same and the correlation should degrade as the representations decrease in similarity (note that all the hidden unit representations are in one quadrant so the cosine is never negative). Thus the network model makes the same predictions as Kroch's Grammar Mixture model for cases in which grammatical similarity corresponds to closely to distributional similarity. It differs in cases where there is a detectable divergence.

3.9   Summary

Based on the discussion in the previous section, a useful table of comparisons can be drawn up. The network encodes two distinct mappings: between hidden-unit representations and outputs, and between inputs and hidden-unit representations. The hidden→output mapping corresponds approximately to the linguistic notion of "syntax" as distinct from "lexicon": it constrains certain units (words and phrases) with different phonologies to behave alike in particular contexts. The input→hidden mapping corresponds closely to linguistic representations of the "lexicon": each word is associated with a set of weights projecting to the hidden layer, which, in effect, determine which contexts it can occur in (as well as how frequently it can occur there). A significant difference between the linguistic and the Connectionist representations is in the coding of ambiguous words. Linguistic grammars usually code ambiguous words as distinct structures which happen to have the same phonological manifestation. The network, on the other hand, associates the same weights with every occurrence of a particular word-string. Differences in behavior are predicted via the interaction of this representation with context. In this regard, the network representation is somewhat like that of a parsing-program which associates a disjunction of lexical entries with each word-string, and replaces the disjunction with a single entry (or a smaller disjunction) on the basis of contextual information. One desirable property the network has is that, while it *can* let the context be constraining enough in the case of truly ambiguous words to significantly rule-out entire classes of readings, it also *can* let the intermediate character of the representation come into play so that hybrid-behaviors emerge. This seems to be the appropriate prediction in the case of certain polysemous elements and syntactic blends. I take this matter up in Chapter 6. Finally, as indicated by the analysis of Frequency Linkage and lexical Q-Divergence given in the previous section, there is a tie between these two phenomena and the first and second weight-blocks of the network as well: Frequency-Linkage is a phenomenon pertaining to the systematic relationships of words and phrases to one-another (i.e. syntax) and it is modeled as weight-change in the hidden→output weights, or as weight-change in the hidden→hidden weights. Lexical Q-divergence is a phenomenon pertaining to the status of an individual word with respect to the grammar

and it is modeled as weight-change in the input→hidden weights. Figure 3.14 summarizes these observations.

Figure 3.14: Summary of the relationships between models and phenomena.

Network Model	Input→Hidden	Hidden→Output Hidden→Hidden
Linguistic Models	Lexicon	Syntax
Diachrony Phenomena	Lexical Q-Divergence	Frequency Linkage

Chapter 4

Frequency Linkage

4.1 Frequency Linkage Effects

"Parametric" theories of grammar hold that the language-representation device is equivalent to a set of switches that can be set either "on" or "off" or to one of a small finite number of positions. A switch, or "parameter", is generally responsible for determining the behaviors of a variety of grammatical elements, so the theories make predictions about which kinds of grammatical constructions can cooccur in a given language (e.g. Hyams 1986, Bresnan and Moshi 1990, Roeper and Williams 1987, Chomsky and Lasnik 1993).

Sometimes particular instances of such theories are violated by the appearance in the same language of two constructions which are supposed to be generated by contrasting settings of some parameter. For example, Pintzuk 1991, adopting a theory in which the base position of Infl can be fixed at one of several possible positions, provides evidence that Infl is nevertheless base-generated in both medial position and final position in Old English clauses. Upon encountering such data, one may be inclined to conclude that the proposed parametric alternatives are incorrect and that a different hypothesis about the nature of the categories is needed. But there is an alternative account which should be considered. Suppose that adult speakers of a language can alter the settings of particular parameters on a construction-by-construction basis. If the changing settings are chosen in some systematic way, then this proposal to weaken the

parametric restrictiveness of Universal Grammar does not eliminate its predictive power altogether. One theory along these lines comes out of the work of Labov (e.g., 1969) and has been explored in some depth by Kroch 1989a, 1989b, Pintzuk 1991 and others. It works as follows. For each parameter, there is a fixed probability distribution over its possible settings. Every time a meaning needs to be expressed, a setting is chosen according to the distribution, and the form corresponding to that setting is used. The model has been referred to as "the Double Base" hypothesis, in cases where the choice is thought to be between exactly two alternative "Base" structures, and the "Competing Grammars" or "Competing Subsystems" model in general. Because it is not my concern here to advance the claim that there are always only two options whenever a parameter is variably set, I use one of the more general terms here: "Competing Grammars".

The choice to talk about real-valued probabilities and category-valued typological alternatives in the same discourse leads to a potential terminological confusion: both kinds of variables are often called "parameters". To avoid confusion, I will refer to the real-valued variables as "r-parameters" and the typological variables as "c-parameters".

The Competing Subsystems model makes a prediction that monolithic models of grammar do not make. When two constructions are generated by the same setting of a particular c-parameter, then they should always have the same relative frequency. Unfortunately, even making allowances for noise in the data, this prediction is not born out: the frequencies of constructions generated by the same c-parameter setting often differ *systematically*. Fortunately, the particular type of systematicity involved is easy to incorporate into the model.

It turns out that in historical episodes where a new construction is replacing an old one, the graph of relative relative frequency versus time is typically S-shaped: change is slow when the new element is first coming in; it speeds up during the time when the two constructions are in roughly equal proportions; then it slows down again as the new construction comes near to saturating the context. Several curves have been suggested as models for such developments (see Osgood and Sebeok 1954, Altman et. al. 1983). One that works well and has been thoroughly taken up in synchronic variation research is the sigmoid introduced as a Connectionist "activation function" in the previous chapter. It

is reproduced here in Figure 4.1.

Figure 4.1: Sigmoid model for historical relative-frequency changes.
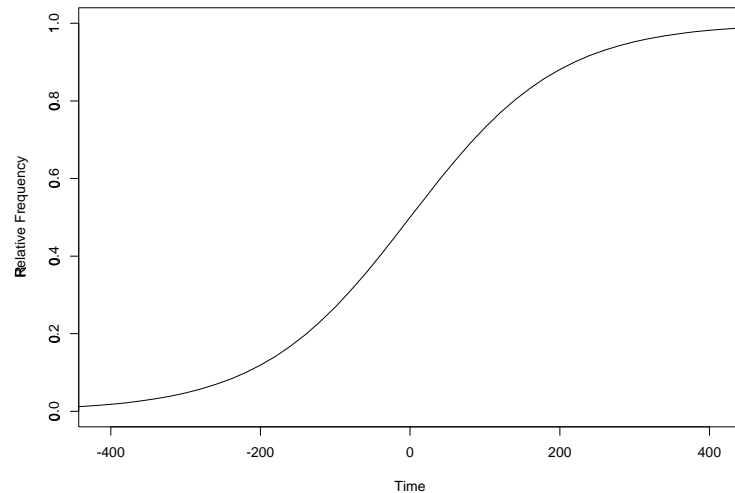
It is convenient to make the input to the sigmoid a linear function of time, with slope s and intercept k. If s is large, then the sigmoid is very steep in the middle and flat in the tails; if k shifts in the positive direction, then the whole curve is translated leftward relative to the zero-point. Thus this r-parameterization provides a way of fitting the sigmoid to a variety of observed frequency-changes.

Kroch 1989a and 1989b make the appealing suggestion that the frequency-curves for constructions generated by the same c-parameter-setting are constrained to have the same slope: it is only the intercepts, k, which differentiate among them. This is an appealingly restrictive claim in the sense that it uses exactly one free r-parameter for each contextual effect and only one free r-parameter for the choice of grammar as a function of time. It is common practice in generative grammar to assume that contextual effects are independent of grammar; it is also common practice not to theorize about constraints relating contextual influences to each other; under these assumptions, assigning a free r-parameter to each context is unavoidable. Therefore, using only one additional

r-parameter to fit the choice of grammar is quite a parsimonious move. Kroch calls this hypothesis the "Constant Rate" hypothesis. I showed in Chapter 3 that a connectionist network predicts Constant-Rate effects under certain ideal conditions (when constructions have essentially identical representations), but it predicts correlated change with divergent slopes if the relevant representations differ. Consequently, I will refer to the whole class of effects involved here with the more general term "Frequency Linkage". Under this terminology, Constant Rate effects are a special case of Frequency Linkage effects.

What is the evidence for the constant-rate hypothesis? Kroch 1989b cites a number of historical variation studies in which sigmoids were fitted to historical data and the slopes were found to be nearly identical. A listing of the cases he reviews, as well as several other studies, is given below.

4.1.1 Prior studies of Constant-Rate effects

Noble 1985. Noble 1985 studies the increasing use by speakers of British English of *have got* in place of *have* for possession from the mid-18th century to the present. She partitions the data in two different ways: concrete possessed object versus abstract possessed object and temporally bounded possession versus permanent possession ((65)–(66)).

(65) a. I've got/I have a new job. [temporally bounded]
   b. I've got/I have brown eyes. [permanent] (Kroch 1989b: 7)

(66) a. She's got/she has a car [concrete]
   b. She's got/she has a careful approach [abstract] (Kroch 1989b: 7)

For each partition, she tabulates relative frequencies for the two environments and finds that the VARBRUL estimates for the context-effects are constant across time.[1] Her data are tabulated in Figure 4.2. Graphs of regression fittings of the linearized data are given in Figure 4.3. Indeed, as also Kroch notes, the slopes are fairly similar.

Oliveira e Silva 1982. Oliveira e Silva studies the rise in the use of the definite article in Portugese in a variety of semantically- and syntactically-defined

---

[1]See Kroch 1989b for a discussion of VARBRUL.

Figure 4.2: Parallel rise of *have got* vs. *have* in several contexts (from Noble 1985, quoted in Kroch 1989b: 8).

Table 1: Effect of possession type on the choice between *have* and *have got*.

Period	Type	% *have got*	Total	Varbrul Prob.
1750–1849	tempor. bdd.	12	83	0.66
	permanent	4	108	0.34
1850–1899	tempor. bdd.	34	99	0.64
	permanent	16	122	0.36
1900–1935	tempor. bdd	89	74	0.66
	permanent	70	43	0.34

Table 2: Effect of concreteness on the choice between *have* and *have got*.

Period	Type	% *have got*	Total	Varbrul Prob.
1750–1849	concrete	13	68	0.66
	abstract	4	123	0.34
1850–1899	concrete	34	74	0.61
	abstract	20	147	0.39
1900–1935	tempor. bdd	86	51	0.58
	abstract	79	66	0.42

contexts: Unique reference, Object of a Preposition, Third Person, Kinship Term. She finds that the relative strengths of the contributions of these contextual factors are constant across time under a VARBRUL analysis. Kroch 1989b asserts that this implies constant slopes across the contexts.

Fontaine 1985. Fontaine examines the loss of subject-verb inversion French and finds parallel falling curves for full NP subjects, pronominal subjects, and pro-drop (empty) subjects. The last category seems especially interesting since it makes reference to quite abstract grammatical structure.

Pintzuk 1991. Pintzuk finds roughly correlated changes in the rate of INFL-medial versus INFL-final base position in two clause environments in OE: matrix and subordinate (p. 334). Also, in this case, the measured distinction is a quite

Figure 4.3: Constant-rate effects in Noble 1985's data on *have* vs. *have got*.

Regression fit for Noble 1985's 'have/have got' data (T)emporally bounded vs. (P)ermanent

T: Intercept -61.48
T: Slope 0.0329
P: Intercept -62.37
P: Slope 0.0327
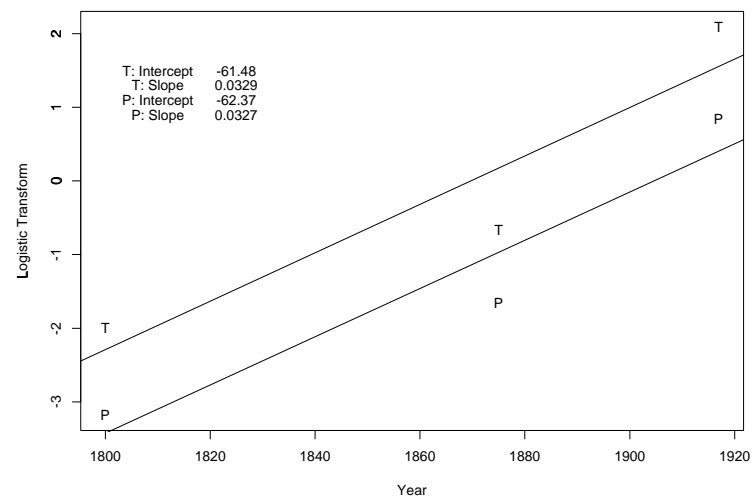
Regression fit for Noble 1985's 'have/have got' data (C)oncrete vs. (A)bstract

T: Intercept -56.12
T: Slope 0.03
P: Intercept -69.62
P: Slope 0.0368
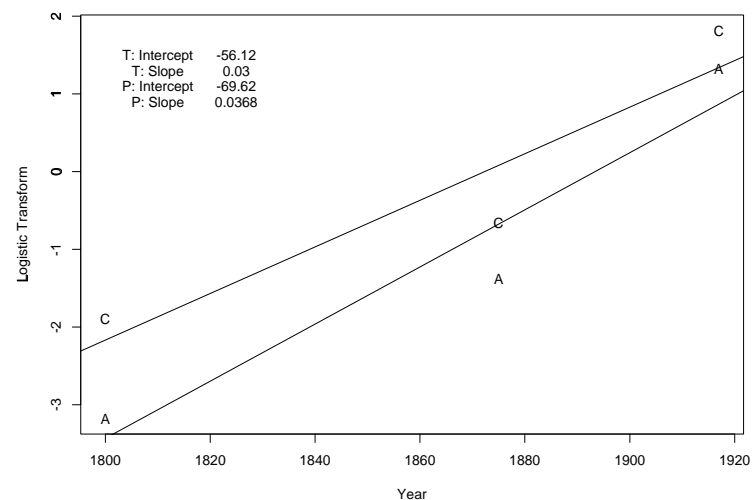
abstract one—it can be obscured by processes such as Extraposition, Heavy-NP Shift and Verb-to-Comp movement. Thus her quantitative correlations also indicate an abstract variable parameter.

Ellegård 1953. Ellegård 1953 collected data on the rise of periphrastic *do* in English. He found that during the period from about 1390 to 1560, periphrastic *do* rose fairly steadily in frequency in a variety of grammatical contexts: affirmative and negative declarative sentences; affirmative and negative questions. After 1560, the frequency of affirmative declarative *do* subsided, eventually becoming ungrammatical in non-emphatic contexts. Meanwhile, the uses of *do* in negative declaratives and in questions, went on to become obligatory in the absence of another auxiliary verb. This case is the focus of both Kroch 1989a and 1989b and is one of the most thoroughgoing of quantitative historical studies. Consequently, I examine it in more detail in Section 4.

4.2 Frequency linkage in a feedforward network

To see how the network models frequency-linkage effects, it will be useful to start with a simple example. For this purpose, I have created an abstract version of Noble's *have/have got* case study which can be modeled in a feedforward network.

Under the paradigm I outlined in Chapter 3, a network develops a grammatical representation by training on a corpus of examples. Concentrating on the interaction of the *have/have got* alternation with the concrete/abstract distinction, I generated a corpus in which two types of nouns ("N1" and "N2") occur in a variety of contexts ("C1", "C2", etc.). I mean the noun-types to correspond respectively to the categories "Concrete" and "Abstract". I mean the contexts to represent a variety of syntactic environments in which there is some systematic contrast between the behaviors of Concrete and Abstract nouns. Thus "C1" corresponds to "object-of-*have/have-got*", "C2" could be "possessor of another noun" which can be marked by the preposition *of* or the possessive *'s*, "C3" could be "theme of the verb *give*" which can either immediately follow a verb or occur after the direct object. I have purposefully avoided trying to make the noun-types and context-types reflect real properties of nouns and contexts in any detail in order to create an example that is very simple for the purpose of studying the network's predictions.

I coded each noun as a bit-vector with one bit "on" so the input representations for the two nouns would be orthogonal and of equal magnitudes (See

Chapter 3, Section 1.1). In order to be able to control the interaction of the nouns with contexts, I pre-specified the context-representations instead of letting the network learn them as one would do in a recurrent-network simulation. I used the distributed, equal-length codes shown in (67) for the context representations.

(67) Context Representations for *have* versus *have-got* simulation:

C1	1	1	0	0
C2	0	1	1	0
C3	0	0	1	1
C4	1	0	1	0
C5	0	1	0	1

For each each noun+context combination, I specified a behavior by giving a probability distribution over output features. For simplicity, I made the outputs associated with different contexts orthogonal as well. The resulting mapping is shown in Figure 4.4. In the simulations, an output was chosen for each input by sampling the relevant probability distribution shown in Figure 4.4. A key property of this mapping is that while N1, N2, and N3 all show similar kinds of behaviors in that they appear in many of the same contexts with roughly the same kinds of associated behaviors, there is a stark contrast between N1 and N2 on the one hand and N3 on the other: in all the contexts except for C1, N1 and N2 have identical behavior while N3 has quite different behaviors. Only in context C1 are the behaviors of the three elements the same. As I noted above, C1 is intended to correspond conceptually to the situation of occurring as the complement of the posession verb, *have/have-got*. The idea is to impose a change on the behavior of one noun, N2, in context C1 and see what effect this has on the behaviors of the other nouns (N1 and N3) in this same context. N2 is intended to correspond to some concrete noun, N1 to another concrete noun whose distribution is nearly identical to that of N2, and N3 to an abstract noun the distribution of which is somewhat differen form that of N2. C1 is intended to correspond to the environment, complement-of-the-possession-verb, *have/have got*. Note that under the mapping in Figure 4.4, all nouns, concrete or abstract, use *have* most of the time and *have got* only very infrequently. Thus this mapping generates a set of cooccurrence behaviors whose distributional properties

are analagous to the English distribution early in the period of emergence of *have got*.

Figure 4.4: A mapping from noun+context instances to distributions over behaviors.

Input	Context	Output-Distribution									
		1A	1B	2A	2B	3A	3B	4A	4B	5A	5B
N1	C1	96	4	0	0	0	0	0	0	0	0
N2	C1	88	12	0	0	0	0	0	0	0	0
N3	C1	96	4	0	0	0	0	0	0	0	0
N1	C2	0	0	40	60	0	0	0	0	0	0
N2	C2	0	0	40	60	0	0	0	0	0	0
N3	C2	0	0	90	10	0	0	0	0	0	0
N1	C3	0	0	0	0	80	20	0	0	0	0
N2	C3	0	0	0	0	80	20	0	0	0	0
N3	C3	0	0	0	0	0	100	0	0	0	0
N1	C4	0	0	0	0	0	0	30	70	0	0
N2	C4	0	0	0	0	0	0	30	70	0	0
N3	C4	0	0	0	0	0	0	70	30	0	0
N1	C5	0	0	0	0	0	0	0	0	100	0
N2	C5	0	0	0	0	0	0	0	0	100	0
N3	C5	0	0	0	0	0	0	0	0	10	90

The distribution in C1 is intended to simulate the distribution of various nouns over the behaviors, cooccurrence-with-*have* and cooccurrence-with-*have-got* in British English around the year 1800 (cf. Noble 1985).

I created a network with 7 input units, 8 hidden units, and 10 output units. None of the units had bias parameters. The activation function was the standard sigmoid (See Chapter 3, Section 1). Weight-update was performed after every pattern presentation.

According to the procedure described in Chapter 3, I trained the network on a corpus generated under the mapping in Figure 4.4 until it had learned the mapping well. Of particular interest at this point in the course of the experiment are the hidden unit representations associated with the three nouns when they occur in context C1, the "*have* versus *have-got*" context. The distances between

pairs of hidden-unit representations are shown in (68). Note that N1 and N2 are much closer to each other than either of them is to N3. This is because N1 and N2 exhibit virtually identical behavior in all contexts, while N3 has quite distinct behavior in 4 out of the 5 contexts.

(68)

Noun-Noun Pair	Separation in Hidden-Unit Space
(N2, N1)	0.370
(N2, N3)	0.724
(N1, N3)	0.810

After initial-training was completed, I post-trained the network[2] on a corpus generated under a distorted version of the mapping. The distorted mapping had all the same properties as the original mapping except that no examples of N1 or N3 in context C1 were included, and when N2 appeared in context C1, its behavior was drawn from the distribution in (69).

(69) Changed N2-in-C1 behavior:

N2 in C1   have   0%
            have-got   100%

The change in the distribution of N2-in-C1 is intended to correspond to a change in the frequency with which some particular concrete noun uses *have got* instead of *have*. The reason for withholding N1 and N3 from the training set while N2's behavior is being changed is that this provides a way of seeing how the change in N2 exerts an influence on N1 and N3 in virtue of the structure of the representation. Under the analysis given in Chapter 3, Section 8, we expect the change in N2 to produce a stronger correlated effect in N1 than in N3.

Figure 4.5 shows the results. Each curve traces the relative frequency at which the network expects (or generates) *have got* for one of the nouns during the course of post-training. For ease of implementation, I used end-goal training rather than stick-and-carrot, so the time-axis has been scaled to compute an approximation of the stick-and-carrot outcome.[3]

---

[2]i.e., trained it additionally—see Chapter 3, Section 6.1.

[3]The scaling is computed as follows: since the weight-changes are linear in the difference

Figure 4.5: Network relative frequency data for the *have* versus *have got* simulation
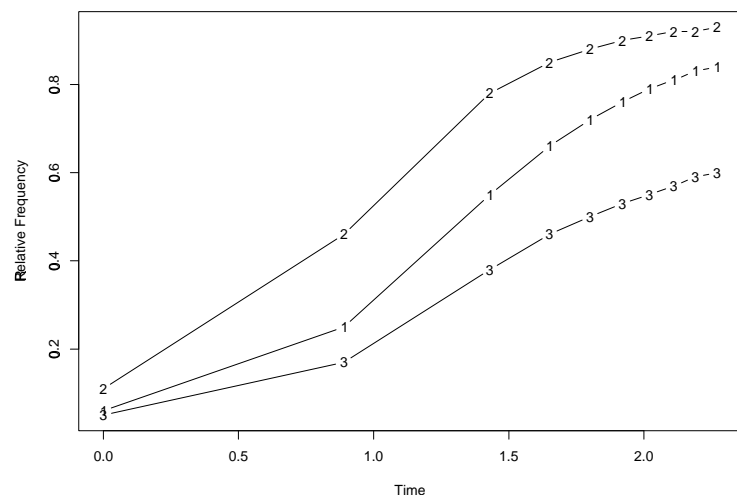
The curve marked "2" shows the relative frequency for N2. N2 is being given a target behavior of 100% *have got* and indeed during the course of post-training it rises up close to this value. N1, on the other hand, is not being trained to change at all. Nevertheless, because its representation is very similar to that of N2, the change imposed on N2 affects N1 in a corresponding fashion and its frequency curve rises almost parallel to N2's curve. N3, also not being trained, has a rising curve too, but its curve is only roughly parallel to N1 and N2's curves. Least square regression fits of the logistic transforms of the three curves are shown in Figure 4.6, along with the slope estimates. Note that the slopes of the curves for N1 and N2 are quite similar while the slope for N3 is quite different.

Thus, the simulation results confirm the analysis given in Chapter 3, Section

---

between target and output, each epoch-increment during post-training is multiplied by a scalar proportional to the increment-initial difference between the output and the target.

8, as indeed they should. The network model predicts a gradation of frequency-linkage effects: identically-distributed elements show perfect constant-rate behavior but contrastingly-distributed elements show frequency-linkage to the extent that their distributions are similar.

Figure 4.6: Fitted Logistic Transforms of the network relative frequency data in the *have* versus *have got* simulation.
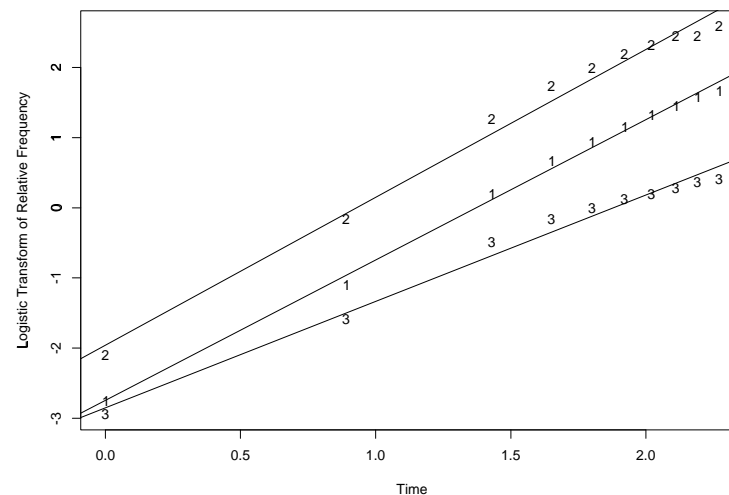
Curve	Slope	Intercept
N1	2.002	-2.747
N2	2.108	-1.960
N3	1.517	-2.851

4.3 Frequency-linkage in a recurrent network

4.4 Case-study: English periphrastic *do*

Ellegård 1953 studied the 14th-18th century rise of English periphrastic *do*. The main data from his study are reproduced in Figure 4.7 and plotted in Figure

4.8. Illustrative examples of the sentence types are given in Figure 4.4.

Figure 4.7: Percentages and tokens of periphrastic *do* from 1400–1700 (from Ellegård, 1953—p. 161).

Period	Aff. Decl.		Neg. Decl.		Aff. Q.		Neg. Q.		Neg. Imp.	
	+do	+/−do	+do	−do	+do	−do	+do	−do	+do	−do
1390–1400	6	45000	0	—	0	—	0	—	0	—
1400–1425	11	4600	0	177	0	10	2	15	0	52
1425–1475	121	45500	11	892	6	136	2	23	3	279
1475–1500	1059	59600	33	660	10	132	3	24	0	129
1500–1525	396	28600	47	558	41	140	46	32	2	164
1525–1535	494	18800	89	562	33	69	34	22	0	101
1535–1550	1564	19200	205	530	93	114	63	21	0	72
1550–1575	1360	14600	119	194	72	56	41	7	4	39
1575–1600	1142	18000	150	479	228	150	83	45	8	117
1600–1625	240	7900	102	176	406	181	89	6	65	119
1625–1650	212	7200	109	235	116	24	32	6	5	16
1650–1700	140	7900	126	148	164	43	48	4	17	16
Swift	5	2800	61	9	53	3	16	0	28	0

Note: The columns "Aff. Q." and "Neg. Q." contain counts only for what Ellegård calls "Adverb questions" and "Verb questions"—see Figure 4.4

4.4.1 Kroch's analysis

Kroch 1989a and 1989b is concerned with showing how the Competing Grammars model, when combined with a trenchant grammatical analysis, can make strong predictions about the relative frequency developments that Ellegård traced. Regarding the grammar, he argues (with Warner 1982, 1983 and Roberts 1985 and contra Lightfoot 1979) that the class AUX(iliary verb) already existed by the beginning of Middle English (ME) (see also Plank 1984, Lightfoot 1991). The main evidence, based on Roberts 1985, is that the modals of that time were used to express the subjunctive mood, and as such must have been analyzed as operators on clause. Moreover, Kroch notes, when periphrastic *do* first appeared it never took complements headed by modals or *have* or *be*, so there must have been a categorical distinction between these verbs and the rest.

Figure 4.8: Graph of the percentages from Ellegard 1953's study.
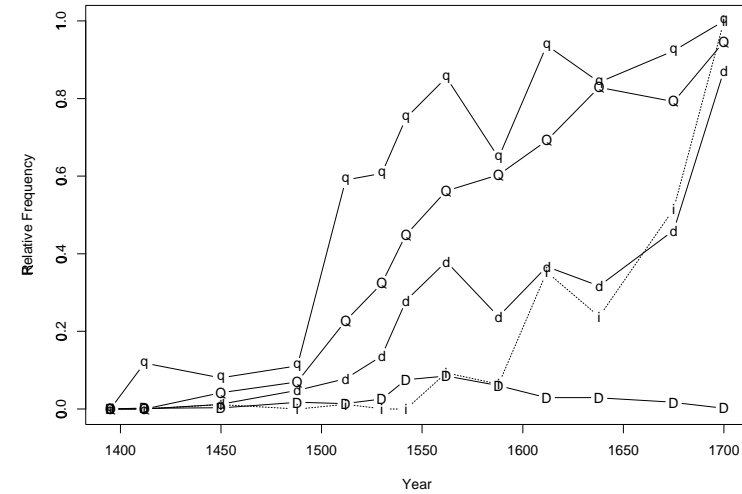
With Roberts, Kroch argues for a rule of Main-Verb to Infl (V→I) movement in ME. The primary evidence is that ME, like Modern French, placed weak sentence adverbs, negation markers, and "floated" quantifiers after the main verb in clauses where no auxiliary element is present, but before the main verb in sentences containing an auxiliary. The fact that ME main verbs inverted with the subject in questions ((72) and (73)) also provides evidence for the V→I analysis on the assumption that questions were formed by a rule of Subject-AUX inversion that applied after V→I movement. The V→I movement analysis is diagrammed in Figure 4.10. Under this analysis, periphrastic *do* in ME must, like the modals, have been an AUX element that moved obligatorily to INFL[4] for it was generally finite, generally occurred prior to weak sentence adverbs, negation, and floated quantifiers, and did not, to my knowledge, occur as the complement to another verb. For convenience, I will refer to weak sentence adverbs, negation markers, and "floated" quantifiers as "Spec(VP)" elements henceforth.

---

[4]Or perhaps was base-generated there.

Figure 4.9: Examples of sentence types for periphrastic *do* study. (Numbers in angle-brackets refer to Ellegård's text-numbering system, which uses the following syntax: <author:work:page:line>.)

(70) **Affirmative Declarative** (non-emphatic)
a. The king met them.
b. 15th c. the kinge, who takinge the hiringe...yn to his owne handes, dyd decyde and ende the controversie *Exeter gilds* <171::304:25> [E 59]

(71) **Negative Declarative**
a. 1564 ...spoile him of his riches by sondrie fraudes, whiche he perceiueth not. *William Bullein* <346::86:23> [Kb 15]
b. 1456 I asked licence to ryde yn to my contree, and my maistr dyd not graunt it. *Paston Letters* <314::67:6> [E 66]

(72) **Affirmative Question**
1. "O(bject) questions":
a. What said he?
b. What did he say?
2. "A(dverb) questions":
a. When came you?
b. 1505 Where doth the grene knyght holde hym? *Valentine and Orson* <304::97:15>
3. "V(erb) questions":
a. Went he?
b. Did he go? [E: 202]

(73) **Negative Question** (presumably mostly "V-questions")
a. 1613 has he not cause to weep...*Beaumont* <365::183:37> [E 206]
b. 1604 You sold me a rapier, did you not? *Ben Jonson* <360:I::1258> [E 205]

(74) **Negative Imperative**
a. Forget not to bring some.
b. 15th c.? doo not forgete to gyue some chalyces for the sacryfyce of god *Life of St. Bridget* <284::53:29> [E 131]

Under this analysis, as Kroch notes, the change that occurred betwen Middle and Modern English is that verb raising to INFL became restricted to auxiliary verbs. This hypothesis explains the changes in the behavior of main verbs in

Figure 4.10: Middle English V→I Movement. [From Kroch 1989b]
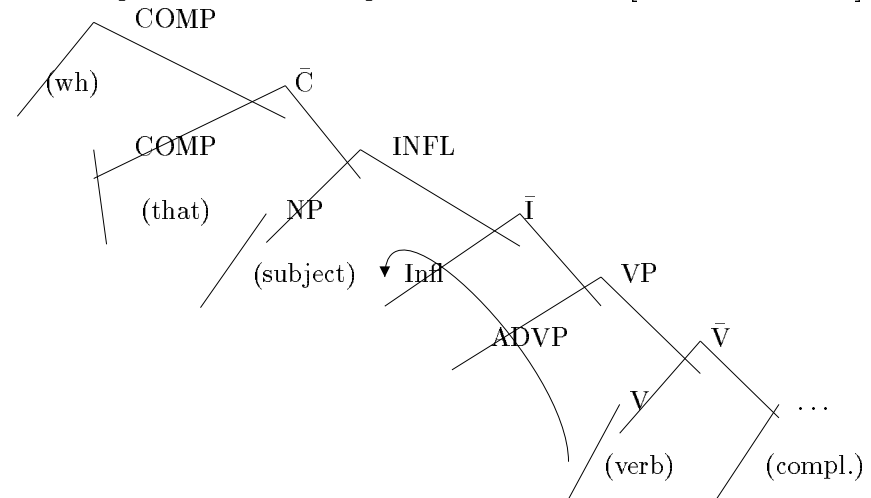
questions and in sentences with Spec(VP) elements. Moreover, it suggests a way of predicting the correlated developments among questions and negated sentences in Ellegård's data: we can posit a competition between a grammar with V→I movement and one without. I'll call the grammar with main-verb movement the "V→I Grammar" and the grammar without movement the "Aux-Support Grammar". In the Aux-Support Grammar, all sentences with main verbs also have an auxiliary element (either *do*, *be*, *have*, or a modal).

Something else will have to be said, of course, about unemphatic affirmative declarative *do* sentences: they rose in frequency with the other contexts until 1560 but then declined afterward and are now ungrammatical. Kroch 1989b does have a story for this part of the development. Before turning to a review of it, it will be useful to see how the Competing Grammars model makes accurate quantitative predictions about the developments in questions and negative sentences up until 1560.

I noted in Section 1 above that, in its purest form, the Competing Grammars model predicts that the relative frequencies of parametrically-linked constructions should be identical. That this claim is not very plausible in the case of

the *do* data can be seen by inspection of Figure 4.8:[5] the systematic ordering of the curves over time (Negative Question > Affirmative Question > Negative Declarative > Affirmative Declarative) would be unlikely if the frequency differences were merely due to noise.

Kroch 1989a argues that there are some plausible reasons for this ordering of the curves. First, a language without case-inflection, as English had recently become at the time of these developments, relies on prepositions and word order to distinguish among arguments of the verb. Kroch suggests that sentences in which each element is adjacent to its case marker are easier to process. Since the case-marking of direct objects is signalled by the verb itself, there should be a preference for having the direct object next to the verb if the grammar permits it.

Consider now the V→I grammar and the Aux-Support Grammar in this light. Under the V→I Grammar, the verb and the direct object are separated from each other by the negative in Negative Declaratives (e.g., *She saw not the raven.*), by the subject in Affirmative Questions (e.g., *Saw you the raven?*), and by both the negative and the subject in Negative Questions (e.g., *Saw you not the raven.*), while in Affirmative Declarative sentences, the verb is next to the object (e.g., *They saw the raven.*). By contrast, under the Aux-Support Grammar, the verb is next to the object in *every* case. Thus we might hypothesize that the extra processing load incurred by the Negative Sentences and Questions under the V→I Grammar created a stronger preference for using the Aux-Support Grammar when they occurred so the Negatives and Questions had higher frequencies. Second, Kroch 1989a notes that in both kinds of questions, the V→I Grammar generates a sequence of clitics one after another when both the subject and the object are clitic pronouns:.[6]

(75) Know ye me nat? <243:975:6> [Ka 232]

Kroch, Pintzuk, and Myhill 1982 show that *do* was used relatively more in sentences with both subject and object pronominal clitics than in sentences

---

[5] I report Kroch's confirming statistical analyses below.

[6] The V→I Grammar seems to have had an additional rule requiring the object to cliticize to the verb or the verb+clitic-subject if the object was pronominal. This may be why *not* appears subsequent to the object in (75).

which did not have this property. Kroch suggests that this may be because stacks of clitics are hard to pronounce and perceive. Since the Aux-Support Grammar does not generate clitic-stacks in Questions, this may have created a preference for using it relatively more in Questions. Kroch does not suggest any processing factors that might have *dis*-favored use of *do* in affirmative declarative sentences, but it is plausible to suggest that the production expense of including an element that did not enhance the information content of a sentence was such a factor. Given the great frequency of affirmative declarative sentences, this factor may have created a substantial burden for speakers by the time the Aux-Support Grammar had risen to its maximal level around 1560.

All of these arguments provide support for the claim that context-specific constraints create the contrasts between the rates of *do*-use in the different environments. Moreover, since the proposed differentiating qualities depend only on locally-computable properties of the sentences (not on their relationships to other sentences), these factors should remain constant through time. This suggests modelling them by including a time-invariant r-parameter in the Competing Grammars model. If we adopt the sigmoid assumption about the way frequencies change over time, this r-parameter could be either the slope or the intercept or some function of the two. Kroch proposes that it is the intercept that varies across contexts and that the slope is constant. This claim is called the "Constant Rate Hypothesis".

As far as I can tell, the choice of the Constant Rate hypothesis over the "Same Intercept" hypothesis or a combination-hypothesis is a stipulation. It does not follow from any of the assumptions made so far. Nevertheless, it is an empirically testable claim and Kroch finds support for it in the data on negative sentences and questions prior to 1560. The results of separate regression fits of the logistic transforms of these curve-segments, as reported in Kroch 1989b, are given in Figure 4.11.

Kroch notes that if a slope is computed for a single curve based on all the data under consideration here, then the probability of finding deviations from this common slope as large as those found here is greater than 95% under the assumption that noise is normally distributed ($\chi^2 = 0.504$). These observations (along with the studies by Oliveira e Silva 1982, Noble 1985, Fontaine 1985, and Pintzuk 1991 cited above) provide some empirical support for the constant-rate

Figure 4.11: Slope-Intercept values for the pre-1560 *do*-data on negatives and questions.

Negative declaratives		Negative questions		Aff. trans. adv. & yes/no questions		Aff. intrans. adv. & yes/no questions		Affirmative wh-object questions	
slope	int.	slope	int.	slope	int.	slope	int.	slope	int.
3.74	-8.33	3.45	-5.57	3.62	-6.58	3.77	-8.08	4.01	-9.26

hypothesis.

But we must return to the case of affirmative declarative *do*. Kroch 1989a cites as evidence in favor of his linkage-in-the-grammar claim the fact that affirmative declarative *do* rises with the other environments prior to 1560. But, as he recognizes, this means he must provide an account for the divergence of this curve after 1560. The obvious account to give is that a reanalysis took place in 1560 that made affirmative declarative *do* no longer grammatically related to the other environments. Kroch suggests that one piece of the putative reanalysis was that the parameter-setting allowing V→I movement was permanently turned off.

Unfortunately, this is a rather stipulative account unless it can be motivated by independent evidence. In fact, given the continued optionality of *do*-support in negatives and questions, there seems to be quite a bit of evidence *against* the claim that V→I movement was no longer employed after 1560. At the very least, banishing the rule at this point means that some other hypothesis has to be found about how the negative sentences and questions without *do* are generated so Kroch proceeds to offer some suggestions along this line. He proceeds by first using the quantitative data from the post-1560 periods to find out which slopes are and are not identical and then seeking distributional arguments to support a Competing Grammars scenario in which the constant-rate curves are generated by the same grammars and the different-slope curves are generated by different grammars. It would, in fact, be more convincing if he could use distributional arguments to motivate the claim that a particular set of competing grammars gives the most parsimonious account, and then show that the slope-values conform to this most-ideal grammar-arrangement. The trouble

with proceeding in the other direction is that the behavior of affirmative declarative *do* looks like prima facie evidence against the Constant-Rate hypothesis. It is therefore essential, if the Constant-Rate hypothesis is to be substantiated, that this case be predicted, rather than assumed, to conform.

Returning to the account, the slope-intercept values Kroch 1989b reports for the post-1560 period are given in (4.12). Statistical significance tests show that the slopes in the Question environments are not significantly different from each other but are significantly different from the slopes in the Declarative environments and the slope in the Negative Declarative environment is significantly different from the slope in the Affirmative Declarative environment.

Figure 4.12: Slope-Intercept values for the post-1560 *do* data.

Negative declaratives		Negative questions		Aff. trans. adv. & yes/no questions		Aff. intrans. adv. & yes/no questions		Affirmative wh-object questions	
slope	int.	slope	int.	slope	int.	slope	int.	slope	int.
0.497	-0.947	1.42	0.870	1.36	0.830	1.30	-0.329	0.743	-0.810

Kroch offers the following suggestions as to what reanalyses might have given rise to the observed correlations and non-correlations:[7]

(a) A grammar in which *not* fails to block Affix-Hopping became active in negative sentences for a period of time starting in 1560. [p. 33]

(b) Negative sentences where *not* followed the main verb were reanayzed as involving cliticization of *not* onto the end of the verb. [p. 34]

(c) In questions, V began moving directly to COMP without stopping in INFL. [p. 34]

But he offers very meagre independent evidence in favor of these analyses:

---

[7] In fact, there is a danger in using evidence of the form that two slopes are "not significantly different" to argue that the corresponding constructions have a particular structural relationship: two estimates of an r-parameter can be "not significantly different" either because they are truly constrained to be that way by the process that generated them or because there is not enough data to be confident about the meaningfulness of an observed difference in their behaviors.

(a) The possibility of placing *not* in front of the main verb was an option through most of Middle and Early Modern English but for a period starting around 1560, it became an especially "live option". [p. 33]

(b) It is around 1560 that *not* begins reducing to *n't* on auxiliaries.

(c) It may be agreement morphology in INFL that blocks V→COMP movement in general (Platzack and Holmberg 1990), and agreement morphology in English was getting pretty weak by 1560.

Apparently, after 1560, we no longer have one competition between two grammars, but two two-way competitions and one three-way competition among four grammars. This set-up is summarized in Figure 4.13.

Figure 4.13: Two two-way competitions and one three-way competition among four grammars under Kroch 1989b's analysis of post-1560 *do*.

Environment	Grammar Competition
Negative Questions	Aux-Support vs. V→C Movement
Affirmative Questions	Aux-Support vs. V→C Movement
Negative Declaratives	Aux-Support vs. Affix-Hopping vs. *not*-enclisis
Affirmative Declaratives	Aux-Support vs. Affix-Hopping

Evidently, the way Post-Verbal *not*-enclisis helps the story is by making the competition in the Negative Declarative environment different from the competition in the Affirmative Declarative environment. If these competitions were the same, then we would expect the slopes to be the same in these two environments. But that the environments of competition are defined essentially intuitively: we say constructions are in competition if they express pretty nearly the same meaning. This means that it is not very hard to find ways of introducing new grammars into the competition to rescue it from making wrong predictions. For example, suppose another quantitative study based on a larger data base shows that Negative (yes/no) Questions have a slope that differs significantly from other Question environments during the later part of the *do* shift. Well, it turns out that there is evidence from around 1700 that Negative Questions may have had a very minor option of allowing Subject-Aux inversion to happen before Neg-Contraction (*Did not you eat the fish?*) instead of after it (*Did you not eat the fish?*) [Swift, *Letters to Stella* 1710]. This optionality implies an additional competing grammar. If there's an extra competing grammar in the Negative Question environment then all bets are off about Constant-Rate behavior. Unfortunately, this makes the theory rather hard to disprove.

One additional stipulation is needed to get the model to working right: to explain why Affirmative Declarative *do* declined from 1560 onward, we must assume that just at the point where V→I Movement was lost, Affix Hopping and *do*-support entered into a competition which was eventually won by Affix Hopping. As Kroch 1989b (p. 31) notes, there is no obvious explanation for this coincidence.

And there is even a possible problem with the model's predictions. The post-1560 slopes of all the Negative and Question environments are quite a bit lower than their pre-1560 slopes. Although Kroch is not explicit about how the slopes are determined, one would think that under the Competing Grammars hypothesis, contexts which favor use of *do* would *increase*, not decrease, their rates of transition once they are delinked from the laggard context.

In sum, the current Competing Grammars account of the *do* episode seems a bit stipulative. Its main stumbling block is the fact that affirmative declarative *do* changes with the other environments before 1560 but diverges after. The problem is that there is not much of the normal kind of evidence for a reanalysis at that point.

Consider, now, the following alternative account:

> Alternative *do* Hypothesis. Affirmative declarative *do* reached a *quantitative* threshold around 1560 which precipitated a reorganization of the grammar.

If unemphatic affirmative declarative *do* is dispreferred from a processing standpoint because of its redundancy, while the negative and question uses are preferred, as Kroch has argued, and if the environments are structurally linked with some kind of measurable-strength bond at the beginning of the history of the periphrasis, then it is not implausible that they should rise together to a point at which the processing burdensomeness of using affirmative declarative *do* became unbearable, and then diverge from that point on. The problem with implementing such a hypothesis in the Competing Grammars model is that the

model provides no way of talking about strengths of structural bonds. By contrast, this is something the Connectionist model is well-equipped to do. In the next section, I present a tentative simulation result showing how the main characteristics of the periphrastic *do* episode are predicted by a network model which makes only two case-specific assumptions: (1) the modals and *do* had a distinct AUX distribution pattern at the beginning of the history of the periphrasis and (2) processing pressures drew Negative and Question environments toward full *do*-use and the Affirmative Declarative environment toward full *do*-absence.

4.4.2 Network Simulation

I used the probabilistic grammar shown in Figure 4.14 to generate a corpus. This grammar is intended as an approximation of key properties of the late-14th century distribution of *do* and the modal verbs. A sample corpus fragment is shown in Figure 4.15.

I trained a recurrent network with 20 input units, 8 hidden units (fully connected), and 20 output units on a 2000 sentence corpus generated by this grammar (3 time-steps of the hidden units were unfolded to obtain an approximation of the total gradient in backpropagation—see Chapter 3, Section 3). I then post-trained the corpus on the output of the distorted grammar indicated by the arrows in Figure 4.14.

A trace of the log likelihoods of the different sentence-types during post-training is shown in figure 4.16. The correct predictions made by this model are:

> (i) The competing grammar-external pressures interact with the structural bond between the four types of *do* environments to produce the pattern in which Affirmative Declarative sentences rise partway with the others and then fall.

> (ii) The structural bond between Affirmative Declaratives and the other environments is not eliminated at the point of turn-around (Epoch 3, in the simulation). Instead it continues to bind the environments but with ever-weakening strength. What happens at the turning-point, is that it becomes more effective from the standpoint of satsifying the processing constraints (i.e. reducing the error) to

Figure 4.14: A grammar approximating the late-14th century distribution of *do* and modal verbs.

S : SD p	1.00		
SD : AD	0.25		
SD : ND	0.25		
SD : AQ	0.25		
SD : NQ	0.25		
NQ : V NP not	0.68	→	0
NQ : do NP not V	0.02	→	70
NQ : Mod NP not V	0.30		
AQ : V NP	0.68	→	0
AQ : do NP V	0.02	→	70
AQ : Mod NP V	0.30		
ND : NP V not	0.68	→	0
ND : NP do not V	0.02	→	70
ND : NP Mod not V	0.30		
AD : NP V	0.68	→	70
AD : NP do V	0.02	→	0
AD : NP Mod V	0.30		
NP : dogs	0.17		
NP : birds	0.17		
NP : barrows	0.17		
NP : rakes	0.17		
NP : plums	0.16		
NP : timbers	0.16		
V : sing	0.17		
V : run	0.17		
V : live	0.17		
V : break	0.17		
V : quiver	0.16		
V : bounce	0.16		
Mod : may	0.25		
Mod : should	0.25		
Mod : can	0.25		
Mod : will	0.25		

begin disintegrating the bond between Affirmative Declarative and

Figure 4.15: A sample corpus fragment for the *do* simulation.

barrows run not p can dogs bounce p quiver barrows not p break barrows p
rakes bounce p barrows do not run p dogs can not bounce p live dogs not p
barrows run not p barrows break p sing rakes p do rakes not live p sing plums p
plums sing not p may dogs sing p timbers quiver not p dogs quiver p sing dogs
p quiver birds p rakes should sing p birds run p can barrows live p live rakes p
dogs run not p barrows should not sing p run dogs not p plums bounce not p
live birds p rakes quiver p plums run p plums break not p quiver birds not p
plums sing not p may birds not sing p. . .

(p stands for "period")

the other environments than to continue moving all the environments
simultaneously toward high *do*-use.

Problems with this model are

(i) It eventually orders the environments as NQ > ND > AQ > AD
rather than NQ > AQ > ND > AD.

(ii) I used equal frequencies among all types of sentences. This prob-
ably makes the affirmative-declarative divergence more pronounced
than it would be if I used realistic proportions of sentence types.

(iii) The predictions of the model will be affected by incorporating
information about how the modals were changing during the same
period of time. It is pretty clear that they were changing during this
time but, as far as I know, nobody has collected any quantitative
data on the matter.

(iv) The model does not predict some of the post-1560 wavering in
the negative sentences which Ellegård and Kroch believe are signif-
icant.

Despite these shortcomings, the network model is appealing in its simplicity,
especially when compared with Kroch's Competing Grammars account. It has
the additional desirable property that it is explicit about how the analyses
are determined from the data. Because it depends on the continuity of the

representation space to model variable-strength grammatical bonds, its success
in predicting the basic properties of the *do*-developments lends support to the
Restrictive Continuity hypothesis.

Figure 4.16: Network simulation of the rise of periphrastic *do* in Affirmative
Declaratives (D), Negative Declaratives (d), Affirmative Questions (Q), and
Negative Questions (q).
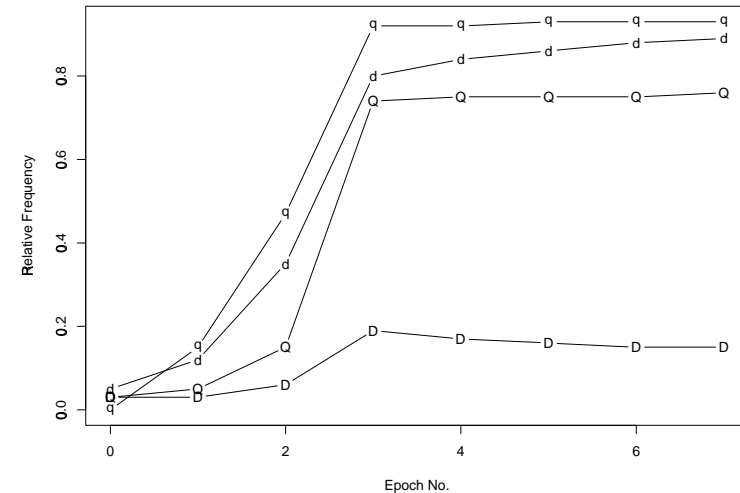
Figure 4.17: Logs of the absolute probabilities for the network *do* simulation.
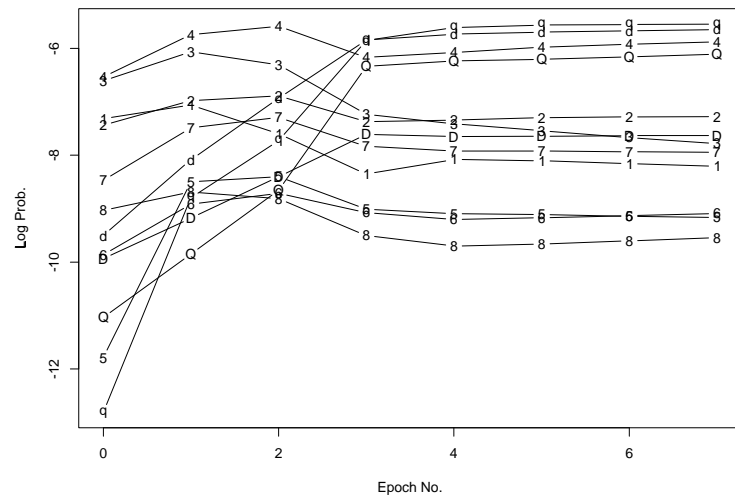
Key

q	Negative Questions with <do>
Q	Affirmative Questions with <do>
d	Negative Declaratives with <do>
D	Affirmative Declaratives with <do>
1	Negative Questions with a main verb only (<run>)
2	Affirmative Questions with a main verb only (<run>)
3	Negative Declaratives with a main verb only (<run>)
4	Affirmative Declaratives with a main verb only (<run>)
5	Negative Questions with a modal (<can>)
6	Affirmative Questions with a modal (<can>)
7	Negative Declaratives with a modal (<can>)
8	Affirmative Declaratives with a modal (<can>)

Chapter 5

Reanalysis-anticipatory Frequency Change (Q-Divergence)

5.1 Q-Divergence

Under current theories of sentence structure, reanalyses are hard to predict. I noted in Chapter 1 that reanalyses are often accompanied by quantitative changes which, from an impressionistic point of view, make the changing elements gradually less like their source types and more like the new types they eventually adopt. I labelled this phenomenon, "Q-Divergence", and noted that if grammars can be made to encode the relevant similarity structure, then grammatical theory may be more useful in making predictions about reanalyses.

Here, I report on two case-studies which provide evidence for Q-Divergence: development of Degree Modifier *sort/kind of* and development of *be going to* as a Future Marker in Middle and Modern English. After reviewing the history for each case I show how the recurrent network trained on word-prediction predicts the observed correlations between quantitative and categorical changes.

5.2   Case-studies

5.2.1   English *sort/kind of*

ModE has a widespread use of the expressions *sort of* and *kind of* in which *sort* and *kind* are plausibly analyzed as nouns and *of* is a preposition:

(76) a. They found some sort/kind of cactus on the rim.

   b. What sort/kind of knife do you need?

   c. I don't really go for that sort/kind of thing.

But there is also a more colloquial usage in which *sort of* and *kind of* are best described as DegMods:[1]

(77) a. We are sort/kind of hungry. [be □ AdjP]

   b. He sort/kind of hemmed and hawed. [NP □ VP]

   c. She ran sort/kind of a shady operation. [V □ NP]

   d. It was a sort/kind of dense rock. [Det □ Adj N]

   e. * Sort/Kind of she laughed. [□ NP VP]

(78) a. We are rather/somewhat/really hungry. [be □ AdjP]

   b. He rather/somewhat/really hemmed and hawed. [NP □ VP]

   c. She ran rather/somewhat *(of)/really a shady operation. [V □ NP]

   d. It was a rather/somewhat/really dense rock. [Det □ Adj N]

   e. * Rather/Somewhat/Really she laughed. [□ NP VP]

A type of particular importance in the historical development is that of (77d), which is ambiguous in ModE between the two parsings (79) and (80).

(79) [a [[kind$_N$] [of [dense rock$_{NP}$]$_{PP}$]$_{N'}$]$_{NP}$]
(80) [a [[[kind of$_{DegMod}$] [dense$_{Adj}$]$_{AdjP}$ rock$_{N'}$]$_{NP}$]

The noun *kind* is a Germanic word and has been in English since the oldest times. The noun *sort* was borrowed from French in the Middle English period (first MED example 1384). *Kind of* as a Noun-Prep(osition) sequence is attested in ME (81) but does not grow common until E(arly) Mod(ern) E(nglish). Noun-Prep *sort of* first appears in earliest EModE (82).

---

[1]Bolinger 1972 (p. 239) notes that adverbial *sort of* and *kind of* are not restricted to use with verbs that accept canonical DegMods: Thus there are contrasts like:

   He sort of/kind of/*somewhat/*rather swam over and took hold of the side.

(81) a. 1382 A nette sent in to the see, and of alle kind of fishis gedrynge. Wyclif *Matt.* xiii. 47 [OED]

   b. C. 1470 He lett for no kind of thyng. K. Estmere 193, in *Percy's Rel.* [OED]

(82) a. 1529 Let vs now see whether sort of these twayn might take most harme. More *Suppl. Souls Wks.* 329/1 [OED] (*whether* = 'which')

   b. 1500–1570 …he cute downe a greate sorte of brakes, and that was my beed for a tyme. *Troubles of Mowntayne*, p. 211 (*brakes* = 'ferns', *beed* = 'bed')[HELS][2]

   c. 1560 I knowe that sorte of men ryght well. *Daus tr. Sleidane's Comm.* 63 [OED]

By the late 16th century, significant percentages of the instances of *sort* and *kind* occur in the Noun-Prep collocations, *sort of* and *kind of*. By contrast, the earliest records of unambiguous DegMod *sort of* and *kind of* do not occur for another two centuries:

(83) a. 1804 I kind of love you, Sal—I vow. T. G. Fessenden *Orig. Poems* 100 [OED] [NP □ VP]

   b. 1830 I was kind of provoked at the way you came up. *Massachusetts Spy*, Jan. 6, 1/5 [OED] [BE □ AdjP]

   c. 1833 It sort o' stirs one up to hear about old times. J. Hall, *Legends West*, p. 50 [OED] [NP □ V]

How did DegMod *sort of* and *kind of* arise? Given the fact that not all sequences of the form Noun+*of* serve a double function as degree modifiers in English, the Q-divergence hypothesis predicts that there must have been a period during which *sort/kind of* diverged quantitatively from the behavior of Noun+*of* constructions in general. In particular, it should have diverged in such a way that its characteristic transition probabilities came to resemble those of canonical Degree Modifiers.

During the 17th–19th centuries, canonical Degree Modifiers like *quite*, *somewhat*, and *rather* behaved fairly similarly to the way they behave today. They

---

[2]HELS = Diachronic Helsinki Corpus of English Texts—See Appendix

could modify verbs, adverbs, and predicate adjectives (84). They could also modify adnominal adjectives in constructions of the form <(Det) DegMod Adj N>. Illustrative examples are given in 85.[3] As in modern English, there is no evidence that Degree Modifiers could modify bare nouns.

(84) a. 1695 I was <u>quite</u> Estrang'd from my Husband. Defoe, *Moll Flanders*, Match 15 [*quite*]. [BE □ AdjP]

b. 1780 Sir George Rodney's success has <u>somewhat</u> lessened their force. *Mirror* No. 82. [OED] [NP Aux □ V]

c. 1855 ...said I, <u>rather</u> digressively. *Moby Dick*, p. 86. [□ Adv]

(85) a. 1695 For they now liv'd in a <u>quite</u> different Place from where they were before. Defoe, *Moll Flanders*, Match 12 [*quite*].

b. 1779 The contempt in which, to a <u>somewhat</u> unreasonable degree, he holds modern refinement. *Mirror* No. 61. [OED]

c. 1855 That's a <u>rather</u> cold and clammy reception in the winter time, ain't it Mrs Hussey? Melville, *Moby Dick*, p. 65.

Thus, a major property distinguishing Noun-Prep collocations from Degree Modifier expressions during this period was the fact that Noun-Prep collocations could take either a following Adjective or following Adjective+Noun, while Degree Modifier expressions had to be followed by an Adjective if they were used in a noun phrase and could also be predicate-adjective, verb, or adverb modifiers. Under the Q-Divergence hypothesis, we might expect that in the environment which Noun-Prep sequences share with Degree Modifier sequences (between Det and N in NP), *sort of* and *kind of* should have shifted quantitatively to become more like Degree Modifiers prior to the late 18th century when they first started showing unequivocal signs of being Degree Modifiers.

The top curve in Figure 5.1 traces the frequency with which sequences of the form [Det *sort/kind of* (Adj) N] appeared with the Adjective present between the mid 16th and the mid 19th centuries. Indeed, there appears to be a substantial upward trend. Examples of cases with the Adjective present are given in (86)

---

[3] I have the impression that constructions in which the Degree Modifier occurred between the Determiner and the Adjective (*a quite different place*) were more common relative to constructions in which the Degree Modifier precedes a <Det Adj N> sequence (*quite a different place*) than they are today.

Figure 5.1: Disproportionate rise in the rate of use of adjectives after *sort/kind of* during the period 1500–1900.

S = *sort/kind of*, O = *of* alone

Work	Year	Det... sort of Adj...N	Det... sort of ...N	Det... kind of Adj...N	Det... kind of ...N	Comb. Percent (S)	Det... Adj...N	Det... of ...N	Percent of (O)
HELS ENE1	c.1535	0	8	1	21	3%	0	117	0%
HELS ENE2	c.1605	0	7	4	35	10%	3	89	3%
HELS ENE3	c.1675	3	28	4	41	10%	5	164	3%
Defoe Moll Fl.	c.1695	1	21	4	43	8%	2	53	4%
Hume Var. Treat.	c.1743	1	2	21	187	12%	5	102	5%
Austen Emma	c.1796	28	132	3	9	22%	2	73	3%
Melville Moby D.	c.1855	27	113	4	11	25%	10	95	11%

through (91). It is important to note that the tabulation and graph in Figure 5.1 include every instance of the sequences *sort of* and *kind of* that occurred in the corpuses I examined. Thus, the observed change was a substantial change in the overall distribution of *sort/kind of*. Because some documents have almost exclusively *sort of* and some have almost exclusively *kind of*, but both participate in a similar trend when substantial numbers are present, it seems likely that the two forms are essentially one item syntactically with two different phonological instantiations. Therefore, I have added the counts for the two types together to

form the graph of Figure 5.1, figuring that this gives a more robust estimate of the quantitative properties of the trend.

(86) 1570–1640 Their finest and best, is a <u>kind of</u> course red cloth... *True Report*, p. 41 [HELS]

(87) 1570–1640 ... the blunt edges of it upon a <u>kind of</u> large Pin-cushion cover'd with a course and black woollen stuff. Boyle, *Electricity and Magnetism.* [HELS]

(88) 1640-1710 ... pretty good except 4 or 5 miles they call the Severalls, a <u>sort of</u> deep moore ground and woody. *Great Journey*, p. 144 [HELS]

(89) 1743 But in such questions as the present, a hundred contradictory views may preserve a <u>kind of</u> imperfect analogy; Hume, *Dialogues Concerning Natural Religion*, Pt. 8, Para 1/12, p. 182.

(90) 1796 "I have no doubt of it." And it was spoken with a <u>sort of</u> sighing animation, which had a vast deal of the lover. Austen, <u>Emma</u>, Vol. I, Chap. 6, p. 43.

(91) 1855 Yet was there a <u>sort of</u> indefinite, half-attained, unimaginable sublimity about it... Melville, *Moby Dick*, p. 11

One might, of course, question the significance of this trend on the grounds that it could be the result of a fortuitous choice of texts. The first three samples are drawn from the Helsinki corpus, an on-line collection of texts that spans a multitude of genres (Law, Handbooks, Science, Philosophy, Fiction, etc.; see Kytö 1991). The later texts are all literary. Perhaps the authors of the later texts happen to use an unusually large number of adjectives, and the rise after *sort/kind of* is merely a subcase of this general trend. Indeed, as indicated by the lower curve in Figure 5.1, there is a slight upward trend in adjective use in the environment [Det...of □ N] (cases in which "..." includes *sort/kind of* excluded) across the texts, but the trend is not nearly as large as in the environment [Det...sort/kind of □ N]. A t-test with 5 degrees of freedom confirms the hypothesis that the difference between the two curves has non-zero slope at

a significance level of 0.05. Therefore it would seem that the upward trend in the use of *sort of* and *kind of* is not merely due to fortuitous text-choice.

These results are encouraging in that they provide some empirical evidence in favor of the Q-Divergence hypothesis. However, the hypothesis will become much more believable if it can be shown that there is a grammatical representation that has the properties we would normally expect of a grammatical representation and also predicts the correlation between the initial quantitative change and the categorical development. At a minimum, a grammatical representation ought to predict the range of possible morpheme-sequencings that a language allows. Here, the network model is useful.

I ran a simulation with a recurrent network of the type discussed in Chapter 3, Section 3 (3 layer, feedforward except complete interconnectivity in the hidden layer). It had 20 input and output units and 8 hidden units. In training, I used an approximation of the total gradient by unfolding 6 time-steps of the recurrent connections.

Figure 5.2: An approximation of the pre-19th century change in the distribution of *sort of*.

S : NP VP p	1.00	VP : V'	0.40	P : of	0.60
		VP : is	0.20	P : from	0.40
NP : N"	0.00	VP : is AP	0.20		
NP : Det N"	1.00 → 0.85	VP : is NP	0.20	Adj : sumptuous	0.34
NP : a sort of Adj N"	0.00 → 0.15			Adj : dense	0.33
		V' : Vint	0.50	Adj : soft	0.33
N" : N'	0.60	V' : Adv Vint	0.50		
N" : AP N'	0.40			Adv : rather	0.34
N' : N	0.80	Det : this	0.34	Adv : merely	0.33
N' : N PP	0.20	Det : a	0.33	Adv : really	0.33
PP : P NP[part]	1.00	Det : the	0.33		
				Vint : melts	0.34
NP[part] : N"	1.00	N : block	0.25 → 0.32	Vint : rolls	0.33
NP[part] : Det N"	0.00	N : cheese	0.25 → 0.31	Vint : dissolves	0.33
		N : marmalade	0.25 → 0.31		
AP : Adj	0.50	N : sort	0.25 → 0.05		
AP : Adv Adj	0.50				

I trained the network on the word-prediction task with a 1000 sentence corpus generated by the grammar in Figure 5.2 (ignoring, for the time being, the arrows and the alternative probability values they point to). Under this grammar, <sort of> only occurs as a Noun-Prep sequence; the frequency of the form <a N of Adj N> is relatively low and is the same whether N is <sort> or any other noun. At the end of 300 passes through the corpus, I compared the network's performance at each juncture between words with the grammar-derived

probability distribution at that juncture. On average, the network was off by only approximately 10.4 percentage points[4] so it seems reasonable to say that it learned the grammar's structure well. I then created a new grammar, G', which was identical to grammar G, except that the frequency of NPs of the form <a sort of Adj N> (a now-ambiguous construction) was greatly increased and the frequency of <sort of> as a Noun-Prep sequence was greatly reduced (as indicated by the arrows in Figure 5.2). I then trained the network additionally for a few epochs on a 1000 sentence corpus generated by this new new grammar (the "post-training" process—see Chapter 3) and tracked the likelihoods of various sentences of interest.

The results are shown in Figure 5.3. Each curve traces the logarithm of the probability of observing a particular string of words when the network is treated as a sentence generator. That is, each curve traces the L-value assigned to the string by the network—see Chapter 3, Section 6.2. The curve in the middle of the graph, marked "N", shows the progression of the ungrammatical but not completely atrocious sentence (93).[5] Note that the likelihood of this sentence does not change much over the course of post-training. I call this sentence a "near-grammatical" sentence because, although it is not generated by the grammar, many of its subparts are well-formed. We can interpret the likelihood of this near-grammatical sentence as a threshold such that sentences with higher likelihood count as grammatical and sentences with lower likelihood count as ungrammatical (see Chapter 3, Section 6.2).

The interest of the simulation is that the curve marked "S" which corresponds to an unambiguous example of DegMod <sort of>, starts out far below the grammaticality threshold at the beginning of post-training and then rises above it as post-training proceeds. This result is encouraging: the network

---

[4]i.e., The *root mean squared error*, given in (92) was 0.104.

(92)

$$RMS = \sqrt{\frac{\sum_n \sum_k (t_k - o_k)^2}{n}}$$

(k indexes output units and n indexes junctures between words.)

[5]<p> stands for "period".

Figure 5.3: Simulation result: rise of DegMod *sort of* in conjunction with increased use of <Adj> in the environment <Det *sort of* (Adj) N>.
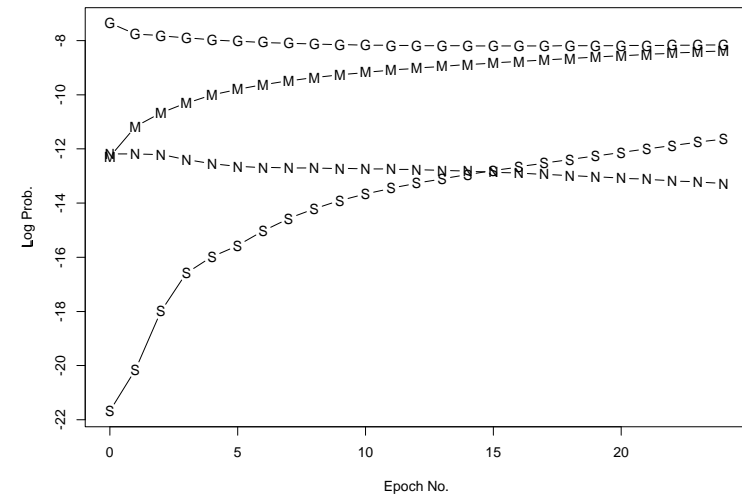
(93) the soft cheese really sumptuous p [(N)ear Grammatical]

(94) a sort of dense cheese rolls p [a(M)biguous]

(95) the block is sort of dense p [degree modifier (S)ort-of]

(96) the block is rather dense p [(G)rammatical]

predicts a correlation between the rise in the frequency of the grammatical ambiguous case ("M") and the innovation of the DegMod use of *sort of*. Note that network never encounters any examples of unambiguous DegMod <sort of> in either initial-training or post-training.

Why is there a correlation between the rise in the frequency of *sort of* in the ambiguous environment and the rise of DegMod *sort of*? Suppose the initially-trained network has been exposed to a sequence of words generated by grammar G and ending with the string, *a block of* as shown in Figure 5.4. Under grammar G, the likelihood that the next word will be a Noun is 60% while the likelihood that it will be an Adjective is 40%. The same is true of a sequence ending with

the string *a sort of*. On the other hand, if the sequence ends with the string *a somewhat*, the likelihood of a following Noun is 0% while the likelihood of a following Adjective is 100%. Thus, in a corpus generated by grammar G, the distribution of successors of *sort of* in this environment is quite different from the distribution of successors of *somewhat*. Now consider these same sequences under grammar G'. The distribution after *a block of* and *a somewhat* stays the same. But the distribution after *a sort of* shifts as shown by the dotted lines in Figure 5.4. In terms of vector distance, *a sort of* has come to look much more like *a somewhat* than *a block of*. Because the network groups corpus junctures into clusters on the basis of the similarity of their next-word behavior, and similarity is an increasing function of vector distance, post-training on the output of G' causes the network to expect increasingly similar behavior from *sort of* and canonical DegMods. Consequently, it increasingly expects *sort of* to appear in unambiguous DegMod environments. Note that this correlation does not follow on a model in which classification is based on categorical combination possibilities alone. From the standpoint of such a model the shift from grammar G to grammar G' has no structural consequences.

Figure 5.4: Change in the distribution following *sort of*.
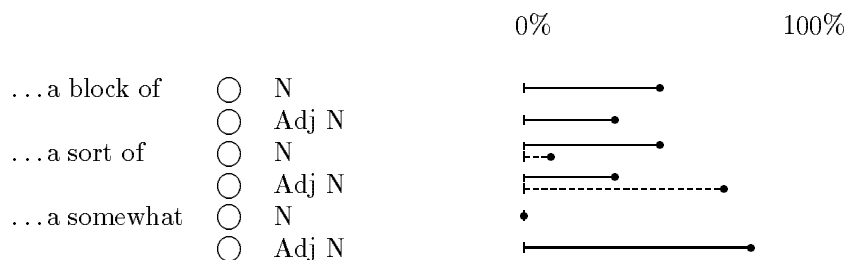
In sum, the model does a reasonable job of predicting the relevant general morpheme-sequencing facts for the period of the language in question. Moreover, the predicted correlation between the rise of the frequency of <sort of > in ambiguous environments and its appearance in the novel Degree Modifier environment, [NP be sort of AdjP], is consistent with the observed Q-divergence effect.

This case is historically-speaking, somewhat unusual. It involves a development from essentially nominal material (Noun and Preposition) into essentially adverbial material (Degree Modifier), which seems to be less common than noun-to-adposition and noun-to-verb transitions (see Hopper and Traugott 1993). I also know of no other language histories which evidence the particular semantic shift involved here (roughly "type of" > "approximately"). Therefore, to lend plausibility to the claim that Q-divergence is a property also of more canonical grammaticalization, I turn next to an examination of a very typical case: the development of the future auxiliary uses of *be going to*.

5.2.2 English *be going to*

Be going to is famous as an example of the cross-linguistically common development from Motion Verb (97) to Future Auxiliary (98) (Danchev *et al.* 1965, PéRez 1990, Bybee, Pagliuca, and Perkins 1991, Hopper and Traugott 1993). In accord with PéRez 1990, I find evidence for a subsidiary development during its Auxiliary period from something like Equi status (98a), where *be going to* plausibly ascribes intention to its subject, to Raising status, where intention is not normally part of the meaning (98b). I use the term "Auxiliary *be going to*" here to refer to precisely those instances in which *be going to* takes a VP complement and the motion meaning is absent.

(97) a. c. 1550 My lord…who was then going to the North…*Jour. Edw.*, p. 353 [HELS]
b. c. 1590 Hark, the kings and princes…are going to see the Queen's picture. Shakespeare, *A Winters Tale*, V ii. [OXF]

(98) a. c. 1695 He was going to reply…but he heard his sister coming, Defoe, *Moll Flanders* (match 8) [AIR]
b. c. 1865 Do you think it's going to rain? Carroll, *Alice Through the Looking Glass*, (match 3) [AIR]

In this section, I first make it explicit what I mean by Equi and Raising status. Then I review the historical development using this distinction and

other standard categorical tools to interpret the data. This review produces several pieces of evidence for the Q-Divergence hypothesis. I then show how the recurrent network again predicts the observed Q-divergence effects while simultaneously predicting the normal grammatical distributional properties of the language.

5.2.2.1   Equi and Raising Verbs

Criteria often considered diagnostic of Raising status for English verbs include ability to take "dummy" subjects (*It seemed/appeared/tended to rain.*, *There seemed to be a thundercloud on the horizon.*) and ability to intervene in idioms (*The cat seems to be out of the bag.*). Equi verbs (e.g., *want to, intend to, try to, yearn to* etc.) contrast in both regards. (See, for example, Rosenbaum 1967, Klein and Sag 1985, McCawley 1988). We can add the fact that Raising verbs permit inanimate subjects while Equi verbs do not, except in an anthropomorphic sense (e.g., *The table seems to be unpainted. # The table wants to be unpainted.*). A good summary of the constraint imposed by Equi verbs is that they require "sentient" subjects. Raising verbs, by contrast, simply put no constraints on the type of their subject (see Pollard and Sag 1993 for an HPSG formalization of an analysis along these lines).[6]

The formal contrast between Equi and Raising is clearly related to the semantic contrast between Root/Deontic and Epistemic modality. Notes Sweetser 1990 (see also Coates 1983, Palmer 1990),

> Linguists have characterized as *root* those meaning which denote real-world obligation, permission, or ability...; and as *epistemic*

---

[6] On this view the idiom criterion is only coincidentally diagnostic of the relevant distinction, and it it is expected to give inconsistent results if there exist idioms with potentially sentient subjects. Indeed, Nunberg, Sag, and Wasow (Forthcoming: 27) note that English has idioms that can both take sentient subjects and combine with Equi verbs:

(99) a. Every dog expects to have its day.
      b. An old dog never wants to be taught new tricks.
      c. Every lion prefers to be bearded in his den.
      d. Birds of a feather like to flock together.
      e. The early bird hopes to get the worm.
      f. They didn't tell me themselves, but they persuaded a little bird to tell me.

those which denote necessity, probability, or possibility in reasoning. [p. 49]

Since obligation, permission, and ability are usually ascribed to sentient beings, root modals, like Equi verbs, occur primarily with sentient subjects. On the other hand, necessity, probability, and possibility are often predicated of events involving non-sentient as well as sentient actors so Epistemic modals, like Raising verbs, occur frequently with nonsentient as well as sentient subjects. But it is clear that the two sets of notions do not coincide, for a modal with Raising properties can have Root meaning (100).

(100) a. I insist: there must be no one on the boat tonight.
       b. Tharp says the curtain may go up now.

Because I was originally thinking in terms of formal syntax, I have adopted the terms "Equi" and "Raising" here. Nevertheless, in the corpuses I have examined, it may be possible to argue that every instance of *be going to* as an Equi predicate is also a Root/Deontic use, and that every instance of a Raising predicate is also an Epistemic use. I leave a careful investigation of this matter for future research.

One other clarification is in order about the Equi/Raising definition. For purposes of classifying lexical items as being Equi and/or Raising in a language where one has unlimited access to examples it seems to be sufficient to make use of diagnostics like "takes Dummy subjects", "takes non-sentient subjects", "can participate in an idiom like *the cat is out of the bag*".[7] However, if we want to ascribe analyses to particular utterances, as we need to do in order to measure things like "rate of use of *be going to* as an Equi verb in year *t*", these diagnositics are inadequate. *Be going to* might be an Equi verb in one utterance and a Raising verb in another. This is a problem even in the analysis of modern corpuses. Fortunately, there is a plausible way of testing particular instances of *be going to*: we can try substituting the verb *intend*, which seems to have nearly the same meaning as Equi *be going to*, and see if the result is grammatical and preserves the meaning. In general such a method would be a dubious one for

---

[7] Although there may be some problems with inconsistency among the tests—see Bolinger 1973, Dowty 1985.

analyzing a language one did not speak natively. But Early Modern and earlier Modern English are so similar to Modern English that most of the examples of relevance to this study can be plausibly evaluated in this way.

Using the *intend*-substitution criterion has important consequences. If one were to use, instead, a simple-minded criterion like "Subjects of Equi verbs must refer to sentient beings while subjects of Raising verbs must refer to non-sentient beings" then one might put examples like (101) and (102) down as Raising instances. But in these cases, it seems that sentience is being attributed to the subject, even though the noun-phrase involved normally refers to non-sentient things. Indeed, substitition with *intend* produces a grammatical result with very nearly the same meaning.

(101) c. 1695 I heard a Man make a Noise to some People to make hast, for the Boat was going to put off, and the Tide would be spent. Defoe, *Moll Flanders*, (match 31) [AIR]

(102) 1990 He did not explain how mail order is going to reduce its long-term loss of retail market share. *Manchester Guardian*, [HECT]

Contrariwise, examples like (103) and (104) involve situations where a human subjects are undergoing an experience over which they have no control:

(103) c. 1796 ...we all felt that we were going to be only half a mile apart, and were sure of meeting every day. Austen, *Emma* (match 3)

(104) c. 1894 "Do you mean that we are going to die too?" asked the child, checking her sobs, and raising her tear-stained face. Doyle, *Sherlock Holmes* I-198 (match 66).

Here, the simple-minded criterion would lead to an Equi classification but substitution with *intend* indicates Raising status: the result of substitution is both bizarre and quite different from the meaning of the original utterance.

An interesting subvariety of these cases involves *be going to* with an embedded Equi verb:

(105) 1990 These customers are going to want the purest, approved original system and not some university mutant. Network News, NEXT-discussions (match 16)

(106) 1990 If I get one, am I going to be glad I spent the $250.00? Network News, NEXT-discussions (match 103)

Again, as we would expect, substitution with *intend* produces a bizarre and semantically divergent result. Therefore, in the study described below, I have treated *intend*-substitution possibility as a litmus indicator of Equi status.

5.2.2.2 Historical Development

I now review the history by considering the first and near-first examples of the different types of constructions that *be going to* has participated in.

In ModE, motion *be going to* occurs in two significantly distinct constructions: *to* can be a preposition, in which case it takes an NP complement; or *to* can be the infinitive marker, in which case it takes a VP complement. The NP-complement type appears to be the oldest of all *be going to* constructions. It has been around since at least the OE period. Examples are given in (107) and (108).

(107)   855   þu ...   bist   gangende   to   Romesbyrig
              you   be-2-sg   going   towards   Rome-gen-city
        'You'll be...going to Rome' *GD–C*, 132.30 [PéRez 1990]

(108) 1450 To abide in prison...wihtout goyng to bayle, abston, or mainprise... *RParl.* 5.201a [MED, PéRez 1990: 56]

Though a number of researchers have scanned the sources (see Danchev and Kytö 1991), the earliest instances anyone seems to have noted of VP-complement behavior are from Middle English. Examples are given in (109). (109a) has a VP reading if Mossé is right in claiming that *þe* is a version of OE *þeon* 'thrive', but Danchev and Kytö are skeptical (p. 3).

(109) a. (early 1300s) Phillip (...) was going too þe ouer Greece, *King Alisaunder*, l.901 [Danchev and Kytö 1991: 3]
      b. 1438 And there vppon the seid persones of the ship of Hull goyng to do the said wrong / yaf to oon henry wales Gentilman duellyng abowte the cost of Develyn x marc3, *Chancery English*, 174 [Danchev and Kytö

1991:4]

c. c. 1590 Hark, the kings and princes...are <u>going to</u> see the Queen's picture. Shakespeare, *A Winters Tale*, V ii. [OXF]

d. c. 1590 I met Lord Bigot and Lord Salisbury / With eyes as red as new-enkindled fire,... <u>going to</u> seek the grave / Of Arthur, Shakespeare, *Historical Plays*. [OXF]

e. c. 1675 When we were <u>going to</u> fight the Dutch, I had such a paine in my right arme that could not use but very litle. *Private Letters*, 3.15.

By contrast, the earliest potential Equi example I have come across is from the 16th century (110a) and the next-earliest are from the 17th (110b–d). There is, of course, no simple marker that distinguishes Equi *be going to* from Motion *be going to* with a VP complement. Nevertheless, it is usually not hard to tell from the context which sense is intended.

(110) a. 1567 when you are <u>going to</u> lay a tax upon the people, Burton, *Parl. Diary* [Danchev and Kytö 1991: 7]

b. c.1675 The council sat upon it, and were <u>going to</u> order a search of all the houses about the town. *History of Charles II*, 1.2.164 [HELS]

c. 1695 He was <u>going to</u> reply...but he heard his sister coming, Defoe, *Moll Flanders* (match 8) [AIR]

d. 1699 Gad, I have forgot what I was <u>going to</u> say to you. Congreve, *The Way of the World* (match 1) [AIR]

Under the *intend*-substitution diagnostic, there is a possible first instance of Rasing *be going to* in 1482 (111).[8]

(111) a. 1482 ...[W]hile thys onhappy sowle by the vyctoryse pompys of her enmyes was <u>goyng to</u> be broughte into helle for the synne and onleful lustys of her body. Loe sondenly anon came done an hye fro heuyn a gret lyght by the whyche bryghtnes and bemys. the forseyde wykyd spiritys and mynystrys of the deuyl. ware dullyd and made onmyghty...*The Revelation to the Monk of Evesham* [p. 43]

---

[8]Before Danchev and Kytö found instance (109b) above, this case was cited by most researchers as the first known case of *be going to* with a VP complement.

If this is Raising *be going to* it is surprisingly early given that the first observed Equi instances do not occur until a century later and the common trend is from Equi/Root to Raising/Epistemic rather than vice versa (see Shepherd 1982, Traugott 1989, Sweetser 1990). But it is possible that this example is not Auxiliary *be going to* but Motion *be going to*. The passage containing it begins with the narrator's description of hearing a great hue and cry and then seeing

a cursyd companye of wyckyd spyrytys and a myghty ['mighty spirit'] ledyng with hem anone as they hopyde ['danced'] to helle a soule of a woman late departyd fro her body. *The Revelation to the Monk of Evesham* [p. 42]

It goes on to describe the tortures the spirits inflicted on the woman's soul as they ushered it along the way to Hell. Thus it is conceivable that "goyng to" above means something like "travelling to" rather than "about to".

It may be significant that the next potential Raising examples, which don't occur until the late 17th century, all have human subjects:

(112) a. c. 1675 my Unckle is <u>going to</u> be married, w=ch= one would wonder at, there being nothing to be liked in him but his fin diamond ring. Private Letters 03, p. 1.240 [HELS]

b. c.1695 How little does he think that having Divorc'd a Whore,...he is <u>going to</u> Marry one that has lain with two Brothers, Defoe *Moll Flanders* (match 24).

Indeed, one might be inclined to call these cases Equi since there is clear intentionality involved in the embedded predicate, but if we apply the *intend*-substitution test, Raising is indicated: If we take an *intend* reading in (112a), the sentence presupposes a truism to the effect that people choose to get married because they are liked. While this may be a truth, it seems unlikely that it was ever a truism among the English so the resulting reading is a bit bizarre. Example (112b) is perhaps possible with an *intend* interpretation because one can assume the "thinker" assigns a *de re* reading (see Dowty *et al.* 1981). Nevertheless, this interpretation is tenuous because the potential is great for assigning the absurd interpretation in which someone intends to do something they have no awareness of.

The first instances I have found of Raising *be going to* with inanimate subjects do not appear until the late 18th century:

(113) a. c. 1796 standing side by side, exactly as if the ceremony were <u>going to</u> be performed. Austen, Mansfield Park, p. 88 [AIR]

b. c.1796 Something <u>is going to</u> happen.... Austen, *Mansfield Park*, p.25 [AIR]

c. c. 1865 she went down to look about her, and to wonder what was <u>going to</u> happen next. Carroll, *Alice in Wonderland*, (match 1) [AIR]

And the first instances of Raising *be going to* with dummy subjects do not appear until the late 19th century:

(114) a. c. 1865 Do you think it's going to rain? Carroll, *Alice Through the Looking Glass*, (match 3) [AIR]

b. 1890 It seems as if it were <u>going to</u> rain. *Chamb. Jrnl* 14 June 370/2 [OED]

c. c. 1894 There is <u>going to</u> be a shooting and somebody is going to get hurt. Doyle, *Sherlock Holmes* v. 1, p. 568. [AIR]

d. c. 1911 Mr. Bloom looked back towards the choir. Not <u>going to</u> be any music. Pity. Joyce, *Ulysses* (match 11)

e. c. 1911 —Who are you laughing at? says Bob Doran. So I saw there was <u>going to</u> be a bit of a dust Bob's a queer chap when the porter's up in him...Joyce, *Ulysses* (match 41)

Except for the possible early case of Raising *be going to* in 1482 these data argue for the following chronology of developments:

(115)

OE or before	Motion Verb + to + NP complement
ME	Motion Verb + to + VP complement
late 16th century	Equi Verb
late 17th century	Raising Verb with human subject
late 18th century	Raising Verb with inanimate object subject
mid 19th century	Raising Verb with dummy subject

This chronology is quite consistent with the hypothesis that Equi/Root > Raising/Epistemic is natural path of grammaticalization. It also provides evidence

for a prototype category representation under the plausible assumption that the last four types constitute progressively more canonical instantiations of Raising verb behavior.[9]

However, it is hard to be very confident in data on first-instances alone: since the absolute frequencies of new forms are generally low, the variance across samples of the observation-times of first forms is high, compared to say, the variance across samples of peak-usage points. Claims about successions of behaviors are likely to be more robust if we can make them about properties of the chronological distribution of behaviors that are relatively insensitive to noise in the sampling process. With this in mind, I have tabulated counts of occurrence of *be going to* in a sequence of historical corpora. The results are shown in Figures 5.5 and 5.6.

In order to facilitate sorting the instances in the sample, I used the (somewhat arbitrary) featural decomposition shown underneath the chart on the left. Therefore, to recover information about numbers of instances in various grammatical classes, it is necessary to sum over columns in the chart. Figure 5.7 shows a plot of the frequency of each of the four main types I have been discussing here (Motion+NP, Motion+VP, Equi, and Raising) relative to the total number number of instances of *be going to*.[10] Note that, in this chart, the ordering of the peaks of maximal relative usage is consistent with the ordering of times of emergence found for the first examples.

Two properties of this graph provide evidence for Q-Divergence:

(i) Correlated with the emergence of the first Equi examples, there is a rise in the relative frequency of Motion *be going to* + VP complement.

(ii) Correlated with the emergence of the first Raising examples, there is a rise in the relative frequency of Equi *be going to*.

The first observation may not be significant because the number of instances is very small in the corpus (Helsinki: 1650–1700) that is responsible for the effect:

---

[9] It is also reasonable to think of Equi verbs as a special subtype of Raising verbs in this regard since their formal contexts of occurrence are a subset of the Raising formal contexts.

[10] I have grouped all the Raising types together in this sample because there do not seem to be enough examples of each subtype to draw strong conclusions.

Figure 5.5: Quantitative tabulation of the development of *be going to* from 1590 to 1990 [Part I]

	1	2	3	4	5	6	7	8	9	10	11	12
V	-	-	-	-	-	-	+	+	+	+	+	+
N	+	+	+	+	+	+	+	-	-	-	-	-
M	+	+	+	+	+	?	+	+	+	+	?	?
H	+	+	+	-	-	+	+	+	-	-	+	+
I	+	?	-	+	-	-	+	+	+	-	+	-
D	-	-	-	-	-	-	-	-	-	-	-	-
1590 shakes.all	17 / 58%	0 / 0%	0 / 0%	1 / 3%	1 / 3%	0 / 0%	1 / 6%	9 / 25%	0 / 0%	0 / 0%	2 / 6%	0 / 0%
1675 hels.ene3	4 / 22%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	2 / 11%	7 / 39%	1 / 6%	0 / 0%	0 / 0%	0 / 0%
1695 defoe.mf	13 / 32%	5 / 12%	0 / 0%	1 / 2%	1 / 2%	0 / 0%	1 / 2%	5 / 12%	0 / 0%	0 / 0%	2 / 5%	1 / 2%
1796 austen.var	52 / 46%	1 / 1%	1 / 1%	0 / 0%	0 / 0%	0 / 0%	1 / 1%	4 / 4%	0 / 0%	0 / 0%	3 / 3%	0 / 0%
1855 melv.mdbb	12 / 32%	0 / 0%	1 / 3%	1 / 3%	2 / 5%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	1 / 3%	0 / 0%	0 / 0%
1894 doyle.sh	19 / 25%	0 / 0%	0 / 0%	0 / 0%	1 / 1%	0 / 0%	0 / 0%	1 / 1%	0 / 0%	0 / 0%	1 / 1%	0 / 0%
1911 joyce.u	15 / 17%	1 / 1%	1 / 1%	0 / 0%	1 / 1%	0 / 0%	1 / 1%	2 / 2%	0 / 0%	0 / 0%	2 / 2%	0 / 0%
1990 hect.g	3 / 3%	0 / 0%	0 / 0%	0 / 0%	1 / 1%	3 / 3%	1 / 1%	1 / 1%	0 / 0%	0 / 0%	0 / 0%	0 / 0%
1991 nn.NEXT	2 / 2%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%
	MN	MN	MN	MN	MN	N	MV	MV	MV	MV	N	N

Feature Symbols	Class Symbols
V = VP complement	MN = Motion + NP
N = NP complement	MV = Motion + VP
M = Motion interpretation	EA = Equi Aux
H = Human subject (literally)	RA = Raising Aux
I = Intend paraphrase possible	N = Uncertain classification
D = Dummy subject	

Figure 5.6: Quantitative tabulation of the development of *be going to* from 1590 to 1990 [Part II]

	13	14	15	16	17	18	19	20	21	N
V	+	+	+	-	+	+	+	+	+	
N	-	-	-	-	-	-	-	-	-	
M	-	-	-	-	-	-	-	-	-	
H	+	+	+	?	-	-	-	-	-	
I	+	?	-	?	+	?	-	-	-	
D	-	-	-	-	-	-	-	?	+	
1590 shakes.all	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	31
1675 hels.ene3	2 / 11%	2 / 11%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	18
1695 defoe.mf	8 / 20%	0 / 0%	2 / 5%	0 / 0%	1 / 2%	0 / 0%	0 / 0%	0 / 0%	0 / 0%	40
1796 austen.var	39 / 34%	6 / 5%	5 / 4%	0 / 0%	0 / 0%	0 / 0%	2 / 2%	0 / 0%	0 / 0%	114
1855 melv.mdbb	15 / 41%	0 / 0%	2 / 5%	0 / 0%	1 / 3%	0 / 0%	2 / 5%	0 / 0%	0 / 0%	37
1894 doyle.sh	38 / 51%	1 / 1%	6 / 8%	0 / 0%	0 / 0%	0 / 0%	6 / 8%	0 / 0%	2 / 3%	75
1911 joyce.u	45 / 52%	2 / 2%	7 / 8%	0 / 0%	0 / 0%	0 / 0%	5 / 6%	0 / 0%	4 / 5%	86
1990 hect.g	28 / 28%	2 / 2%	14 / 14%	1 / 1%	4 / 4%	3 / 3%	31 / 31%	4 / 4%	5 / 5%	101
1991 nn.NEXT	43 / 37%	2 / 2%	16 / 14%	0 / 0%	9 / 8%	1 / 1%	38 / 33%	3 / 3%	1 / 1%	115
	EA	N	RA	N	EA	N	RA	RA	RA	

Feature Symbols	Class Symbols
V = VP complement	MN = Motion + NP
N = NP complement	MV = Motion + VP
M = Motion interpretation	EA = Equi Aux
H = Human subject (literally)	RA = Raising Aux
I = Intend paraphrase possible	N = Uncertain classification
D = Dummy subject	

if this corpus were removed from the data there would be a fall, rather than a rise in the frequency of Motion+VP-complement *be going to* coinciding with the first appearance of Equi *be going to*. Nevertheless, I will make the assumption that there was a rough correlation to this effect and see if the network predicts it. The second observation is probably significant because it is consistent with a long-term parallel trend in the frequencies of the Equi and Raising uses.

Assuming they are significant, how do these observations provide evidence for Q-Divergence? Regarding observation (i), note that auxiliary verbs are verbs that occur primarily with VP complements. Therefore a change in Motion *be going to* from taking no or few VP complements to taking a relatively large number makes it more similar, distributionally, to an auxiliary verb. Moreover, since Motion *be going to* was used primarily (though not exclusively) with human subjects during the early part of the history here tabulated (Figures 5.5 and 5.6), and intention seems generally to have been ascribed to those subjects, it is expected that the first auxiliary instances of *be going to* should be Equi uses, paraphrasable by *intend*, not Raising uses. However, the more *be going to* is

Figure 5.7: Quantitative shift in the use of *be going to* [1590–1990]. (N = Motion NP; V = Motion VP; E = Aux Equi; R = Aux Raising)
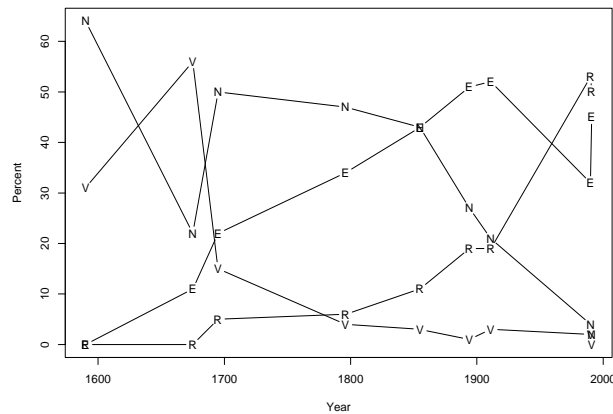
Year:	1590	1675	1695	1796	1855	1894	1911	1990
Source:	Shakes.	Hels	Defoe	Aust.	Melv.	Doyle	Joyce	Hector
Motion NP	61%	22%	50%	47%	43%	27%	21%	4%
Motion VP	32%	56%	15%	4%	3%	1%	3%	2%
Aux Equi	0%	11%	22%	34%	43%	51%	52%	32%
Aux Raising	0%	0%	5%	6%	11%	19%	19%	53
Tokens:	31	18	40	114	37	75	86	101

used as an Equi verb, the more its distribution comes to resemble that of a Raising verb, because the set of formal contexts in which Equi verbs can occur is a subset of the set in which Raising verbs occur. This fact alone, does not predict that the rise in Equi use should be accompanied by a rise of Raising use—for if it did we should expect all Equi verbs to show Raising behavior. But Motion *going to* also exhibits a low rate of use with non-human, non-sentient subjects. Examples of this behavior are shown in (116).

(116) a. c. 1590 If he be of any reasonable stature, he may creep in here; and throw foul linen upon him as if it were <u>going to</u> bucking. Shakespeare, *The Merry Wives of Windsor*, [OXF]

b. c. 1695 It was no surprize to me, for I had foreseen *sometime,* that all was <u>going to</u> Wreck...Defoe, *Moll Flanders*, (match 11) [AIR]

c. c. 1855 This same sperm was carefully manipulated ere <u>going to</u> the

try-works. Melville, *Moby Dick* [AIR]

d. c. 1894 Have you a note of anything <u>going to</u> a person named Holmes? Doyle, *Sherlock Holmes* v. 1, p. 759 [AIR]

Admittedly, there are only these two early instances in my corpuses. Nevertheless, this behavior makes the distribution of early *be going to* less Equi-like and more Raising-like than a canonical Equi verb. So, under the Q-Divergence Hypothesis, the combination of this pattern with the trend toward greater Equi usage may be expected to draw *going to* into the Raising class.

In the remainder of this section, I show how these predictions obtain in a network implementation.

5.2.2.3 Network simulation 1: Advent of Equi *be going to*

I consider the advent of Equi *be going to* first.

I used the grammar shown in Figure 5.8 to generate a corpus. To make it clearer what the important properties of this grammar are, I give a schematized description of the distribution of different kinds of auxiliary verbs in sentence-types in Figure 5.9. The frequency distributions for the different types of verbs across sentence types are based on data from the Hector corpus.

I trained a network with 25 input and output units, 5 hidden units, and 3 time-steps of the recurrent the hidden layer unfolded until it was predicting the transition-likelihoods well. I then post-trained under the distorted values indicated by the arrows in in Figure 5.8. The likelihoods of selected sentences during post-training are shown in Figure 5.10. Indeed, as expected, the likelihood of an Equi-verb usage ("E") rises from below the near-grammaticality threshold to above it in conjunction with the divergence of the frequency of the Motion+VP use of *be going to* ("A") away from the values associated with canonical Motion verbs. This shows that the network predicts the observed Q-divergence effect for a simple approximation of the early *be going to* data.

Figure 5.11 gives a hierarchical clustering of the hidden unit locations associated with a sample of inputs to the network in contexts at the beginning of post-training and at the end of it. The clustering was done using the "maximum" method (see Chapter 3, Section 4). There are clusters for the subjects,

Figure 5.8: An approximation of the pre 17th century change in the distribution of *be going to* (Verb-spectra based on selections from the Guardian section of the Hector corpus).

```
S : MotionS              0.27
S : GoingtoS             0.07
S : EquiS                0.33
S : RaisingS             0.33

MotionS : Hum Vmotion Place     0.61    Fred is-moving-to the-West
MotionS : Thing Vmotion Place   0.25    The-storm is-moving-to the-West
MotionS : Hum Vmotion Agt       0.09    Fred is-moving-to fell-the-tree
MotionS : Thing Vmotion Agt     0.05    The-storm is-moving-to fell-the-tree

EquiS : Hum Vequi Agt           0.76    Fred intends-to fell-the-tree
EquiS : Hum Vequi NonAgt        0.04    Fred intends-to reply
EquiS : Thing Vequi Agt         0.20    The-storm intends-to fell-the-tree

RaisingS : Hum Vraising Agt      0.18   Fred will fell-the-tree
RaisingS : Hum Vraising NonAgt   0.18   Fred will reply
RaisingS : Thing Vraising Agt    0.17   The-storm will fell-the-tree
RaisingS : Thing Vraising NonAgt 0.19   The-storm will reply
RaisingS : Thing Vraising VThing 0.25   The-storm will happen
RaisingS : Pleo Vraising VThing  0.03   It will happen

GoingtoS : Hum Vgoingto Place    0.61  →  0.07 Fred is-going-to the-West
GoingtoS : Thing Vgoingto Place  0.25  →  0.03 The-storm is-going-to the-West
GoingtoS : Hum Vgoingto Agt      0.09  →  0.71 Fred is-going-to fell-the-tree
GoingtoS : Thing Vgoingto Agt    0.05  →  0.19 The-storm is-going-to fell-the-tree

Vmotion : vm1            0.25
Vmotion : vm2            0.25
Vmotion : vm3            0.25
Vmotion : vm4            0.25

Vgoingto : vgt          1.00

Vequi : ve1             0.20
Vequi : ve2             0.20
Vequi : ve3             0.20
Vequi : ve4             0.20
Vequi : ve5             0.20

Vraising : vr1          0.20
Vraising : vr2          0.20
Vraising : vr3          0.20
Vraising : vr4          0.20
Vraising : vr5          0.20
```

Figure 5.9: Grammar Scheme for Motion > Equi *be going to* simulation.

Subject–Complement Combination	Type of Aux permitted
Human Place	motion
Thing Place	motion
Human VP[Agt]	motion, equi, raising
Thing VP[Agt]	motion, equi, raising
Human VP[NonAgt]	equi, raising
Thing VP[NonAgt]	raising
Thing VP[Thing]	raising
Pleo VP[Thing]	raising

of Equi verbs at the end of post-training. It starts showing novel Equi behaviors because the frequency-change in its Motion-use makes it more like an Equi Verb than like a Motion Verb.

5.2.2.4 Network simulation 2: Advent of Raising *be going to*

To simulate the appearance of Raising behavior, I used a grammar (Figure 5.13) to generate a body of sentences that approximate the pre-late-17th-century distribution of *be going to*. For the sake of simplicity, this grammar, like the one of the previous section, treats all sequences of supporting verbs as single words. But it encodes a set of contrasts between types of supporting verbs in distributional terms (See Figure 5.14).

The crucial property of this scheme, which is intended to be an approximate reflection of the actual distribution of these forms (assumed not to have changed much in relevant respects over the centuries 16–20), is that "Equi verb" contexts are a subset of "Raising verb" contexts so there is a good deal of similarity between the classes. But there is also a tie between "Motion verbs" and "Raising verbs" that excludes "Equi", in that "Motion verbs" occur sometimes with inanimate subjects.

I generated a 1000-sentence corpus using the grammar of Figure 5.13 and trained a network (8 hidden units, 4 layers of unfolding) on the word-prediction task until it was making good predictions on the grammar-derived likelihoods.

for the two different types of supporting verbs (Equi and Raising), for the motion verbs, and for the supporting-verb complements.[11] As we should expect under the analysis given in Chapter 3, Section 8 of how the network predicts Q-Divergence effects, the representation of <be-going-to> (marked "vgt") is in the cluster of Motion verbs at the beginning of post-training and in the cluster

---

[11] In this dendogram, the second- and higher-order clusters are different from Elman's results reported in Chapter 1 in that the noun/verb (or NP/VP) division is not fundamental. Instead, there is a basic division between subjects and non-subjects and then a subsidiary division between equi/raising verbs on the one hand, and motion verbs/other complements/period on the other. Although this subsidiary division seems to reflect roughly the difference between auxiliary verbs and more contentful verbs and verbal-complements, the effect is probably unmeaningful, since the relevant properties of the generating grammar only crudely reflect the properties of natural language in relevant regards.

Figure 5.10: Simulation result: appearance of equi <be going to> in conjunction with rise of VP complements with motion verbs.
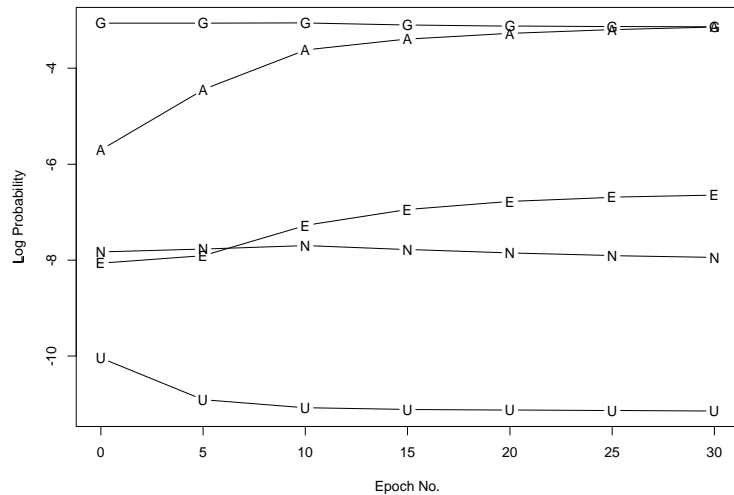
a.	Fred	is-going-to	fell-the-tree	(A) Motion + VP
b.	Fred	is-going-to	reply	Unambiguous (E)qui
c.	The-storm	intends-to	reply	(N)ear grammatical
d.	Fred	intends-to	fell-the-tree	(G)rammatical
e.	Fred	will	the-West	(U)ngramatical

I then used a new grammar to generate a second 1000-sentence corpus in which the frequency of <be going to> as an "Equi verb" was significantly elevated relative to the first corpus (see the arrows in Figure 5.13). The point was to see if the network would start expecting <be going to> to show "Raising" behavior (for example, occurring with a non-human subject and a non-agentive verbal (Vthing) complement).

Figure 5.15 traces the logs of likelihoods with which the network expects various sentences (117) to occur during the process of post-training.

Figure 5.11: Hierarchical clustering of average hidden unit vectors by input behavior after initial training. [Inter-cluster distances computed by the "maximum" method.]
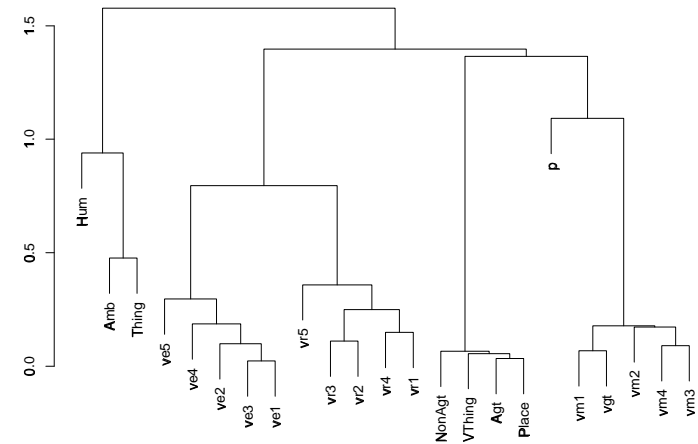
(117)	a.	the-cabinet	seems-to	like-to-dance	(N)ear Grammatical
	b.	she	is-going-to	fall	(E)qui be-going-to
	c.	the-cabinet	is-going-to	fall	(R)aising be-going-to
	d.	he	intends-to	live-alone	(G)rammatical
	e.	the-cabinet	intends-to	fall	(C)anonical Equi in Raising context

The curve marked by the letter "N" corresponds to the "near-grammatical" sentence (117a). The graph shows that as the likelihood of *be going to* in "Equi" contexts (like example (117b)—"E") grows, the likelihood of *be going to* in "Raising" contexts rises from below threshold to above it ((117c)—"R"). This is in contrast to the likelihood of canonical "Equi" verbs in "Raising" contexts ((117e)—"C") which remain near threshold throughout. Thus the network predicts a correlation between a frequency-shift (the rise of "Equi" *be going to* from low frequency to higher frequency) and an innovation event (the rise of "Raising" uses of *be going to* above threshold). Again, this makes it plausible that a

Figure 5.12: Hierarchical clustering of average hidden unit vectors by input behavior after post-training training. [Inter-cluster distances computed by the "maximum" method.]
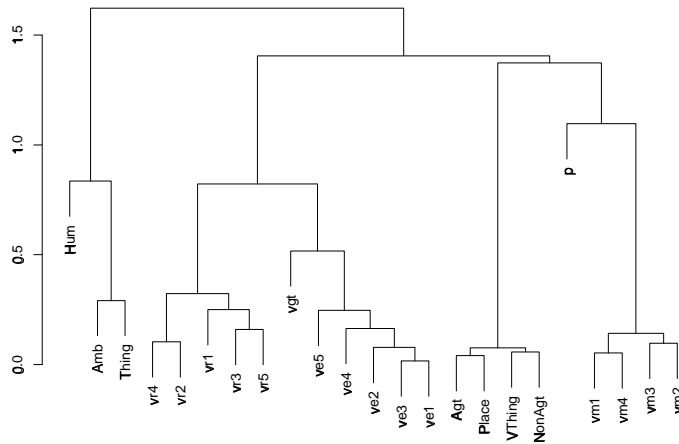
grammar be both restrictive and predict Q-divergence effects.

5.3 Problems for Competing Grammars Models

None of the Q-divergence effects observed in the two case studies just described are easily handled within the Competing Grammars framework. The problem is that these Q-divergence effects involve diachronic linkage between what are traditionally treated as unrelated lexical items. Competing grammars models only predict correlated changes between elements that are controlled by the same parameter-setting. It is certainly not plausible to claim that Noun-Prep *sort/kind of* and Degree Modifier *sort/kind of* are generated by the same parameter setting in Universal Grammar, given that the former construction existed so long before the latter one. It also seems dubious that a single parameter setting could be responsible for generating Motion, Equi, and Raising uses of *be*

Figure 5.13: An approximation of the late 17th-century change in the distribution of *be going to*.

|	
S : Smain	0.50
S : Saux	0.50
Smain : NPhum Vlocagt	0.16
Smain : NPhum Vnonlocagt	0.48
Smain : NPhum Vthing	0.16
Smain : NPthing Vthing	0.10
Smain : NPamb Vamb	0.10
Saux : Smotion	0.20
Saux : Sequi	0.40
Saux : Sraising	0.40
Smotion : NPhum AUXmotion NPloc	0.60
Smotion : NPthing AUXmotion NPloc	0.20
Smotion : NPhum AUXmotion Vlocagt	0.20
Sequi : NPhum AUXequi Vlocagt	0.20
Sequi : NPhum AUXequi Vnonlocagt	0.60
Sequi : NPhum AUXequi Vthing	0.20
Sraising : NPhum AUXraising Vlocagt	0.16
Sraising : NPhum AUXraising Vnonlocagt	0.48
Sraising : NPhum AUXraising Vthing	0.16
Sraising : NPthing AUXraising Vthing	0.10
Sraising : NPamb AUXraising Vamb	0.10
NPhum : she	0.34
NPhum : he	0.33
NPhum : Harry	0.33
NPthing : themoney	0.50
NPthing : thecabinet	0.50
NPamb : it	1.00
NPloc : Nantucket	0.50
NPloc : Delhi	0.50
Vlocagt : delivermail	0.50
Vlocagt : visitthequeen	0.50
Vnonlocagt : leave	0.17
Vnonlocagt : come	0.17
Vnonlocagt : livealone	0.17
Vnonlocagt : wearahat	0.17
Vnonlocagt : liketodance	0.16
Vnonlocagt : sayyes	0.16
Vthing : fall	0.50
Vthing : last	0.50
Vamb : rain	0.50
Vamb : beeasytosleep	0.50
AUXmotion : comesto	0.34
AUXmotion : travelsto	0.33
AUXmotion : isgoingto	0.33
AUXequi : intendsto	0.49 → 0.33
AUXequi : wantsto	0.49 → 0.33
AUXequi : isgoingto	0.02 → 0.33
AUXraising : may	0.50
AUXraising : seemsto	0.50

going to. All motion verbs ought to exhibit Equi and Raising behaviors if that were the case. Moreover, as I noted in Chapter 2 (Section 1.1.2), the Competing Grammars account does not predict succession-of-lobes frequency curves. But the frequency curves in the *be going to* episode seem to have this property (Figure 5.7).

The case-studies in this section have implications for the theory of language change. In particular, they suggest that the role of Indirect Transmission may not be as great as earlier theorists have proposed (e.g. King 1969, Andersen

the assumption that younger language speaker's attempts to situate themselves socially may be the primary factor in change (see Eckert 1988). I discuss this issue further in Chapter 7, Section 2.1. The case-studies presented here also suggest a revision of one of the currently widely-held assumptions about the structural mechanisms of change—that there is a fundamental division between two types of structural change: reanalysis, which changes underlying representations but does not alter surface forms, and analogy, which alters surface forms to make them more consistent with underlying representations. The model and case-studies examined here suggests that we may not need this division, for all events of change can be treated as minor, analogically motivated reanalyses. I discuss this point further in Chapter 7, Section 2.2.

Figure 5.14: Grammar Scheme for Equi > Raising *be going to* simulation.

NPhum NPloc	motion
NPthing NPloc	motion
NPhum Vlocagt	motion, equi, raising
NPhum Vnonlocagt	equi, raising
NPhum Vthing	equi, raising
NPthing Vthing	raising
NPamb Vamb	raising

Figure 5.15: Simulation Result: appearance of "Raising" <be going to> in conjunction with rise in the frequency of "Equi" <be going to>.
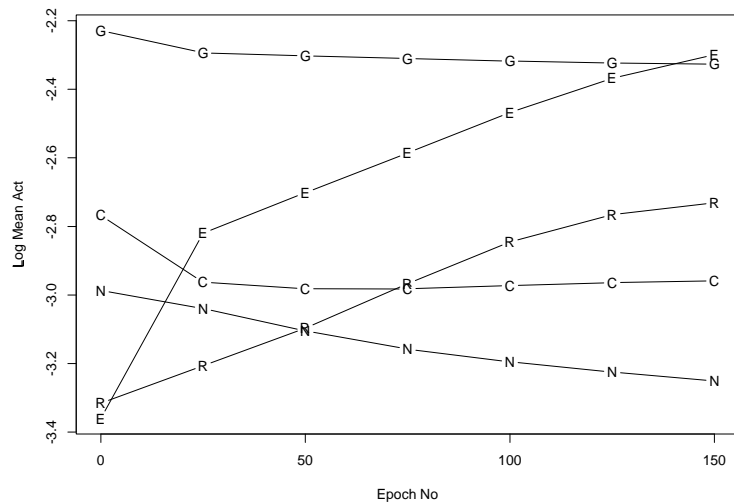
N = Near Grammatical, E = Equi <be going to>,
R = Raising <be going to>, G = Grammatical, C = Canonical Equi Verb

1973).[12] Instead, the Restrictive Continuity model is much more consistent with

parents grammars, but must guess at them based on partial evidence. It is often assumed in diachronic generative linguistics that younger speakers "errors" at guessing the nature of their parents' grammars play a major role in structural change.

---

[12] "Indirect Transmission" refers to the fact that children do not have direct access to their

Chapter 6

Hybrid Structures

6.1 Ambiguity versus blending

The third prediction made by the Restrictive Continuity model is that when an element is in-transit between an old structural status and a new one it may show intermediate behavior. In fact, there are two senses in which this can be true. An element can go through a period of exhibiting behaviors associated with its old category some of the time and behaviors associated with its new category the rest of the time. I'll call such alternating behavior "probabilistic hybrid behavior". Alternatively, an element may participate in a single construction which seems best analyzed as a splicing-together of a construction associated with its old behavior and one associated with its new behavior. I'll call such splicing behavior "structural hybrid behavior" or simply "hybrid structure". In the linguistic literature, "hybrid structures" are sometimes referred to as "blends".

Probabilistic hybrid behavior is probably a universal property of elements in evolutive category transition: that is, it essentially never happens that an element switches from one (traditional) grammatical status to another without going through a period where it it shows one behavior some of the time and the other behavior the rest of the time (cf. Hopper and Traugott 1993: 36). Moreover, probabilistic hybrid behavior is extremely common among elements

not in transition as well. In that case, it is usually called "ambiguity". Hybrid structures are rarer, or at any rate, much less obvious when they occur. Nevertheless, a number of cases have been recorded in the historical literature. It turns out that the recurrent neural network implementation of Restrictive Continuity predicts one observed type of structural hybrid.

I'll review several of the cases noted in the historical literature, show a toy simulation that exhibits blending, and then review a network simulation of a natural language case in which a hybrid occurs (development of Degree Modifier *sort/kind of* in English).

6.2 Hybrids noted in the historical literature

Perhaps the most famous examples of structural hybrids are cases of morphological double- (or triple-) marking which involve the simultaneous presence of several synonymous marker. Thus, ME eventually adopted *brethren* over *brothre*, *brethre*, and *brothren*. ModE children's speech shows double tense marking on irregular verbs (*sanged*) as well as both standard and non-standard single tense-marking (*sang*, *singed*) (Kuczaj 1977; see also Rumelhart and McClelland 1986, Pinker and Prince 1988, Plunkett and Marchman 1989). Similarly ModE children's speech has examples like *mostest* and *leastest* as well as of singly marked superlatives.

There are also syntactic examples. Preposition stranding was prohibited in *wh*-relatives and in questions in OE. But when, in ME, it began to penetrate these environments, certain *resumptive preposition* constructions (118) appeared [Allen 1983: 230].

(118) a. Til that the knight <u>of</u> which I speke <u>of</u> thus,. . .
    'Until the knight that I spoke of thusly.' [Ch F Frank. 807]
    b. And eek <u>in</u> what array that they were <u>inne</u>.
    'And also what array they were in.' [Ch Pro. 41]

In his study of the changes in clitic object-pronouns in Spanish (See Chapter 2, Section 1), Fontana 1993 argues that the 15th century construction shown in (119) is a hybrid.

(119) 15th c.:

Dixo: Le   yo dare   a esta villana   los tornos
said.3rdsg her$_i$ I   will-give to this lowly-woman$_i$ the run-around
'He said: I will give this lowly woman the run-around.' [F 93: 271]

In this example, the object pronoun *le* is separated from the verb whose argument it refers to by a syntactic constituent (*yo*). This phenomenon, labelled "interpolation" by traditional Spanish grammarians, was common in Old Spanish but is absent in Modern Spanish (see discussion in Chapter 2). Also, in this construction, the pronoun *le* doubles an overt Prepositional Phrase argument of the verb (*a esta villana*). This phenomenon was minimally present in Old Spanish but is now obligatory with most Indirect Objects in all varieties of Spanish. Under Fontana's analysis, Interpolation comes about when a clitic pronoun is left-adjoined to an NP that has been topicalized to Spec(IP), the position immediately to the left of Infl, where the verb resides in these examples. He argues, on the other hand, that doubled objects are lexically attached to their governing verbs. Lexical attachment is inconsistent with the assumption that the object pronouns are clitics and can be separated from the verb by a syntactic constituent (at least under normal assumptions about lexical attachment). Consequently his analysis fails to generate these constructions. But he notes that

> ...these examples ...are extremely rare, and generally restricted, as
> far as I know, to the XVth century texts. [p. 271]

Thus example (119) has the properties typical of a diachronic hybrid: it mixes properties of an old construction and a new construction and is associated with a period of transition.

Cohen 1987 provides a list of 1,993 syntactic blends that he has heard in everyday American speech. Although many of them sound, to my ear, like slips of the tongue, and hence seem unlikely to become part of acceptable usage (120), quite a few others sound much more natural, as though they are at least on the border of becoming acceptable (121) (The proposed parent constructions in these examples are due to Cohen.)

(120) a. Shall I check back with you in a little bit later?

   i. in a little bit
   ii. a little bit later [C 87: No. 882]

  b. from a financial stand-point of view

   i. from a financial point of view
   ii. from a financial stand-point [C 87: No. 648]

(121) a. He wouldn't buy them out of his own pocket.

   i. wouldn't buy them with his own money
   ii. wouldn't pay for them out of his own pocket
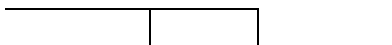
  b. [We'll finish in] a couple of more days.

   i. in a couple of days
   ii. in a few more days [C 87, No. 872]

6.2.1   Formal characterization of hybrids

A useful way of analyzing a hybrid is to write the hybrid down flanked by its proposed parent constructions and to mark the points of coincidence of the hybrid with each parent (Figure 6.1).

A number of these cases have the property that there is some element (a phoneme, phoneme-cluster, word, or word-sequence) which serves as a *bridge* between pieces of the two parent constructions. Thus some part of the diagram has the structure:

(122)

In Section 5 I show how the structures with bridges are naturally predicted by the recurrent Connectionist network under conditions of diachronic shift.

Figure 6.1: Hybrid Diagrams.

```
s    a    ng
┌─┬─┬────┐
│s│a│ ng │ ed
└─┴─┴────┘
s    i    ng   ed
and   in    what array they were
┌───┬──┬───────────────────┬────┐
│and│in│ what array they were │ in │
└───┴──┴───────────────────┴────┘
and    0     what array they were    in
Dixo:   le   yo dare          0          los tornos
┌──────┬──┬────────┬────────────────┬──────────┐
│Dixo: │le│yo dare │ a esta villana │ los tornos │
└──────┴──┴────────┴────────────────┴──────────┘
Dixo:   0    yo dare    a est villana   los tornos
from a financial   stand   point   0
┌─────────────┬──────┬───────┬─────────┐
│from a financial │ stand │ point │ of view │
└─────────────┴──────┴───────┴─────────┘
from a financial   0        point    of view
in a    couple of        0        days
┌────┬─────────┬──────┬─────┐
│in a│couple of│ more │ days │
└────┴─────────┴──────┴─────┘
in a     few         more    days
```

6.3 Problems for standard models

Languages containing hybrids are not problematic descriptively, for standard models of grammar in the way that, for example, languages consisting only of strings of the form $a^n b^n c^n$ are problematic for context-free grammars (e.g., Partee *et al.* 1990, §18.3). But they are an embarrassment for current categorical theories, including Competing Grammars models, because they seem to require stipulating ad-hoc new types for forms whose relationships to existing types is systematic. Although Competing Grammars models generate probabilistic hybrid behavior, they exclude hybrid structures.

6.4 The suitability of the recurrent architecture

The hybrids defined in Section 2.1 above crucially involve the sequential structure of speech. They can be characterized as strings that satisfy local sequential harmony constraints but violate global ones. Consequently, they are most naturally modelled in a system in which sequentially adjacent correlations exert a more powerful influence than sequentially distant correlations. The Simple

Recurrent Net, trained using backpropagation, has this property because of the diffusion of the learning signal with the depth of layering (Mozer 1988). Consequently, I focus on the recurrent architecture in this chapter.

6.5 Simulation of hybrids with a recurrent net

Can a net predict a hybrid? Consider the grammar in Figure 6.2.

Figure 6.2: Grammar with Two Fully-Bracketed Contexts.

S : Pre1 Head1 Post1 p	0.50		
S : Pre2 Head2 Post2 p	0.50		
Head1 : h11	0.20		
Head1 : h12	0.20		
Head1 : h13	0.20		
Head1 : h14	0.20		
Head1 : U	0.20		
Head2 : h21	0.20	$\rightarrow$	0.167
Head2 : h22	0.20	$\rightarrow$	0.167
Head2 : h23	0.20	$\rightarrow$	0.167
Head2 : h24	0.20	$\rightarrow$	0.167
Head2 : h25	0.20	$\rightarrow$	0.167
Head2 : U	0.00	$\rightarrow$	0.167

This grammar establishes two classes of elements, *Head1*, and *Head2* which occur in two contexts whose boundaries are unambiguously marked by surrounding words (*Pre1* and *Post1* for the context of *Head1*; *Pre2* and *Post2* for the context of *Head2*). I coded each of the terminals this grammar generates as a bit vector with only one bit "on" and trained a recurrent network with 4 hidden units and two time-steps unfolded to predict next-words in a 500-sentence corpus generated by the grammar. When the maximal pattern error was less than 20 percent of the minimum distance between the grammar-derived targets, I began post-training the network on a 500-sentence corpus generated by the altered version of the grammar indicated by the arrows in Figure 6.2. Figure

6.3 shows log likelihoods for several strings of interest during the process of post-training.

Figure 6.3: Simulation Result: Temporary Emergence of a Hybrid

1	Pre1 U Post1 p	Grammatical
2	Pre2 U Post2 p	Ungrammatical > Grammatical
3	Pre1 U Post2 p	Hybrid
4	Pre2 U Post1 p	Reverse Hybrid
a	Pre1 h11 Post1 p	Grammatical
b	Pre2 h11 Post2 p	Ungrammatical
c	Pre1 h11 Post2 p	Non-transitional Hybrid
d	Pre2 h11 Post1 p	Non-transitional Reverse Hybrid
w	Pre1 h21 Post1 p	Ungrammatical
x	Pre2 h21 Post2 p	Grammatical
y	Pre1 h21 Post2 p	Non-transitional Reverse Hybrid
z	Pre2 h21 Post1 p	Non-transitional Hybrid

The case of greatest interest is the curve marked "3" which corresponds to the string,

(123) <Pre1 U Post2 p>

This string begins with the left bracket of a *Head1* element and ends with the right bracket of a *Head2* element, followed by <p>, the sentence-end marker. In the middle is <U>, the element which is being induced to alter its behavior during post-training. Thus string (123) has the key properties associated with a transitional hybrid element in human language change. This string is never generated by either the initial-training or the post-training grammar. Nor, at the end of initial-training, is it predicted to occur with higher than sargasso likelihood.[1] But during the process of post-training, there is a period early-on when it rises above sargasso level briefly as shown in Figure 6.3. Moreover this rise is correlated with the initial rise of the string,

(124) <Pre2 U Post2 p>

which is the string whose likelihood has been elevated in the post-training corpus relative to the initial corpus and whose new behavior the network is learning to predict during post-training.

I performed this experiment several times with a variety of random initial weights, learning rates, and training-termination criteria. The likelihood of the hybrid didn't exceed the sargasso level in all cases but it always trended upward in conjunction with the rise of the *Head2* behavior of <U>.

In this simple corpus setting, then, the network predicts a correlation between the rise of a new, categorically well-defined behavior and the temporary appearance of a hybrid.

Why does the network make the hybrid prediction? I cannot offer a sure analysis of this case, but I'll give a speculative account. From the perspective of local symbol-to-immediately-following-symbol transitions, the target behavior of <U> during post-training is an average of the behaviors of <Head1> elements and <Head2> elements. Consequently, the learning algorithm, being most strongly influenced by error signals stemming from local correlations, initially adjusts the representation of the <U> element to a position intermediate

---

[1] In this highly artificial scenario, there is no obvious way of drawing on the intuitive notion of "near grammaticalness" in order to choose a plausible grammaticality threshold. Therefore, in this case, I have measured the maximum probability assigned to any non grammar-generated 4-symbol sequence at the end of initial training and taken this as a threshold, called the *sargasso level*, which corresponds to the notion of near-grammaticality. In the current simulation, the sargasso level is -5.46.

between the <Head1> and <Head2> elements. When it has this intermediate representation, <U> can do a reasonable job of serving simultaneously as a <Head1> element and a <Head2> element. In other words, it can serve as a *bridge* so the hybrid is predicted to occur with higher than Sargasso frequency. As a result of this change, however, the error-signal stemming from local transitions diminishes and the signal stemming from longer-distance dependencies becomes relatively stronger. Eventually, this results in the unlearning of the structural hybrid followed by the emergence of probabilistic hybrid behavior. Thus the all the features of a transitional hybrid episode are predicted.

With nothing more said, this analysis makes a prediction that is probably undesirable: the analysis posits no asymmetry between correlational relationships to the right and left of the bridge element. Therefore, we should expect all hybrids to come in pairs, consisting of a "forward hybrid" and a "reverse hybrid". The reverse hybrid can be computed from the diagram of the forward hybrid by reversing the direction of the bridge as shown in (125):

(125)

The paired hybrid prediction is probably not accurate in its general form. Although in several of the cases described above, the reverse hybrid is a grammatical expression (125), it happens to be a normal, independently motivated grammatical expression. Thus such cases do not provide strong support for the claim that hybrids always come in pairs. More convincing would be a case in which the reverse hybrid is not independently motivated but nevertheless is judged grammatical or is prone to occur as a speech error. But such cases seem rare. In the examples given above, the non-independently-motivated reverse hybrids are bizarre and seem unlikely to occur even as speech errors (125), although admittedly this last point needs more research to be verified. Also, if reverse hybrids were common, one would have expected prior researchers to have made special note of the fact. I have not found this to be the case (see Bergström 1906, Bolinger 1961, Cohen 1987). Thus, as a simple, broad claim, the paired hybrid prediction is questionable. I raise this matter because it may be that some hybrids occur in paired form, and that a more refined theory of

hybrids could predict this fact. I leave this as a matter for future research.

t
  a.   **Forward Hybrid:**   sanged
       Reverse Hybrid: sing
  b.   **Forward Hybrid:**   a couple of more days
       Reverse Hybrid: a few days

t
  a.   **Forward Hybrid:**   and in what array they were in
       Reverse Hybrid: and what array they were
  b.   **Forward Hybrid:**   from a financial standpoint of view
       Reverse Hybrid: from a financial point

It turns out that the network model correctly fails to predict reverse hybrids in the general case. This is because the unidirectional character of the word-prediction task introduces a left/right assymetry. For the most part, the more information one has about the past, the more constrained one's predictions about the future can be. This tends to be particularly true for within-phrase and within-sentence processing in the word-prediction task: as one moves further rightward in a phrase or sentence, the total probability of transition becomes distributed over a smaller and smaller set of elements. When probability is concentrated on only a few elements, the choices are very certain and the network learns quickly.[2]

This effect is visible in the simple hybrid simulation results shown in Figure 6.3. The reverse hybrid, (126), actually decreases in relative probability with the rise of the forward hybrid, and although it gains strength again as the new target structure begins to emerge, it never exceeds sargasso level.

(126) Pre2 U Post1 p

In the next section I consider the case of a real syntactic hybrid that emerged temporarily during the evolution of Degree Modifier *sort/kind of.*

---

[2]This follows from the fact that the sum squared error is likely to be large when the entropy of the target probability distribution is small and visa versa.

6.6 Case-study: *Its a fine ewnin but its a sort a caad*

It turns out that the history of Degree Modifier *sort/kind of* discussed above in Chapter 5 has an added twist. Recall that the first instances of unambiguous Degree Modifier *sort/kind of* first appeared in the early 19th century. Intriguingly, the following construction emerged at almost exactly the same time:[3]

(127) a. 1790 Its a fine ewnin but its a sort a caad. (= 'It's a fine evening but it's sort of cold.') Mrs. Wheeler, *Westm. Dial.* (1821) 63 [OED]

b. 1796 I conceive it to be a sort of necessary; for, let a woman have ever so many resources, it is... Austen, *Emma*, 356 [AIR][4]

c. 1839 I bees a sorter courted, and a sorter not; reckon more a sorter yes than a sorter no. Capt. Marryat. *Diary Am.* Ser. i. II. 218 [OED]

d. 1855 ...he's been a kind of moody—desperate moody, and savage sometimes; but that will all pass off. Melville, *Moby Dick*, 80 [AIR]

e. 1898 Ah'm a sort o'done up wi' walkin' so much, *Leeds Merc Suppl.* (Jan. 8, 1898) [EDD]

All of the examples in (127) contain a constituent of the form [BE a sort/kind of AdjP$_{VP}$]. It seems that not *sort/kind of* but *a sort/kind of* are being used as DegMods. Evidence for this view comes from the fact that *a sort/kind of* appears soon after in a range of DegMod environments:

(128) a. 1855 It seems a sort of foolish to me, tho', Melville, *Moby Dick*, p. 128 [V[stative] □ Adj]

b. 1855 Coming afoul of that old man has a sort of turned me wrongside out, Melville, *Moby Dick*, p. 125 [NP Aux □ V]

c. 1855 The head looks a sort of reproachfully at him. Melville, *Moby Dick*, p. 125 [□ Adv]

d. 1889 Ou, losh, ay! it made me a kind o' queery to look at her. Barrie, *Thrums*, vi [EDD] [V[stative]... □ Adj]

e. 1897 So dey akinda quail'd doon *Sh. News*, Oct. 23, [EDD] [NP □ V]

The type [BE a sort/kind of AdjP$_{VP}$] seems to be a blend of the Noun-Prep and DegMod uses of *sort/kind of*: the sequences *sort of* and *kind of* are preceded by a determiner as though they are Noun-Preposition sequences but they are followed by a phrase-final adjective as though they are DegMods.[5]

This construction seems to have existed marginally during the 19th century. I have found a few instances from the early 20th century. Thus, this hybrid behavior is clearly coincidental with the transition from Noun-Prep to DegMod *sort/kind of*.

Figure 6.4: An approximation of the pre-19th century change in the distribution of *sort of*.

S : NP VP p	1.00	VP : V'	0.40	P : of	0.60
		VP : is	0.20	P : from	0.40
NP : N'	0.00	VP : is AP	0.20		
NP : Det N''	1.00 → 0.85	VP : is NP	0.20	Adj : sumptuous	0.34
NP : a sort of Adj N''	0.00 → 0.15			Adj : dense	0.33
		V' : Vint	0.50	Adj : soft	0.33
N'' : N'	0.60	V' : Adv Vint	0.50		
N'' : AP N'	0.40			Adv : rather	0.34
N' : N	0.80	Det : this	0.34	Adv : merely	0.33
N' : N PP	0.20	Det : a	0.33	Adv : really	0.33
PP : P NP[part]	1.00	Det : the	0.33		
				Vint : melts	0.34
NP[part] : N''	1.00	N : block	0.25 → 0.32	Vint : rolls	0.33
NP[part] : Det N''	0.00	N : cheese	0.25 → 0.31	Vint : dissolves	0.33
		N : marmalade	0.25 → 0.31		
AP : Adj	0.50	N : sort	0.25 → 0.05		
AP : Adv Adj	0.50				

Does the network model predict it? I return to an examination of the simulation discussed in Chapter 5. The grammar which generated the corpus on which the network was trained is repeated in Figure 6.4. Note that the hybrid behavior is not generated by the grammar. Further details of the simulation result are shown in Figure 6.5. As the figure shows, in conjunction with the imposed change in the frequency of the now-ambiguous construction (curve "M"), the frequency of the <a sort of> Degree Modifier construction (curve "A") rises initially and is briefly higher than the curve for the unambiguous Degree Modifier <sort of> construction (curve "S"). This result is only minimally satsifactory,

---

[3] It is probably not coincidental that all of these cited forms appear in language that is considered "dialectal".

[4] AIR = Academic Information Resources electronic text database—see Appendix.

[5] The possibility of empty nominal heads as in *the beautiful (people), the meek (beings)* is not implausible—one could argue that the empty head gets reference via the subject NP and that the generic reference requirement (* *a/this beautiful*, * *a/this meek*) is satisfied because Noun-Prep *sort of* and *kind of* support (and in fact, require) an object with generic reference. But this can't be the whole story: the coincidence in the timing of the appearance of DegMod *sort of* and *a sort of* is not explained. Nor are ungrammaticalities like * *He's been a type/variety of moody*.

though. The "A" curve never exceeds the near-grammatical level and also never exceeds the level of the reverse hybrid ("R"). These predictions are incorrect inasmuch as we have evidence that the *a sort of* degree modifier construction was grammatical for a period, at least in certain dialects, and there seems to be no evidence for the occurrence of the reverse hybrid.

Figure 6.5: Simulation result: rise of the hybrid, DegMod *a sort of*, in conjunction with the rise of DegMod *sort of*.
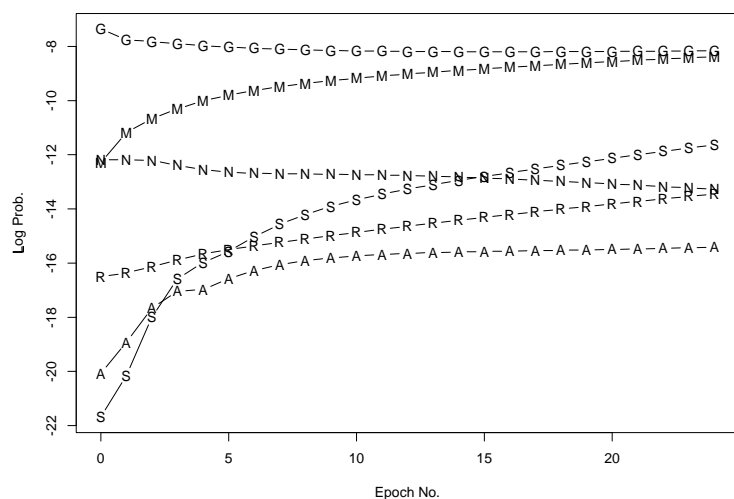
An improvement in these predictions can be made if we normalize the value L for sentence length.[6] A graph of the normalized L-values is shown in Figure 6.6. Here, appropriately, the curve corresponding to the <a sort of> degree modifier construction (curve "A"), rises above the near-grammaticality threshold. Moreover, the reverse hybrid is lower in status than both the <a sort of> construction and the <sort of> through much of the post-training process. Unfortunately, this change in the hypothesis about the relationship between network-predictions and grammaticality adds some complexity to the picture. It means it is no longer possible to view the network's likelihood estimations as direct approximations of observed likelihoods. This makes it harder to theorize about grammaticality predictions. On the other hand, it does not seem implausible to suppose that speakers judge grammaticality by monitoring sentences as they are being produced and and that they equate global well-formedness with continuous or conjunctive local well-formedness. Normalizing L-values approximates this hypothesis more closely than does evaluating the string likelihoods.

This chapter has shown how the network model may be useful in analyzing certain hybrid structures that present problems for most current theories of grammar. The advantage that the network model brings to this domain is its capacity for generating novel forms by interpolation. An important question for future research is whether this interpolative mechanism can be used to provide a more principled account of the intermediate structures that are quite pervasive in language and central in linguistic theorizing: for example, *clitics*, which seem intermediate between syntax and lexicon, *auxiliary verbs* which seem intermediate between main verbs and inflectional elements, *clause boundary* elements which show ambivalence between membership in matrix and embedded clauses.

(129) the soft cheese really sumptuous p [(N)ear Grammatical]

(130) a sort of dense cheese rolls p [a(M)biguous]

(131) the block is sort of dense p [degree modifier (S)ort-of]

(132) the block is rather dense p [(G)rammatical]

(133) the block is a sort of dense p [degree modifier (A)-sort-of]

(134) the block is sort of marmalade p [(R)everse hybrid]

---

[6]Recall that L gives the natural log of the likelihood of observing a string if the network is interpreted as a string generator (Chapter 3, Section 6.2).

Figure 6.6: Simulation result: rise of the hybrid, DegMod *a sort of*, in conjunction with the rise of DegMod *sort of*. [*L*-values normalized for sentence-length]
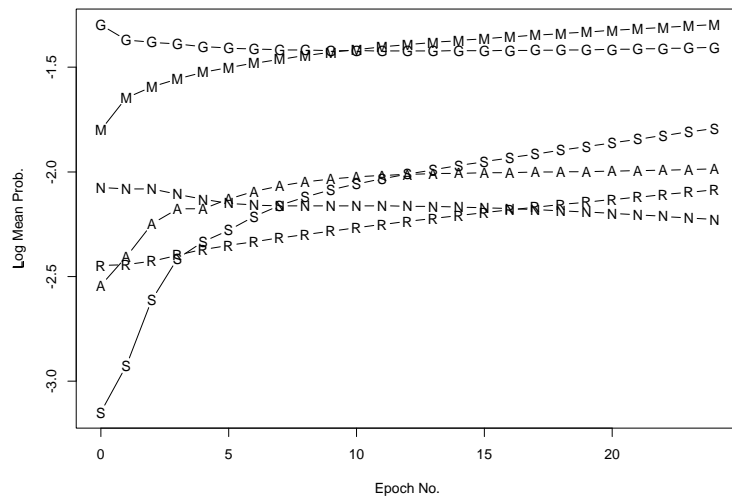
Chapter 7

Conclusion

7.1 Summary

This chapter summarizes the main arguments of the thesis (Section 1), remarks on some implications for the theory of language change (Section 2), and points out interesting new lines of research suggested by the current one (Section 3).

7.1.1 Overview

The aim of this study has been to learn about language structure by examining language change. I have focussed on morpho-syntactic change as it is reflected in successive, closely-spaced historical texts in order to learn something about the dynamical properties of language structure. I have argued for a representation (called *Restrictive Continuity*) that is both diachronically continuous under evolutive grammar change and representationally restrictive in the sense that grammatically-related elements are constrained to change in a correlated manner. A key step in developing this representation was to let quantitative properties of corpuses have a bearing on grammatical structure. A major advantage of the model is that it provides a more constrained theory of structural revision or *reanalysis*.

7.1.2 Motivations

Two bodies of historical work provided the major motivation.

Kroch 1989a, 1989b, Santorini 1989, 1992,, Pintzuk 1991, Taylor 1992, Fontana 1993, and others working in the **Competing Grammars** framework have shown that major syntactic changes can be accurately modelled by positing a gradual shift in the probabilistic weighting of two alternative grammatical systems. Their work suggests that some of the major correlations in diachronic change may be predicted by setting the parameters of Principles and Parameters theory (e.g., Koopman 1984, Chomsky 1986, Chomsky and Lasnik, To appear) probabilistically. The major shortcoming of their work is that it appears to be necessary to posit significant reanalyses at certain times in order to predict the right correlations, and although the distributional evidence in favor of the reanalyses seems to have accumulated critically by the time they occur, the representations adopted under the Competing Grammars framework are insensitive to this evidence. (Chapter 2, Section 1)

Work in the domain of **Grammaticalization** (e.g., Givon 1971, Heine and Reh 1984, Traugott and Heine 1991, Hopper and Traugott 1993) has also provided a major motivation for it indicates that there are strong structural constraints on the changes any particular language can undergo. The historically inherited morpheme-ordering rules seem to constrain the set of reanalyses that can occur, but do not define sufficient conditions, for it seems to be necessary for a grammaticalizing element to undergo a certain kind of semantic/pragmatic change, often described as *abstraction*, before it can start to take on qualitatively new characteristics. Explicitly characterizing these semantic developments in expressive terms has proved difficult. But it appears that quantitative distributional analysis can provide a useful alternative way of making the conditions explicit: when elements become more abstract, their relative frequency behaviors change in systematic and measurable ways (cf. Bybee 1985, Bybee, Pagliuca and Perkins 1991, Bybee (forthcoming)). (Chapter 2, Section 2)

7.1.3 Proposal

I propose a model, called the **Restrictive Continuity** model, which takes advantage of the insights just described. Its main properties are that it employs

a continuous mapping from representations to behaviors; the mapping itself evolves continuously with time; and yet the representation is very compact compared to the space of behaviors, so strong correlations among behaviors are predicted. Behaviors are modeled as probabilities of sequences of elements drawn from a finite vocabulary. (Chapter 1, Section 6; Chapter 3, Sections 6–8).

The model is implemented in a recurrent Connectionist network. The network has one layer of input units, one layer of output units, and one intervening layer of hidden units. The recurrence consists in complete interconnectivity among the units of the hidden layer. The network is trained on the task of predicting next-words in a corpus on the basis of previous words, by unfolding the recurrent part of the architecture in time and backpropagating an error signal. Such a network is particularly suitable for capturing the strong constituent well-formedness constraints which characterize natural language: it represents contextual independence by driving the relevant representations onto orthogonal dimensions. (Chapter 3, Sections 1-4)

Change is modeled as retraining a once-trained network to take on new targets. This technique implements the Restrictive Continuity hypothesis by guaranteeing that representational change is approximately continuous. It permits one to model persistent external pressures on the language as distortions in the function that defines the target behavior of the network. (Chapter 3, Section 6)

7.1.4 Predictions

The Restrictive Continuity model makes three empirical predictions which I have labelled **Frequency Linkage**, **Q-Divergence**, and **Hybrid Structures**.

Frequency Linkage refers to the prediction that elements with similar distributional structure should undergo correlated changes in proportion to the degree of their similarity. The prediction is thus related to Kroch 1989a and 1989b's **Constant Rate Hypothesis**, which holds that elements in the same grammatical class should undergo perfectly correlated frequency changes. The two predictions coincide in cases where elements in the same grammatical class have essentially identical distributional behavior. But they differ in cases where elements are classed together but some are more typical members of the class

than others (Chapter 3, Section 8; Chapter 4).

Q-Divergence refers to the prediction that quantitative changes may be correlated with qualitative or "categorical" changes. In general, correlated change predictions in the network model follow from the restrictiveness of the representation: since similar behavioral properties are mapped to nearby locations in the hidden unit space, change in one behavior is correlated with change in a related one. Since the network is sensitive to quantitative properties of language (i.e., how frequently particular words are used in particular contexts), change in these quantitative characteristics of already-grammatical elements can have the effect of moving the likelihood of a related element across the grammaticality threshold. Thus there can be a correlation between a quantitative and qualitative development. This prediction is not made by the Competing Grammars model because quantitative changes are taken to be independent of categorical status in that model. (Chapter 3, Section 8; Chapter 5)

Hybrid Structures are structures that seem best described as splicings-together of pieces of independently motivated constructions. It turns out that hybrid structures tend to be associated particularly with episodes of historical transition, when the parent structures are in alternation with one another. The network model predicts this effect because when an element (called a "bridge") is alternating between two different behaviors, the network first approximates its distribution by averaging over its contrasting behaviors. Consequently, it chooses a representation that is intermediate between the parent representations. This makes the element reasonably suitable for filling both roles simultaneously. Eventually, the network can learn to detect the contextual conditioning factors that distinguish the two alternatives so it develops two separte representations for the item in question, and the hybrid prediction goes away. The recurrent network is especially prone to predicting the appearance of a hybrid structure in cases where the conditioning features that distinguish the two parent structures are displaced from the bridge in the symbol sequence. This is because displacement in the symbol sequence is implemented as occurrence on contrasting layers of the unfolded network and the detectability of the error signal diminishes as the distance between layers increases. The Competing Grammars model does not make hybrid structure predictions for it posits only probabilistic, not structural, mixture. (Chapter 6)

7.2 Implications for the theory of language structure

The effectiveness of the Restrictive Continuity model at making appropriate diachronic predictions stems fundamentally from the synchronic representation it is based on. I remark here on three properties of the included theory of synchronic representation (at the morpho-syntactic level) that are of particular interest because they contrast with standard assumptions or methodologies.

7.2.1 Morpho-syntax is sensitive to quantitative contrast

The studies of Frequency Linkage and Q-divergence presented in Chapters 4 and 5 indicate that quantitative contrast in language use is pertinent to morphosyntactic representation. For example, in the study of the development of periphrastic *do*, I argued that similarity between the representation of *do* in the affirmative declarative context and its representation in questions and negative sentences could only be tolerated when the frequencies of the periphrastic forms were low. As the frequencies of negatives and questions rose, the cost of keeping the wordy affirmative declaratives in the same category became too great and so the representations diverged. Similarly, in the study of the development of *sort/kind of* as a degree modifier, the rise in the frequency of *sort/kind of* before adjectives made it too costly to keep the representation of this collocation identical to that of all other Noun+Preposition sequences. A convenient and less-costly alternative was to move the representation for this category over in the direction of the representations for the Degree Modifiers. Again, a pure frequency contrast had implications for the categorical representation.

It is probable that the picture I have painted for some of these cases is too simple in the sense that major structural differences are taken to hinge on one-dimensional frequency contrasts. Given that real language use seems only rarely to repeat itself verbatim for more than a couple of phrases in sequence, a large Connectionist network trained on real data will be prone to form very high-dimensional representations, with different dimensions corresponding to each of the many contextual features that are pertinent to the distribution of words. Similarity between representations will thus arise because certain collocations

share a wide variety of distinct contexts rather than because they appear with high frequency in exactly the same sentences. At first glance, this observation seems to imply that quantitative information is not so important after all: perhaps a suitably elaborate categorical model can do the job. But this is an illusion. First, to establish that something occurs in a "wide variety" of contexts requires a quantitative evaluation. One may avoid counting instances of word-use but one will nevertheless have to count types of contexts. Second, though it may not be necessary to employ real numbers in the grammatical representation—under the proposed scenario, all distinctions can be made in binary terms—the case studies in this thesis indicate that it is crucial to be able to compare observed behavior with potential behavior. Such comparison is outside the descriptive capability of a standard generative grammar but constitutes the essence of quantitative representation.

7.2.2 Morpho-syntax computes intermediate representations

A second implication for synchronic linguistics is that an efficient grammar must be able to compute intermediate representations. By "intermediate representations", I mean representations that can be specified as a weighted sum of other representations with positive weights. Intermediacy in this sense implies that there is a distance measure associated with the representation space and hence that the space can be interpreted as a metric space (see Chapter 3, Section 7). In fact, a standard categorical parametric representation does have one natural interpretation as a metric space. For example, in the case in which all parameters are assumed to be binary, one can define the distance between two parameter settings as the number of bits on which the settings differ (this measure is called *Hamming distance*). Although Hamming distance does give rise to a notion of intermediacy which makes some correct diachronic predictions, the standard representations are too coarse to encode many diachronically relevant similarity relationships.

Various hybrid structure phenomena discussed in earlier chapters support this point. As I noted in Chapter 6, hybrid structures seem to be splicings together of pieces of independently motivated constituents. They are clearly

difficult to generate under standard analyses because the standard analyses take the relevant constituents as indivisible. On the other hand, they are naturally modelled with intermediate representations in a metric-space representation because such a representation takes similarity to be inversely related to distance and such constructions are partially similar to each of their parent constructions.

The succession-of-lobes data reported by Craig 1991 (see Chapter 2, Section 2) is a case where a well-motivated categorical parametric model seems to be in a position to make some diachronic predictions. Recall that Craig posited elements undergoing a three-stage process: postposition > clitic preverb > lexical preverb. It is generally acknowledged that clitics are somehow structurally intermediate between full-fledged words and affixes. For example, Inkelas 1989 proposes to capture the range of prosodic types observed within the lexicon by positing two binary parameters: prosodic dependence and morphological dependence. The four combinations of the settings of these two parameters yield the four observed types as shown in Figure (7.1). If we map the parameter

Figure 7.1: Four types of sublexical units specified by two binary parameters. [based on Inkelas 1989]

	Morphologically Dep.	Morphologically Ind.
Prosodically Dep.	affixes	clitics
Prosodically Ind.	roots	stems

settings into binary values, then the Hamming distance between elements that are catercorner in the table is 2 while the Hamming distance between elements adjacent in the table is 1. We could use this fact to predict, for example, that stems will become clitics before they become affixes. This is consistent with a number of observed cases. Unfortunately, the same principle implies that stems will become roots before they become affixes. I know of no evidence that this occurs with any regularity.

A case like the development of *be going to* from Motion Verb to Equi Verb to Raising Verb is also hard for a categorical parametric model to make accurate predictions about. One could posit the features,

(135) +/- Main Verb

+/- Equi

and let the featural assignments be as in (136).

(136)

	Main Verb	Equi
Motion *be going to*	+	+
Equi *be going to*	-	+
Raising *be going to*	-	-

This makes the Hamming distance between Motion and Raising equal to 2 while the other two distances are both 1. Again, if we insist on gradual change in Hamming location, the possible paths of change are consistent with the observed change discussed in Chapter 5. However, the claim that Motion *be going to* is [+ Equi] is an unmotivated stipulation. It is not clear that the Equi/Raising distinction makes sense when applied to main verbs. If it does make sense, then it seems most reasonable to define it in terms of the range of types of subjects that the main verb permits. But on this view, it might well be argued that Motion *be going to* is [- Equi] since it has been able to occur with inanimate, non-anthropomorphized subjects from the earliest times (see Chapter 5, Section 2.2.2). The Connectionist model makes appropriate predictions about this case because Motion *be going to* had a *quantitative* distribution that was much more like that of an Equi verb than like that of a Raising verb during the period of initial auxiliary emergence. Thus this example argues in favor of a representation that takes quantitative information into acount in defining intermediacy.

7.2.3   Methodological principle: assess global harmony

The Connectionist model employed here is trained by random sampling across an entire "language". Representations for different pieces of the language are built up simultaneously and properties of one part can, in principle, influence the representation of any other part. Thus it is important to sample across the spectrum of the data.

An example of a case in which this principle plays an important role is the development of *sort/kind of* as Degree Modifiers. If one were to naively examine *sort/kind of* early on, when they were not yet showing all manifestations

of being Degree Modifiers, one might be inclined to analyze them merely as Noun+Preposition sequences with a few peculiar properties. One would miss the fact that the peculiar properties made them more like Degree Modifiers. Of course, it is assumed that standard generative methodology avoids such pitfalls because one does not fail to notice correlations that might be relevant to the analysis of a particular form. But this does not prevent one from making mistakes. The troublesome thing about the standard practice is that it is rarely specified how one decides which parts of the language have a bearing on particular analyses. It is thus an advantage of the Connectionist network that the impact of every observation on the representation is explicitly computed. Admittedly, by using as training data the output of intuitively-designed toy grammars which are intended to approximate only "relevant facts" about the historical language states in question, I have nullified this potential advantage in the experiments reported above. Nevertheless, it seems worthwhile to emphasize the worth of such explicitness for it is desirable and may well be practicable in larger projects than the current one.

## 7.3   Implications for the theory of Language Change

This section considers some of the implications of the Restrictive Continuity Model for the theory of language change. I first examine the role of the indirectness of grammar transmission from generation to generation in grammatical change, arguing that children's lack of information about the nature of their parents' grammar(s) may not play as great a role as has generally been assumed. I then consider the contrast in historical theory between *reanalysis* and *analogy*, arguing that we can do without this distinction under Restrictive Continuity, and the result is a simpler theory. Finally, I review the implications this thesis has for syntactic innovation, noting that, although it does not discuss cases of radical syntactic newness of the kind that some earlier studies have been concerned with (e.g. Lightfoot 1979, 1991), it offers insight into such cases nevertheless.

7.3.1 The Role of Indirect Transmission in Language Change

7.3.1.1 The Indirect Transmission Model

In the generative era it has become common to assume the Indirect Transmission model of language change shown in Figure 7.2. The figure is intended to emphasize the fact that language-learners do not have direct access to the grammars that are presumed to exist in speakers' heads: all they can do is observe the output of those grammars and try to adopt a generating mechanism that matches their observations.

Figure 7.2: The Indirect Transmission Model (based on Andersen 1973: 767—see also King 1969: 85, Anttila 1992: 19).
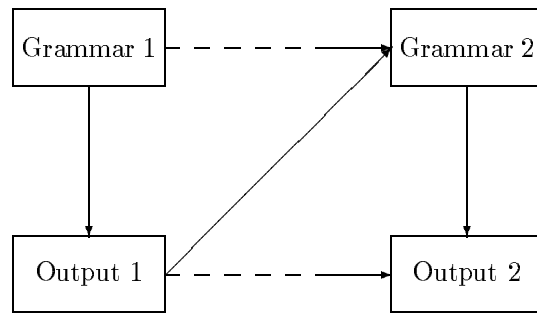
The fact of Indirect Transmission has inspired two assumptions whose validity is questionable:

(i) Change happens because the output that learners are exposed to does not uniquely determine the generating grammar. In other words, change is due to *information loss* stemming from the indirectness of transmission.

(ii) A theory of change must characterize the solid arrows in the figure rather than the dotted arrows.

Assumption (i) is certainly possible *a priori* but it does not have to be the case simply because transmission is indirect: it could be that learners have plenty of information available but some other influence motivates them to adopt a grammar different from that of their forebears. I suggest a plausible alternative, called *Sociolinguistic Projection*, in Section 7.3.1.3 below.

Assumption (ii) only makes sense if it is a comment about methodological feasibility: perhaps it is easier to model historical change if we break the task into two subparts—modeling generation of speech by a grammar and modeling discovery of a grammar on the basis of examples. If we succeed at both of these tasks, however, then we can trivially model direct grammar-to-grammar transmission by combining the two submodels into a single model. In other words, making a model of both types of solid arrows in the figure is equivalent to making a model of the dotted arrows. Thus the mere existence of indirect transmission is not an argument against there being a direct model. Moreover, given the results described in the previous chapters, which assume direct transmission, it seems like a direct transmission model may be an easier first-step.

7.3.1.2 Deduction, Induction, Abduction

Because I am skeptical about Assumption (i) above, I am skeptical of Andersen 1973's claim that the distinction between three modes of inference—*deduction*, *induction*, and *abduction*[1]—is useful for thinking about the role of Indirect Transmission in diachronic change.

Deduction matches a universal LAW (e.g., *All men are mortal*) with an observed CASE (e.g., *Socrates was a man*), to derive a RESULT (e.g., *Socrates was mortal*). Induction extrapolates from multiple CASEs and RESULTs to derive a general LAW. Abduction combines a RESULT with a LAW to produce a conjecture about what CASE might be responsible for the RESULT: e.g., Given that we know *All men are mortal* and we have observed that *Socrates died*, we may note that we could predict the RESULT indirectly by making the

---

[1] These terms and their initial clarification are due to the philosopher Charles Peirce.

assumption that *Socrates was a man.*

According to Andersen, abduction plays a crucial role in language change. The linguistic correlate of a LAW is rule made available by Universal Grammar. The linguistic correlate of a CASE is a language-particular classification choice (e.g., the choice to assign a particular phoneme sequence to a particular lexical class). The linguistic correlate of a RESULT is an observed language behavior. Grammar change occurs when a language learner matches some utterance she has heard (a RESULT) with a different set of rules of grammar (or LAWs) from the one that her predecessors used to generate it, thereby assigning different classifications (or CASEs) from the ones her predecessors applied. Such grammar change (or "reanalysis") can result in the appearance of a new construction if the new rule or new classification is capable of generating constructions that are not licensed under the old grammar.

In the case studies of morpho-syntactic innovation discussed above, the abduction hypothesis seems quite implausible. For example, to explain the innovation of Degree Modifier *sort of* by abduction, we would presumably argue as follows. Earlier speakers used the phrase structure rules (or "LAW"s) shown in Figure 7.3. to generate Noun Phrases like (137).

(137) c. 1675 …the blunt edges of it upon a <u>kind of</u> large Pin-cushion cover'd with a course and black woollen stuff. Boyle, *Electricity and Magnetism.* [HELS]

But there is a different set of LAWs, that can generate the same string of words with a very similar meaning (i.e., the same RESULT) (Figure 7.4). We hypothesize (following Andersen) that at some point prior to the 19th century (when the first evidence for DegMod *sort/kind of* appeared), certain speakers, acting abductively, attributed analysis (7.4) to certain strings of the form <Det *sort/kind of* Adj N>. This meant *sort/kind of* were now classified in the lexicon as DegMods. When this new classification was employed under a rule like (138), it resulted in the appearance of a novel construction like (139).

(138) **VP → BE AdjP**

(139) We are sort of amazed.

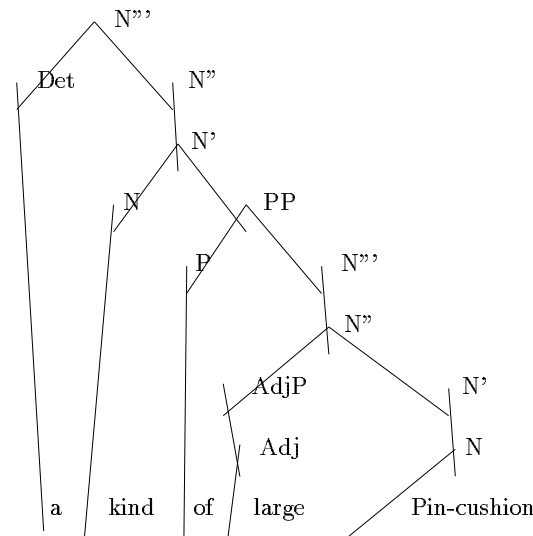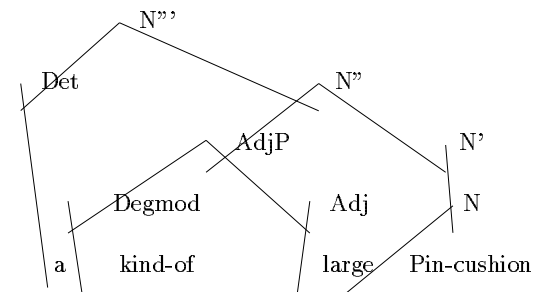Figure 7.3: Noun Preposition *kind of.*

Figure 7.4: Degree Modifier *kind of.*

There are a number of problems with this account. It presupposes that the words around *sort/kind of* receive the analyses standardly assigned to them (e.g. *a* is a **Det** *large* is an **Adj**, etc.). But if these elements receive their standard analyses, why shouldn't *sort of* receive its standard analysis, which was presumably the Noun-Prep analysis before there was any evidence to the contrary? Also, abduction is supposed to make a hypothesis that is consistent

with what the learner has seen up to some point. It seems very dubious that a learner more than a few days old would manage to avoid hearing *sort/kind of* in the context <Det ...N>, and also avoid hearing *sort* and *kind* separately as Nouns, and *of* separately as a Preposition. We must assume, then, that a learner will not have abduced *only* the new analysis for this construction, but that she will have posited *two* analyses for *sort/kind of* by abduction. But if learners are permitted to freely posit multiple analyses for particular words, we should expect change to run rampant. Even if semantic equivalence is invoked to constrain the range of possible changes, we must wonder why the analyses in (140) did not occur.

(140) a[Det] kind[DegMod] of-large[Adj] pin-cushion[N]

a-kind[Det] of[DegMod] large[Adj] pin-cushion[N]

Andersen 1973 proposes that although the output of abduction may project a massively different grammar from the original generator, the effects of such radical underlying change are masked initially by "adaptive rules" which patch-up the output of the new grammar to make it look like the output of the old grammar. An adaptive rule which changed DegMod *sort/kind of* into a Noun-Prep sequence would have to be implemented as a post-lexical-insertion transformation that replaced one piece of tree structure with a wholly different type. This kind of tree-rewriting transformation is radically divergent from the current constraints that are generally agreed to hold on transformations. The textual evidence indicates that unambiguous *sort/kind of* first came out in the environment <BE *sort/kind of* Adj>. Even a subcategorization constraining the rule **DegMod** → *sort/kind of* to occur only in this environment would require completely unmotivated features under a non-transformational theory like GPSG (Gazdar et al. 1984) or HPSG (Pollard and Sag 1987).

In sum, the abduction approach to the *sort/kind of* case raises a host of problems. These problems are not specific to the case at hand but are problems for abduction accounts of most syntatic and morpho-syntactic reanalyses. Thus pinning this kind of change on information loss due to Indirect Transmission is not very satisfactory.

7.3.1.3 An alternative to Change by Information Loss: Sociolinguistic Projection

If information loss is not responsible for structural change, what is? Variation researchers have long been concerned with an apparently distinct problem: What drives quantitative change? That is, when it makes sense to view two (or more) linguistic expressions as being in probabilistic competition for expression of a meaning, and we see that over time, one of them gradually edges out the other one (e.g., Ellegard's data on the history of *do* discussed in Chapter 4; the spread of *going to* in place of *will* as a marker of in recent English discussed in Chapter 5), what is it that drives the change along? Eckert 1988 notes that sociological considerations may provide an effective explanation, at least in the case of sound change:

> Comparison of speech patterns between parent-aged adults and immediatley older peers establishes for each cohort of emerging adolescents the age-salience of innovative phonology and the established direction of change. The simple continuation in this direction, or exaggeration of changes already in progress, thus has the potential to symbolize age-group identity and autonomy from the parental age group. [p. 198]

I call this hypothesis the **Sociolinguistic Projection** hypothesis and suggest that it may apply to morphological and syntactic change as well. Taken in conjunction with standard variable-rule and variable-parameter models of grammar, this hypothesis only explains how quantitative properties of usage are modified once the competing structures are made available by the grammar. Thus it seems not to offer any insight into how structural change occurs. But under the Restrictive Continuity model, quantitative change and structural change are underlyingly the same. Trends in quantitative change produce structural revision if they proceed sufficiently far. All the case studies in this thesis have illustrated this effect (periphrastic *do*—Chapter 4; Degree Modifier *sort/kind of*—Chapter 5, Section 2.1; Auxiliary *be going to*—Chapter 5, Section 2.2). Thus, in conjunction with the Restrictive Continuity hypothesis, Sociolinguistic Projection becomes a much more powerful explanation. We no longer need to attribute a major role to information loss. In fact, we can say that speakers could, if

they "wanted to", reproduce their parents' grammars exactly with respect to at least the significant grammatical contrasts like the ones studied here. That they don't do so is hypothesized to be a matter of social choice, not logical inevitability.

Indeed, this view is consistent with one of Eckert's findings about a sound change taking place in the region around Detroit. Two different socially-defined groups of adolescents ("Jocks" and "Burnouts") are respectively less and more progressive with respect to the sound change—backing and lowering of (uh). The less progressive group, the Jocks, are culturally more cooperative with, and similar to, the group of adults as a whole. Since people evidently choose to be Jocks or Burnouts for a host of culturally generated reasons, it seems likely they are positioning their language usage on the scale of progressiveness with respect to the sound-change in order to help define their identities. They are not positioning themselves by accident because they misguessed the grammars of their parents. Otherwise, why would their grammatical behavior line up so well with their behavior in so many language-independent domains? Thus, Eckert's observations lend support to the theory that linguistic change is exploratory, occurring by extension from the received system, rather than haphazard, occurring by (mis-)abduction of an earlier grammar.

7.3.1.4   Local versus Global Abduction

I noted in Chapter 3, Section 6.2 that trainable network models are naturally suited for modeling language acquisition, even though this capability does not play a role in the Change Model I have outlined here. Given this fact, it is worth commenting on a closely related change model in which Indirect Transmission is explicitly simulated: a network is trained on the received state of the language at some point and then its output becomes the input for another network (see Denaro and Parisi 1993).[2] Under this scenario, the model may exhibit a grammar-distorting effect which bears some resemblance to Andersen's abduction mechanism, although it is not the same thing. I noted in Chapter 3 that feedforward networks trained with backpropagation can be thought of

---

[2]Or, even more plausibly, a pool of finite-lifespan networks can provide each other with inputs in analogy with an evolving society.

as performing gradient descent in an error space (See Chapter 3, Section 2.1). One "problem" with gradient descent, often noted, is that the error space may contain local minima which trap the search mechanism. Whether the network gets trapped in such a local minimum depends on whether its random initial position is in the "watershed" of such a minimum. If each new network in the Indirect Transmission paradigm just described is started in a different random starting position, then, assuming there are local minima in the error space, it is possible that one of the successively trained networks will fall into one. This will create a change in the "grammar" that generates the outputs for successive networks. The distinct error space associated with this new grammar may, of course, also have local minima which some successor network will fall into. Thus a certain kind of information loss (loss of information about the random starting position) gives rise to structural change. In this sense, the mechanism can be counted as a type of abduction. But it is different from the type of abduction Andersen 1973 discusses, for the "wrong solutions" the networks arrive at are based on global information about the structure of the language (as gleaned from exposure to many examples) rather than on local attempts to match individual constructions with one of the rules that universal grammar provides. I will call this variety of abduction "global abduction" to contrast it with the variety Andersen discusses, which I'll call "local abduction". Global abduction is more constrained than local abduction in that it needs to posit no adaptive rules. Also, the misapproximations it leads to are not arbitrary misapproximations but rather particularly natural ones in the sense that they largely recapitulate the effects of the earlier grammar.

7.3.2   Reanalysis and Analogy

I remarked at the end of Chapter 5 that the hypothesis of underlying continuity of representational change may make it possible to do away with the distinction between *reanalysis* and *analogy* as separate mechanisms of change.

Reanalysis, as defined by Langacker 1977, involves

> change in the structure of an expression or class of expressions that does not involve any immediate or intrinsic modification of its surface manifestation. [p. 58]

For example, in one type of reanalysis, called "boundary loss", a sequence of two or more morphemes is reclassified as being only one morpheme. Langacker describes the case of the Proto-Uto-Aztecan absolutive suffix, *-ti*, which occurred in sequence with an accusative suffix *-a* on accustive nouns. The *a* conditioned deletion of the preceding *i* so the resulting form was *-t-a*. When absolutive inflection was lost generally in the descedent Uto-Aztecan language, Yaqui, the *t* in such forms was not lost but rather was treated as part of the accusative suffix. This account suggests a model in which a covert reanalysis gave the *t* a different status prior to the change that eliminated all elements of its original status, thereby permitting it to survive. Presumably, the case of Degree Modifier *sort/kind of* discussed in Chapter 5 would be treated as a case of boundary loss on this theory, as well as a case of what Langacker calls "semantic (or syntactic) shift" (pp. 84-5). Reanalysis is radical, on this view, in the sense that there is a fair degree of arbitrariness to the change: there is no telling from the structure itself whether it is going to undergo boundary loss, boundary shift, boundary creation, semantic or syntactic shift, or no change at all. This concept of radical reanalysis is related to Indirect Transmission if it is assumed that radical reanalyses are due to lack of information content in the data learners are exposed to.

The term "analogy" is from the Neo-grammarians. They used it to speak both of the fact that much observed linguistic material can be organized into systematic combinatory paradigms and of the active tendency of speakers to create new expressions by recombining old forms in new ways. Neo-grammarian analogy is now mostly rememberd in the second sense, as a tendency that distorted the effects of sound change and thus produced apparent violations of the Neo-Grammarian hypothesis that sound change is regular.

Under the New-grammarian view, analogy was not opposed to reanalysis, but with the rise of Structuralism it became so. For example, Meillet 1912 points out that analogy alone does not seem capable of providing an explanation for all types of structural change:

> Whereas analogy may renew forms in detail, usually leaving the overall plan of the system untouched, the *grammaticalisation* of certain words creates new forms and introduces categories which had

> no linguistic expression. It changes the system as a whole. [p. 133, translation from Hopper and Traugott 1993, p. 22]

Although Meillet does not speak explicitly of 'reanalysis', it is clear that his *grammaticalisation* is closely related to one current way of using the term "reanalysis". His position is echoed in Lightfoot 1979 who claims that

> in asserting (correctly) the importance of surface analogies, one should not deny any role to abstract derivations. The seeds for [the] re-analysis [of Portugese *se*-impersonals and ME impersonals] lay in the complexity of certain derivations, and that complexity is an essential ingredient of any explanation of why the re-analysis took place. There are other changes, e.g. that of the English modals, where surface analogies play no detectable role. [pp. 358–9]

It will clarify things if I note that the terms "reanalysis" and "analogy" are here being used to label several distinct oppositions. The first sense, the one I attributed to Langacker 1977 above, can be summarized as follows:

Def. 1 (Reanalysis vs. Analogy):

> **Reanalysis** makes an invisible change in the analysis of a form.
> **Analogy** extends the change to produce surface differences.

If we translate Meillet's *grammaticalisation* as 'reanalysis', then the following definition fits Meillet's usage, and is indeed a current sense of the term:

Def. 2 (Reanalysis vs. Analogy):

> **Reanalysis** creates new grammar.
> **Analogy** extends existing devices to new uses.

Lightfoot, while agreeing with Meillet's point, is using the words slightly differently:

Def. 3 (Reanalysis vs. Analogy):

Reanalysis is grammar change.

Analogy is a motivation for grammar change: change is taken to be in aid of making form/meaning pairings more systematic.

My purpose in this section is to note that if local abduction does not, after all, play a significant role in structural change, then reanalysis as defined in Definition 1 is not a particularly useful concept. In place of "covert reanalysis", we can posit "tiny reanalysis": structural change occurs via a sequence of small adjustments that approach continuous change in the limit. The reason it looks like there is covert reanalysis is that most of the tiny representational changes make only quantitative adjustments in the behavior of the language, so from the standpoint of standard grammatical analysis they are invisible. Moreover, the tiny changes involve correlations among a variety of surface behaviors because of the reduced dimensionality of the representation space. Consequently, when the changes do bring about qualitative shift, a large number of behaviors shift together. This makes it look as though a radical structural change has occurred. In fact, a radical reanalysis has occurred, relative to some earlier state of the representation. But the representation has never changed abruptly.

I take up Meillet's distinction between truly innovative change and analogically motivated change (Definition 2) in the next section.

7.3.3   Syntactic Innovation

What sort of "syntactic innovation" is this thesis about? Unlike some past studies of syntactic change (e.g. Lighfoot 1979, 1991; Kroch 1989a, 1989b; Roberts 1985) it is concerned with fairly local innovations rather than resettings of global parameters. Would it be more appropriate to say that it is about "lexical innovation" instead? Not really. There are several reasons that the results here are of relevance to the study of syntactic change proper, as opposed to merely lexical reclassification.

In the case of the reanalysis of *sort/kind of* as Degree Modifiers, the result of the change may be reasonably described as involving the addition of a new element to the set of lexical items classed as Degree Modifiers.[3] However, the starting point of the change, in which *sort/kind of* are analyzed as Noun-Prep sequences, clearly involves an instance of productive word-combination, and thus must be considered a syntactic phenomenon. In order to state constraints on the process whereby some combination of words gets reclassified as a single word, as is the case here, and indeed is very common in grammaticalization, it is necessary to have a representation system in which some kind of mapping between syntactic and lexical units is expressible. It is an appealing feature of the learning mechanism described here that all classification is based uniformly on efficacy with respect to the next-word (or speech-stream-element) prediction task. Therefore, if two (or several) words in sequence have essentially the same distribution as one individual word, then the hidden-unit trajectories pass through nearly the same points before and after the multi-word sequence as before and after the single word in context. In this regard, the network representation is consistent with the thesis of Construction Grammar (e.g., Fillmore, Kay, and O'Conner 1988) which distinguishes itself by "not requiring strict separation between lexicon and grammar, or between the idiomatic and the general" (Fillmore 1994; see also Section 3.3 below.)

The reclassification of *be going to*, though it was almost certainly influenced by the presence of raising verb complexes like *seem to*, *be likely to*, and *be prone to* as well as lexical raising verbs like *will*, has led to a form that is syntactically distinct from all of these. In particular, it seems plausible to claim that current *going to* with future meaning is almost a single word on the grounds that it undergoes idiosyncratic phonological reduction in contrast with its still-surving purposive/locative antecedents (143) and is essentially intolerant of intervening material (144).

(143) a. She's going to/gonna be 32 tomorrow.

---

[3] It is worth noting, however, that the distribution of degree-conveying *sort/kind of* does not coincide perfectly with that of any other Degree Modifier, or meta-linguistic hedge as the following examples indicate:

(141)   He sort of/kind of/*somewhat/*rather/really/actually swam over and took hold of the side. (based on Bolinger 1972).

(142)   Really/actually/*sort of/*kind of, they mean it.

b. She's going to/*gonna visit her grandfather. (motion sense)

c. She's going to/*gonna Sioux St. Marie.

(144) a. It is going *now/*soon/??incidentally to rain.

b. Fred is going now/soon/incidentally to visit his grandfather. (motion sense)

c. He is going now/soon/incidentally to Evening Mass.

It also contrasts with other raising verb complexes involving *to* in these regards:

(145) a. It's prone to/*pronna snow here on alternate days.

b. They seem to/*seema like it here.

(146) a. It's prone now to be dewy by 7.

b. They seem now to have learned the habit of changing their clothes when they first arrive.

And although it can be said that the distribution of current *be going to* is very similar, though not identical to that of *will*,[4] which is a lexical modal verb, it cannot be claimed that the whole phrase has simply been reclassified as a lexical modal, given the continued inflection of *be*. It appears, therefore, that *be going to* with future meaning is syntactically novel. I must point out, however, that the simulations I describe in Chapter 5 do not predict auxiliary *be going to* as a novel syntactic structure in this sense because of the way I approximate the historical data. It seems likely that a more complex model, in which the sequence <be going to> is not treated as a single lexical item, will exhibit the desired effect, especially given the comparable result in the *sort/kind of* simulation. I leave this as a matter for future research.

I chose, in this study, to focus on these local, "relatively lexical" developments because they are easier to understand and easier to collect data on than some of the major syntactic revisions that have also been studied (e.g., the evolution of the English modals as a class [Lightfoot 1979, 1991], the development of the English verbal gerund [Abney 1987], the regulation of English *do*-support, the switch from Ergative to Active case-marking [Harris 1985]). Because the

---

[4]See Coates 1983 for discussion and examples.

network model does not make an *a priori* distinction between syntactic and lexical class membership, its predictions extend to these kinds of cases as well. For example, the model predicts that major syntactic change can only happen gradually. It seems likely, in fact, that contrary to earlier claims for the existence of *radical reanalysis* (e.g., Lightfoot 1979), the well-studied cases of major syntactic revision involve protracted, incremental processes: either incremental change of the frequency of a form (e.g., Kroch 1989a, 1989b, Santorini 1989, Pintzuk 1991, Fontana 1993) or incremental addition of lexical items to a class (e.g., Plank 1984; Fischer and van der Leek 1983; Israel 1993). All of the studies in this thesis have indicated that incremental change in the *frequency* of a form can produce a restructuring of the network's representation (see Chapters 4 and 5). It turns out that incremental change in lexical class-membership can also lead to restructuring under the Connectionist representation. This is because, in the network, similarity of behavior is determined by proximity of representation in the vector sense. In the case of a frequency-change preceding a restructuring, one vector becomes more similar to another by becoming longer or shorter. In the case of incremental lexical change leading to a restructuring, one vector becomes more similar to another by changing its direction (and also, potentially, its length). In either case, restructuring hinges on relative distance between vectors, a kind of similarity. In a preliminary study of the development of the English gerund I have shown how incremental addition of verbs to the class of participants in a simplified "verbal-gerund construction" (e.g. *after seeing the bobcat*) foments a series of novel-construction predictions by the network (Tabor 1992).

In the previous section (2.2) I argued that the distinction between reanalysis as covert structural change and analogy as surface manifestation of structural change ("Definition 1") is not a particularly useful one. Here I would like to suggest that Meillet 1912's contrast between innovative grammar change and mere extension of existing devices ("Definition 2") is also formally unnecessary, although it is useful as a descriptive distinction. According to Meillet, there are certain kinds of grammatical changes which can be motivated analogically on the basis of existing forms, but there are other types, involving the innovation of new grammatical categories, which cannot. A case which seems relevant to the point is the development of French *pas*, originally merely a noun, 'step' (from

Latin *passum*), into a negation marker. Hock 1986 identifies the following stages (p. 194):

(147)   Stage I   il ne vait/va   'he doesn't go'
                      he not go
                      il ne sai   'he doesn't know'
                      he not know
        Stage II   il ne vait/va pas   'he doesn't go a step'
                      il ne sai rien   'he doesn't know a thing'
        Stage III   il ne va pas   'he doesn't go'
                      il ne sait pas   'he doesn't know'
        Stage IV   il va pas   'he doesn't go'
                      il sait pas   'he doesn't know'

At Stage II, the nouns *pas* and *rien* (as well as a number of other such terms) were still being consistently used with with verbal heads appropriate to their semantics as nouns. However, they had probably come to function as emphasizers, as their translations currently do in English. Speaking in categorical terms, we can say that *pas* underwent a significant structural shift in getting from Stage II to Stage III for at that point it seems to have made the transition from being an emphasizer/noun to being a purely abstract negation marker. Although OFr shows some evidence of doubling *ne* with nominal and adjectival negative particles (e.g. *Je* **ne** *cherche* **ni** *lui* **ni** *son frère, Je* **n'ai** **nulle** *envie de le voir* [Price 1971]), this usage is quite distinct from the verbal double negation with *pas*, so it seems dubious that these usages could have formed an analogical basis for the creation of the verbal negation. It is also not clear that the nominal and adjectival doubling preceded the verbal doubling. Therefore, following Meillet, one might say that this is a case of innovative *grammaticalisation* that cannot be treated as an instance of analogy.

The traditional way of formalizing analogy is to draw a *proportion*. For example, Hock 1986 notes that a proportion like

        stone   :   stone-s
(148)   cow   :   X
            X = cow-s

can be used to derive the innovative form (*cow-s*) that replaced the inherited *kine*. It is not hard, in fact, to construct a proportion that derives the appropriate prediction in the case of French *pas*. For example:

        Il ne va   :   Il ne va pas
(149)   Il ne sait   :   X
            X = Il ne sait pas

Perhaps, then, this is not a good case of a non-analogical innovation after all. But Meillet can object that analogical proportions cannot be allowed to compare any arbitrary bits of speech-stream that one cares to lay down on either side of a colon. In (148) we surely want to insist that any unit we compare with *stone* must be, like *stone*, a singular count noun in order not to predict absurdities like **the-s, *quickly-s, *dogs-s, *serenity-s*. The corresponding move in the case of the French negation would seem to be to give a very specific analysis like *Il ne va pas* as <NP Neg Verb[walking] NP[walking-subunit]> because if we cut in at a more abstract level, for example <NP Neg Verb NP>, then we will make all sorts of absurd predictions like

        They climbed   :   They climbed trees
(150)   They slept   :   X
            X = They slept trees

And yet if we make the specific analysis, the proportion is too limited; it cannot predict the observed extension. Unfortunately, it does not seem that the tools of grammatical description can provide us with a representation of any useful intermediate generality. Thus Meillet's objection seems to stand.

It turns out, however, that in the earliest attested French, a number of particles were coming to be used as emphasizers. Besides *pas*, there were also *mie* < L. *mica* 'crumb', *point* < L. *punctum* 'point', *rien* < L. *rem* 'thing', and several others (Möhren 1943, Price 1971). If we try to apply the proportional analysis individually to each of them, the same results obtain: the proportion must either be too restrictive and fail to predict the innovation, or too permissive and predict rampant, absurd innovation. But here the Restrictive Continuity model offers some insight. Because such a variety of particles associated with diverse verbs were coming to be used as emphasizers, it must have become increasingly efficient to allocate a region of hidden unit space to the emphasizer category.

This must have happened gradually, with the region being initially a cloudy mixture of pieces of the regions associated with the different emphasizer particles in their original senses. As the emphasizers became more normal (actually *less* emphatic) and hence more general, the distinctness of the region must have become more pronounced. This distinctness must have developed without there being a special genetically-encoded "parameter" in the representation space, specially dedicated to activating a second negation marker should a language happen to provide positive evidence for one (cf. Lightfoot 1991). Instead, the distinctive properties of the second verbal negation category must have come about as the linguistically idealized imprint of a trend in Vulgar Latin toward using an abundance of emphasizers.

This is a speculative indication of the way in which the network model may be useful in predicting "truly innovative" innovation in Meillet's sense. Although none of the simulations I have presented here show this kind of effect directly, it may be noted that the presented simulations illustrate induction of categories like Noun, Verb, Adverb, Auxiliary, etc. from language-like data sources with no special pre-encoding of these categories into the architecture. Consequently, it is quite plausible that a network encountering a progression of distributions like that of early French could converge on the representation of an emphasizer category without special pre-coding for this category. If so, then the work the model will have done is to make a connection between a low-level, plausible societal tendency (the tendency to generalize particles with meanings like 'a step', 'a crumb', 'a thing' to more and more of the cases in which they are semantically appropriate) and the eventual abstraction over all these cases to a special, grammatically-distinct class of negation-emphasizers, which ultimately become restricted to a small set of negation-doublers. The insight it provides is that complex, but systematic structure may arise from the interaction of a persisting social force, whose actual properites may not be very linguistically "well-formed", with a general form-idealizing mechanism like a Connectionist network. It thus recommends rejection of strong functionalism and literalist nativism alike and replaces them with an elastic mechanism which chooses something formally "nice", fitting as optimally as possible the characteristics of its environment.

7.4 Future Research

Three projects for research hence seem especially worth mentioning. (a) The natural analogy between networks and individual people can be exploited to build a more sophisticated account of language transmission in a society; (b) the question of how to map between linguistic phrase-structure representations and network metric-space representations can be addressed; (c) the question of how an architecturally homogeneous network can make predictions about the phenomena that, under traditional linguistic analysis, motivate the supposition that grammar is *modular* can be investigated. A case of particular relevance to the last project is the putatively modular distinction between syntax and lexicon.

7.4.1 Diachronic network chains

Letting the output of one network provide the input to another network is a way of adjusting the network change model to make it consistent with the Indirect Transmission fact. As I noted in Section 2.1.4 above, such a framework could be used to study the possibility that global abduction gives rise to structural change. Experimenting in this framework might also shed some light on "truly innovative" structural change in Meillet's sense, discussed in Section 2.2 above. In fact, Denaro and Parisi (ms) have done a study which hints at the possible fruitfulness of this approach. These authors used a task in which feedforward networks learned to map each of the integers from 0 to 14 to its successor. The inputs and outputs were encoded in binary format using four units. They set up a chain of ten networks that were all trained simultaneously, with the first network getting as its target values the task-defined targets, the second network getting the outputs of the first, the third getting the outputs of the second, etc. A graph of the performance error for each of the ten networks during the training process is shown in Figure 7.5. What is interesting about this result are the ripples that appear in the error curves for the networks somewhat removed from the initial model. Presumably these ripples arise in the intermediate networks because of random distortions in their starting configurations. But in several cases they become enhanced as they are transmitted to networks further down the line. Presumably the enhancement comes about because there

Figure 7.5: Sequential imitation for 10 nets [from Denaro and Parisi (ms)]

is some structure in the networks that happens to jibe in an effective way with the structure of the task. If the networks were actually ideal imitators, they should show no tendency to magnify a distortion, only a tendency to preserve it or, given randomness, to distort it arbitrarily. This suggests that the chained-training set-up might provide a way of modelling a case like the development of French measure-particles into the second negation marker. What it could do for a case like that is show how a small initial distortion, or a semi-persistent perturbational pressure could be taken up and formed into a full-fledged category, perhaps one corresponding to a limited set of attractor states that the network is capable of entering. One might, in this case, think of the evolutionary process itself as a kind of pattern completion on noisy inputs (see Hertz, Krogh, and Palmer 1991, pp. 11–24).

7.4.2 Constituenthood in the recurrent network representation

A question which really ought to be answered in order to make it easy to compare the model discussed here to previous linguistic work is how to map from linguistic constituent representation to the network metric-space representation. Using hierarchical clustering analysis to probe the structure of hidden unit space (as in Elman 1990 and Chapter 5, Section 2.2) is a helpful first cut at the problem for the simple case of lexical constituents. The case of phrasal constituents is more difficult. Elman 1991 proposes one approach to this problem. He reports the trajectories a word-prediction network traces out in Principal Component space (see Chapter 3, Section 7.2; Weigend 1993) as it processes clauses with some fairly sophisticated embedded structure. An example trajectory is shown in Figure 7.6. The fact that the network makes nearly the same circuit with each repetition of the relative clause suggests that constituents may be assigned to subspaces of hidden unit space. One might suppose that context independence of constituents is encoded as orthogonality between the relevant subspaces. If this can be confirmed, there is an important question as to how partial context-independence is interpreted on this account, as for example in the case of multiple prepositional phrases modifying their predecessors objects (e.g., *the peach in the box on the seat of the car...*). Such constructions are pretty endlessly grammatical but the longer they get, the higher the likelihood that the repetition will terminate at the next phrase-end. Presumably this gives rise to some slippage in the representation away from the perfectly repetitive representation of each successive PP-sequence that a context-free grammar gives.[5] Such slippage may be capable of leading to constituent structure revision (as apparently happened in the case of *sort of* and *kind of*). What needs better understanding is what kind of structural claim the network is making when it is employing one of these slipped constituent units.

---

[5] It may be that the displacement of the third and fourth instances of *boy* from the second in Figure 7.6 reflects such slippage.

Figure 7.6: Trajectory in hidden unit space for the sentence *Boy chases boy who chases boy who chases boy.* Principal Component 1 is displayed along the horizontal axis and Principal Component 11 is displayed along the vertical axis [From Elman 1991, p. 113]

7.4.3 Modularity

Most current linguistic theories are highly modular in the sense that they divide the task of generating language up into a set of subtasks that can be performed by processors acting independently of one another and sometimes involve distinct kinds of processing. A case in point is the modularization of the syntax and lexicon. It is a basic architectural assumption of almost all current theories that grammar starts with a lexicon and builds sentences by binding words together in an essentially concatenative fashion. Even theories, like Baker 1986's theory of Incorporation, which allow syntactic transformational processes to

revise word-internal structure, depend on the assumption that there is a vocabulary of atomic morphemic units which form the building blocks. However, some suspicion is cast on the assumption about atomicity by the observation that certain phrases, usually counted as *idioms*, seem to show varying degrees of compositionality or "syntactic productivity" (e.g., Fillmore, Kay, and O'Conner 1988, Nunberg, Sag, and Wasow (ms)). This suspicion is adumbrated by the observation that elements seem able to transit diachronically from being morphologically complex to monolithic, and moreover, do so in a gradual fashion. The data on the separability of *going* and *to* in future *going to* provide a case in point. Hardly anything intervenes in currently observed usage. Indeed hardly anything seems to be able to grammatically intervene, but a scan of a large corpus produces a few examples which are not altogether infelicitous. For example:

(151) There is just now a curious unease about Australian cricket, almost as if the country is <u>going</u> suddenly <u>to</u> wake up and find that England was a dream after all. *Henry Blofeld in Perth* [HECT.Indept.51.corp]

This evidence urges a representation in which there is a notion of intermediacy between status as an atomic morphemic unit and the lack of it. But at first it is hard to imagine what sort of representation could fill the bill. After all, our notions about morphological and syntactic structure depend rather fundamentally on the assumption that there are units which can be concatenated in various orders. Nevertheless, there is plenty of motivation for positing submorphemic linguistic units (syllables and phonemes). Moreover, this thesis has provided some indication that the network representation can portray gradual shift from sequence-of-categories status to single-category status (the case of Noun-Prep *sort/kind of* spawning a mono-morphemic Degree Modifier). An intriguing possibility, therefore, is to inquire what the network correlates of syntax and lexicon are and see if analyzing its continuous representation sheds light on the question how there can be syntax without fundamental atomicity.

OXF = Oxford Shakespeare Corpus

Appendix A

Corpora

AIR = Academic Information Resources online database at Stanford University.
Cited works from AIR's database:

> Austen, J. *Emma, Mansfield Park, Northanger Abbey*
>
> Defoe, D. *Moll Flanders*
>
> The Diachronic Helsinki Corpus of English Texts (HELS: see below)
>
> Doyle, A. C. *Sherlock Holmes* (collected volumes)
>
> Hume, D. *Enq. Hum. Und., Dial. Nat. Rel., Enq. Princip. Morals,*
> *Treatise H. N., My Own Life, History of Religion, Passions,* and
> various Essays.
>
> Joyce, J. *Ulysses*
>
> Melville, H. *Billy Budd, Moby Dick*
>
> The Oxford Shakespeare Corpus (OXF)
>
> The Network News Corpus of email exchanges about NEXT com-
> puters (NN)

EDD = The English Dialect Dictionary

HECT[G] = Hector Corpus, part of the Oxford Corpus of British English (se-
lections for this study taken from the *Manchester Guardian*)

HELS = Diachronic Helsinki Corpus of English Texts (see Kytö 1991)

OED = The Oxford English Dictionary

References

Abbott, E. A. 1966. *A Shakespearian Grammar*. New York: Dover.

Allen, C. L. 1983. *Topics in Diachronic Syntax*. Ann Arbor: University Microfilms International.

Altmann, G., H. v. Buttlar, W. Rott, and U. Strauβ. 1983. A law of change in language. In B. Brainerd (Ed.), *Historical Linguistics*, 104–115. Bochum: Studienverlag Dr. N. Brockmeyer.

Andersen, H. 1972. Diphthongization. *Language* 48:11–50.

Andersen, H. 1973. Abductive and deductive change. *Language* 49(4):765–793.

Andersen, H. 1989. Understanding linguistic innovations. In L. E. Breivik and E. H. kon Jahr (Eds.), *Language Change: Contributions to the Study of its Causes*, 5–28. Berlin: Mouton de Gruyter.

Bailey, C.-J. 1973. *Variation and Linguistic Theory*. Washington: Center for Applied Linguistics.

Bergström, G. A. 1906. *On Blendings of Synonymous or Cognate Expressions in English*. Lund: Håkan Ohlsson.

Bolinger, D. 1961. Syntactic blends and other matters. *Language* 37(3):366–81.

Bolinger, D. 1972. *Degree Words*. The Hague: Mouton.

Bolinger, D. 1973. Ambient *it* is meaningful too. *Journal of Linguistics* 9:209–383.

Bresnan, J., and S. Mchombo. 1987. Topic, pronoun, and agreement in Chichewa. *Language* 63(4):741–782.

Bresnan, J., and L. Moshi. 1990. Object asymmetries in comparative bantu syntax. *Linguistic Inquiry* 21:147–85.

Brill, E., D. Magerman, M. Marcus, and B. Santorini. n.d. Deducing linguistic structure from the statistics of large corpora. Manuscript, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104.

Brown, P. F., V. J. D. Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. 1990. Class-based n-gram models of natural language. In *Proceedings of the IBM Natural Language ITL*, 283–298.

Bybee, J., and W. Pagliuca. 1985. Cross-linguistic comparison and the development of grammatical meaning. In J. Fisiak (Ed.), *Historical Semantics and Historical Word Formation*, 59–83. de Gruyter.

Bybee, J., W. Pagliuca, and R. Perkins. 1991. Back to the future. In E. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 2*, 17–58. John Benjamins.

Campbell, L. 1991. Some grammaticalization changes in Estonian and their implications. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 1*, 285–300. John Benjamins.

Carlson, R. 1991. Grammaticalization of postpositions and word order in senufo languages. In E. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 2*, 201–224. John Benjamins.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton and Co.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Chomsky, N. 1986. *Barriers*. Cambridge: MIT Press.

Chomsky, N., and H. Lasnik. To appear. Principles and parameters theory. In J. Jacobs, A. van Stechow, W. Sternefeld, and T. Vennemann (Eds.), *Syntax: An International Handbook of Contemporary Research*. Walter de Gruyter.

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

Cohen, G. L. 1987. *Syntactic Blends in English* Parole. Frankfurt am Main: Verlag Peter Lang.

Craig, C. 1991. Ways to go in Rama: a case study in polygrammaticalization. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 2*, 455–492. John Benjamins.

Danchev, A., and M. Kytö. Forthcoming. The construction *be going to + infinitive* in Early Modern English. In D. Kastovsky (Ed.), *Papers from the Early Modern English Conference (EMEC), Tulln, 1991.* Mouton de Gruyter.

Danchev, A., A. Pavlova, M. Nalchadjan, and O. Zlatareva. 1965. The construction *going to* + inf. in Modern English. *Zeitschrift für Anglistik und Amerikanistik* 13(4):375–386.

Dell, G. S., C. Juliano, and A. Govindjee. 1993. Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science* 17:149–195.

di Sciullo, A.-M. 1990. On the properties of clitics. In A.-M. di Sciullo and A. Rochette (Eds.), *Binding in Romance; Essays in Honor of Judith McA'Nulty*, 209–232. The Canadian Linguistics Association.

Dowty, D. 1985. On recent analyses of the semantics of control. *Linguistics and Philosophy* 8:291–331.

Dowty, D. R., R. E. Wall, and S. Peters. 1981. *Introduction to Montague Semantics.* Dordrecht: D. Reidel Publishing Co.

Eckert, P. 1988. Adolescent social structure and the spread of linguistic change. *Lang. Soc.* 17:183–207.

Ellegård, A. 1953. *The Auxiliary* DO. Stockholm: Almqvist & Wiksell.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science* 14:179–211.

Elman, J. L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7:195–225.

Finch, S. P. 1993. Finding structure in language. Doctoral Dissertation, Cognitive Science Department, University of Edinburgh.

Fontaine, C. 1985. Application de methodes quantitatives en diachronie: L'inversion du sujet en français. M.A. Thesis, Université du Québec à Montréal.

Fontana, J. M. 1993. Phrase structure and the syntax of clitics in the history of Spanish. Ph.D. Dissertation, Linguistics, University of Pennsylvania.

Franco, J. 1991. Conditions on clitic doubling: The agreement hypothesis. Paper read at the LSRL XXI, UC Santa Barbara.

Franco, J. 1993. On object agreement in spanish. Ph.D. dissertation, University of Southern California.

Genetti, C. 1986. The development of subordinators from postpositions in bodic languages. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society*, Vol. 12, 387–400. Berkeley Linguistics Department.

Genetti, C. 1991. From postposition to subordinator in rama. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 2*, 227–256. John Benjamins.

Givón, T. 1971. Historical syntax and synchronic morphology: An archaeologist's field trip. *Proceedings of the 7th Regional Meeting of the Chicago Linguistic Society* 7:394–415.

Givón, T. 1976. Topic, pronoun, and grammatical agreement. In C. N. Li (Ed.), *Subject and Topic*, 149–88. Academic Press.

Givón, T. 1979. *On Understanding Grammar.* New York: Academic Press.

Givón, T. 1984. *Syntax I.* Amsterdam: John Benjamins.

Givón, T. 1991. Serial verbs and the mental reality of 'event': grammatical vs. cognitive packaging. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 1*, 81–127. John Benjamins.

Goossens, L. 1982. On the development of the modals and of the epistemic function in english. In A. Ahlqvist (Ed.), *Papers from the 5th International Conference on Historical Linguistics*, 74–84. John Benjamins.

Graudina, L. N. 1964. Razvitie nulevoj formy forditelńogo monožestvennogo u suščestvitelńych—jedinic izmerenija. In *Razvitie grammatiki i leksiki sovremennogo russkogo jazyka*, 210–221. Moskva: Nauka.

Greenberg, J. H. 1978. How does a language acquire gender markers? In J. H. Greenberg, C. A. Ferguson, and E. Moravcsik (Eds.), *Universals of Human Language, v. 3*, 47–82. Stanford: Stanford University Press.

Harris, A. C. 1990. Alignment typology and diachronic change. In W. P. Lehman (Ed.), *Language Typology 1987: Systematic Balance in Language; Papers from the Linguistic Typology Symposium*, 67–90. John Benjamins.

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

Hiltunen, R. 1983. *The Decline of the Prefixes and the Beginnings of the English Phrasal Verb*. Turku: Turun Yliopisto. Cited in van Kemenade 1987.

Hook, P. E. 1991. The emergence of perfective aspect in Indo-Aryan languages. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 2*, 59–90. John Benjamins.

Hopper, P. 1991. On some principles of grammaticization. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 1*, 17–36. John Benjamins.

Hopper, P. J., and E. C. Traugott. 1993. *Grammaticalization*. Cambridge, England: Cambridge University Press.

Hyams, N. M. 1986. *Language Acquisition and the Theory of Parameters*. Dordrecht: D. Reidel.

Janda, R. D. 1980. On the decline of declensional systems: The overall loss of OE nominal case inflections and the ME reanalysis of *-es* as *his*. In E. C. Traugott, R. Labrum, and S. Shepherd (Eds.), *Papers from the 4th International Conference on Historical Linguistics*, 243–252. John Benjamins.

Jespersen, O. 1918. *Chapters on English*. London: Allen and Unwin, Ltd.

Kiparsky, P. n.d. The phonological basis of sound change. Paper given at the Workshop on Sound Change held at Stanford University, February 15–16, 1993.

Kiparsky, P. 1982a. Analogical change as a problem for linguistic theory. In *Explanation in Phonology*, chapter 11. Foris.

Kiparsky, P. 1982b. Remarks on analogical change. In *Explanation in Phonology*, 199–216. Foris.

Klein, E., and I. Sag. 1985. Type-driven translation. *Linguistics and Philosophy* 8:163–202.

Koopman, H. 1984. *From Verb Movement Rules in the Kru Languages to Universal Grammar*. Dordrecht: D. Reidel.

Kroch, A., S. Pintzuk, and J. Myhill. 1982. Understanding *do*. In K. T. *et al.* (Ed.), *Papers from the 18th Regional Meeting of the Chicago Linguistic Society*, Chicago. University of Chicago Press.

Kroch, A. S. 1989a. Function and grammar in the history of English: Periphrastic *do*. In R. W. Fasold and D. Schiffrin (Eds.), *Language Change and Variation*, 134–169. Philadelphia: John Benjamins. Published as Vol. 52 of the series *Current Issues in Linguistic Theory*.

Kroch, A. S. 1989b. Reflexes of grammar in patterns of language change. *Journal of Language Variation and Change* 1(3):199–244.

Kroeger, P. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. Stanford, CA: CSLI.

Kytö, M. 1991. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts*. Helsinki: Department of English, University of Helsinki.

Labov, W. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45:715–62.

Lambrecht, K. 1981. *Topic, Antitopic and Verb Agreement in Non-Standard French*. Amsterdam: John Benjamins.

Langacker, R. W. 1988. A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics*, 127–164. John Benjamins.

Langacker, R. W. 1987. *Foundations of Cognitive Grammar, v. 1.* Stanford, California: Stanford University Press.

Lausberg, H. 1962. *Romanische Sprachwissenschaft, v. 1–3.* Berlin: de Gruyter.

Li, C. N., and S. A. Thompson. 1973. Serial verb constructions in Mandarin Chinese: Subordination or coordination? In *You Take the High Node and I'll Take the Low Node.* Chicago: Chicago Linguistic Society.

Lichtenberk, F. 1991. On the gradualness of grammaticalization. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 1*, 37–80. John Benjamins.

Lightfoot, D. 1979. *Principles of Diachronic Syntax.* London: Cambridge University Press.

Lightfoot, D. 1982. *The Language Lottery.* Cambridge, Massachusetts: MIT Press.

Lightfoot, D. 1991. *How to Set Paramters: Arguments from Language Change.* Cambridge, Massachusetts: MIT Press.

Lord, C. 1973. Serial verbs in transition. *Studies in African Linguistics* 4(3):269–296.

Lord, C. 1976. Evidence for syntactic reanalysis: from verb to complementizer in Kwa. In *Papers from the Parasession on Diachronic Syntax, Chicago Linguistic Society*, 179–91. University of Chicago Press.

Magerman, D. M., and M. P. Marcus. 1990. Parsing a natural language using mutual information statistics. In *Proceedings of AAAI-90*, Boston, Massachusetts.

Marchand, H. 1966. *The Categories and Types of Present-Day English Word-Formation: A Diachronic Approach.* Alabama: University of Alabama Press.

McCawley, J. D. 1988. *The Syntactic Phenomena of English, v. 1–2.* Chicago: The University of Chicago Press.

McClelland, J. L., D. E. Rumelhart, and the PDP Research Group. 1986. *Parallel Distriubted Processing. Explorations in the Microstructure of Cognition. Volume 2.* Cambridge, Massachusetts: MIT Press.

Möhren, F. 1943. Le renforcement affectif d el négation par l'expression d'une valeur minimale en ancien français. *ZRPH Beiheft* 175:1–264.

Noble, S. 1985. To have and have got. Paper presented at NWAVE 14, Georgetown University.

Oliveira e Silva, G. 1982. Estudo da Regularidade na Variação dos Possessivos no Português do Rio de Janeiro. Ph.D. dissertation, Universidade Federal do Rio de Janeiro.

Osgood, C., and T. Sebeok. 1954. Psycholinguistics: A survey of theory and research problems. *Journal of Abnormal and Social Psychology* 49(4, part 2):1–203.

Palmer, F. R. 1990. *Modality and the English Modals, 2nd ed.* London: Longman.

Pérez, A. 1990. Time in motion: Grammaticalisation of the *be going to* construction in English. *La Trobe University Working Papers in Linguistics* 3:49–64.

Pintzuk, S. 1991. *Phrase Structures in Competition.* Ph.D. Dissertation, University of Pennsylvania.

Plank, F. 1984. The modals story retold. *Studies in Language* 8(3):305–64.

Platzack, C., and A. Holmberg. 1990. The role of AGR nad finiteness in some European VO languages. Manuscript, University of Lund.

Prince, A., and P. Smolensky. 1993. Optimality: Constraint interaction in generative grammar. Ms., Rutgers Cognitive Science Center and Institute of Cognitive Science at the University of Colorado at Boulder.

Roberts, I. 1985. Agreement parameters and the development of the English modal auxiliaries. *Natural Language and Linguistic Theory* 3(1):21–58.

Roberts, I. 1992. *Verbs and Diachronic Syntax.* Dordrecht: Kluwer.

Roberts, I. 1993. A formal account of grammaticalisation in the history of romance futures. *Folia Linguistica Historica* XIII/1–2:219–58.

Roeper, T., and E. Williams (Eds.). 1987. *Parameter Setting.* Dordrecht: D. Reidel.

Rosch, E., C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8:382–439.

Rosenbaum, P. S. 1967. *The Grammar of English Predicate Complement Constructions.* Camnbridge, Massachusetts: MIT Press.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986a. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (Eds.), *Parallel Distributed Processing, Volume I*, 318–362. MIT Press.

Rumelhart, D. E., J. L. McClelland, and the PDP Research Group. 1986b. *Parallel Distriubted Processing. Explorations in the Microstructure of Cognition. Volume 1.* Cambridge, Massachusetts: MIT Press.

Santorini, B. 1989. The generalization of the verb-second constraint in the history of Yiddish. PhD Dissertation, University of Pennsylvania.

Santorini, B. 1992. Variation and change in Yiddish subordinate clause word order. *Natural Language and Linguistic Theory* 10:595–640.

Schachter, P. 1974. A non-transformational account of serial verbs. *Studies in African Linguistics* Supplement 5:253–270.

Schütze, H. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92.*

Schütze, H. 1993a. Distributed syntactic representations with an application to part-of-speech tagging. Paper given at the International Conference on Neural Networks.

Schütze, H. 1993b. Word space. In S. J. Hanson, J. D. Cowan, and C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 5*, 895–902. San Mateo, California: Morgan Kaufmann Publishers.

Shepard, R. N., and J. D. Carroll. 1966. Parametric representation of nonlinear data structures. In P. R. Krishnaiah (Ed.), *Multivariate Analysis*, 561–592. Academic Press.

Shepherd, S. 1982. From deontic to epistemic: An analysis of modals in the history of English, creoles, and language acquisition. In A. Ahlqvist (Ed.), *Papers from the 5th International Conference on Historical Linguistics*, 316–323. John Benjamins.

Shibatani, M. 1991. Grammaticalization of topic into subject. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization, v. 2*, 93–134. John Benjamins.

Suner, M. 1988. The role of agreement in clitic-doubled constructions. *Natural Language and Linguistic Theory 6* 6:391–434.

Sweetser, E. 1990. *From Etymology to Pragmatics.* Cambridge, England: Cambridge University Press.

Tabor, W. 1994. The gradual development of degree modifier *sort of*: A corpus proximity model. In K. Beals, G. Cooke, D. Kathman, K.-E. McCullough, S. Kita, and D. Testen (Eds.), *Proceedings of the 29th Regional Meeting of the Chicago Linguistic Society.* University of Chicago.

Tajima, M. 1985. *The Syntactic Development of the Gerund in Middle English.* Tokyo: Nan'un-do.

Taylor, A. 1992. The change from verb-final to verb-medial in Ancient Greek. In the papers packet for the 2nd Diachronic Generative Syntax Workshop, held from November 5–8, 1992 at the University of Pennsylvania.

Traugott, E. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 65(1):31–55.

Traugott, E. C. Forthcoming. Subjectification in grammaticalization. In D. Stein and S. Wright (Eds.), *Proceedings of the Seminar on Language, Subjectivity and Subjectification.* Cambridge, England: Cambridge University Press.

Traugott, E. C., and E. König. 1991. The semantics-pragmatics of grammaticalization revisited. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization.* John Benjamins.

Trutenau, H. M. J. 1973. The verbal status of the NP-linker in gã. *Studies in African Linguistics* 4(1):71–86.

von Neuman, J. 1941. Distribution of the ration of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 12:153–162.

Warner, A. 1982. *Complementation in Middle English and the Methodology of Historical Syntax.* University Park: The Pennsylvania State University Press.

Warner, A. 1983. Review of D. Lightfoot, *principles of diachronic syntax. Journal of Linguistics* 19:187–209.

Weigend, A. S. 1994. On overfitting and the effective number of hidden units. In M. C. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, and A. S. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School*, 335–342. Erlbaum Associated.

Weinreich, U., W. Labov, and M. Herzog. 1968. Empirical foundations for a theory of language change. In W. P. Lehmann and Y. Malkiel (Eds.), *Directions for Historical Linguistics*, 95–188. University of Texas Press.

Zipf, G. K. 1943. *Human Behavior and the Principle of Least Effort.* Hafner [1965].